İSTATİSTİK BÖLÜMÜ

MAKİNE ÖĞRENMESİ TEKNİKLERİYLE OYUNCULARIN KÜMELENMESİ VE POZİSYONLARA GÖRE ROLLERİN İNCELENMESİ

Ahmet Yiğit Bartık 19023003

Danışman: Dr. Öğr. Üyesi Erhan ÇENE

ÖZET

Bu çalışmada, Avrupa'nın 5 büyük ligi olan Premier Lig, La Liga, Serie A, Ligue 1 ve Bundesliga'daki 2021-2022 sezonu futbolcu istatistikleri kullanılarak, oyuncuların mevkilere göre temel bileşen analizi ve kümeleme analizi teknikleriyle pozisyonlarındaki rollerinin incelenmesi amaçlanmıştır. Kalecileri 3, savunma oyuncularını, orta saha oyuncularını ve forvetleri ise 5 farklı kümeye ayırarak, farklı oyuncuların benzer performans gösterebileceği mevkileri ve oyun stillerini belirlemeye çalışılmıştır. Bulgular, futbolcularda çok yönlü performans ve farklı mevkilerdeki oyuncuların benzer rolleri üstlenebileceği konusunda önemli ipuçları sunmuştur. Ayrıca, regresyon modeli kurularak mevkilere ve rollere göre ayrılan kümelerin değişkenleri detaylı bir şekilde incelenmiştir. Bu çalışma, oyuncu izleme alanında transfer süreçlerinde kulüplere yardımcı olabilecek bir kaynak niteliği taşımaktadır.

KEŞİFSEL VERİ ANALİZİ

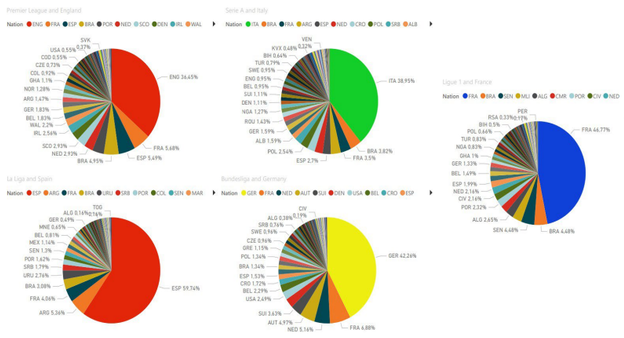
Keşifsel veri analizi ile bu veri seti veri analizleri uygulanabilir hale getirildi. 2500'den fazla oyuncu ve 143 değişkenden oluşmaktaydı. Analiz sürecinde aşağıdaki adımlar izlendi:

1. Veri Temizleme ve Filtreleme: İlk olarak, oyuncuların maç sayısı ve oynadıkları dakikaları temel alarak alt sınırlar belirlendi. 1. çeyrek değerlerinin altında kalan oyuncular veri setinden çıkarıldı. Bu adım, yeterli veri bulunmayan oyuncuların analizden çıkarılmasını sağladı.

2. Değişkenlerin Düzenlenmesi: Pozisyon değişkeni oldukça karmaşık ve çeşitliydi. Bu karmaşıklığı azaltmak amacıyla, oyuncuları temel dört mevkiye (örneğin, kaleci, defans, orta saha, forvet) indirgendi. Bu sınıflandırma, veriyi daha anlaşılır hale getirdi ve analizlerin daha odaklı olmasını sağladı.

3. Özet İstatistiklerin Hesaplanması ve Görselleştirme: Verinin genel dağılımını ve özelliklerini anlamak için özet istatistikler hesaplandı ve çeşitli grafiksel yöntemler kullanıldı.

(1)



(1) Örnek olarak bu grafiklerde her ligdeki oyuncu grubunun hangi yüzde ile yerel ülkelerden geldiğini görüyoruz. İspanya Ligi'nde yerel oyuncuların yüksek bir yüzde ile temsil edildiği dikkat çekerken, Futbolun NBA'i olarak nitelendirilen Premier Lig'de ise her ülkeden oyuncunun daha fazla katılım gösterdiği ve İngiliz oyuncularının yüzdesinin diğer yerel liglerdeki ülkelerin oyuncularına kıyasla daha düşük olduğu gözlemlenmektedir.

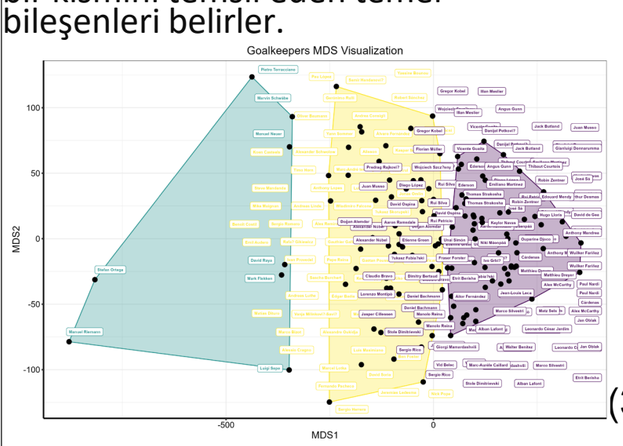
(2)

groups	Goals	Shots	SoT	PasTotCmp	PasTotAtt	PasTotDist	PasTotPrgDist	PasShoCmp	PasShoAtt	PasMedCmp	PasLonCmp	Assists
1	0.0385	0.6238	0.1709	29.4183	39.9360	540.9652	212.8774	12.8271	14.9443	12.5701	3.5046	0.0551
2	0.0540	0.6091	0.1710	67.9328	76.0388	1451.6493	450.0209	20.4457	21.9299	35.0000	11.9506	0.0513
3	0.0417	0.5677	0.1496	36.6606	46.6770	709.3282	258.2291	14.4887	16.4548	16.5784	5.0483	0.0542
4	0.0513	0.6044	0.1729	54.2685	62.6766	1104.9629	359.2589	18.5698	20.1817	26.6992	8.3794	0.0470
5	0.0467	0.5263	0.1452	43.9198	53.7751	876.7819	306.4147	16.0344	17.8738	20.8215	6.5118	0.0593

(2) Bu tabloda, savunma oyuncularının oluşturduğu beş ayrı kümenin ortalamaları görülmektedir.

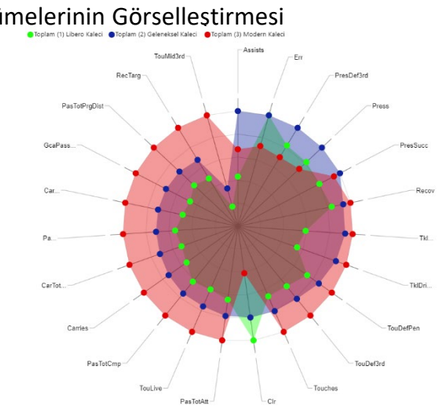
PCA (TEMEL BİLEŞEN ANALİZİ)

PCA, çok boyutlu veri setlerindeki değişkenlikleri daha az boyutta bir yapıya indirgeyen bir veri analizi tekniğidir. Bu teknik, veri setindeki değişkenler arasındaki ilişkileri anlamak ve veri setinin boyutunu azaltmak için kullanılır. PCA, veri setindeki ana desenleri ve varyansın büyük bir kısmını temsil eden temel bileşenleri belirler.



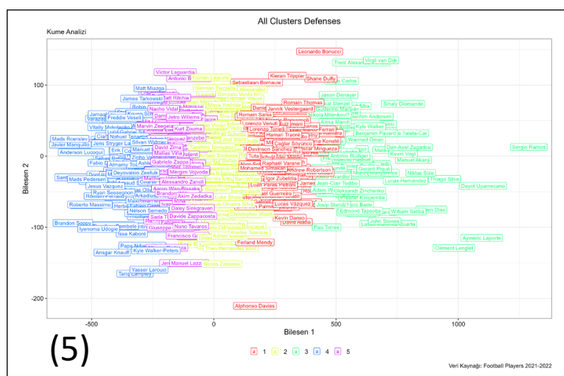
(3)

(4) Çok Boyutlu Ölçekleme ile Kalecilerin Kümelerinin Görselleştirmesi

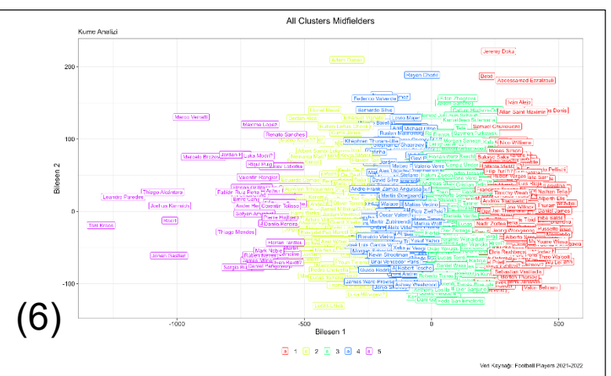


K-MEANS CLUSTERING (KÜMELEME ANALİZİ)

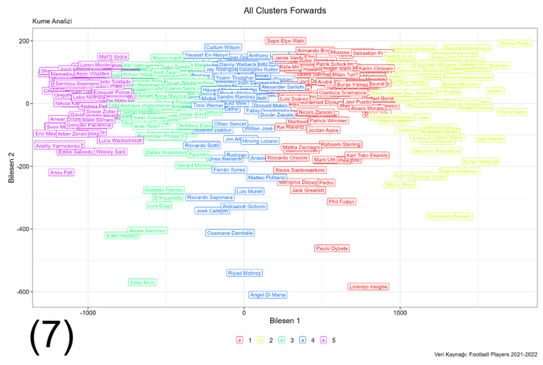
K-means kümeleme analizi, veri noktalarını belirli bir sayıda k kümesine ayıran ve her kümenin merkezini hesaplayan bir algoritmadır. Başlangıçta rastgele seçilen k merkezle başlayarak, her veri noktası en yakın merkeze atanır ve ardından her kümenin yeni merkezi hesaplanır.



(5)



(6)



(7)

(3),(5),(6) ve (7) görselleri Kaleci, Savunma, Orta Saha ve Forvet pozisyonlarının kümeleme görselleridir.

MDS (ÇOK BOYUTLU ÖLÇEKLEME)

Çok boyutlu ölçekleme çok boyutlu verileri daha az boyutta temsil etmek için kullanılan bir tekniktir. Genellikle görselleştirme amacıyla kullanılır ve benzerlik veya mesafe ölçümlerini korur.

Yapılan kümeleme analizinde, 4 farklı pozisyon (Kaleci, Savunma, Orta Saha ve Forvet) için ayrı ayrı kümeler oluşturulmuştur. Kaleci pozisyonu için 3, diğer pozisyonlar için ise 5 farklı küme belirlenmiştir. Her bir küme, ilgili pozisyonun önemli değişkenleri kullanılarak tanımlanmıştır. Bu durum, farklı pozisyonlardaki oyuncuların benzer performans sergileyebileceğini işaret etmektedir.

Örneğin, Savunma pozisyonunda yapılan kümeleme analizinde, stoper ve bek oyuncularının aynı kümeye dahil olduğu gözlemlenmiştir. Bu durum, çoklu mevkide oynayabilen ya da daha önce bu pozisyonlarda oynamamış oyuncuların da belirli kümelerde performans gösterebileceğini göstermektedir.

Her bir kümenin özellikleri, değişken seçimi yapılarak ve kümeleme analizi tekrar gerçekleştirilerek belirlenmiştir. Bu yöntem, oyuncu seçimi ve transfer çalışmalarında yardımcı olabilecek bir araç sunmaktadır.

REGRESYON ANALİZİ VE MODELİN PERFORMANSI

Regresyon Modeli Denklemi

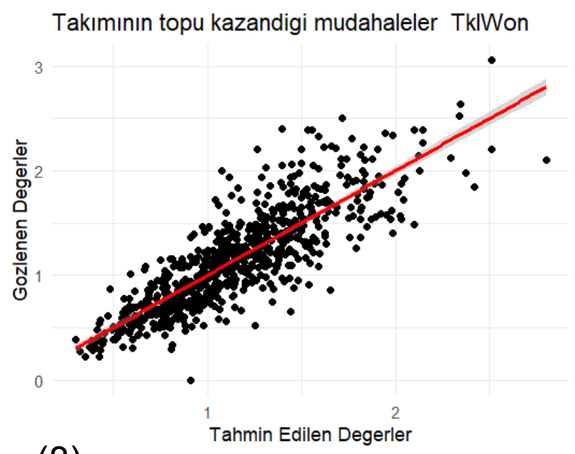
$$TklWon=0.0379\cdot Press+0.2780\cdot TklDri+0.4619\cdot TklDef3rd+0.0037\cdot PasGround+\varepsilon$$

Model Sonuçları

Modelin performansı, bağımlı değişkendeki varyansın %70.66'sını açıkladığını göstermektedir.

(R^2 =0.7066, Düzeltilmiş R^2 = 0.705).
F-istatistiği 445.5 - p-değeri < 2.2e-16
Modelin anlamlı olduğunu belirtir.

- **Press**: 0.0379 (p < 2e-16)
- **TklDri**: 0.2780 (p = 3.48e-12)
- **TklDef3rd**: 0.4619 (p < 2e-16)
- **PasGround**: 0.0037 (p = 2.71e-06)



(8)

KAYNAKÇA

- Abdi, H., & Williams, L. J. (2010). Principal Component Analysis..
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
- Borg, I., Groenen, P., & Mair, P. (Eds.). (2018). *Applied Multidimensional Scaling.* Springer.

İSTATİSTİK BÖLÜMÜ

KONJOİNT ANALİZİ İLE ÜRÜN TERCİHİ VE PERAKENDE SEKTÖRÜNDE BİR UYGULAMA

Ali Efe Çınar 18023038

Danışman: Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK

ÖZET

Konjoint analizi, tüketicilerin ürün ve hizmet özelliklerine verdiği değerleri ölçen bir tekniktir. Bu analizde, çeşitli ürün profilleri tüketicilere sunularak tercihleri alınır. Analiz, ürün özelliklerinin tüketiciler için ne kadar değerli olduğunu fayda değerleriyle belirler. Özellikle bir araştırmada annelerin sürülebilir peynir tercihleri incelenmiş, 402 anne ile görüşülerek marka, çeşit ve fiyat özelliklerine göre tercihleri değerlendirilmiştir. Bulgu olarak, marka %58 ile en önemli tercih kriteri olup, ardından çeşit %25 ve fiyat gelmiştir. Labne peyniri en yüksek pozitif fayda değerine sahipken, krem peynir en düşük değeri almıştır. Fiyatla ilgili olarak, en düşük fiyat en yüksek fayda değerine sahipken, en yüksek fiyat en düşük fayda değerini almıştır. 60,45 TL ve 70,95 TL arasında önemli fiyat kırılmaları yaşanmış, 90,45 TL'den sonra tercihler düşmüştür.

KONJOİNT ANALİZİ

Konjoint analizi, tüketicilerin çeşitli ürün ve hizmet özelliklerine verdikleri değerleri ve tercihleri ölçen istatistiksel bir tekniktir. Bu analiz, ürünlerin farklı özellik ve özellik kombinasyonlarına göre oluşturulmuş çeşitli profillerin tüketicilere sunulması ve bu profiller arasında tercih yapmalarının istenmesi ile gerçekleştirilir. Bu yöntem, tüketicilerin gerçek dünyada yapacakları seçimlere benzer şekilde, belirli özellikler ve bu özelliklerin seviyeleri arasında psikolojik ödünleşmeleri tahmin etmeyi amaçlar.

KONJOİNT ANALİZİNİN TEMEL İLKELERİ

Ürün Profilleri: Tüketicilere sunulan çeşitli ürün profilleri, farklı özellik kombinasyonlarından oluşur. Her profil, tüketicinin belirli bir ürünü tercih ederken dikkate aldığı farklı özellikleri (örneğin marka, fiyat, kalite) içerir.

Tercih Verileri: Tüketiciler, kendilerine sunulan bu profiller arasından tercih yaparak hangi özelliklerin kendileri için daha önemli olduğunu belirtirler. Bu tercihler, analiz için ham veri sağlar.

Fayda Değerleri: Elde edilen veriler, her bir özelliğin ve seviyesinin tüketici açısından ne kadar değerli olduğunu gösteren fayda değerlerine dönüştürülür. Bu değerler, tüketicinin belirli özellikleri ne kadar önemsediğini ve hangi özellik kombinasyonlarının daha çekici olduğunu ortaya koyar.

KONJOİNT ANALİZİNİN TÜRLERİ

Seçime Dayalı Konjoint Analizi (CBC): Tüketicilere çeşitli ürün profilleri sunularak en çok tercih ettikleri profili seçmeleri istenir. Bu yöntem, gerçek hayattaki alışveriş davranışlarını taklit ederek, tüketicilerin hangi özellikler arasında ödünleşim yaptığını anlamamızı sağlar.

Uyarlamalı Konjoint Analizi (ACA): Bu yöntemde, tüketicilerin daha önceki tercihlerine göre değişen sorular sorulur. Böylece daha özelleştirilmiş ve ilgi çekici bir deneyim sunularak, daha doğru veriler elde edilir.

Tam Profilli Konjoint Analizi: Tüketicilere tam ürün profilleri sunulur ve bu profiller arasında değerlendirme yapmaları istenir. Bu yöntem, her bir ürün kombinasyonunun ne kadar tercih edildiğini detaylı bir şekilde analiz eder.

KONJOİNT ANALİZİNİN AVANTAJLARI

Psikolojik Ödünleşimleri Tahmin Eder: Tüketicilerin birden fazla özelliği değerlendirirken yaptıkları ödünleşimleri anlamamızı sağlar.

Bireysel Tercihleri Ölçer: Tüketici tercihlerini bireysel düzeyde inceleyerek, daha kişiselleştirilmiş ürün ve hizmet tasarımı yapılmasına olanak tanır.

Gizli İtici Güçleri Ortaya Çıkarır: Tüketicilerin kendilerinin bile farkında olmadığı tercihler ve öncelikleri belirler.

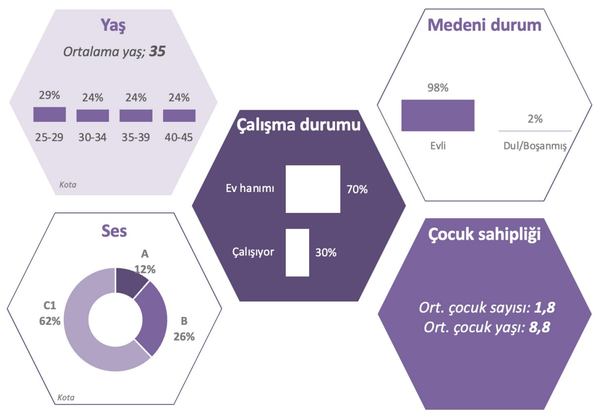
Gerçekçi Alışveriş Deneyimini Taklit Eder: Tüketicilerin gerçek alışveriş davranışlarını simüle ederek daha doğru veriler sağlar.

Konjoint analizi, pazarlama ve ürün geliştirme süreçlerinde tüketici tercihlerini anlamak ve ürünleri buna göre optimize etmek için güçlü bir araçtır.

UYGULAMA

Annelerin sürülebilir peynir tercihlerini incelemek üzerine yapılan araştırmada 402 anne ile görüşülmüştür. 6 adet peynir markası, 3 adet çeşit ve 50,95-120,95 aralığında olmak üzere bir fiyat skalası annelere gösterilmiştir.

DEMOGRAFİ



Yapılan araştırmada ağırlıklı olarak C1 grubu annelerle görüşülmüş, ortalama 35 yaş ve 2 çocuk skalaları yakalanmıştır. Her 10 anneden 7'si ev hanımıdır.

ÖNEM DERECELERİ

Marka, Çeşit ve Fiyat değişkenlerinin, tüketicilerin tercihlerine etkilerine baktığımızda;

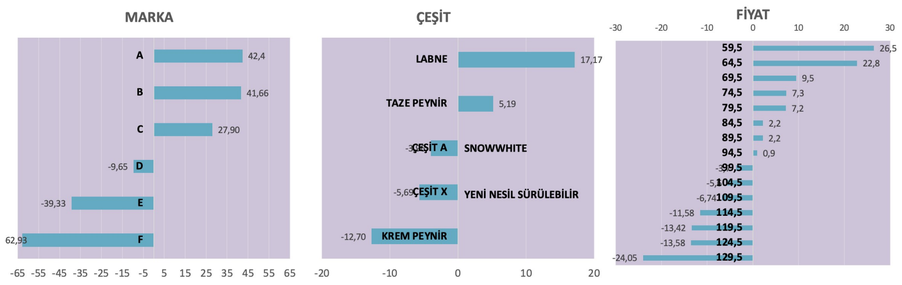
-Sürülebilir peynir tercihinde Marka en önemli kriterdir. %58 seviyesindeki önem derecesi sürülebilir peynir tercihinde tüketicinin marka bağlılığı eğiliminin yüksek olduğunu göstermektedir.

-Çeşit ise %25 önem derecesi ile markanın peşinden ikinci önemli kriterdir.

-Fiyat, marka ve çeşitten sonra gelen kriter olmuştur.

FAYDA DEĞERLERİ

Pozitif fayda değeri olan markalardan A ve B, C'nin önünde yer almaktadır. Görüşülen kitlede D, E ve F markalarının fayda değerleri çok düşüktür. Sürülebilir peynirler içerisinde Labne peynir tercihi açık ara öndedir. Taze Peynir de pozitif tercih değerine sahiptir. Öte yandan Krem Peynir tercih açısından en düşük fayda değerini almıştır. Çeşit A ve Çeşit X yeni ürün senaryolarıdır. Çeşit A ve Çeşit X negatif fayda değerine sahip olmakla birlikte tercihte Çeşit A görece olarak Çeşit X'in önünde yer almıştır. Genel tüketici eğilimi olarak beklenti, en iyi markanın en düşük fiyata alınmasıdır. En düşük fiyatın fayda değeri en yüksek pozitif değeri almışken, en yüksek fiyat da en düşük fayda değerini almıştır. 60,45 TL ilk kırılmanın olduğu fiyat seviyesidir. Bir sonraki kırılma 70,95 TL'de yaşanmaktadır. 90,45 TL sonrası ise tercih payı dramatik şekilde düşmektedir.



SONUÇ

Annelerin sürülebilir peynir tercihlerindeki kriterlerini anlamak amacıyla yapılan bu araştırma sonucunda annelerin büyük markalar olan A,B ve C markalarını daha çok tercih ettiği görülmüştür. Çeşitlerde Labne anneler için en çok tercih edilen üründür. Çeşit A ve X senaryoları negatif fayda değerine sahip olmuş, Snowwhite olarak adlandırılan Çeşit A diğer senaryoya göre daha başarılı olmuştur. 60,45 ve 70,95 aralığı annelerin satın alım skalasındaki markaların tercih etmesi gereken aralıktır.

KAYNAKÇA

[1] Hague, P. (2006). Market research in practice: A guide to the basics. Kogan Page Publishers.

[2] OpenAI. (2024). ChatGPT (GPT-4) [Yapay Zeka Modeli]. OpenAI. https://www.openai.com/chatgpt

[3] Rao, V. R. (2014). Applied conjoint analysis. Springer.

İSTATİSTİ KBÖLÜMÜ

UZUN SÜRELİ EKRAN KULLANIMININ İNSAN SAĞLIĞINA ETKİSİ

BAHATTİNCAN SEZER 18023059

Danışman:Doç.Dr.Fatma Sevinç KURNAZ

ÖZET

Günümüzde teknolojinin yaygınlaşmasıyla birlikte, bireyler ekran başında geçirdikleri süreyi artırmakta ve bu durumun sağlık üzerindeki potansiyel etkileri giderek daha fazla ilgi çekmektedir. Bu çalışma, uzun süreli ekran kullanımının insan sağlığına olan etkilerini anlamak amacıyla yapılmıştır. Araştırmanın odak noktası, teknolojik cihazların uzun süreli kullanımının fiziksel ve psikolojik sağlık üzerindeki olası sonuçlarını belirlemektir. Araştırmalarda sağlık konusundan daha az bahsedilirken yaşanılan bu sorunların kaç kişi tarafından hissedildiği ve kaç kişinin bunlara maruz kaldığını tespit etmek için istatistiki veriler kullanılarak açıklama yapılmıştır.

LİTERATÜR TARAMASI

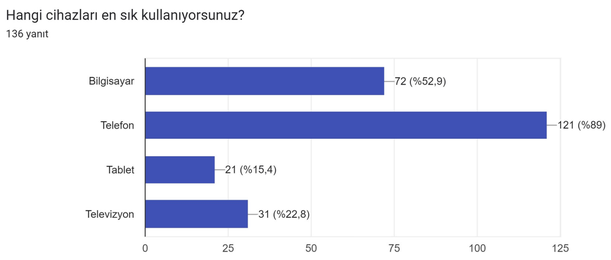
Burçin ŞENEL (2021) yayımladığı bir makalede yaygın olarak kullanılan teknolojik cihazların ve özellikle bilgisayar gibi sabit kullanılan bir cihazın oldukça uzun sürelerde hareketsizlik durumunda sakat kalmaktan kalbin durmasına kadar uzanan geri dönüşü olmayan hastalıklara yol açma riski oluşturduğunu belirtmiştir. Uzun süre ekran başında kalmak göz yorgunluğuna, gözde kuruluğa, ağrıya, acımaya ve buna benzer rahatsızlıklara neden olduğunu da eklemiştir.

Parikshit Muley (2023) yapmış olduğu araştırmada uzun süreli ekran kullanımında baş ağrısı, yanma hissi, kaşıntılı ve kuru gözler, ışığa duyarlılık, yorgunluk, konsantrasyon eksikliği, kaygı, uyku sorunları gibi şikayetler, ortalama 10 saat ekran süresine sahip bireyler arasında yaygın olduğunu raporlamıştır.

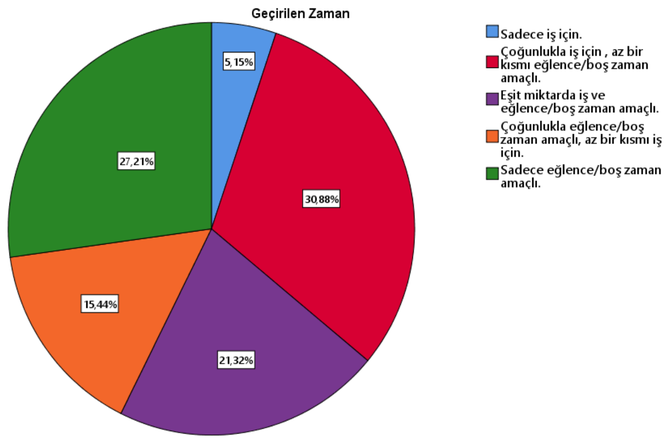
1.GİRİŞ

Günümüzde teknolojinin hızla gelişmesiyle birlikte, insanların günlük yaşamlarında ekran tabanlı cihazlara maruz kalma süreleri de artmaktadır. Bilgisayarlar, akıllı telefonlar, tabletler ve televizyonlar gibi cihazlar, iş, iletişim, eğlence ve bilgi edinme gibi birçok amaç için yaygın olarak kullanılmaktadır. Ancak, bu artan ekran kullanımının insan sağlığı üzerindeki etkileri giderek daha fazla araştırılmaktadır. Göz rahatsızlıkları, uyku kalitesinin azalması, fiziksel rahatsızlıklar ve duygusal durumda değişiklikler gibi pek çok olumsuz etki, uzun süreli ekran kullanımıyla ilişkilendirilmektedir.

2.Açıklayıcı İstatistikler



Grafikte görüldüğü üzere Telefonların **(%89)** en yaygın kullanılan cihaz olduğu, ardından Bilgisayarların **(%52,9)** geldiği ve Televizyon **(%22,8)** ile Tabletlerin**(%15,4)** daha az tercih edildiği anlaşılmaktadır.



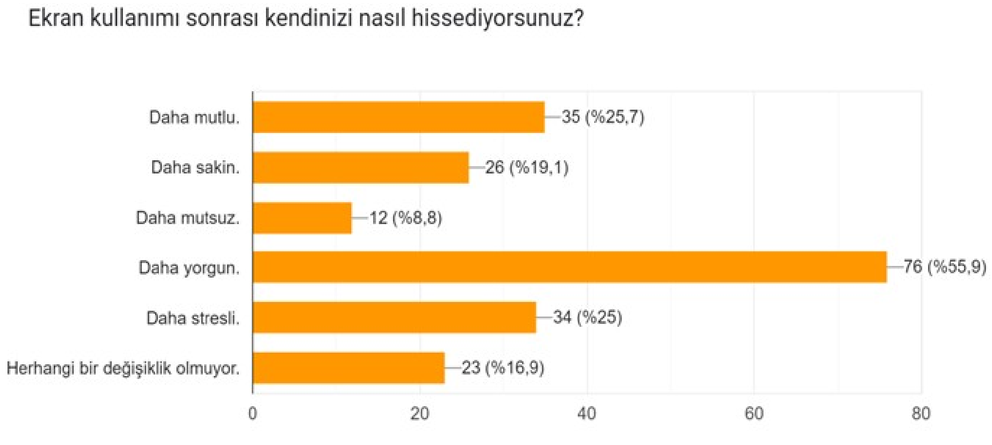
Grafik sonuçlarına göre, katılımcıların **%36,03**'ü elektronik cihazları ağırlıklı olarak iş amaçlı kullanmaktadır. Katılımcıların **%42,65**'i ise bu cihazları çoğunlukla eğlence amaçlı Kullanmaktadır. Elektronik cihazları eşit miktarda iş ve eğlence amaçlı kullanan katılımcıların oranı ise **%21,32**'dir.

OrtalamaSaat * Yorgunluk

Count

		Yorgunluk		
		Hayır	Evet	Total
OrtalamaSaat	1-2 saat	15	0	15
	2-6 saat	33	22	55
	6-9 saat	9	28	37
	9+ saat	3	26	29
Total		60	76	136

Yorgunluk ve ekran süresi verilerini çarpraz tablo yaptığımızda ekran kullanımı sonrası yorgun olan kişiler 6 saat ve 6 saatten fazla ekran süresi olan kişiler olduğu gözlemlenmektedir. .



Ekran kullanımının ardından insan üzerindeki psikolojik etkiler sorulduğunda, katılımcıların büyük bir kısmı kendini daha yorgun **(%55,9)** ve stresli **(%25)** hissettiğini belirtmiştir. Ayrıca, ekran kullanımı sonrası kendini daha sakin **(%19,1)** ve daha mutlu **(%25,7)** hissedenlerin oranları da dikkate değer düzeyde yüksektir.

Göz Rahatsızlığı

	Frekans	Yüzde
Evet , hissediyorum.	84	61,8
Hayır, hissetmiyorum.	52	38,2

Anketimize katılan 84 kişi**(%61,8)** ekran kullanımı sonrasında gözlerinde rahatsızlık hissettiğini belirtirken, 52 kişi **(%38,2)** ise ekran kullanımı sonrasında herhangi bir rahatsızlık hissetmediğini ifade etmiştir

Uyku Kalitesi

	Frekans	Yüzde
Evet , hissediyorum.	75	55,1
Hayır, hissetmiyorum.	61	44,9

Anket sonucunda, 75 kişi **(%55,1)** uzun süreli ekran kullanımı sonrasında uyku kalitesinde azalma olduğunu belirtirken, 61 kişi **(%44,9)** uyku kalitesinde herhangi bir değişiklik olmadığını ifade etmistir.

Fiziksel Rahatsizlik

	Frekans	Yüzde
Evet , hissediyorum.	57	41,9
Hayır, hissetmiyorum.	27	19,9
Bazen hissediyorum, bazen hissetmiyorum	52	38,2

Ankete katılanlardan 57 katılımcı **(%41,9)** ekran kullanımı sonrasında bel ağrısı, boyun ağrısı gibi fiziksel rahatsızlık hissettiklerini belirtmiştir. 27 katılımcı **(%19,9)** fiziksel bir rahatsızlık hissetmediğini ifade ederken, 52 katılımcı **(%38,2)** ise bazı durumlarda fiziksel rahatsızlık hissettik --lerini bazen ise hissetmediklerini belirtmiştir.

Aktivite Engeli

	Frekans	Yüzde
Evet , vaktim kalmıyor.	46	33,8
Hayır , vakit ayırabiliyorum.	90	66,2
Total	136	100,0

Katılımcılarımızdan 90 kişi **(%66,2)** ekran kullanımı nedeniyle dışarıdaki aktivitelere katılmada herhangi bir kısıtlama hissetmediğini belirtirken 46 kişi **(%33,8)** bu aktivitelere vakit ayıramadığını belirtmiştir.

3.Korelasyon Analizi

Korelasyon

		OrtalamaSaat	GozRahatsizlik	UykuKalitesi	FizikselRahatsizlik	AktiviteEngeli
OrtalamaSaat	Pearson Correlation	1	,427**	-,343**	,436**	-,460**
	Sig. (1-tailed)		,000	,000	,000	,000
	N	136	136	136	136	136

Yaptığımız korelasyon testinde önceliklik olarak p değerlerine (Sig.) baktığımızda ilişkilerin anlamlı olduğunu görebilmekteyiz.Korelasyon katsayılarına baktığımızda ise sırasıyla ortalama ekran süresi ile göz rahatsızlığı arasında orta düzeyde pozitif ilişki**(0,427**) bulunmaktadır. Bu, günlük ekran süresi arttıkça göz rahatsızlığının arttığını gösterir.Sonrasında ise ortalama ekran süresi ile uyku kalitesi arasındaki ilişkiye baktığımızda orta düzeyde negatif ilişki**(0,343)** olduğunu görmekteyiz.Buradan çıkarılabilecek sonuç ise ekran kullanımının süresinin artması uyku kalitesini azalttığını söyleyebiliriz.Ortalama ekran süresi ile fiziksel rahatsızlık arasındaki ilişki orta düzeyde pozitif**(0,436)** ilişkidir.Bu , günlük ekran süresi arttıkça fiziksel rahatsızlık hissinin arttığını gösterir.Ortalama ekran süresi ile aktivite engeli arasındaki ilişkide de orta düzeyde negatif**(0,460)** ilişki olduğunu görmekteyiz.Bu durum ekran süresinin sosyal aktivitelerle ters orantılı olduğunu veya ekran süresinin bazı bireyler için sosyal aktiviteleri artırıcı bir rol oynayabileceğini söylemek mümkün olabilir.

4.SONUÇ

Bu araştırmada, anket verileri kullanılarak ortalama ekran süresi ile çeşitli sağlık ve yaşam kalitesi faktörleri arasındaki ilişkiler incelenmiştir. Günümüz dünyasında teknolojik cihazların yaygınlaşması, hem iş hem de eğlence amaçlı olarak hayatımızın ayrılmaz bir parçası haline gelmiştir. Araştırmalardan elde edilen bulgulara göre, bazı bireyler bu cihazları yoğun bir şekilde kullanırken, diğerleri ise kullanım sürelerini yalnızca birkaç saat ile sınırlamıştır.

Anket ve testlerin sonuçları incelendiğinde, uzun süre boyunca ekranlara maruz kalmanın, insanların büyük bir bölümünde göz sağlığına olumsuz etki ettiği gözlemlenmektedir. Bu olumsuz etkilerin yanı sıra, ekranlara uzun süre maruz kalmanın uyku kalitesini de düşürdüğü söylenebilir. Bu nedenle, katılımcıların bir kısmının gece yatmadan önce ekran kullanımını tamamen kesmeyi tercih ettikleri gözlemlenmektedir.

Ekran kullanımının yol açtığı bir diğer olumsuz etki, fiziksel rahatsızlıklardır. Ekranlara yanlış açılardan bakma, bulunduğumuz ortamın konfor seviyesi gibi faktörler zaman zaman bel ağrısı, boyun ağrısı, baş ağrısı gibi fiziksel rahatsızlıklara neden olabilmektedir. Ekranlara bakma zorunluluğu olduğunda, bu tür etmenlere dikkat edilmesi ihmal edilebilir ve bu rahatsızlıklar saatler geçtikten sonra kendini gösterebilir.

Fiziksel sağlığımızın yanı sıra duygusal ve sosyal yönden de etkilenmekteyiz. Ekran kullanımı sonrasında, katılımcıların büyük çoğunluğu daha stresli ve daha yorgun hissetmektedir. Bu durumun birden fazla sebebi olabilir; ancak, ekran süresinin artmasıyla yorgunluk ve stresin de arttığı görülmektedir. İnsanlar, bu tür durumlardan kurtulmak için belirli aralıklarla mola vermeyi tercih ettiği gözlemlenmektedir.

Ekran kullanımının sosyal etkilerine değinmek gerekirse, ekran kullanımının fazla olması, insanların çoğunda sosyal aktivitelere katılımda herhangi bir kısıtlama hissetmediklerini göstermektedir. Bu durum, gelişen teknoloji ile birlikte sosyal ağların daha yaygın hale gelmesinden kaynaklanıyor olabilir. Ayrıca, insanlar ekran kullanımı alışkanlıklarını değiştirmek amacıyla aile ve arkadaşlarla daha fazla vakit geçirmeyi veya kitap okumak gibi aktivitelere yönelmeyi de tercih edebilmektedir.

Elde edilen bulgular, ekran kullanımının sağlık ve yaşam kalitesi üzerindeki potansiyel etkilerini vurgulamakta ve bu konuda daha bilinçli kullanım stratejileri geliştirilmesi gerektiğini işaret etmektedir

KAYNAKÇA

[1] Burçin ŞENEL , Mayıs 2021, ‘’*Uzun Süreli Ekran Maruziyeti Nelere Yol Açıyor?*’’ , YEŞİLAY , 1048

[2] George, D. and Mallery, M. (2010) *SPSS for Windows Step by Step: A Simple Guide and Reference,* Boston.

[3] Parikshit MULEY , July 2021, ‘’ *Excessive Screen Time and Effects on Human Health & Development* ‘’ , ResearchGate

İSTATİSTİK BÖLÜMÜ

K-MEANS VE HİYERARŞİK KÜMELEME ANALİZİ İLE MÜŞTERİ SEGMENTASYONU VE TABLEAU İLE GÖRSELLEŞTİRİLMESİ

Begüm YAĞCI 20023029

Danışman: Doç. Dr. Gülder KEMALBAY

ÖZET

Bu çalışmada, perakende sektöründe önemli bir yere sahip olan müşteri segmentasyonunu k-means ve hiyerarşik kümeleme analizi yöntemleri ile uygulamak ve elde edilen kümeleme sonuçlarını Tableau kullanarak bir dashboard üzerinden analiz etmek amaçlanmaktadır. Amacımız, müşterilerin satın alma geçmişine, aldıkları ürün miktarına, firmaya sağladıkları kazanca ve alışveriş sıklıklarına göre segmentlere ayrılmasıdır. Bu segmentasyon sayesinde müşteri tabanını oluşturmak, belirli müşteri kümelerini hedef alarak şirketin planlama yapmasına ve kampanya gibi girişimlerinde bu kümelemelerden yararlanmasına yardımcı olunması hedeflenmektedir.

VERİ

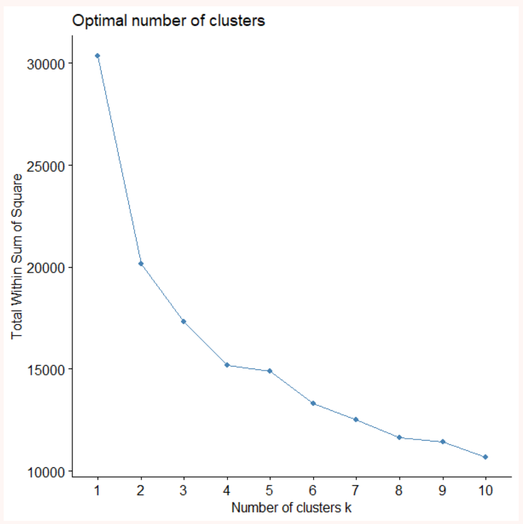
Bu araştırmada kullanılan veri bir perakende şirketinin 2022 yılının ilk 6 ayını kapsayan Müşteri ID, alışveriş tarihi, mağaza numarası, mağazanın il ve ilçesi, müşterinin cinsiyeti, alışverişteki indirim tutarı, nakit tutarı ve kredi kart tutarı, alınan ürünün miktarı, alışverişin gerçekleştiği satış kanalı, alınan ürünlerin ana kategorik grupları ve toplam alış veriş tutarından oluşmaktadır. Bu ham veriye ilk olarak aynı müşterinin birden fazla kez alışveriş yaptığını görüp her alışverişini toplayıp fiş sayısı olarak yeni bir değişken oluşturulmuştur. Alınan ürün kategorileri her bir kategori bir değişken olacak şekilde toplam 6 yeni değişken oluşmuştur. En son olarak Müşteri_ID değişkenini unix hale getirirken müşteri bazında alınan ürün kategorilerinin kaç kere alındığının, ürün miktarının, indirim tutarının, indirimin kaç kere uygulandığının , kredi kart tutarının ve alış verişin toplam tutarının toplamları alınmış diğer kategorik değişkenlerin ise modları alınarak başta 10000 gözlem içeren veri 2533 gözlem ve 19 değişken haline getirilmiştir.

Kategorik Değişkenler	Nümerik Değişkenler
Musteri_ID	Indirim_Kategorik_toplam
Mgz_No	Indirim_toplam
Il_mod	Kredi_kart_tutari_toplam
Ilce_mod	Miktar_Adet_toplam
Cinsiet_mod	Fis_sayisi_toplam
Satis_Kanali_mod	Toplam_Tutari_toplam
Nakit_tutari_mod	Kategori_Grup_A
	Kategori_Grup_B
	Kategori_Grup_C
	Kategori_Grup_D
	Kategori_Grup_E
	Kategori_Grup_F

KÜMELEME ANALİZİ

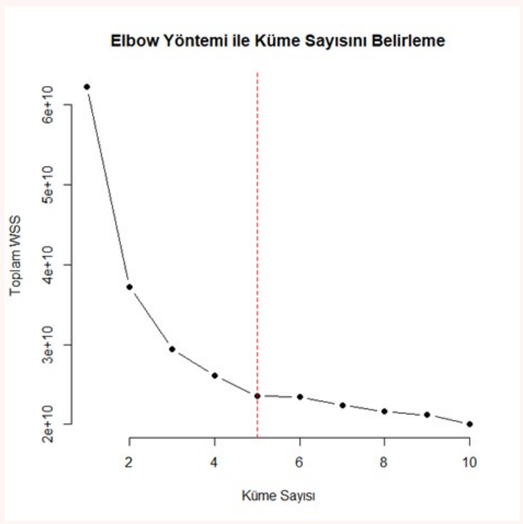
Nesnelerin temel özelliklerini dikkate alarak benzerliklerine göre gruplayan bir çok değişkenli istatistiksel analiz yöntemidir. Amaç birbirine benzeyen birimlerin aynı grupta olduğu, kümelerin kendi içlerinde homojen ama kendi aralarında ise heterojen yapıda olduğu sınıfları oluşturmak hedeflemektedir.
En uygun küme sayısının belirlenmesi için kmeans yöntemi ile birlikte 'wss (Within-Cluster Sum of Squares)' yöntemi kullanılır. Elbow (dirsek) yöntemi uygulanarak her bir küme sayısı için küme içi toplam kareler toplamına bakılır.

K-MEANS YÖNTEMİ



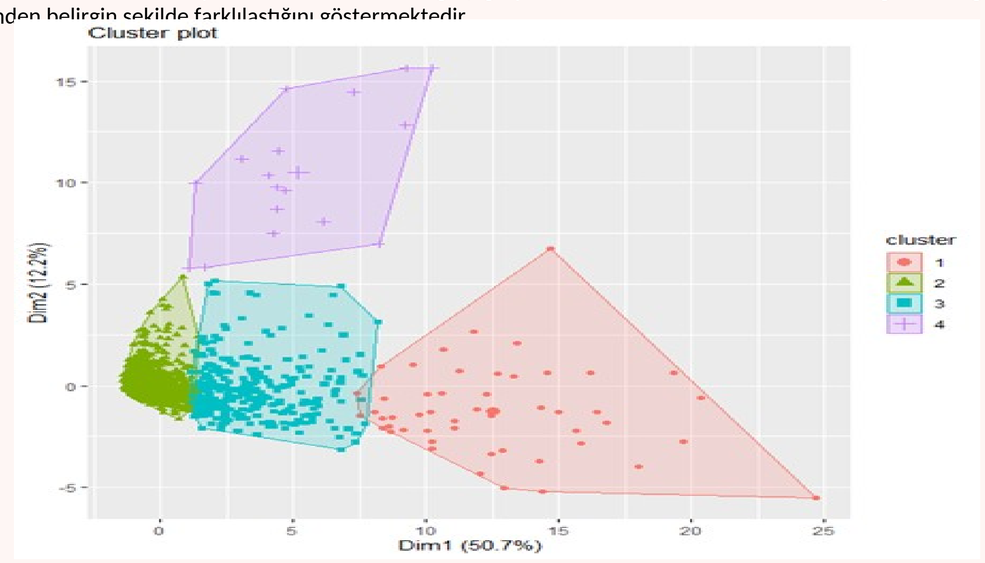
Yandaki grafikte wss'nin küme sayısına göre azalışını göstermektedir. Bu grafiği oluşturmak için verideki nümerik değişkenler üzerinden kmeans yöntemi uygulanmıştır.
Burada WSS'nin hızla azaldığı ve ardından yavaşladığı nokta k=4 olarak gözükmektedir.

K-PROTOTYPE YÖNTEMİ

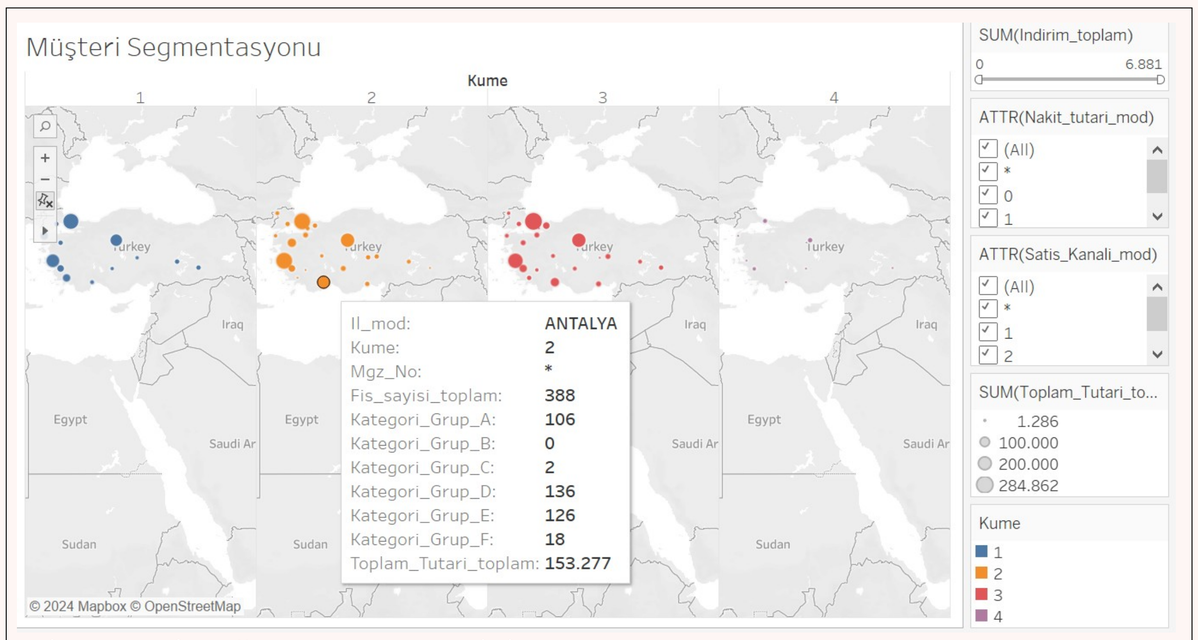


Yukarıdaki grafikte ise veri hem nümerik hem kategorik değişkenlere sahip olduğu için sadece nümerik değişkenleri almadan karma şekilde uygulanan 'k-prototypes' algoritması kullanılmıştır. Optimum küme sayısı olarak k=5 olduğu gözükmektedir.

Her iki yöntemle de bakıldığında küme sayısı 4 olarak belirlenmiştir. Kmean modeli veriye uygulanmıştır. Uygulanama sonuçlarını ve kümelerin ayrımını daha net görmek için aşağıdaki görsel oluşturulmuştur. Toplam varyansın %49.9'u gruplar arası farklardan kaynaklanmaktadır. Bu değer, gruplar arasındaki farkların önemli olduğunu ve grupların birbirinden belirgin şekilde farklılaştığını göstermektedir.

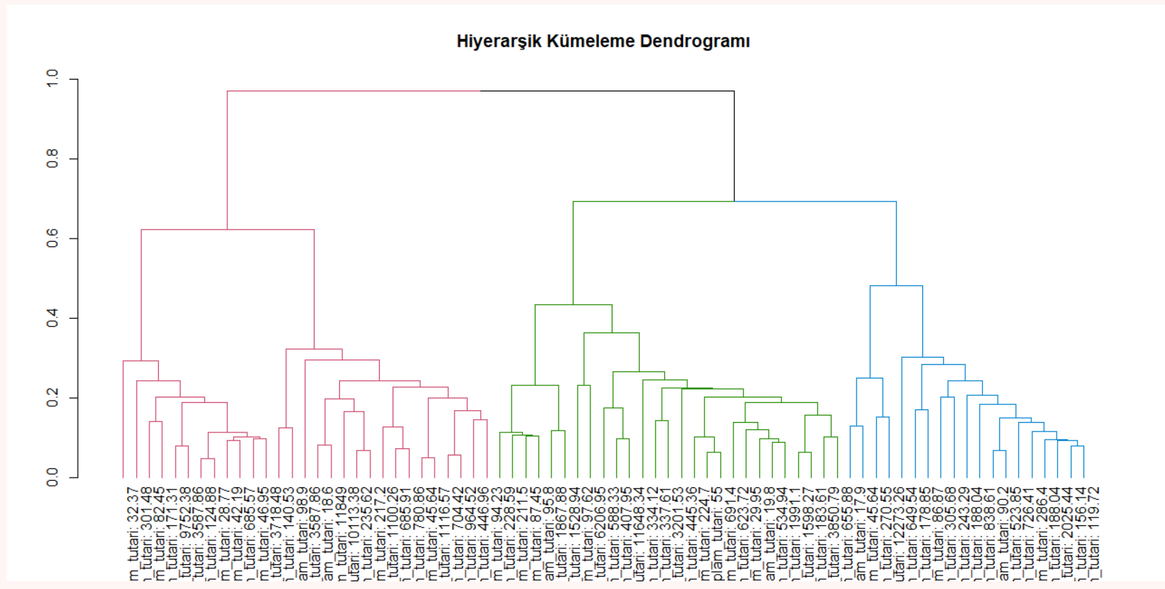


Kümeleme sonucu Tableau uygulamasında bir dashboard olarak hazırlandığında; 3. kümede bulunan müşterilerin daha sık alışveriş yaptığını, en çok indirimin bu kümede ve İstanbul ilinde uygulandığını, ve daha çok E kategori grubundaki ürünlerden aldıkları ortaya çıkmıştır. En büyük fark gösteren 4. kümede ise müşterilerin en az alışveriş yaptığı, indirim tutarının en düşük tutarda ve sayıda uygulandığı, online alışverişin yapılmadığını ve en çok C kategori grubundaki ürünlerden aldıkları sonucuna varılmıştır. 1. küme için indirimin yüksek tutarda uygulandığı ikinci küme olduğu ve 2. küme için ise A kategori grubundaki ürünlerden alışverişin daha fazla gerçekleştiği tek küme olduğu söylenilebilir.

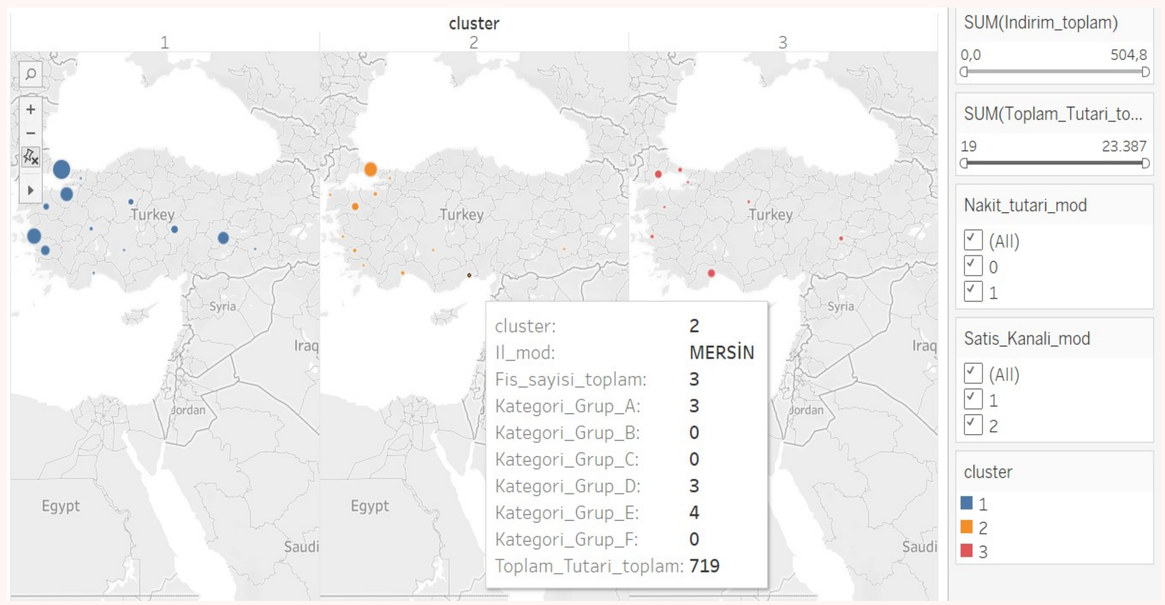


HİYERARŞİK KÜMELEME ANALİZİ

K-means yöntemindeki küme sayısının önceden belirlenmesi dezavantajını ortadan kaldırmak için ortaya çıkan hiyerarşik kümeleme analizi tümevarım ve tümdengelim olarak ikiye ayrılır. Bu veride tümdengelim yaklaşımı yani her aşamada uzaklık matrisine göre nesneler ana kümeden ayrılarak alt kümeler oluşturur.
Oluşturulacak dendrogramının düzgün çıkması için her bir değişkenin içinden rastgele %3'lük bir kısım alınmıştır. Veride hem kategorik hem de nümerik değişkenler bulunduğu için 'Gower' mesafe matrisi kullanılarak hiyerarşik kümeleme analizi yapılmıştır.



Yukarıdaki hiyerarşik kümeleme dendrogramında her alt kümenin ucuna alışverişin toplam tutar değeri getirilmiştir.



Hiyerarşik kümeleme dashboardında ise 2. kümede bulunan müşterilerin daha sık alışveriş yaptığı, en az indirim tutarının ve online alışverişin gerçekleşmediği grubun 3. kümede olduğunu, en fazla kazancın 1. kümede olduğunu buna bir sebep olarak en fazla indirimde aynı şekilde 1. kümede gerçekleştiği sonucuna varılmıştır.

KAYNAKÇA

[1] https://www.veribilimiokulu.com/hiyerarsik-kumeleme/
[2]https://medium.com/deep-learning-turkiye/k-means-algoritmas%C4%B1-b460620dd02a#:~:text=KMeans%20algoritmas%C4%B1%20bir%20unsupervised,Temel%20olarak%20etiketlenmemi%C5%9F%20verilerle%20ilgilenir.
[3] https://www.questionpro.com/blog/tr/kume-analizi-nedir-ve-nasil-kullanilir/#hierarchical_clustering

İSTATİSTİK BÖLÜMÜ

SÜRDÜRÜLEBİLİR ÜRÜNLERİN TERCİH FAKTÖRLERİNİN İNCELENMESİ

Begüm YILDIRIM 19023023

Danışman: Doç. Dr. Öyküm Esra YİĞİT

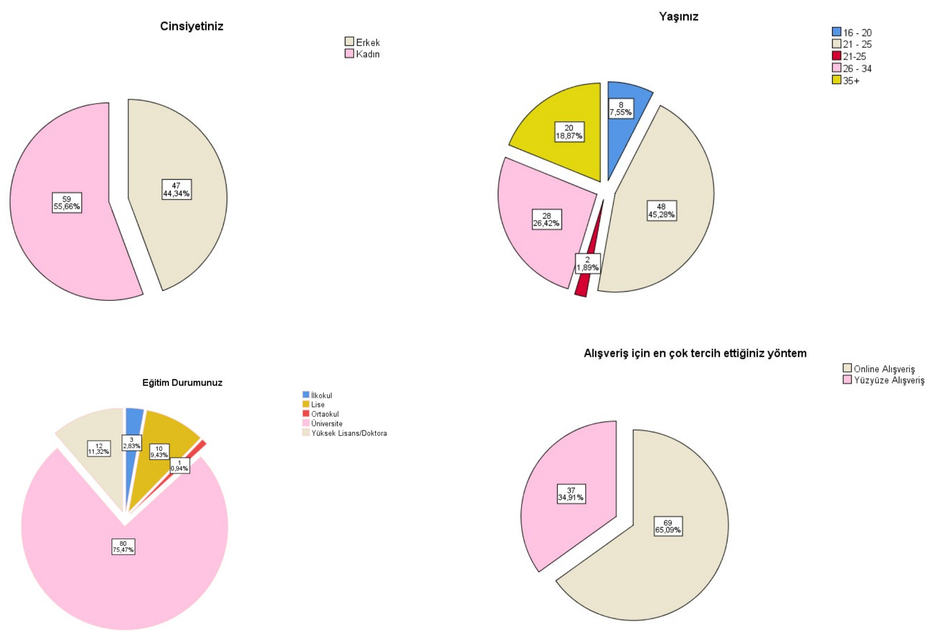
ÖZET

Son yıllarda küresel ısınmanın ve azalan yeryüzü kaynaklarının negatif tesirleri bireysel yaşantımızda ve bütün dünyada hissedilmektedir. Oluşan bu farkındalık sürdürülebilirlik teriminin öne çıkmasını sağlamaktadır. Sürdürülebilirlik ekolojik, ekonomik ve sosyal koşulları içinde barındırır, kaynakların etkin yönetilmesi sağlanamadığında kaynakların kullanımı düzenlenememektedir. Sürdürülebilirlik bir farkındalığı gerektirir. Hem toplumsal hem de bireysel olarak bir hedefi belirtir. Amaç bu hedefi yerine getirmektir. Yapılan bu araştırmada kişilerin cinsiyetlerinin, yaşlarının, eğitim durumlarının ve aylık gelir düzeyi gibi bağımsız değişkenlerin sürdürülebilir ürünleri tercih etme eğilimi üzerinde nasıl bir etkisi olduğu lojistik regresyon aracılığıyla incelenmiştir. Elde edilen sonuçlar tüketicilere yönelik sürdürülebilir ürünlerin pazarlanmasına yönelik yapılan çalışmalar için bir bakış açısı sunmaktadır. Bu bulgular, pazarlamacıların daha başarılı pazarlama kampanyaları oluştururken tüketici demografiklerini dikkate almalarının gerekliliğini göstermektedir.

Verilerin Açıklanması ve Analizi

Veriyi oluşturan anket sorularımız, katılımcıların satın alma alışkanlıklarını, çevre bilincini ve tercihlerini anlamak için önemli veriler sağlayacaktır. Bu çalışmamıza farklı yaş ve demografik özelliklere sahip 106 kişi katılmıştır. Bu veriler, farklı demografik özellikler ile satın alma alışkanlıkları ve çevre bilinci arasındaki ilişkiyi analiz etmemize yardımcı olur. Bireyin satın alma sıklığı, aylık harcaması gibi sorular, satın alma alışkanlıklarını belirlemek için önemli veriler sağlayacaktır. Aynı şekilde ürün tercihleri, satın alma yöntemleri gibi sorular da tüketici tercihlerini anlamada önemlidir. Asıl incelediğimiz sürdürülebilir ürünleri tercih etmek sorusunun hangi özelliklere bağlı olduğunu da bu anketle beraber incelemiş olduk. Çevre bilinci ve sürdürülebilirlik söz konusu olduğunda ambalaj kullanımına dikkat etmek, sürdürülebilir ürünleri tercih etmek, yerel mağazaların çevresel sürdürülebilirlik konusunda bilgi verip vermediği şeklinde sorular tüketici farkındalığını ve çevre davranışlarını ölçmemize destek olacaktır.

Veri hazırlama aşaması, verilerin daha anlaşılabilir olması için araştırılmış olduğu yerdir. Veriler temizlendi, dönüştürüldü ve entegre edildi. Aşağıda, verilerin bir parçası bulunmaktadır.



Yukarıda sırasıyla; cinsiyet, yaş, eğitim durumu ve alışveriş için en çok tercih edilen yöntem ile ilgili pasta grafikleri bulunmaktadır. Ankete katılan 106 kişinin 59'u kadın, 47'si erkektir. Yaş grupları arasında 21-25 yaş grubunun katılımcı oranının diğer yaş gruplarına oranla daha fazla olduğunu görürüz. Katılımcıların büyük çoğunluğu üniversite düzeyinde eğitim düzeyine sahiplerdir. Katılımcılara alışveriş için en çok tercih ettikleri yöntem sorulduğunda %65,09'u online alışveriş tercih ederken %34,91'i yüzyüze alışverişi tercih ettiği yanıtını almıştır.

HİPOTEZ TESTLERİ

Ki-Kare Testleri

	değer	sd	p(2-taraflı)
Pearson Ki-Kare	5,135	4	0,274
Olabilirlik Oranları	8,151	4	0,086
Geçerli Örneklem Sayısı	106		

Ki Kare testi sonucuna bakılırsa anlamlılık değeri 0,274 çıkmıştır. Bu durumda H0 hipotezi %5 anlamlılık düzeyinde reddedilemez, buna göre eğitim seviyesi ve sürdürülebilir ürünleri tercih etme eğilimi arasında anlamlı bir ilişki olmadığına karar verilir.

Korelasyon

			Ailedeki aylık gelir düzeyiniz nedir?	Ailedeki kişi sayısı
Spearman's rho	Aileddeki aylık gelir düzeyiniz nedir?	Korelasyon katsayısı	1,000	-0,284
		P		0,003
		N	106	106
	Ailedeki kişi sayısı	Korelasyon katsayısı	-0,284	1,000
		P	0,003	
		N	106	106

Spearman korelasyon katsayısı (-0.284), iki değişken arasında güçlü bir negatif ilişki olduğunu göstermektedir. Anlamlılık değerini incelediğimizde p<0.01 olduğunu gördük yani bu ilişki anlamlıdır. Bu negatif korelasyon, ailedeki kişi sayısının artmasıyla birlikte aile gelir düzeyinin azaldığını gösterir. Yani, daha fazla kişinin bakımı için ailenin geliri kişi başına azalabilir veya ailede daha fazla gelire ihtiyaç duyan kişiler olabilir. Sonuç olarak, bu çıktı, ailedeki kişi sayısı ile aile gelir düzeyi arasında anlamlı ve negatif bir ilişki olduğunu göstermektedir.

Ki-Kare Testleri

	değer	sd	p(2-taraflı)
Pearson Ki-Kare	4,340	4	0,362
Olabilirlik Oranları	4,551	4	0,337
Geçerli Örneklem Sayısı	106		

Ki Kare testi sonucuna bakıldığında anlamlılık değeri 0,362 çıkmıştır. Bu durumda H0 hipotezi %5 anlamlılık düzeyinde reddedilemez ve yaş ve Sürdürülebilir ürünleri tercih etme eğilimi arasında ilişki olmadığı görülür.

LOJİSTİK REGRESYON

Regresyon yöntemleri bir bağımlı değişken ile bir veya birden fazla bağımsız değişken arasındaki ilişkiyi açıklamada kullanılan çok değişkenli tekniklerdendir. En yaygın örneği bağımlı (sonuç) değişkeninin sürekli olduğu doğrusal regresyon modelidir. Bir regresyon modeli olan lojistik regresyonu lineer regresyondan ayıran farklılık sonuç değişkenin iki veya ikiden fazla kategoriye sahip olmasıdır. Bu fark hem varsayımlarda hem de model seçiminde kendini göstermektedir.

Modelin Özeti

Adım	-2 Log Olabilirlik	Cox & Snell R Kare	Nagelkerke R Kare
1	86,382	0,225	0,342

Bu tablo Cox & Snell R kare ve Nagelkerke R Kare değerlerini içerir. Buradaki değerler modelin bağımlı değişkendeki varyansının %34.2'sinin açıklandığını gösterir.

Sınıflandırma Tablosu

		Tahmin edilen		
		Sürdürebilir ürünleri tercih etmeye çalışır mısınız?		Yüzde düzeltme
		Evet	Hayır	
Sürdürebilir ürünleri tercih etmeye çalışır mısınız?	Evet	78	4	95,1
	Hayır	15	9	37,5
Genel yüzde				82,1

Bu tablo Sürdürülebilir ürünleri tercih etmenin meydana gelme olasılığını tahmin eder. Modelin toplam doğruluk oranı %82,1 olarak hesaplanmıştır. Bu sonuçlar modelin Evet diyenleri doğru şekilde tahmin edebildiğini gösterir. Modelde yalnızca sabit bir terimin olduğu görülmüştür.

Model Katsayılarının Omnibus Testi

	Ki-kare	sd	p
Adım	27,018	12	0,008
Blok	27,018	12	0,008
Model	27,018	12	0,008

Eşitlikteki Değerler

		B	S.E.	Wald	sd	p	Exp(B)	%95 Güven aralığı	
Adım	Cinsiyetiniz	0,475	0,556	0,73	1	0,36	1,608	0,54	4,779
	Yaşınız			0,61	4	0,96			
	Yaşınız(1)	0,371	1,216	0,09	1	0,76	1,449	0,13	15,70
	Yaşınız(2)	19,85	7824,966	0,00	1	0,99	419213179,05	0,00	
	Yaşınız(3)	0,385	1,280	0,09	1	0,76	1,470	0,12	18,06
	Yaşınız(4)	-0,30	1,442	0,04	1	0,83	0,737	0,04	12,43
	Eğitim Durumunuz			0,98	4	0,91			
	Eğitim Durumunuz(1)	-0,222	21903,13	0,00	1	1,00	0,801	0,00	
	Eğitim Durumunuz(2)	0,290	44342,36	0,00	1	1,00	1,337	0,00	
	Eğitim Durumunuz(3)	19,73	18728,86	0,00	1	0,99	373788643,82	0,00	
	Eğitim Durumunuz(4)	20,52	18728,86	0,00	1	0,99	817734664,0	0,00	
	Ailedeki aylık gelir düzeyiniz nedir?			6,76	3	0,08			
	Ailedeki aylık gelir düzeyiniz nedir?(1)	-35,8	11066,17	0,00	1	0,99	0,000	0,00	
	Ailedeki aylık gelir düzeyiniz nedir?(2)	1,827	1,014	3,247	1	0,072	6,215	0,85	45,34
	Ailedeki aylık gelir düzeyiniz nedir?(3)	0,010	0,930	0,000	1	0,992	1,010	0,163	6,247
	Sabit	-21,6	18728,863	0,000	1	0,999	0,000		

Wald sütunu bağımsız değişkenlerin her biri için istatistiksel anlamlılığı gösterir. Sig. sütununu incelediğimizde sonuçların modele önemli ölçüde katkı sağlamadığını görürüz. Sürdürülebilir ürünleri tercih etme yönelimi üzerindeki en anlamlı sonucu veren değişkenin ailedeki aylık gelir düzeyiniz nedir sorusunun olduğunu görmekteyiz. Tablo sürdürülebilirliğin ('evet' kategorisi) erkeklerde kadınlara göre 1,608 kat daha fazla olduğunu göstermektedir.
Yaşın, cinsiyetin, eğitim durumunun ve ailedeki aylık gelir düzeyinin sürdürülebilir ürünleri tercih etme eğilimi üzerindeki etkilerini belirlemek için bir lojistik regresyon gerçekleştirildi. Model istatistiksel olarak anlamlıydı, $\chi 2\ (4) = 27,018$, $p < 0,0005$. Model Sürdürülebilir ürünlerdeki varyansın %34.2'sini açıklandığını gösterir. Kişilerin %82,1'ini doğru şekilde sınıflandırdı.







SONUÇ

Bu çalışmanın amacı tüketici davranışını etkileyen faktörler arasındaki ilişkiyi sürdürülebilir ürünlere olan bakış açısıyla incelemektir. Test ettiğimiz hipotez sonuçlarına göre, yaş değişkeni ve eğitim seviyesinin sürdürülebilir ürünleri tercih etme eğilimi arasında anlamlı bir ilişkinin bulunmadığı sonucuna varılmıştır. Baktığımız korelasyona göre ailedeki kişi sayısı arttıkça aile gelir düzeylerinin azaldığını ve bunun da tüketici davranışını etkileyebileceğini göstermektedir. Aile geliri ile aile büyüklüğü arasındaki ilişkinin anlaşılması, tüketicinin satın alma gücünü belirleyen faktörlerin karmaşıklığını ortaya çıkarmaktadır. Lojistik regresyon analizine bakıldığında sürdürülebilir ürünleri tercih etme eğilimi üzerindeki test sonucu olarak bağımsız faktörleri incelediğimizde en anlamlı etkiyi ailedeki aylık gelir düzeyi değişkenin verdiği sonucuna vardık. Tüm sonuçları incelediğimizde sürdürülebilir ürünleri tercih etmenin birçok demografik faktörle ilişkisi olduğunu gözlemlemekteyiz.

KAYNAKÇA

[1] https://sites.education.miami.edu/statsu/wp-content/uploads/sites/4/2020/07/Logistic-Regression-Webinar.pdf
[2] David W..Hosmer,Lemeshow,S.,&Rodney X.. Sturdivant. (2000).Applied logistic regression.New York:Wiley
[3] https://stock.adobe.com/tr/search/images

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

İÇERİK BAZLI FİLTRELEME SİSTEMİ İLE FİLM ÖNERİ UYGULAMASI

Hazırlayan: Berker Yalçın - 17023041

Danışman: Prof. Dr. Fatma NOYAN TEKELİ

ÖZET

Günümüzde dijital medya platformlarının iyice çoğalması ve yaygınlaşması ile üretim de doğru orantılı şekilde artmıştır. Bu nedenle film izlemek isteyen bir kullanıcı seçimini binlerce film arasından yapmak zorunda kalıyor ve bu durumda seçim yapması zorlaşıyor. Buna çözüm olarak içerik bazlı filtreleme sistemi ile kullanıcının daha önce izleyip beğendiği filme içerik olarak benzer öneriler yapmak mümkün. Bu öneri uygulamaları kullanıcının daha önceden izlediği filmleri analiz ederek bu kullanıcıya benzer özelliklerde başka filmler önermektedir. Bunu yaparken de içerik bazlı filtreleme sistemi kullanılmaktadır. Bu makalede öneri sistemlerinin nasıl çalıştığı incelenip benzer bir uygulama yapılmıştır. Bunu yapmaktaki amacımız ise öneri sistemlerinin tüketeceğimiz içerik seçiminde bizim işimizi nasıl daha çok kolaylaştırdığını ve kullanıcı deneyimini nasıl iyileştirebileceğini ortaya koymaktır.

Öneri sistemleri nedir?

Öneri sistemleri ya da tavsiye sistemleri kullanıcının özellikleri, geçmiş tercihleri ve diğer kullan[1]ıcların tercihleri gibi verileri analiz ederek kullanıcı için en uygun ürünü bulmak için kullanılan algoritmalardır. Öneri sistemlerinin genellikle kullanıldığı alanlar filmler, müzikler, kitaplar, haberler, sosyal medya içerikleri, bilimsel içerikler ve satış sitelerindeki önerilen ürünlerdir.

İçerik Bazlı Filtreleme Sistemi

İçerik bazlı filtreleme sisteminde ürünün açıklaması ve öneril yapılan kullanıcının profillerinin analizi ile öneri yapılmaktadır. Bu sistemde öğeleri tanımlamak için anahtar kelimeler kullanılır ve öneri yapılacak kullanıcının beğendiği öğenin türünü belirtmek için kullanıcı profili oluşturulur. Kullanıcıya önerilmek için belirlenen aday ürünler kullanıcı tarafından daha önce değerlendirilmiş ürünler ile karşılaştırılır ve en iyi eşleşen ürün kullanıcıya önerilir

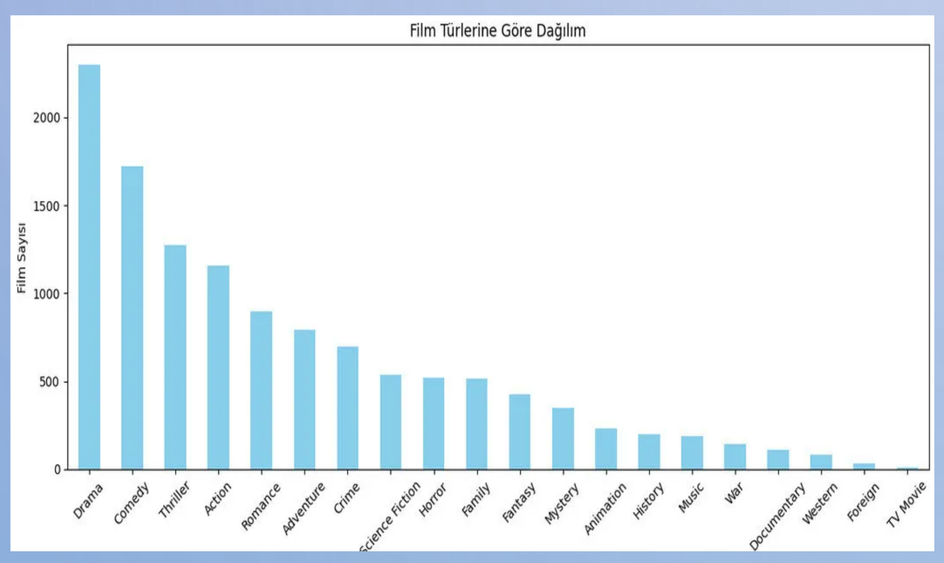
Veri Seti

The Movie Database tarafından sağlanan TMDB 5000 veri seti 4803 adet filmin bilgilerini içermektedir.

Veri setinde budget, genres, homepage, id, keywords, original_language, original_title, overview, popularity, production_companies, production_countries, release_date, revenue, runtime, spoken_languages, status, tagline, title, vote_average, vote_count değişkenleri olmak üzere tam 20 değişken bulunmakta. Bu değişkenlerden budget filmin bütçesini, genres filmin türü ve alt türlerini, homepage bu filmin resmi web sitesini, id her bir filmin kendine has kimlik numarasını, keywords ise filmi tanımlayan önemli kelimeleri, original _language filmin orijinalinde konuşulan dili, original_title ve title filmin ismini, overview filmin konusunu anlatan kısa bir özeti, popularity her film için TMDB tarafından oluşturulmuş temsili bir popülerlik skorunu, production_companies filmin yapımcı şirketini, production_countries yapımcı şirketin hangi ülkeye ait olduğu, release_date filmin yayınlanma tarihi, revenue gişede yaptığı kazancı, runtime filmin süresi, spoken_languages filmde konuşulan dilleri, status ise filmin durumunu yani yapım aşamasında mı yoksa yayınlanıp yayınlanmadığı gibi bilgileri, tagline filmin sloganını, vote_average filmin izleyicilerden aldığı ortalama puan ve son olarak vote_count ise film için kullanılan oy sayısını temsil etmektedir.

```
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   budget                4803 non-null   int64
 1   genres                4803 non-null   object
 2   homepage              1712 non-null   object
 3   id                    4803 non-null   int64
 4   keywords              4803 non-null   object
 5   original_language     4803 non-null   object
 6   original_title        4803 non-null   object
 7   overview              4800 non-null   object
 8   popularity            4803 non-null   float64
 9   production_companies  4803 non-null   object
 10  production_countries  4803 non-null   object
 11  release_date          4802 non-null   object
 12  revenue               4803 non-null   int64
 13  runtime               4801 non-null   float64
 14  spoken_languages      4803 non-null   object
 15  status                4803 non-null   object
 16  tagline               3959 non-null   object
 17  title                 4803 non-null   object
 18  vote_average          4803 non-null   float64
 19  vote_count            4803 non-null   int64
dtypes: float64(3), int64(4), object(13)
memory usage: 750.6+ KB
```

Veride yer alan filmleri türlerine göre ayırıp bar grafiği olarak görselleştirdiğimizde ise sonuç aşağıdadır.



Öneri Sisteminin Uygulanması

İlk olarak python'a kullanılacak kütüphaneler olan numpy ve pandas 'import' komudu ile ekleniyor ve daha sonra 'as' özelliği il bunlara birer kısaltma veriliyor. Daha sonra pandas özelliği olan 'read_csv' komudu ile iki farklı csvde olan veri setini cretdits v movies değişkenlerine DataFrame olarak atanıyor.

```
import numpy as np
import pandas as pd

credits = pd.read_csv("G:/My Drive/tez/veri seti/tmdb5000/credits.csv")
movies = pd.read_csv("G:/My Drive/tez/veri seti/tmdb5000/movies.csv")
```

İki veriyi de inceledikten sonra fark ediliyor ki filmlerin kimlik numaraları movies veri setinde 'id' ile isimlendirilirken, credits veri setinde ise 'movie_id' olarak isimlendiriliyor. Bu iki veri setini filmlerin kimlik numaraları üzerinden birleştirebilmek için credits'deki 'movie_id' değişkeninin ismini movies veri setindeki değişkenle aynı, yani 'id' olarak değiştiriliyor

```
credits = credits.rename(columns={"movie_id": "id"})
df_merge = pd.merge(credits, movies, left_on=['title','id'], right_on=['title','id'])
```

Bu adımda TfidfVectorizer sınıfından bir örnek oluşturuluyor. TF-IDF, bir metindeki bir terimin önemini belirlemek için kullanılan bir istatistiksel ölçüdür. Terim frekansı (Term Frequency - TF), bir terimin belirli bir belgedeki göreceli sıklığını temsil eder. Belge frekansı tersine çevrilmiş (Inverse Document Frequency - IDF), bir terimin kaç belgede geçtiğini ve bu terimin yaygınlığını temsil eder. Yani, bir terim belgede sıkça geçiyorsa ve diğer belgelerde nadiren geçiyorsa, bu terimin önemi daha yüksek olacaktır.

```
tfidf_vectorizer = TfidfVectorizer()
overview_tfidf_matrix = tfidf_vectorizer.fit_transform(df_merge["overview"].values.astype('U'))
```

Gözlemler arasındaki benzerlik hesaplandıktan sonra kullanıcıdan bir film adı girmesi isteniyor. Kullanıcının girdiği film ismi film_adi değişkeninde tutulduktan sonra data frame içerisinde bu filmin olup olmadığını kontrol ediliyor ve eğer varsa bu filmin adı film değişkenine atanıyor.

```
film_adi = input("Filmin Adını Giriniz:")
film = df_merge[df_merge["title"] == film_adi]
```

Else komutunun içine geçtikten sonra ilk önce kullanıcı tarafından verilen filmin türleri film_turu değişkeni içerisine atanıyor. Ardından İlgili filmdeki "overview" sütunundaki metni içeren bir NumPy dizisi alır ve astype('U') ifadesiyle bu metinleri Unicod olarak işlemek için dönüştürür. Unicode formatına dönüştürülme nedeni ise bu formatın çok sık kullanılıyor olmasıdır. tfidf_vectorizer.transform() fonksiyonu, TfidfVectorizer sınıfının bir metodudur ve belirli bir metni , yani burada overview sütununu, TF-IDF matrisine dönüştürmek için kullanılır. Bu durumda, film["overview"].values.astype('U') ifadesi, seçilen filmin özetini içeren metni temsil eden bir dizidir. Filmin overview verisini TF_IDF matrisine dönüştürüldükten sonra bunu film_tfidf değişkenine atıyor.

Bu işlemden sonra sırada kullanıcının girdiği filmin diğer filmlerle olan benzerlik skorlarını hesaplamaya geçiliyor.

```
if film.empty:
    print("Belirtilen film bulunamadı.")
else:
    film_turu = film["genres"].values[0]
    film_tfidf = tfidf_vectorizer.transform(film["overview"].values.astype('U'))
    benzerlik_skorlari = cosine_similarity(film_tfidf, overview_tfidf_matrix)
    oneriler = df_merge.iloc[np.argsort(benzerlik_skorlari[0])[-6:-1]][::-1]

    print("Seçilen film: ",film_adi)
    film_turu_str = ", ".join([genre["name"] for genre in eval(film_turu)])
    print("Film Türü:", film_turu_str,"\n")
    print("Benzer filmler:")
    print(oneriler["title"])
```

Uygulama çalıştıktan sonra kullanıcı Interstellar filmini girdi olarak kullanıyor. Ardından Seçilen filmin adı ve türü ekrana yazdırılıyor ve bu filmin listede yer aldığı onaylanmış oluyor. Daha sonra ise veri setinde bu filme overview değişkeni olarak en çok benzeyen beş film sırasıyla id'leri ile birlikte yazdırılıyor. Kullanıcının girdiği film olan Interstellar yok olmakta olan bir dünyada kurtuluşumuzun arandığı uzay temalı bir bilim kurgu ve macera filmi. Önerilen filmlere baktığımızda ise yine uzay ve uzaylı temalı, günümüz dünyasının yok olmakta olduğu bilim kurgu filmlerinin yer aldığı gözlemleniyor

```
Filmin Adını Giriniz:Interstellar
Seçilen film:  Interstellar
Film Türü: Adventure, Drama, Science Fiction

Benzer filmler:
1709        Space Pirate Captain Harlock
220                         Prometheus
300                   Starship Troopers
4176    Battle for the Planet of the Apes
634                          The Matrix
Name: title, dtype: object

Process finished with exit code 0
```

```
Filmin Adını Giriniz:Batman Begins
Seçilen film:  Batman Begins
Film Türü: Action, Crime, Drama

Benzer filmler:
299             Batman Forever
1359                    Batman
65             The Dark Knight
3        The Dark Knight Rises
428             Batman Returns
Name: title, dtype: object
```

İSTATİSTİK BÖLÜMÜ

ARAÇ (KASKO) SAHIBI MÜŞTERILERIN RISK PROFILLERININ MODELLENMESI VE TALEP TUTARLARININ ÖNGÖRÜLMESI: MAKINE ÖĞRENMESI YAKLAŞIM

Berna Nur ÇETİNKAYA 19023074

Danışman: Assoc. Prof. Gülder KEMALBAY

ÖZET

Bu çalışma, araç sahibi müşterilerin risk profillerinin modellenmesi ve talep tutarlarının öngörülmesi amacıyla makine öğrenmesi tekniklerinin kullanılmasını ele almaktadır. Veri seti olarak, araç sahibi müşterilerin poliçe bilgilerini içeren bir veri kümesi kullanılmıştır. Veri seti, müşterilerin demografik bilgilerini, poliçe detaylarını, talep tutarlarını ve diğer ilgili değişkenleri içermektedir.Çalışmada, öncelikle veri ön işleme adımları gerçekleştirilmiş ve veri keşfi yapılmıştır. Daha sonra, farklı makine öğrenmesi algoritmaları kullanılarak modelleme işlemi gerçekleştirilmiştir. Kullanılan algoritmalar gradient boosting , XGBoost , Gradient Boosting,Random Forest'dir.. Modellerin performansı, doğruluk ve tahmin hataları gibi metrikler kullanılarak değerlendirilmiştir.Sonuçlar, müşterilerin risk profillerini ve talep tutarlarını öngörmek için kullanılabilecek etkili modellerin oluşturulabileceğini göstermektedir. Bu modeller, sigorta şirketlerinin risk yönetimi ve fiyatlandırma stratejilerini geliştirmelerine yardımcı olabilir. Ayrıca, daha fazla araştırma ve geliştirme için temel oluşturabilir.Bu çalışma, sigorta endüstrisinde makine öğrenmesi tekniklerinin kullanımının önemini vurgulayarak, gelecekteki çalışmalar için bir çerçeve sağlamaktadır.

ARAÇ (KASKO) SİGORTASI NEDİR?

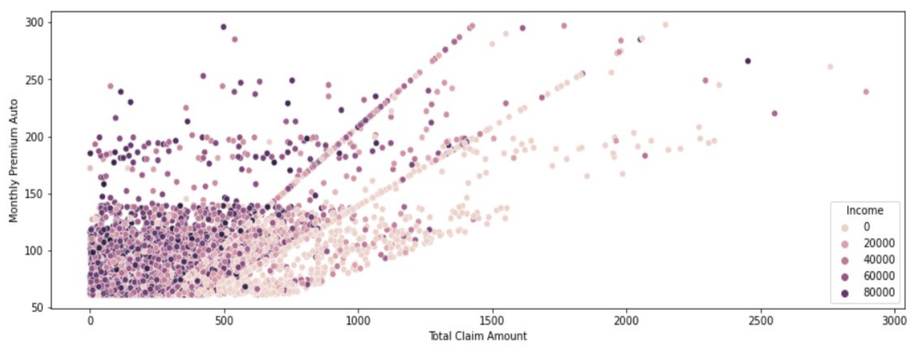
Kara taşıtları sigortası veya kısaca kasko, bir aracın kaza geçirmesi, yanması, çalınması veya çalınmaya teşebbüs edilmesi sonucu oluşabilecek zararları güvence altına alan bir sigorta sözleşmesidir. Trafik sigortasından farklı olarak, poliçe sahibinin kendi aracının hasarını da kısmen veya tamamen karşılar. Bu, araç sahibine aracının kendi hasarlarını kapsama alma seçeneği sunar. Kasko sigortası trafik sigortasının aksine zorunlu değildir, ancak birçok araç sahibi tarafından tercih edilir çünkü aracın beklenmedik hasarlarına karşı ek bir güvenlik sağlar.

Sigorta talepleri, sigorta poliçesinde belirtilen risklerin gerçekleşmesi durumunda sigortalının tazminat talep etmesidir. Örneğin, araba kazası, ev yangını veya sağlık sorunları bu talepleri tetikleyebilir. Sigorta şirketi, talebin geçerliliğini değerlendirir ve geçerli bulursa tazminat öder. Geçersiz bulunursa ödeme yapılmaz.

Sigorta kapsamı, sigortalıyı belirli risklere karşı korur. Örneğin, araba kazasında araç hasarı veya ev yangınında yeniden inşa masrafları sigorta kapsamındadır. Sigortalı, bu koruma karşılığında sigorta primi öder. Bu primler, sigorta şirketinin riski üstlenmesi ve talepleri karşılaması için gereklidir

VERİ SETİ VE KEŞİFCİ VERİ ANALİZİ (EDA)

Veri seti Amerika'daki bir araç sigorta(kasko) şirketine ait müşterilerin bilgilerini içermektedir.26 değişkenden oluşmaktadır. Veri setini %80 eğitim ve %20 test olarak ayırılmıştır. Sigorta şirketleri, sürdürülebilir kârlılıklarını ancak doğru ve tutarlı öngörülerle sağlayabilirler. Bu analizde, "Total Claim Amount" (Toplam Talep Tutarı) değişkeni hedef değişken olarak ele alınmıştır. "Total Claim Amount" değişkeni, sigorta poliçesi kapsamında yapılmış olan toplam talep miktarını ifade etmektedir. Bu, müşterinin sigorta şirketine hasar, kayıp veya başka bir olay nedeniyle yaptığı tüm taleplerin toplam değeridir. Başka bir deyişle, bu rakam sigorta şirketinin, müşterinin talep ettiği toplam zarar tutarını göstermektedir.



Grafikte, "Monthly Premium Auto" (Aylık Prim Tutarı) ile "Total Claim Amount" (Toplam Talep Tutarı) arasındaki ilişki, gelir seviyelerine göre farklı renklerle gösterilmektedir.. Genel olarak, aylık prim tutarı ile toplam talep tutarı arasında negatif bir ilişki gözlemlenmekte ve yüksek gelir seviyesine sahip müşterilerin daha düşük talep tutarlarına sahip olma eğiliminde olduğu görülmektedir. Belirgin dış noktalar, yüksek risk taşıyan müşterileri belirlemek için kullanılabilir.Bu bulgulara dayanarak, sigorta şirketi primlerini daha adil bir şekilde belirleyebilir ve düşük gelirli, yüksek talep tutarına sahip müşterilere yönelik prim politikalarını gözden geçirebilir. Yüksek risk taşıyan müşteriler için özel primler veya poliçe şartları oluşturulabilir.



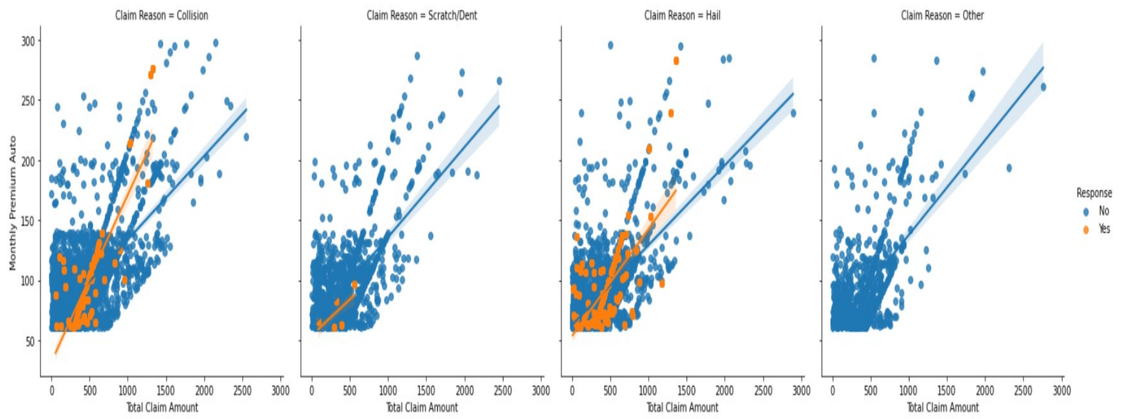
- Total Claim Amount ve Monthly Premium Auto arasında pozitif ve yüksek korelasyon var.
- Total Claim Amount ve Income arasında negatif ve düşük korelasyon var.
- Total Claim Amount ve Claim Amount arasında pozitif ve düşük korelasyon var.

KAYNAKÇA

[1] Artificial Intelligence and Machine Learning Applications in Big Data Analysis , Muhammet Atalay1 , Enes Çelik2

[2] Araç Sahiplerinin Kasko Sigorta Poliçesi Edinmesinde Etkili Olan Faktörlerin Araştırılması , Hakan Eygü , Fatih Soğukpınar

[3] Saldırı Tespit Sistemlerinde Topluluk Öğrenme Yöntemlerinin Kıyaslanması İlya Kuş1* , Sinem Bozkurt Keser2 , Esra Nergis Yolaçan3



Tüm talep türlerinde aylık primlerin artmasıyla birlikte onaylanan taleplerin toplam tutarında bir azalma gözlemlenmektedir. Bu durum, daha yüksek prim ödeyen müşterilerin ortalama olarak daha az talepte bulunduğunu göstermektedir.Kaza taleplerinde aylık prim ve onaylanan talep tutarı arasındaki korelasyon en güçlüdür. Aylık primlerdeki artış, onaylanan kaza talep tutarında önemli bir düşüşe neden olmaktadır. Çizik/çentik ve dolu taleplerinde de bu korelasyon zayıf olsa da, aylık primlerdeki artışla onaylanan talep tutarlarında hafif bir azalma görülmektedir. Diğer talep türlerinde ise aylık prim ve onaylanan talep tutarı arasında belirgin bir eğilim gözlemlenmemektedir.Bu bulgular, sigorta şirketlerinin risk profillerini ve prim fiyatlandırmalarını geliştirmeleri için kullanılabilir. Ayrıca, daha yüksek prim ödeyen müşterilerin taleplerinin onaylanma olasılığının daha yüksek olduğunu ve reddedilme olasılığının daha düşük olduğunu da göstermektedir.

MAKİNESİ ÖĞRENMESİ UYGULAMALARI

GRADİENT BOOSTİNG

Bu algoritma, hata oranını minimize etmek için zayıf tahmin edicileri (genellikle karar ağaçları) ardışık olarak ekler ve her yeni model, önceki modellerin hatalarını düzeltmeye çalışır.

XG BOOST

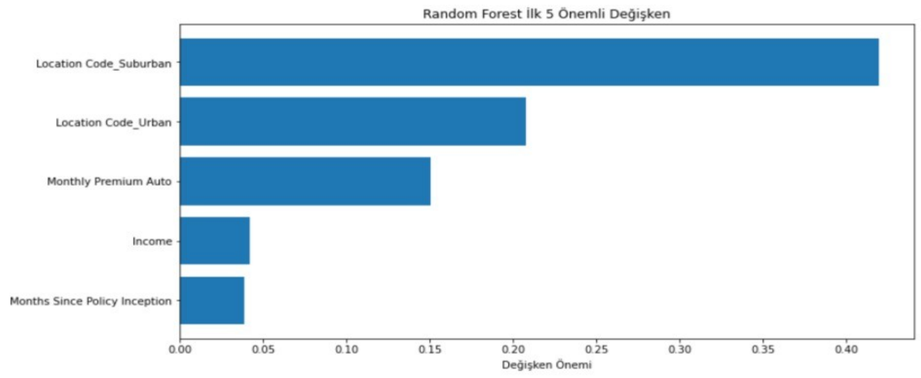
Gradient Boosting algoritmasının optimize edilmiş bir versiyonudur. Yüksek performansı ve hızlandırılmış hesaplama özellikleri ile öne çıkar, özellikle büyük veri setlerinde etkilidir.

RANDOM FOREST

Random Forest, birçok karar ağacının (decision tree) oluşturulup bir araya getirilerek, her ağacın verdiği tahminlerin ortalamasının alınmasıyla çalışan bir makine öğrenmesi algoritmasıdır. Bu, daha doğru ve güvenilir tahminler elde etmeyi sağlar.

	R-squared (Train)	MSE (Train)	MAE (Train)
XGBoost	0.895446	0.088297	0.180889
Gradient Boosting	0.785921	0.180791	0.250980
Random Forest	0.965886	0.028810	0.092370

	R-squared (Test)	MSE (Test)	MAE (Test)
XGBoost	0.729880	0.224033	0.262696
Gradient Boosting	0.731330	0.222830	0.271857
Random Forest	0.736609	0.218452	0.256605



SONUÇ

- Model performansları, R-kare, MSE ve MAE gibi metrikler kullanılarak değerlendirilmiş ve en başarılı modelin Random Forest olduğu tespit edilmiştir. Bu modelin başarısında önemli değişkenler arasında lokasyon kodu (Suburban ve Urban), aylık prim, gelir ve poliçe başlangıcından itibaren geçen süre öne çıkmaktadır. Sonuç olarak, geliştirilen bu modeller, sigorta şirketlerinin risk yönetimi ve fiyatlandırma stratejilerini iyileştirerek daha etkili kararlar almasına olanak tanımaktadır..

İSTATİSTİK BÖLÜMÜ

Makine Öğrenmesi Algoritmaları İle Müşteri Terk Analizi İncelemesi

Bilgesu Miray Karakoç Öğrenci No : 18023060

Danışman: Prof.Dr.Ersoy Öz

Bu çalışmada, **makine öğrenmesi algoritmalarının müşteri terk oranlarını tahmin etme yetenekleri incelenmiştir**. Telekomünikasyon şirketine ait 7043 müşterinin demografik bilgileri, hizmet kullanım detayları ve hesap bilgileri içeren veri seti kullanılmıştır. Veri analizi ve özellik mühendisliği sonrası, Lojistik Regresyon, KNN, Karar Ağaçları, Rastgele Ormanlar, SVM, XGBoost, LightGBM ve CatBoost algoritmalarıyla modeller oluşturulmuş ve performansları değerlendirilmiştir. **En yüksek performansı CatBoost ve XGBoost algoritmaları göstermiştir. Sözleşme süresi, ödeme yöntemi, çevrimiçi güvenlik, teknik destek ve toplam hizmet sayısı en önemli özellikler olarak belirlenmiştir.** Sonuçlar, makine öğrenmesi algoritmalarının müşteri terkini tahmin etmede etkili olduğunu ve stratejik kararlar alınmasına yardımcı olabileceğini göstermektedir.

Veri Setinin Detayları

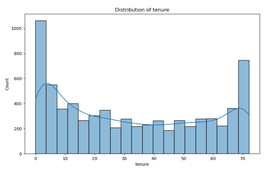
Çalışmada kullanılan **değişkenlerimiz, yapılarına göre iki gruba** ayrılabilmektedir :

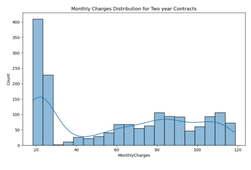
1-Müşterilerin kişisel demografik bilgilerini sağlayanlar
2-Müşterinin hesap ve hizmet kullanım detaylarını sağlayanlar

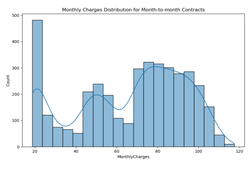
Kişisel demografik bilgiler olarak değerlendirilebilecek değişkenlerimiz ise **müşterinin cinsiyeti, yaşı (SeniorCitizen), medeni durumu (Partner) ve bakmakla yükümlü olduğu kişiler (Dependents) gibi bilgileri içermektedir**.
Müşterinin hesap ve hizmet kullanım detaylarını sağlayan değişkenler ise şunlardır:

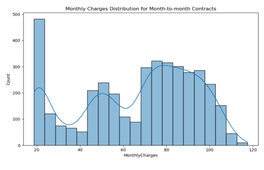


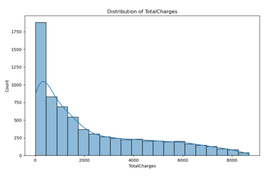
CustomerId: Her müşteri için benzersiz kimlik
Tenure: Müşterinin şirkette kaldığı ay sayısı.
MultipleLines: Müşterinin birden fazla hattı olup olmadığı.
InternetService: Müşterinin internet servis sağlayıcısı (DSL, Fiber optik, Yok)
OnlineSecurity: Müşterinin çevrimiçi güvenliği olup olmadığı
PhoneService:Müşterinin telefon hizmeti olup olmadığı
OnlineBackup: Müşterinin çevrimiçi yedeği olup olmadığı
DeviceProtection: Müşterinin cihaz korumasına sahip olup olmadığı
TechSupport: Müşterinin teknik destek alıp almadığı
StreamingMovies: Müşterinin film akışı olup olmadığı
Contract: Müşterinin sözleşme süresi (Aydan aya, Bir yıl, İki yıl)
PaperlessBilling: Müşterinin kağıtsız faturası olup olmadığı
PaymentMethod: Müşterinin ödeme yöntemi (Elektronik çek, Posta çeki, Banka havalesi, Kredi kartı)
MonthlyCharges: Müşteriden aylık olarak tahsil edilen tutar
TotalCharges: Müşteriden tahsil edilen toplam tutar
Churn: Müşterinin hizmeti kullanıp kullanmadığı (Evet veya Hayır)



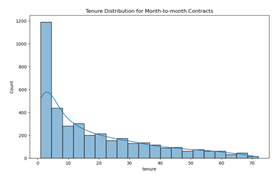


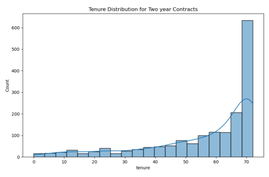




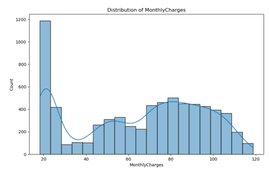


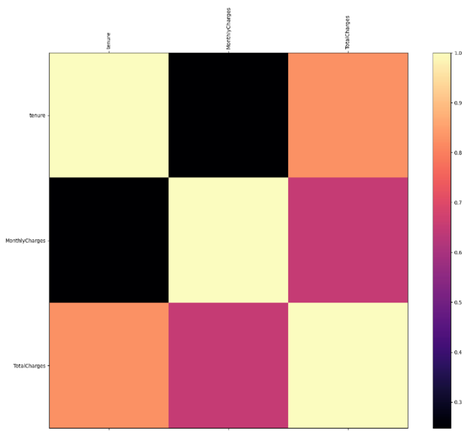
Numerik ve kategorik değişkenlerin analizi, müşteri davranışları ve ödeme kalıpları hakkında önemli bilgiler sağlar. Yapılan analizlerimizde , tenure (şirkette kalınan süre) ve MonthlyCharges (aylık ödemeler) değişkenlerinin dağılımı incelendi.





Aylık sözleşmesi olan müşterilerin genellikle kısa süreli ve yüksek maliyetli hizmetleri tercih ettiği, iki yıllık sözleşmesi olan müşterilerin ise daha uzun süre kalma eğiliminde olduğu ve daha düşük maliyetli hizmetleri seçtiği gözlemlendi.





Korelasyon matrisi, churn analizimizde veri setindeki özellikler arasındaki doğrusal ilişkileri anlamamıza yardımcı olur. Özelliklerin birbirleriyle ve hedef değişken `Churn` ile ilişkilerini belirlemek, veri desenlerini anlamamızı sağlar. Örneğin, `TotalCharges` değişkeninin `MonthlyCharges` ve `tenure` ile yüksek korelasyona sahip olması, bu özelliklerin benzer bilgiler taşıdığını gösterir. Bu bilgi, özellik seçimi ve mühendisliğinde daha bilinçli kararlar almamızı sağlar, böylece modelimizin performansını artırır. Ayrıca, hedef değişkenle yüksek korelasyona sahip özellikler, churn tahmininde daha etkili olabilir.

TotalChargers'in aylık ücretler ve tenure ile yüksek korelasyonlu olduğu görülmekte

Veri setimizi incelediğimizde , TotalChargers değişkenimizde 11 tane boş değerler olduğunu gördük . **TotalCharges sayısal bir değişken çünkü toplam tutarları ifade ediyor ve sürekli bir sayısal veri türüdür . Bu nedenle boş değerleri medyanla doldurmaya karar verdik .**
Veri setindeki eksik değerleri medyanla doldurmak, uç değerlerden etkilenmeyen ve veri kümesinin merkezi eğilimini daha güvenilir bir şekilde temsil eden bir yöntemdir. **Medyan, çarpık dağılımlarda ortalamaya göre daha dayanıklıdır ve veri setinin yapısını bozmadan eksik değerleri doldurur.**
Ardından base modellerimizi kurup , modelleri kıyaslamak için tablomuzu oluşturduk

Method	Accuracy	AUC	Recall	Precision	F1
LR	0.805	0.853	0.655	0.744	0.696
KNN	0.765	0.81	0.603	0.685	0.642
CART	0.729	0.748	0.577	0.639	0.606
RF	0.793	0.837	0.637	0.724	0.677
SVM	0.765	0.81	0.603	0.685	0.642
XGBoost	0.807	0.852	0.658	0.746	0.698
LightGBM	0.813	0.857	0.664	0.752	0.703
CatBoost	0.815	0.859	0.668	0.756	0.707

Bu aşamada baktığımızda (*aykırı değer analizi, encoding, özellik çıkarımı gibi aşamalar içeren "veri ön işleme aşaması"nı yapmadan*)

Catboost, doğruluk, AUC, recall, precision ve F1 skorları açısından diğer modellere göre üstün performans gösteriyor.

Modelleme Süreci

Veri setimizdeki sayısal değişkenler için **aykırı değerleri belirledik** ve sürecimiz şu şekilde ilerledi:
Çeyreklik Değerlerin Hesaplanması:
Her bir sayısal değişken için, %5 (Q1) ve %95 (Q3) çeyreklik değerlerini hesapladık.
Eşik Değerlerin Belirlenmesi:
Bu çeyreklik değerler kullanılarak, interçeyrek aralığı (IQR) hesapladık: **IQR = Q3 - Q1.**
Alt eşik değeri: **low_limit = Q1 - 1.5 * IQR.**
Üst eşik değeri: **up_limit = Q3 + 1.5 * IQR.**
Aykırı Değerlerin Tespiti:
Her bir sayısal değişken için, değerlerin belirlenen alt ve üst eşiklerin dışında olup olmadığını kontrol ettik.
Eğer bir değer bu eşiklerin dışında kalıyorsa, bu değeri aykırı olarak belirledik.
Aykırı Değerlerin Değiştirilmesi:
Aykırı değerleri, ilgili değişkenin alt veya üst eşik değeri ile değiştirdik.
ve bu değerleri alt ve üst eşiklerle değiştirdik. İlk olarak, her bir sütunun %5 ve %95 çeyreklik değerlerini kullanarak alt ve üst eşikleri hesapladık. Daha sonra, bu eşiklerin dışında kalan aykırı değerleri tespit ettik ve bu değerleri ilgili eşiklerle değiştirdik. Bu işlemle, veri setimizdeki aykırı değerlerin etkisini azaltarak model performansını ve genel veri doğruluğunu artırmayı hedefledik.

Özellik Çıkarımı ,
veri setindeki mevcut değişkenlerden yeni ve daha anlamlı değişkenler oluşturma sürecidir. Bu işlem, modelin öğrenme kapasitesini ve performansını artırmak için yapılır. Bu işlem, genellikle veri ön işleme ve modelleme sürecinin başında gerçekleştirilir.

1. **Tenure Değişkeninden Yıllık Kategorik Değişkenler Oluşturma:
2. **Müşteri Bağlılığının Belirlenmesi
3. **Destek Hizmetleri Almayan Müşteriler
4. **Aylık Sözleşmesi Bulunan ve Genç Müşteriler:
5. **Toplam Hizmet Sayısı
6. **Streaming Hizmeti Alan Müşteriler
7. **Otomatik Ödeme Yapan Müşteriler
8. **Ortalama Aylık Ödeme
9. **Güncel Fiyatın Ortalama Fiyata Göre Artışı
10. **Servis Başına Ücret

Bu adımlarla, veri setine yeni ve anlamlı değişkenler ekleyerek model performansını artırmayı amaçladık

Encoding Aşaması

Değişkenlerin Tiplerine Göre Ayrılması:

Veri setindeki değişkenleri kategorik, numerik ve kardinal olarak ayırdık.

Neden?*: Hangi türdeki dönüşümlerin uygulanacağını belirlemek için.*

-Label Encoding:
İkili (binary) kategorik değişkenleri sayısal değerlere (0 ve 1) dönüştürdük.
Neden?*: Makine öğrenmesi modellerinin sayısal verilerle çalışabilmesi için*.

-One-Hot Encoding:
Çoklu kategorik değişkenleri one-hot encoding ile kodladık.

Neden?: *Kategorik verilerin sayısal olarak temsil edilmesini ve modelin bu verilerden bilgi öğrenmesini sağlamak için.*

Sonuç: Tüm kategorik değişkenler sayısal forma dönüştürüldü. ***Neden?****: Modellerin verileri daha iyi anlamasını ve performansını artırmasını sağlamak için*

Yukardaki işlemlerimizin ardından , düzenlediğimiz veri setimiz ile tekrar modelleme gerçekleştirdik ve ilk base modelde kullandığımız yaklaşımları tekrar kullandık. CatBoost ve LightGBM modelleri oldukça benzer performans sergiliyor. Her iki model de yüksek doğruluk (Accuracy), AUC ve F1 skorlarıyla öne çıkıyor.

Model	Accuracy	AUC	Recall	Precision	F1
LR	0.8038	0.8458	0.5131	0.6724	0.5807
KNN	0.7704	0.754	0.465	0.5861	0.5176
CART	0.727	0.6586	0.5094	0.4864	0.4975
RF	0.7933	0.827	0.5062	0.6408	0.5652
SVM	0.7681	0.7255	0.2579	0.6657	0.3708
XGB	0.7811	0.8235	0.496	0.6082	0.5459
LightGBM	0.7974	0.8353	0.5297	0.6448	0.5814
CatBoost	0.7999	0.8416	0.5265	0.6529	0.5828

Sonuç:
CatBoost ve LightGBM modelleri oldukça benzer performans sergiliyor, ***ancak CatBoost modeli, Accuracy, AUC, Precision ve F1 skoru açısından hafif bir üstünlük sağlıyor.*** *Bu nedenle, CatBoost modelini en iyi model olarak değerlendirebiliriz.*

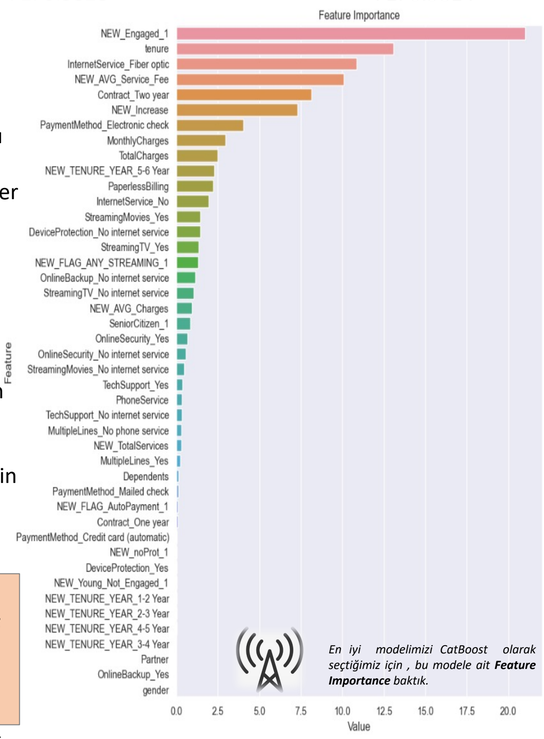
CatBoost:
Accuracy: 0.7999
AUC: 0.8416
Recall: 0.5265
Precision: 0.6529
F1: 0.5828

LightGBM:
Accuracy: 0.7974
AUC: 0.8353
Recall: 0.5297
Precision: 0.6448
F1: 0.5814

NEW_Engaged_1 ve ***tenure*** özellikleri **en yüksek öneme sahiptir.** Bu, müşterilerin uzun vadeli sözleşmelerde bulunmalarının ve hizmette geçirdikleri sürenin churn tahmininde kritik rol oynadığını gösterir.
InternetService_Fiber optic ve ***NEW_AVG_Service_Fee*** gibi özellikler de yüksek öneme sahiptir. Bu, fiber optik internet hizmeti kullanıcılarının churn olasılığının düşük olduğunu ve hizmet başına ortalama ücretin churn tahmininde önemli bir faktör olduğunu belirtir. ***Orta Derecede Önemli Özellikler:***

Contract_Two year ve ***NEW_Increase*** gibi özellikler .İki yıllık kontratlar ve aylık ücretlerdeki artışlar, churn olasılığını etkileyen faktörler arasında yer alır.
PaymentMethod_Electronic check ve ***MonthlyCharges*** da churn üzerinde belirgin bir etkiye sahiptir. Bu özellikler, müşterilerin ödeme yöntemleri ve aylık ödemelerinin churn kararlarını nasıl etkilediğini gösterir.

Özetle, *grafikteki en önemli özellikler müşterilerin kontrat süreleri, hizmette kaldıkları süre, internet hizmeti türü ve ödeme yöntemleri gibi faktörlerdir. Bu özellikler churn tahmininde kritik rol oynar ve telekomünikasyon şirketlerinin müşteri kaybını azaltmak için bu alanlara odaklanmamızı gös*terir.



En iyi modelimizi CatBoost olarak seçtiğimiz için , bu modele ait ***Feature Importance*** *baktık.*

KAYNAKÇA

[1]https://medium.com/@OlorunwaFagbenro/predictive-customer-churn-analysis-in-python-uncovering-insights-for-effective-retention-57ec8e1d55c4
[2] https://www.kaggle.com/code/jsaguiar/churn-prediction-tutorial-with-sklearn
[3] http://esjournal.cumhuriyet.edu.tr/en/download/article-file/370495



İSTATİSTİK BÖLÜMÜ

Türk Dizilerinin Demografik Yapısı ve İzleyici Analizi

Burak SAKAR 19023014

Danışman: Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK

ÖZET

Bu çalışma, Türk dizilerinin demografik yapısını ve izleyici analizini ele almaktadır. Google Forms üzerinden Anket yöntemiyle farklı sosyoekonomik sınıfa mevcut 855 kişiden toplanan veriler, Türkiye'de televizyonun demografik yapısı ve izlenme alışkanlığının farklı yaş, cinsiyet ve sosyoekonomik gruplar üzerindeki dağılımını değerlendirmektedir.
Analiz sonuçları, aşk ögesinin yüksek olduğu dizilerin genç izleyiciler arasında popüler olduğunu, yaşlı izleyici grubunun ise daha çok geleneksel temalara yönelik dizileri tercih ettiğini göstermektedir. Erkek izleyiciler aksiyon ve politik içerikli dizilere, kadın izleyiciler ise romantik ve dramatik dizilere daha fazla ilgi göstermektedir. Çalışma, Türk dizilerinin nasıl geniş bir izleyici kitlesine hitap ettiğini ve bu dizilerin toplumsal cinsiyet, yaş ve sosyal sınıf gibi demografik faktörlerle nasıl etkileşimde bulunduğunu ortaya koymaktadır. Sonuç bölümünde, dizilerin daha fazla izleyici çekebilmesi ve çeşitli demografik gruplardan insanların ilgisini sürdürebilmesi için içerik çeşitliliğinin artırılması gerektiği vurgulanmaktadır.

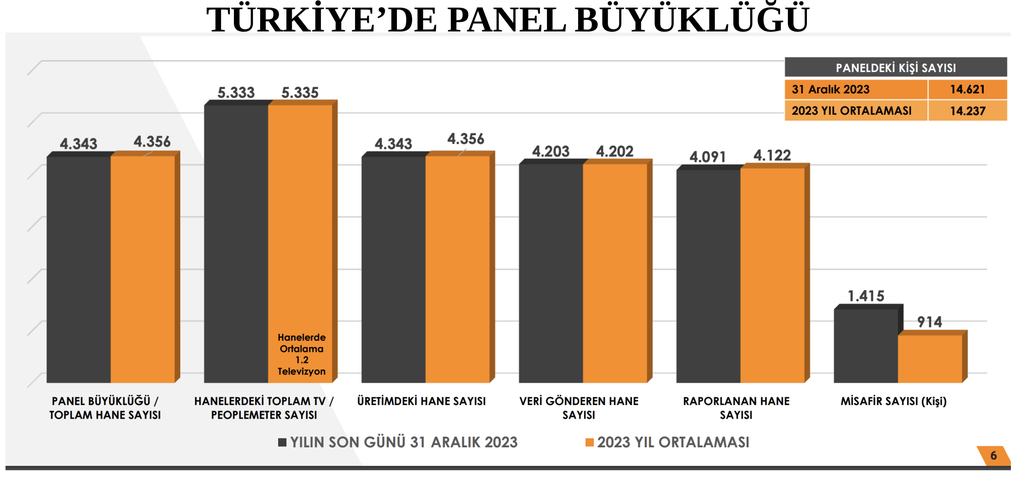
Türk Dizilerinin Demografik Yapısı ve İzleyici Analizi

Türkiye'de Televizyon İzleyicisi ve Rating Ölçümü

Türkiye'de televizyon yayıncılığı, 1968 yılında İstanbul'da kurulan ve sınırlı sayıda programla başlayan TRT ile temellerini atmıştır. Zamanla yayın ağı genişleyen TRT, Türk toplumunu hem eğiten hem de eğlendiren çeşitli programlar sunmuştur. 1980'lerin sonlarına doğru özel televizyon kanallarının ortaya çıkışıyla birlikte Türkiye'deki televizyonculukta büyük bir dönüşüm yaşanmıştır. Bu dönemde çeşitlilik ve rekabet artarken, yerel diziler ve yarışma programları gibi içerikler popülerlik kazanmıştır. İnternetin yaygınlaşmasıyla birlikte dijital platformlar da önem kazanmış, televizyon izleme alışkanlıkları değişim göstermiştir. Günümüzde Türkiye televizyonculuğu, global trendlere uyum sağlayarak hem geleneksel hem de çevrimiçi platformlarda geniş bir yelpazede içerik sunmaktadır.

Türkiye'de reyting ölçümü, televizyon izleyici kitlesinin büyüklüğünü ve demografik özelliklerini belirlemek amacıyla yapılan bir süreçtir. Bu ölçümler, genellikle bağımsız reyting ölçüm şirketleri tarafından gerçekleştirilir. Ölçüm için seçilen hanelere yerleştirilen özel cihazlar, izlenen programları ve izleme sürelerini kaydeder. Ölçümü KANTAR Medya yaparken bu sonuçları TİAK (Televizyon İzleme Araştırmaları Anonim Şirketi) günlük, aylık ve yıllık periyotlarda halka açık paylaşır. Şirket aynı zamanda her yılın sonunda panel ölçümü, evren değerleri ve izleyici alışkanlıklarını içeren "X Yılı İzleme Davranışları" raporunu yayınlamaktadır.

TÜRKİYE'DE PANEL BÜYÜKLÜĞÜ



Türkiye'de rating ölçümü 40 ilde bulunan 4.356 hane ve 14.237 kişi üzerinden gerçekleştirilmektedir. Yıl süresince 914 kişi misafir olarak kabul edilmekte, 4.122 hane üzerinden raporlama yapılmaktadır. Bu örneklem değeri, 17 milyon 728 bin haneyi takiben 58 milyon 367 bin kişilik evren büyüklüğünü oluşturmaktadır. Evinde en az bir çalışır durumda televizyonu olan, konut olarak kullanılan tüm haneler ve bu hanelerde yaşayan 5 yaş ve üzerindeki bireyler bu çalışmanın örneklemeni oluşturmaktadır.
Anket soruları bu rapordaki demografik yapılara göre hazırlanmış, özet sonuçlar aşağıda yer almaktadır.

ÖZET SONUÇLAR

ANKET SORUSU	CEVAP	ÖRNEKLEM	YÜZDE
Cinsiyetiniz nedir?	Kadın	768	90,2%
	Erkek	83	9,8%
Yaş aralığınız nedir?	18 yaş ve altı	65	40,1%
	18-24 yaş	97	11,3%
	25-34 yaş	341	39,8%
	35-44 yaş	301	35,1%
	45-54 yaş	73	8,5%
	55 yaş ve üzeri	32	3,7%
Öğrenim durumunuz?	İlkokul	3	0,3%
	Ortaokul	4	0,5%
	Lise	67	7,8%
	Üniversite	566	66%
	Yüksek Lisans	174	20,3%
	Doktora	44	5,1%
İkamet ettiğiniz bölge?	Akdeniz	58	6,8%
	Doğu Anadolu	7	0,8%
	Ege	133	15,6%
	Güney Doğu Anadolu	14	1,6%
	İç Anadolu	134	15,7%
	Karadeniz	35	4,1%
	Marmara	473	55,4%
Çalışma durumunuz?	Tam zamanlı çalışıyorum	471	55%
	Yarı zamanlı çalışıyorum	84	9,8%
	Çalışmıyorum	207	24,2%
	Öğrenci	95	11,1%
Medeni durumunuz?	Evli	489	57,1%
	Bekar	327	38,2%
	Boşanmış / Dul / Eşini kaybetmiş	41	4,8%
Hanenizde en yüksek gelire sahip kişi?	Ben / Kendim	291	34%
	Eşim - beyim	344	40,2%
	Babam	136	15,9%
	Annem	35	4,1%
	Kardeşim / Ablam / Abim	29	3,4%
	Diğer	21	2,8%
Oturduğunuz evde ikamet eden birey sayısı?	1-2 birey	316	36,9%
	3-4 birey	488	56,9%
	6 birey veya daha fazlası	53	6,2%
Hane halkı gelir düzeyi?	0-20.000 TL	3	0,3%
	20.000-40.000 TL arasında	4	0,5%
	40.000-60.000 TL arasında	67	7,8%
	60.000-80.000 TL arasında	566	66%
	80.000 TL ve üstü	174	20,3%

ANKET SORUSU	CEVAP	ÖRNEKLEM	YÜZDE
Günlük ortalama televizyon izleme süreniz?	0-1 saat	270	31,5%
	2-4 saat	414	28,3%
	4-8 saat	74	8,6%
	8-12 saat	16	1,9%
	12 saat veya daha fazla	5	0,6%
	Televizyon izlemiyorum	74	8,6%
Televizyon izleme sürenizin en yoğun olduğu gün?	İzleme alışkanlığım günlere göre değişmiyor	499	58,7%
	Cuma	98	11,5%
	Salı	84	9,9%
	Pazar	65	7,6%
	Pazartesi	30	3,5%
	Diğer	74	8,7%
Televizyonda hangi program türleri için izliyorsunuz? (Birden fazla işaretleyebilirsiniz.)	Yerli Dizi	583	69,6%
	Yabancı Dizi	346	41,3%
	Türk Filmleri	137	16,3%
	Yabancı Filmler	256	30,5%
	Çizgi Filmler	35	22%
	Canlı Spor Müsabakaları	175	20,9%
	Haber Bültenleri	229	27,3%
	Yarışma Programları	183	21,8%
	Tartışma Programları	123	14,7%
	Gerçek yaşamlar	83	9,9%
Bir gün içerisinde en çok televizyon izlediğiniz zaman dilimi hangisidir?	Sabah Kuşağı (06:00-11:00)	24	2,9%
	Gündüz Kuşağı (12:00-18:00)	37	4,5%
	Prime Time (19:00-23:00)	729	88%
	Gece Kuşağı (24:00-05:00)	38	4,6%
Televizyon dizilerini hangi platformdan takip ediyorsunuz?	Televizyon	399	47,5%
	Youtube	550	65,5%
	Televizyon kanalının sitesinden	79	9,4%
	Puhutv	50	6%
	Diğer	171	13,6%
Aşağıda yer alan dizilerden hangisini/hangilerini düzenli olarak izliyorsunuz? (Birden fazla işaretleyebilirsiniz.)	Bahar	613	71,8%
	Gönül Dağı	24	2,8%
	İnci Taneleri	77	9%
	Kızılcık Şerbeti	358	41,9%
	Kızıl Goncalar	237	27,8%
	Kuruluş Osman	8	0,9%
	Hiçbiri	159	18,6%
Aşağıda yer alan dizilerden. hangisinin/hangilerinin adını daha önce duydunuz? (Birden fazla işaretleyebilirsiniz.)	Altın Kafes	27	3,2%
	Aşka Düşman	97	11,6%
	Bir Derdim Var	214	25,5%
	Kader Bağları	50	6%
	Senden Önce	80	9,5%
	Yan Oda	224	26,7%
	Hiçbiri	485	57,9%

HİPOTEZLER VE Kİ-KARE TESTİ YORUMU

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	995.012[a]	36	<.001
Likelihood Ratio	407.125	36	<.001
N of Valid Cases	897		

a. 24 cells (49.0%) have expected count less than 5. The minimum expected count is .07.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	11.919[a]	6	.064
Likelihood Ratio	11.585	6	.072
N of Valid Cases	854		

a. 3 cells (21.4%) have expected count less than 5. The minimum expected count is .28.

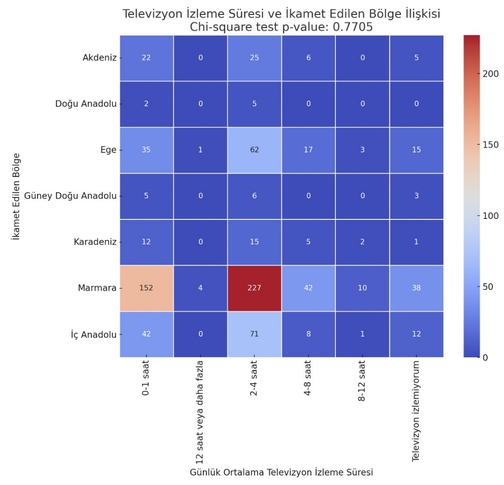
Ki-Kare İlişki Hipotezleri:



H0: Televizyon izlenme süresiyle arasında ilişki yoktur.
H1: Televizyon izlenme süresiyle arasında ilişki vardır

TELEVİZYON İZLEME SÜRESİNİ DEĞİŞTİREN FAKTÖRLER

- Cinsiyet (Kadınlar erkeklerden daha fazla izliyor)
- Yaş (Yaş arttıkça izleme süresi artıyor)
- Öğrenim durumu (Öğrenim derecesi düştükçe izleme süresi artıyor)
- Evde bulunan televizyon sayısı (Televizyon sayısı arttıkça izleme süresi artıyor)
- Hane halkı toplam geliri (Toplam gelir düştükçe televizyon izleme süresi artıyor)
- Çalışma durumu (Çalışmayan kitle çalışan kitleden daha fazla televizyon izliyor)
- Medeni durum (Evli bireyler bekar bireylerden daha fazla televizyon izliyor)
- Öncelikli program türü (Yerli diziye öncelik verdikçe televizyon izleme süresi artıyor)



SONUÇ

Bu çalışma, Türk dizilerinin seyirci profilini tespit etmeyi amaçlamaktadır. Elde edilen bulgular, kadın katılımcıların, erkeklere göre daha fazla televizyon izlediğini ortaya koymaktadır. Yaş arttıkça televizyon izleme süresi de artmaktadır. Televizyonu en fazla 45 yaş ve üzeri kadın seyirci seyretmekte, çalışmayan kitle çalışan kitleye kıyasla daha fazla dizi takip etmektedir.

Bölgesel kategoriler ile televizyon izleme arasında bir bağlantı saptanmamıştır. Katılımcıların yaşadığı bölge, tercih ettiği içerik ve formatlarda anlamlı bir değişikliğe sebep olmakta ancak ortalama televizyon izleme süresini etkilememektedir. Benzer şekilde, hanede yaşayan birey sayısı ve öğrenim durumu da televizyon izleme oranı etkilemeyen faktörler arasındadır.

Hanedeki gelir düzeyinin artması, yani C2 sınıfından C1 sınıfına yükselmesi televizyon izleme süresinin düşmesine neden olmaktadır. Evli insanların televizyon izleme süre ortalaması, bekarlara göre daha yüksektir. Ana gelir kaynağı eşi olan kişiler, eşi olan kişilere kıyasla daha az televizyon izlemektedir. Bu kişilerin öncelikli program tercihi dizi değil haber bültenidir. Katılımcıların büyük çoğunluğu televizyonu Prime Time'da (19:00-23:00) izlemekte, günler arasında ayrım yapmamaktadır.

Sonuç olarak, TİAK'ın people meter cihazıyla yaptığı ölçüme benzer profilin televizyonu izlediği saptanmıştır. Lisansüstü (yüksek lisans veya doktora) öğrencilerin televizyonu Kızıl Goncalar ve Bahar dizileri için açtığı gözlenmiştir. Bu anket sonucuna göre Bahar dizisi tüm demografik yapılarda izlenme alabilmektedir. Ancak diğer diziler kimi sosyoekonomik statülerdeki yüksek izlenmelerine karşılık kimilerinde az izlenme almaktadır. Bu dizilerin demografik yapılardaki eksikleri saptayıp onlara göre yeni stratejiler izlemesi daha fazla reyting almalarını sağlayabilir.

KAYNAKÇA

[1]TİAK (Televizyon İzleme Araştırmaları A.Ş.)
[2]İzleyici Tanıma Anket Sonuçları

İSTATİSTİK BÖLÜMÜ

E-TİCARET VERİ SETİ ÜZERİNDE MÜŞTERİ KAYBI TAHMİN MODELLERİNİN GELİŞTİRİLMESİ

BUSE NUR BARLAS 20023045

Danışman: DOÇ. DR. GÜLDER KEMALBAY

ÖZET

Pazarlama sektöründe önemli bir rol oynamakta olan “Yeni müşteri kazanmak, mevcutta var olan müşteriyi tutmaktan çok daha maliyetlidir” düşüncesi, işletmeleri müşteri sadakatini kazanma çalışmalarına itmektedir. Bugünün rekabetçi koşullarında, müşterilerin aldıkları ürün ve hizmetler hakkında olan beklentilerini karşılamak işletmeler için önemli bir konu haline gelmiştir. Rakip firmaların sayısının düzenli olarak arttığı günümüz şartlarında işletmelerin devamlılıklarını sağlayabilmek için müşterilerin memnuniyet ve şikayetlerini göz ardı etmeyerek müşteriler ile iletişim içerisinde olması büyük ölçüde önem arz etmektedir. Bu kapsamda, bu çalışma ile işletmenin gerekli aksiyonların alınmasını sağlayacak ilgili kararları verebilmesi hakkında yardımcı olacak bilgiler sunarak müşteri sadakatini kazanma çalışmalarına katkıda bulunulması amaçlanmıştır. Çalışma, müşterilerin Singapur’daki bir E-Ticaret sitesini kullanarak 2012 yılından itibaren yapmış olduğu alışverişler hakkında detaylı bilgi içeren bir veri seti üzerinde uygulanan makine öğrenmesi sınıflandırma algoritmaları ile gerçekleştirilmiştir.

MÜŞTERİ KAYIP (CHURN) ANALİZİ

Kelime olarak müşterinin mevcuttaki hizmet sağlayıcısını değiştirerek rakip işletmeye geçmesi anlamına gelen “Churn”, “Change” (değişim) ve “Turn” (dönüş) kelimelerinden türemiştir. Müşteri Kayıp Analizi, işletmelerin ellerindeki verilerinin detaylıca incelemesini yaparak müşteri kayıplarının neden ve ne ölçüde gerçekleştiğinin analizini yapabilmelerini sağlamaktadır. Böylece yüksek riskli müşterilerin belirlenmesiyle birlikte gelecekte olası müşteri kayıplarının minimum seviyede tutulması sağlanarak işletmelerin karlılık hedeflerine ulaşmalarına yardımcı olur.



MAKİNE ÖĞRENMESİ SINIFLANDIRMA ALGORİTMALARI

Makine öğrenmesinin yaygın problemlerinden birisi olan sınıflandırma problemlerinde veri noktalarının belirli yöntemlerle sınıflara ayrılması amaçlanır. Kategorik hedef değişkenler üzerinde denetimli öğrenme algoritmaları kullanılarak oluşturulan modeller ile sınıflandırma gerçekleştirilmektedir.

Lojistik Regresyon: Bağımlı ve bağımsız değişkenler arasındaki ilişkiyi doğrusal bir modelle inceleyerek sınıflandırma yapar. Sigmoid fonksiyonu ile bağımlı değişkeni, 0 ile 1 arası olasılık değerlerine dönüştürür.

Naive Bayes: Kullandığı Bayes Teoremi ile her bir örneğin ait olduğu sınıfı koşullu olasılıklar üzerinden belirleyen olasılık temelli bir algoritmadır.

K-En Yakın Komşu: Veri noktalarını, belirli ölçüm kriterleri ile belirlenen en yakın k komşusuna göre sınıflandırır.

Destek Vektör Makineleri: Veri setini sınıflara ayıracak optimum hiperdüzlemin bulunması amaçlanmaktadır. Optimum hiperdüzlem, marjin mesafesinin maksimum tutulmasıyla bulunur. Doğrusal olmayan durumlar için farklı Kernel değerleri kullanılmaktadır.

Yapay Sinir Ağları: İnsan sinir hücrelerinden esinlenerek geliştirilmiş olup giriş, gizli ve çıkış katmanlarından oluşan bir yapıya sahiptir.

Karar Ağaçları: Veri ilişkilerini belirli karar kuralları yardımı ile hiyerarşik yapıda modelleyerek sınıflandırma yapmaktadır. Birbirlerine dallar ile bağlanan kök, iç ve yaprak düğümlerinden oluşur.

Rasgele Orman: Veri setinin farklı alt kümeleri üzerinde eğitilmiş birden fazla karar ağacı yapısı ile çalışmaktadır. Her bir ağaç tarafından elde edilen tahminler genellenerek sonuca oy çokluğu ile karar verilir.

GBM: Ağaç temelli bir algoritma olan GBM, ardışık modeller ile çalışan topluluk öğrenmesi algoritmasıdır. Her bir model kendinden bir önceki modelin hatalarına odaklanarak öğrenme işlemini gerçekleştirmektedir.

XGBoost: GBM algoritmasının optimize edilmiş bir versiyonu olan XGBoost, GBM’e kıyasla daha hızlıdır. Kayıp değerler ile çalışabilme özelliğine sahiptir.

LightGBM: GBM tabanlı bir karar ağacı algoritması olan LightGBM, yüksek gradyana sahip örneklere odaklanarak çalışır. Böylece daha düşük hata oranlarına sahip modeller kurulmaktadır.

CatBoost: Kategorik veriler için geliştirilen GBM algoritmasıdır. Sahip olduğu otomatik veri işleme özelliği sayesinde ön işleme aşamasında One-Hot Encoding yönteminin kullanımını gerektirmemektedir.

SINIF DENGESİZLİĞİ PROBLEMİ

Makine öğrenmesi algoritmaları, veri setinin eşit dağıldığı varsayımı üzerinden çalışmaktadır. Bu nedenle algoritmalar, veri setinin dengesiz sınıf dağılımına sahip olduğu durumda yeterli öğrenim yapılamaması nedeniyle sağlıklı sonuçlar verememektedir. Sınıf dengesizliğinin varlığı durumunda Veri Toplama, Tek-Sınıf Sınıflandırma, Aşırı Örnekleme, Eksik Örnekleme gibi çeşitli yöntemler uygulanarak bu problem çözülebilmektedir. Bu çalışmada, SMOTE ve ADASYN aşırı örnekleme yöntemleri kullanılmıştır. SMOTE ve ADASYN, azınlık sınıfı üzerinden gerçekleştirdiği sentetik veri üretimi ile azınlık sınıfı örnek sayısının artmasını sağlamaktadır.

DEĞERLENDİRME ÖLÇÜTLERİ

Sınıf dengesizliği probleminin varlığı durumunda, modelin performansını yalnızca doğruluk oranı ile ölçmek yanıltıcı sonuçlar vermektedir. Sonuçların diğer değerlendirme ölçütleri ile birlikte yorumlanması gerekmektedir.

Gerçek Değerler	Tahmin Edilen Değerler: Pozitif	Tahmin Edilen Değerler: Negatif
Pozitif	TP: Doğru, Pozitif	FN: Yanlış, Negatif
Negatif	FP: Yanlış, Pozitif	TN: Doğru, Negatif

Doğruluk: Kurulan modelin ne kadar doğru çalıştığını belirtir. Doğru değerlerin bütün değerlerin toplamına oranı ile bulunur.

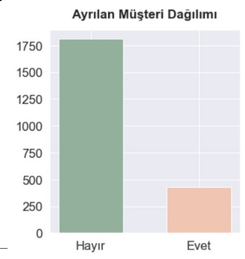
Kesinlik: Pozitif tahminlerin gerçekten ne kadar pozitif olduğunu belirtir. Doğru pozitif tahminlerin pozitif tahminlere oranı ile bulunur.

Duyarlılık: Gerçek pozitiflerin ne kadarının doğru tahmin edildiğini belirtir. Doğru pozitif tahmin edilen örneklerin gerçek tüm pozitif miktarına oranı ile bulunur.

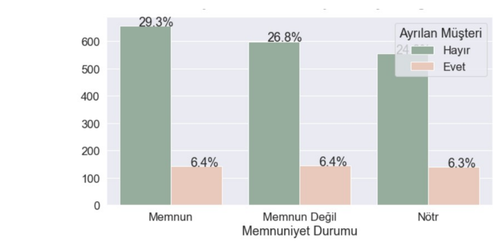
F1 Skoru: Kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması ile hesaplanır.

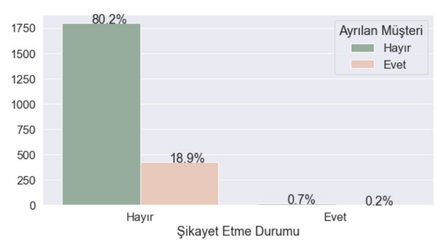
VERİ SETİ

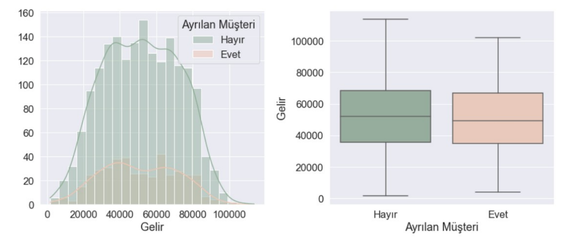
24 özniteliğe sahip 2240 müşterinin detaylı alışveriş bilgilerini içeren veri seti kullanılmıştır. Veri seti, aykırı ve eksik değerler bulundurmakta olup çarpık dağılıma sahip değişkenler bulundurmaktadır. Ayrılan müşteri grubunu belirten “1” grubunun 428 müşteriye ve “0” grubunun toplamda 1812 müşteriye sahip olduğu gözlemlenerek sınıf dengesizliği probleminin varlığı da bulunmuştur.

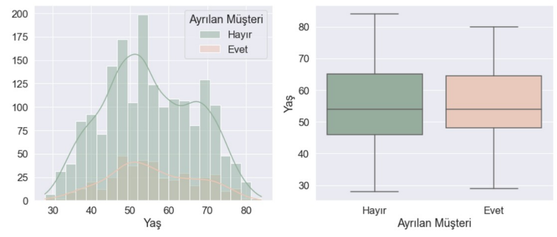












Ön işleme aşamasında, önemsiz öznitelikler ve aykırı değerlere sahip olduğu bilinen örnekler silinerek 22 özniteliğe ve 2236 müşteriye sahip veri seti üzerinden çalışmaya devam edilmiştir. Eksik değerler ortalama ile doldurulmuş ve logaritma transformasyonu yapılarak çarpıklık problemi çözülmüştür. Sınıflandırma algoritmalarının kullanılabilmesi için One-Hot Encoding yöntemi uygulanarak verinin kodlanması sağlanmıştır. Kullanılan Standart Scaler standardizasyon yöntemi ile özellik ölçeklendirme yapılmıştır.

BULGULAR

ALGORİTMALAR	DEĞERLENDİRME ÖLÇÜTLERİ						
	DOĞRULUK ORANI (%)	KESİNLİK (%)		DUYARLILIK (%)		F1 SKORU (%)	
		0	1	0	1	0	1
LOJİSTİK REGRESYON	0.80	0.80	0.00	1.00	0.00	0.89	0.00
NAIVE BAYES	0.80	0.80	0.50	0.89	0.00	0.89	0.03
KNN	0.75	0.80	0.16	0.91	0.07	0.85	0.10
SVC	0.80	0.80	0.30	0.99	0.02	0.89	0.04
RBF SVC	0.80	0.80	0.00	1.00	0.00	0.89	0.00
YAPAY SİNİR AĞLARI	0.76	0.80	0.16	0.93	0.05	0.86	0.08
CART	0.70	0.82	0.26	0.82	0.26	0.82	0.26
RASGELE ORMAN	0.80	0.80	0.43	0.99	0.02	0.89	0.04
GBM	0.79	0.80	0.27	0.98	0.03	0.88	0.05
XGBOOST	0.78	0.81	0.26	0.95	0.07	0.87	0.11
LIGHTGBM	0.78	0.80	0.25	0.96	0.05	0.88	0.09
CATBOOST	0.80	0.80	0.25	0.99	0.01	0.89	0.01

Çeşitli veri ön işleme adımlarından geçen veri seti üzerinden kurulmuş olan model sonuçlarına bakıldığında, çoğu algoritmanın %80 doğruluk oranlarına sahip olduğu ve en düşük oranın CART algoritması tarafından verildiği görülmüştür. Ancak bu oranlar kesinlik, duyarlılık ve F1 Skoru değerlendirme ölçütleri ile birlikte incelendiğinde verideki dengesizlik problemi nedeniyle “1” grubu (müşterinin terk etmesi durumu) için başarılı sonuçlar elde edilemediği gözlemlenmiştir. Fazla örneğe sahip olan “0” grubu için öğrenme işleminin daha başarılı geçmesi ve buna karşılık “1” grubu için yeterli öğrenimin gerçekleştirilememesi nedeniyle elde edilen rasgele tahmin başarıları, genel doğruluk oranları üzerinde yanılsamaya neden olarak yüksek sonuçların elde edilmesine sebep olmuştur.

ALGORİTMALAR	DEĞERLENDİRME ÖLÇÜTLERİ						
	DOĞRULUK ORANI (%)	KESİNLİK (%)		DUYARLILIK (%)		F1 SKORU (%)	
		0	1	0	1	0	1
LOJİSTİK REGRESYON	0.48	0.81	0.20	0.46	0.50	0.58	0.30
NAIVE BAYES	0.42	0.80	0.20	0.37	0.62	0.50	0.30
KNN	0.48	0.78	0.17	0.50	0.43	0.61	0.25
SVC	0.55	0.80	0.20	0.59	0.41	0.68	0.27
RBF SVC	0.46	0.79	0.19	0.44	0.53	0.56	0.28
YAPAY SİNİR AĞLARI	0.69	0.80	0.20	0.82	0.18	0.81	0.19
CART	0.64	0.81	0.23	0.72	0.34	0.76	0.27
RASGELE ORMAN	0.78	0.81	0.34	0.95	0.11	0.88	0.16
GBM	0.78	0.81	0.32	0.95	0.10	0.87	0.15
XGBOOST	0.74	0.80	0.17	0.90	0.08	0.85	0.11
LIGHTGBM	0.75	0.80	0.22	0.91	0.10	0.86	0.14
CATBOOST	0.78	0.81	0.31	0.95	0.10	0.87	0.15

Sınıf dengesizliği problemini çözmek amacıyla uygulanan SMOTE aşırı örnekleme yöntemi ile toplamda 3212 örneğe sahip veri seti üzerinde çalışılmıştır. SMOTE, doğruluk oranlarında genel bir düşüşe neden olmuştur. Rasgele Orman ve GBM, CatBoost algoritmaları en yüksek doğruluk oranlarına sahip iken, en düşük oran Naive Bayes algoritmasına aittir. Kesinlik oranlarında “1” grubu için genel bir artış sağlanamamış, aksine bazı algoritmalarda düşüş gözlemlenmiştir. “0” grubu için ise ham veri üzerinden alınan benzer sonuçlar alınmaya devam edilmiştir. Duyarlılık ve F1 Skorlarında “1” grubu için genel bir yükseliş gözlemlenirken “0” grubu için bu ölçütlerin değerlerinin düştüğü gözlemlenmiştir.

ALGORİTMALAR	DEĞERLENDİRME ÖLÇÜTLERİ						
	DOĞRULUK ORANI (%)	KESİNLİK (%)		DUYARLILIK (%)		F1 SKORU (%)	
		0	1	0	1	0	1
LOJİSTİK REGRESYON	0.52	0.81	0.21	0.52	0.53	0.64	0.31
NAIVE BAYES	0.41	0.78	0.18	0.38	0.56	0.51	0.28
KNN	0.43	0.78	0.18	0.40	0.54	0.53	0.27
SVC	0.65	0.80	0.20	0.75	0.26	0.77	0.22
RBF SVC	0.57	0.80	0.20	0.61	0.40	0.69	0.27
YAPAY SİNİR AĞLARI	0.69	0.80	0.20	0.82	0.19	0.81	0.19
CART	0.61	0.80	0.19	0.69	0.29	0.74	0.23
RASGELE ORMAN	0.78	0.81	0.35	0.95	0.11	0.87	0.17
GBM	0.78	0.81	0.32	0.95	0.09	0.87	0.14
XGBOOST	0.75	0.80	0.22	0.91	0.11	0.85	0.14
LIGHTGBM	0.78	0.81	0.26	0.95	0.07	0.87	0.11
CATBOOST	0.77	0.81	0.29	0.94	0.09	0.87	0.14

ADASYN yöntemi ile toplamda 3146 örneğe sahip veri seti üzerinde çalışılmıştır. Uygulanan ADASYN yöntemi sonrasında doğruluk oranlarında düşüş gözlemlenmiştir. Rasgele Orman, GBM ve LightGBM algoritmalarının en yüksek doğruluk oranlarına, en düşük doğruluk oranının ise Naive Bayes algoritmasına ait olduğu görülmüştür. Kesinlik oranlarında “1” grubu için genel bir artış sağlanamamıştır, “0” grubu için ise yüksek oranlar alınmaya devam edilmiştir. Duyarlılık ve F1 Skorlarında “0” grubu için genel bir düşüş gözlemlenirken “1” grubu için bu değerlerin yükseldiği gözlemlenmiştir.

Sınıf dengesizliği problemine çözüm olarak uygulanan SMOTE ve ADASYN aşırı örnekleme yöntemlerinin, ayrılan müşteri grubunu belirten “1” grubu için yapılan tahmin sonuçlarını benzer şekilde iyileştirdiği gözlemlenmiştir. Ancak, bu iyileştirmelere rağmen kurulan modeller yeterince başarılı sonuçlar elde edilememiştir. Genel olarak en iyi sonuçlar, kullanılan ağaç tabanlı makine öğrenmesi algoritmaları tarafından elde edilmiştir.

KAYNAKÇA

[1] James, G., Witten, D., Hastie, T., & Tibshirani, R., Taylor J. (2023). An Introduction to Statistical Learning with Applications in Python. Springer.
[2] Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow. O'Reilly Media.
[3] Alpaydin, E. (2014). Introduction to Machine Learning. MIT Press.

YTÜ FEN EDEBİYAT FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ

MAKİNE ÖĞRENMESİ VE SPORDA VERİ ANALİTİĞİ: YAPAY ZEKA TABANLI METRİKLERİN MAKİNE ÖĞRENMESİ İLE PERFORMANS ANALİZİ VE SENARYO TAHMİNİNDE KULLANILMASI

Can İPEK – 20023072
Danışman: Arş. Gör. Dr. Coşkun PARİM

ÖZET

Modern analiz metotları dendiğinde akla ilk gelen kavram olan makine öğrenmesi, günümüzde sürekli olarak değişen oyun yapısına rağmen başarıya giden yolun en hızlı tahminleri yaparak en erken kararları almaktan geçtiği spor sektörünün önemli bir parçası haline gelmiştir. Kısa bir tarihe sahip olan spor analitiği ortaya çıktığı ilk dönemden itibaren yükselen ivme ile gelişmişken, son dönemlerde yapay zekânın gelişimiyle birlikte ortaya çıkan metriklerle yeni bir anlam kazanmaktadır. Bu çalışmada kullanılan veri, İngiltere Premier Ligi 2023-2024 sezonunda oynanmış 380 maçı kapsamaktadır. Veri seti sezon geneli oyuncu ve takım istatistiklerinden dakikaya kadar değişen granülitlerde veri içerir. İçerdiği değişkenler arasında şut tipi, atak şekli gibi kategorik veya gol sayısı, defansif aksiyon başına pas sayısı, gibi sayısal istatistiklerin yanı sıra, son dönemlerde popülaritesi artan xG, xA gibi yapay zekâ tabanlı yeni nesil metrikler de bulunmaktadır. Görselleştirmeler ile veri setindeki bazı değişkenlerin farklı kırınımlardaki ilişkileri ile ilgili çeşitli çıktılar elde edilmiş, performans analizleri ile ilgili uygulanabilir yenilikler tartışılmış ve üzerine yorumlar yapılmıştır. Sonrasında ise makine öğrenmesi kullanılarak lineer regresyon, rastgele orman, destek vektör makinesi ve yapay sinir ağları modelleri kullanılmış ve elde edilen verilere göre en doğru sonuçları veren modeller seçilmiştir. Tüm bu analizlerin teknik anlamda karar almaya sağlayabileceği katkılar üzerine tartışılmıştır. Uygulamadaki tüm çalışmalarda Python kullanılmıştır.

Sporda Veri Analitiği

Spor veri analitiği, sporcular, antrenörler ve hakemler hakkında saha düzeyinde analizleri, yöneticilerin kararlarını ve ekonomi ve psikoloji alanındaki analizleri içerir. [2] Bazı popüler spor branşlarında, saha düzeyinde analiz ile ilgili gelişmeler aşağıdaki gibidir.

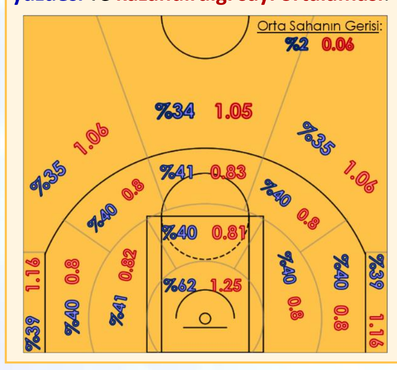
Beyzbol

Oyun yapısı itibariyle veri analitiğinin belirleyici konumda olduğu bir spor dalı olan beyzbol, aynı zamanda veri analitiğinde 1960'lı yıllara dayanan geçmişi ile bu alanda en eski spor dallarından biridir. 2003 yılında Michael Lewis'in kaleme aldığı, beyzbol kulübü Oakland Athletics takımında görev yapan Billy Beane'in, veri odaklı yönetimini konu alan '*Moneyball: The Art of Winning an Unfair Game kitabı*' ve bu kitabın 2011 yılında sinemaya uyarlanmasının yarattığı etki, sporda veri analitiği üzerine yapılan çalışmaların artmasında etkili olmuştur. Beane, oyundaki bazı ilişkileri dikkate alarak kurduğu düşük bütçeli kadrosuyla büyük başarılara imza atmıştır [5].

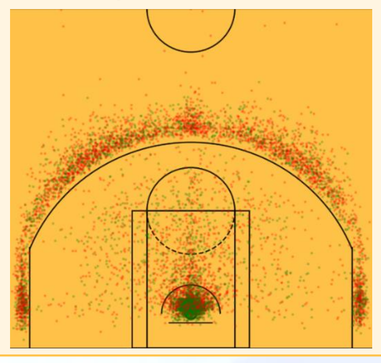
Basketbol

Basketbolda 90'lı yıllarda başkalaşım geçiren ve hız kazanan veri analitiği çalışmalarının, özellikle son yıllarda etkisi artmıştır. Houston Rockets takımında uzun yıllar yönetici görevi üstlenmiş Daryl Morey'in, orta mesafeli atışlardan kaçınmayı öneren veri odaklı yaklaşımı bazı yorumcular ve taraftarlar tarafından eleştirilmiş olsa da Rockets önemli başarılar elde etmiştir [7]. 2017-2018 sezonunda Rockets, sakatlıklara rağmen Batı Konferansı finallerine yükselmiş ve aynı sezon 3 sayılık atış rekorunu kırarak Batı Konferansı'nı 1. sırada tamamlamıştır.

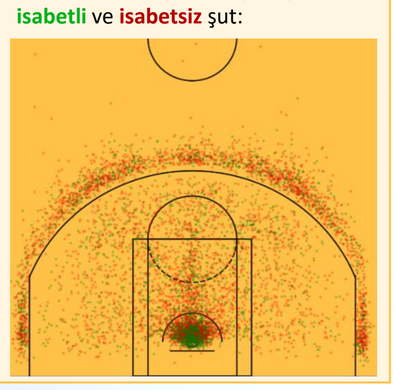
1997 – 2020 yılları arasında NBA'de atılmış tüm şutların bölgesel **isabet yüzdesi** ve **kazandırdığı sayı ortalaması**:



2017-18 sezonunda Houston Rockets'ın attığı 8 bin **isabetli** ve **isabetsiz** şut:



2017-18 sezonunda diğer takımlar arasından rastgele seçilmiş 8 bin **isabetli** ve **isabetsiz** şut:



'Sıcak El' Tartışması

1985 yılında yapılan bir çalışmada, basketbol oyuncularının ardışık başarılı atışlar yapmasının sonraki atışların başarı olasılığını artırmadığı öne sürülmüştür [8]. Bu çıkarım bazı basketbol takımlarının şutları ile ilgili toplanan verileri analiz ederek yapılmış ve aksi uzun yıllar bilimsel olarak ispat edilemese de, izleyici ve araştırmacılar arasında tartışma konusu olmuştur. Bu çalışma, psikoloji ve davranış ekonomisi gibi farklı alanlardaki çalışmalara dahi konu olmuştur [10].

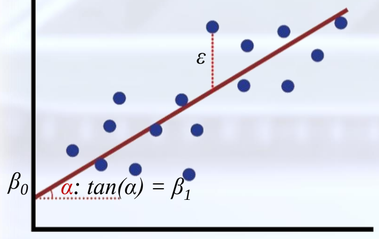
Futbol

1950'lerde Charles Reep'in atak öncesi yapılan pas sayısı ile atakların gol ile sonuçlanması arasındaki ilişkiyi saptayarak kurguladığı oyun modelinin geçerliliğini uzun yıllar sürdürmesiyle başlayıp, 90'lı yıllarda ve sonrasında çok kez değişip gelişen futbolda veri analitiği, geçmişte olduğu gibi bugün de farklı fikirlerin geçerli olduğu dönemler geçirmeye devam ederken, son yıllarda yapay zeka tabanlı metriklerin kullanılmaya başlamasıyla yeni bir anlam kazanmıştır [9].

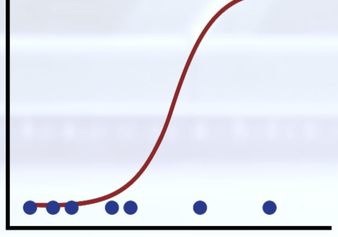
Makine Öğrenmesi

Makine öğrenmesi, bilgisayar bilimi ile istatistiğin kesiştiği noktada, yapay zekâ ve veri biliminin merkezinde yer alan, günümüzün en hızlı büyüyen teknik alanlarından biridir. Makine öğrenmesi, bilgisayara bir işlem için özel olarak kodlanmadan bu işlemi yapabilme becerisinin kazandırılmasını sağlayan çalışma alanıdır [6]. Makine öğrenmesi, denetimli, denetimsiz, yarı denetimli ve pekiştirmeli öğrenme olarak dörde ayrılır. Aşağıda bir denetimli öğrenme sınıfı olan regresyonda kullanılan bazı modellere örnekler verilmiştir.

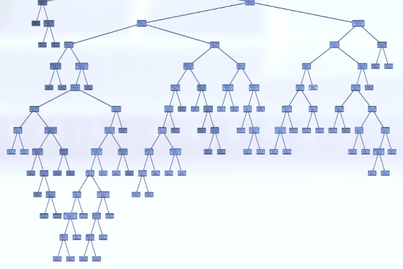
Lineer Regresyon



Lojistik Regresyon



Karar Ağaçları

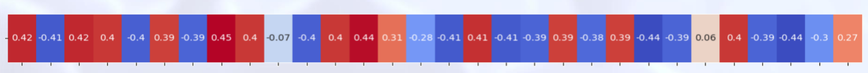


Uygulama

Uygulama bölümünde, İngiltere Premier Ligi 2023-2024 sezonuna ait veriler kullanılarak, makine öğrenmesi modelleri kullanılarak olabildiğince yüksek güçte çıkarımsal, tahmine yönelik sonuçlar elde edilmeye çalışılacak, xG istatistiği üzerine tartışılacak ve performans analizi için ne gibi çıkarımlar yapılabileceği incelenecektir.

1) Makine Öğrenmesi ile Senaryo Tahmini

Aşağıda '**Puan Beklentisi'** değişkeninin tüm değişkenler ile arasındaki korelasyon katsayıları ve değişkenlerden 'iç/dış saha', 'rakip lig geneli puan beklentisi', 'rakip lig geneli gol beklentisi' ve 'takım lig geneli puan beklentisi' değişkenleri ile tahmin edilmesi için kurulan regresyon modellerinin sonuçları mevcuttur.



MODEL	LİNEER REGRESYON		REGRESYON AĞACI		DESTEK VEKTÖR MAKİNESİ		YAPAY SİNİR AĞLARI	
Veri	Test	Eğitim	Test	Eğitim	Test	Eğitim	Test	Eğitim
Ortalama Kare Hatası	0.38	0.42	0.47	0.4	0.4	0.42	0.37	0.42
Ortalama Mutlak Hata	0.5	0.54	0.55	0.51	0.49	0.5	0.49	0.54
Karekök Ortalama Kare Hatası	0.62	0.65	0.69	0.63	0.63	0.65	0.61	0.65
R^2	**0.49**	**0.46**	**0.38**	**0.49**	**0.48**	**0.46**	**0.51**	**0.46**

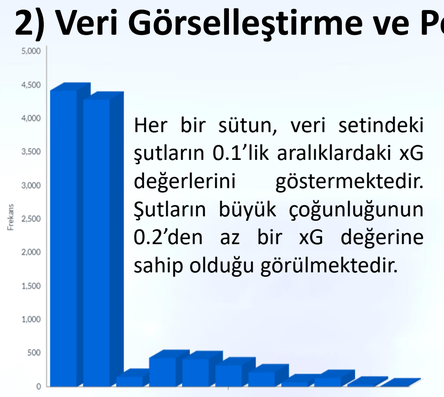
Diğer değişkenlerin tahmini için aynı metotla seçilen modeller ve test verisinden aldıkları R^2 değerleri:

Bağımlı Değişken	Seçilen Model	Test R^2
Kaleye Yakın Pas Sayısı	Yapay Sinir Ağları	**0.49**
Gol Beklentisi **(xG)**	Destek Vektör Makinesi	**0.35**
Gol Beklentisi Toleransı **(xGA)**	Lineer Regresyon	**0.34**

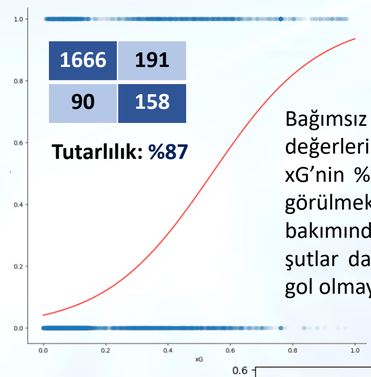
xG İstatistiği

'xG', bir futbol takımının veya oyuncunun belirli bir pozisyonda gol atma olasılığını ifade eder. Bu model, bir şutun gol olma olasılığını değerlendirmek için şutun mesafesi, açısı, şutun yapıldığı pozisyon gibi faktörleri kullanır. Hesaplama, genellikle lojistik regresyon veya diğer makine öğrenimi modelleriyle yapılır [3][4]. Bir şutun 0.2 xG değerine sahip olması, bu şut 100 kez denendiğinde ortalama 20 tanesinin gol ile sonuçlanacağının tahmin edildiği anlamına gelir.

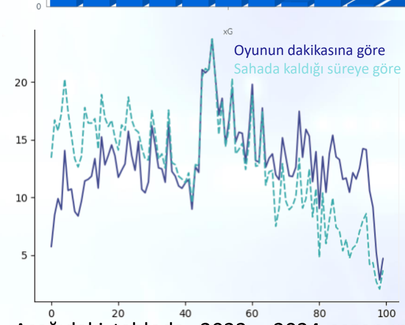
2) Veri Görselleştirme ve Performans Analizi



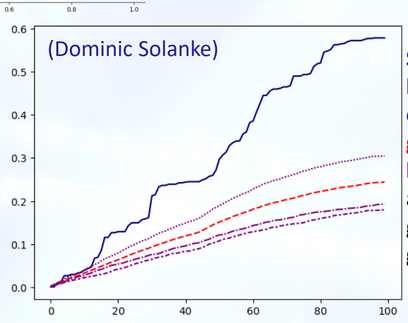
Her bir sütun, veri setindeki şutların 0.1'lik aralıklardaki xG değerlerini göstermektedir. Şutların büyük çoğunluğunun 0.2'den az bir xG değerine sahip olduğu görülmektedir.



Bağımsız ve bağımlı değişken olarak sırasıyla xG ve gol değerlerinin kullanıldığı lojistik regresyon modeliyle, xG'nin %87 oranında gol olma durumunu tanımladığı görülmektedir. Buna karşın, gol olmayan şutlar xG bakımından sağa çarpık dağılım gösterirken, gol olan şutlar daha tekdüze dağılmışlardır. Bu yüzden model, gol olmayan şutları tahmin etmede daha başarılıdır.



Oyuncuların xG katkısının maçın dakikasına ve oyunda kalma süresine göre değişimi yer yandaki grafikte yer almaktadır. Oyuncuların sahada kalma süresi arttıkça ortalama xG katkısının genel ortalamanın altına düştüğü görülmektedir.

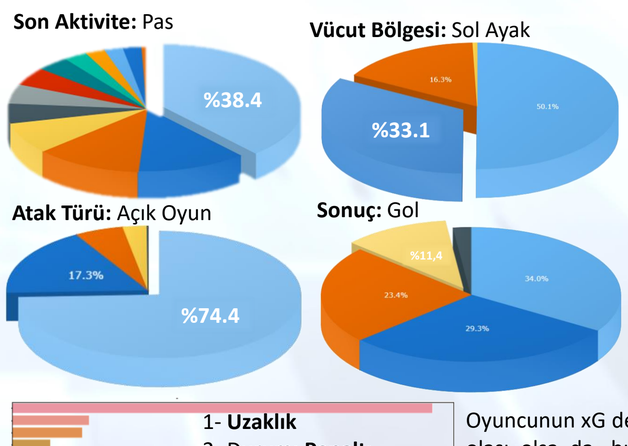


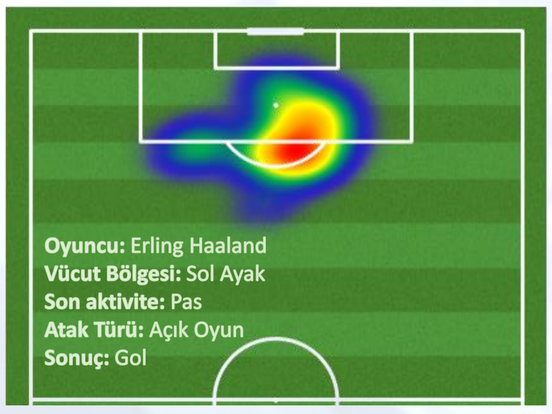
Seçilen herhangi bir oyuncunun ortalama xG artışı, genel ve **mevki bazında** ortalama artışlar yandaki grafikte görselleştirilmiştir.

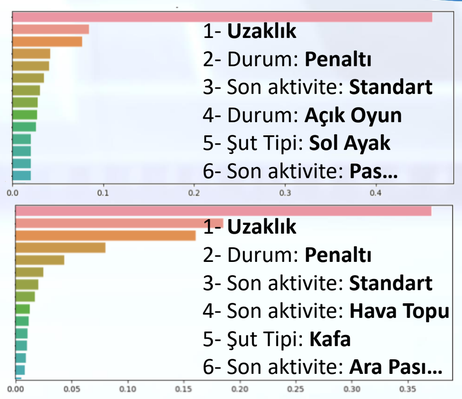
Aşağıdaki tabloda, 2023 - 2024 sezonunu 27 gol ile en fazla gol atan oyunu olarak tamamlayan Erling Haaland'ın şutlarının toplam xG değerinin, gollerinin toplamından 4,65 gol daha fazla olduğu görülmektedir.

Oyuncu Adı	Gol	xG	Takım	Gol - xG Farkı
Erling Haaland	27	31.65	Manchester City	**-4.65**

Bu, oyuncunun elde ettiği şut imkanlarını ortalamanın altında bir başarıyla değerlendirdiğini gösterir. Yani yüksek gol sayısı, pozisyonları gole çevirebilme kabiliyetinden daha çok, fazla pozisyona girme becerisiyle ilişkilidir. Bu durumun sebepleri çeşitli analiz ve görselleştirme metotlarıyla incelenebilir. Aşağıda örnek olarak ligdeki tüm şutların, 4 farklı kategori grubundan seçilmiş birer kategorinin, tüm kategorilerin frekansına göre dağılımının pasta grafiği ve ilgili oyuncu Erling Haaland'ın bu kategorilere ait şutlarının ısı haritası ile grafiği mevcuttur.







Oyuncunun xG değerinin gol sayısından daha yüksek olması oyuncu ile ilgili olması olası olsa da, bu sebep birçok farklı tipte oyuncunun attığı şutların verisi ile eğitilmiş xG modelinin ilgili oyuncunun gol olasılıklarını tahmin etmede genelde olduğu kadar başarılı olamaması da olabilir.

Yandaki grafiklerde ilgili oyuncu Erling Haaland'ın attığı şutların yukarıdakine benzer çeşitli özelliklerini bağımsız değişken olarak kullanarak oluşturulmuş, xG istatistiğini tahmin etmek için rastgele orman regresyonu ve gol değişkenini tahmin eden rastgele orman sınıflandırması algoritmalarının bağımsız değişkenlerinin önem sıralaması mevcuttur. Buradan, Haaland'ın attığı şutların gol olma olasılığını etkileyen faktörlerin, xG istatistiğinin derecesini etkileyen faktörlerden farklı olduğu görülmektedir. Sonuç olarak antrenman verileri gibi verileri kullanarak xG metriklerinin oyunculara özel olarak tasarlanması, oyuncu bazında tutarlılığı arttıracak ve analizleri geliştirecektir.

KAYNAKÇA

[1] Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.). O'Reilly Media.

[2] Morgulev, E., Azar, O.H. & Lidor, R. Sports analytics and the big-data era. Int J Data Sci Anal 5, 213–222 (2018). https://doi.org/10.1007/s41060-017-0093-7

[3] Hewitt, J. H., & Karakuş, O. (2023). A Machine Learning Approach for Player and Position Adjusted Expected Goals in Football (Soccer). *arXiv preprint*. https://doi.org/10.48550/arXiv.2301.13052

[4] Mead, J., O'Hare, A., & McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. *PLOS ONE, 18*(4), e0282295. https://doi.org/10.1371/journal.pone.0282295

[5] Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. W. W. Norton & Company.

[6] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3(3), 210-229.

[7] web.archive.org/web/20150325075449/http://www.houstonpress.com/2007-11-01/news/rocket-science/

[8] Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*(3), 295–314. https://doi.org/10.1016/0010-0285(85)90010-6

[9] Pollard, R. (2002). Charles Reep (1904-2002): pioneer of notational and performance analysis in football. *Journal of Sports Sciences*, *20*(10), 853–855. https://doi.org/10.1080/026404102320675684

[10] Rabin, Matthew and Vayanos, Dimitri, The Gambler's and Hot-Hand Fallacies: Theory and Applications (February 2007). CEPR Discussion Paper No. 6081, Available at SSRN: https://ssrn.com/abstract=1004563

İSTATİSTİK BÖLÜMÜ

MAKİNE ÖĞRENMESİ İLE MEME KANSERİ TAHMİNİ

Ceren ACAR 18023603

Danışman: Doç. Dr. Reşit ÇELİK

ÖZET

Meme tümörü bulunan hastaların özelliklerinin bulunduğu bir veri seti ile çalışılmıştır. Çalışmanın amacı hastalardaki tümörün kötü huylu (kanser) olup olmadığını doğru bir şekilde tahmin etmektir. Lojistik Regresyon, Naive Bayes, K-En Yakın Komşu (KNN), Destek Vektör Sınıflandırıcısı (SVM), Yapay Sinir Ağları (ANN), Sınıflandırma Ağaçları (CART), Random Forest, Gradient Boosting Machines (GBM), XGBoost, LightGBM ve CatBoost algoritmaları kullanılmıştır. Tüm modeller karşılaştırılıp en iyi model sıralaması yapılmıştır. Belirlenen en iyi model ile hastada bulunan tümörün kötü huylu (kanser) olup olmadığını tahmin eden bir fonksiyon yazılmıştır.

Meme Kanseri

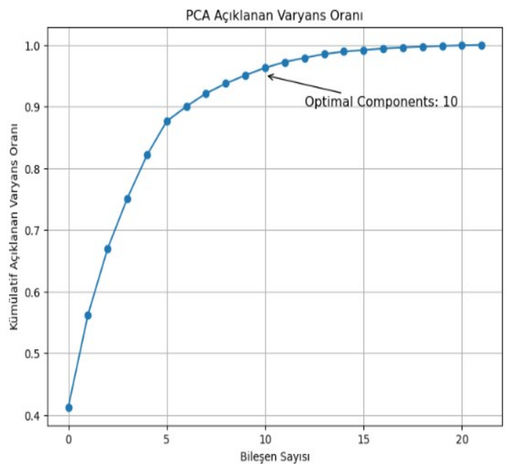
Meme kanseri, meme dokusunun içinde veya çevresinde büyüyen kötü huylu bir tümördür. Kadınlarda en yaygın görülen kanser türü meme kanseridir. Nadiren erkeklerde de görülebilir.
Yapılan araştırmalar, her yıl ABD'deki kadınlarda görülen tüm yeni kanser vakalarının %30'unun meme kanseri olarak teşhis edildiğini göstermektedir, buna karşın ABD'de tüm meme kanserlerinin %1'inden azı erkeklerde görülmektedir.

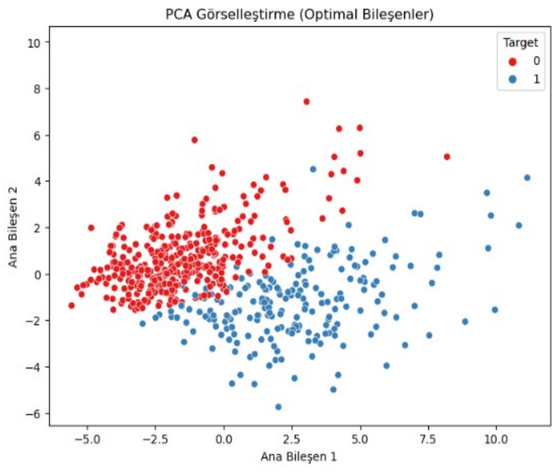
Veri Seti Hakkında

Veri seti meme tümörü bulunan hastaların özelliklerinden meydana gelmektedir.
Çalışmada kullanılan veri seti "Kaggle" web sitesinden alınmıştır.
Veri setinde 569 gözlem, 31 değişken vardır.

Temel Bileşen Analizi (PCA)

Temel Bileşen Analizi (PCA), veri setinin boyutunu azaltmak için kullanılan istatistiksel bir tekniktir. PCA'nın temel amacı veri setindeki karmaşıklığı azaltmaktır. Bu teknikte veri setinin varyansı maksimize edilerek yeni değişkenler (bileşenler) elde edilir. Bu yeni bileşenler orijinal değişkenlerin birleşimidir ve veri setindeki bilgiyi daha az sayıda bileşenle temsil etmeye çalışır. PCA sürecinde veri setindeki değişkenler arasındaki korelasyonlar dikkate alınır ve bu korelasyonlar temel alınarak yeni bileşenler oluşturulur. Temel Bileşen Analizi veri seti hakkında diğerlerinden daha az bilgi taşıyan bileşenleri atar. PCA çok boyutlu bir veride temel özellikleri yakalar ve bilgi kaybını en aza indirerek veriyi daha az sayıda özellikle temsil eder. Bu, veri setindeki karmaşıklığı azaltır ve verinin daha iyi anlaşılmasını sağlar.

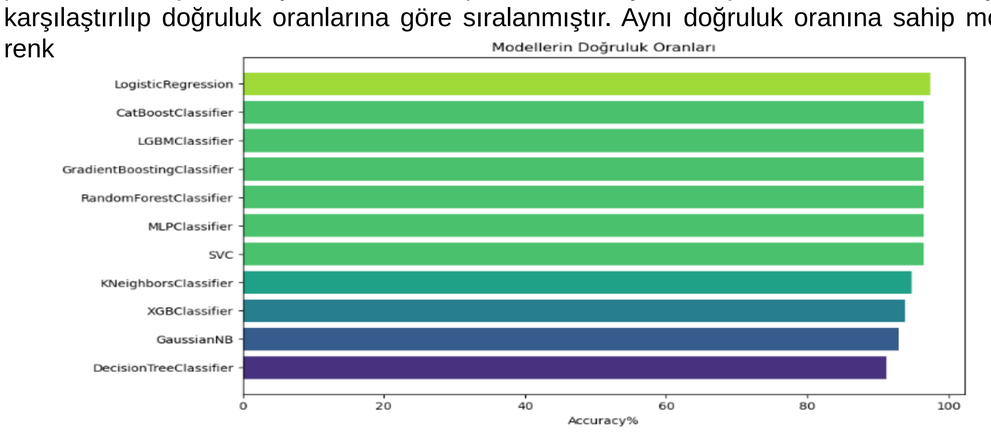




PCA'ya göre veri setinde en iyi bileşen sayısının 10 olduğunu görmekteyiz. Bu, veri setindeki değişkenliğin büyük bir kısmının ilk 10 bileşen tarafından açıklandığını göstermektedir. İlk 10 bileşenin toplam varyansın büyük bir kısmını açıklaması, veri setindeki önemli özelliklerin büyük bir kısmını kapsadığı anlamına gelir. Grafiğe bakıldığında kanserli olan ve olmayan tümörlerin belirgin şekilde ayrıldığı görülmektedir. Kırmızı noktalar iyi huylu tümörü gösterirken mavi noktalar kötü huylu tümörü (kanser) göstermektedir. Grafikte kırmızı noktaların birbirine daha yakın ve mavi noktaların da birbirine daha yakın olduğu gözlemlenmiştir. Bu durum PCA bileşenlerinin farklı sınıfları birbirinden ayıran önemli değişkenleri yakaladığını ve sınıflar arasındaki farklılıkları temsil ettiğini gösterir.
569 gözlem ve 31 değişkenden oluşan veri setinde aşırı uyumu önlemek amacıyla Temel Bileşen Analizi (PCA) ile boyut indirgeme işlemi gerçekleştirilmiştir. Bu işlem sonucunda veri seti 23 değişken içerecek şekilde yeniden yapılandırılmıştır.

Modeller

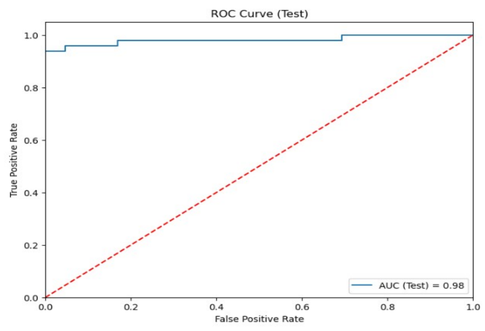
PCA sonrası oluşan yeni veri seti ile tüm hastaları %80 eğitim ve %20 test olarak ayırdıktan sonra aşağıda bulunan grafikteki modeller eğitilmiş ve test edilmiştir.
Bu çalışmada toplam 11 farklı model kullanılmıştır. Modellerin performansını değerlendirmek için parametreler ayarlanmış ve modeller optimize edilmiştir. En iyi modeli belirlemek için modeller karşılaştırılıp doğruluk oranlarına göre sıralanmıştır. Aynı doğruluk oranına sahip modeller aynı renk

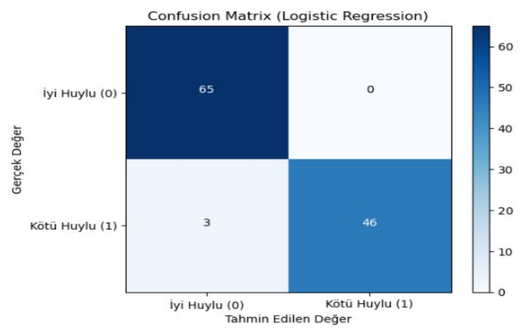


Bu grafiğe bakıldığında en yüksek doğruluk oranına sahip olduğu için Lojistik Regresyon modelinin en iyi model olduğu sonucuna varılmıştır. Bu bulgu Lojistik Regresyon modelinin daha iyi tahminler yapma eğiliminde olduğunu göstermektedir. Bu nedenle hastaların kanser olup olmadığı bu model ile tahmin edilmiştir.

Lojistik Regresyon Modeli

Lojistik Regresyon, bağımlı değişkenin iki kategorisi arasındaki ilişkiyi modellemek ve tahmin etmek için kullanılan bir makine öğrenimi tekniğidir. En iyi modelin 0.9736 doğruluk oranı ile Lojistik Regresyon olduğu saptanmıştır. Bu model kullanılarak Hata Matrisi (Confusion Matrix) ve ROC eğrisi grafiği oluşturulmuştur. ROC eğrisi modelin sınıflandırma performansını değerlendirmek için kullanılır. ROC bir olasılık eğrisidir ve altında kalan alan olan AUC ayrılabilirliğin derecesini veya ölçüsünü temsil eder. AUC değerinin yüksek olması ise modelin performansının iyi olduğunu gösterir. Grafik incelendiğinde modelin başarılı olduğu sonucuna varılmaktadır.





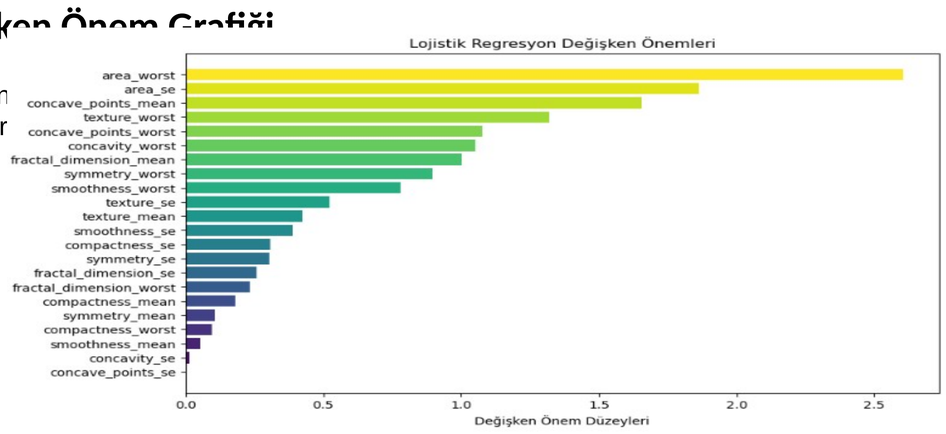
Öğrenme Eğrisi (Learning Curve)

Learning Curve grafiği, x ekseninde eğitim örneklerinin sayısını, y ekseninde ise modelin performansını (doğruluk oranı) gösterir. Learning Curve grafiği genel olarak modelin veri seti boyutunu artırdıkça eğitim ve test setlerindeki performansını ve ikisi arasındaki ilişkiyi gösterir. Grafikte görüldüğü irbirine yaklaşıyor. Dolayısıyla veride modelin veri setine iyi genelleme yaptı .



Değişken Önem Grafiği

Lojistik önemli değişken ts_mean özellikler



Sonuç

Bu çalışmada meme kanseri tahmini için hastaların %80'i eğitim için kullanıldı ve %20'si test için ayrıldı. Bu işlem 11 farklı model üzerinde gerçekleştirildi. Modellerin performansını iyileştirmek için parametre ayarlamaları yapıldı ve tüm modeller çalıştırıldı. Modeller karşılaştırıldığında en iyi modelin 0.9736 doğruluk oranı ile Lojistik Regresyon olduğu tespit edilmiştir. Bu model kullanılarak ROC eğrisi, öğrenme eğrisi ve değişken önem grafiği incelendi. Sonrasında Lojistik Regresyon modeli ile kanser tahmini yapmak için bir fonksiyon yazıldı.

KAYNAKÇA

[1] Uddin, S., Khan, A., Hossain, M. et al. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19, 281 (2019). https://doi.org/10.1186/s12911-019-1004-8
[2] Hajian-Tilaki K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian journal of internal medicine, 4(2), 627–635.
[3] Perlich, C. (2010). Learning Curves in Machine Learning.

İSTATİSTİK BÖLÜMÜ

VERİ MADENCİLİĞİ TEKNİKLERİ İLE DİYABET HASTALIĞININ TAHMİN EDİLMESİ

Çağla ÖZDEMİR 18023075

Danışman: Prof. Dr. Fatma NOYAN TEKELİ

Bu çalışmada, Veri Madenciliği Tekniklerinden sınıflandırma tekniği olan Lojistik Regresyon ve Karar Ağacı yöntemleri kullanılmıştır. 8 farklı değişkene bağlı olarak kişinin diyabet hastası olup olmadığı tahmin edilmeye çalışmıştır ve yöntemlerin karar vermedeki performansları değerlendirilmiştir.

VERİ MADENCİLİĞİ

VERİ MADENCİLİĞİ TANIMI

Veri madenciliğini, toplanan verilerden anlamlı bilgiler çıkarma, veri içerisinde gizli olan örüntüleri ve eğilimleri tespit etme ve çeşitli değişkenler arasında ilişkiler bulma amacıyla kullanılan bir yaklaşım olarak tanımlanır.

VERİ MADENCİLİĞİ MODELLERİ

Veri madenciliği uygulamalarında kullanılacak algoritmanın seçimi, verinin özelliklerine bağlı olarak değişebilir. Bu süreçte, veri madenciliği modelleri, kullanıcıların model seçimlerine yol göstererek istedikleri sonuca ulaşmalarına yardımcı olur.
Sınıflama ve Regresyon Modelleri :gelecekteki veri eğilimini öngörebilen analiz yöntemleridir.

-LOJİSTİK REGRESYON

Lojistik regresyon, bağımsız değişkenler arasındaki ilişkiyi ve bu ilişkinin bağımlı değişkene etkisini değerlendirir. Bu yöntem, bir lojistik fonksiyon kullanarak bağımlı değişkenin olasılığını tahmin etmek için verileri analiz eder. Lojistik fonksiyon, giriş verilerini alarak bir olasılık değeri üretir ve bu değeri bir sınıf etiketine dönüştürür.

-KARAR AĞACI

Bir veri kümesindeki öznitelikleri kullanarak bir hedef değişkeni tahmin etmek veya sınıflandırmak için bir ağaç yapısı oluşturur. Analiz aşamasında, karar ağaçları veri kümesindeki özniteliklerin değerlerine göre bir dizi karar kuralı oluşturur. Bu karar kuralı ağacın dallarını ve yapraklarını oluşturur. Her iç düğüm, bir öznitelik değerine göre veri kümesini bölerek bir karar verirken, yaprak düğümleri ise sonuçları temsil eder.

Metrikler: Karar ağacı ve lojistik regresyon yöntemlerinde karar verme için kullanılan bazı metrikler vardır.

Accuracy: Sistemde doğru olarak yapılan tahminlerin tüm tahminlere oranıdır. = (TP+TN) / (TP+TN+FN+FP)

Recall: Pozitif durumların ne kadar başarılı tahmin edildiğini gösterir.= TP / (TP + FN)

Precision: Pozitif olarak tahmin edilen bir durumdaki başarıyı gösteren durum. TP / (TP + FP)

F-Meaure: Recall ve Precision'ın harmonik ortalamasıdır .= 2*Precision*Recall / (Precision + Recall

VERİ SETİ

Değişkenler

-Pregnancies:Hamilelik sayısı (integer)
-Glucose:glukoz tolerans testindeki 2 saatlik plazma glukoz konsantrasyonu (integer)
-Blood Pressure: Diyastolik kan basıncı (mm Hg) (integer)
-Skin Thickness: Triseps cilt kat kalınlığı (mm) (integer)
-İnsulin: 2 saatlik serum insülini (mu U/ml) (integer)
-BMI: Vücut kitle indeksi (kg / (boyut m)^2) (float)
-Diabetes Pedigree Function: Diyabet soy fonksiyonu (float)
-Age: Yaş (yıl) (integer)
-Outcome: Sınıf değişkeni (1=diyabet hastası,0=diyabet hastası değil) (integer)

-TANIMLAYICI İSTATİSTİK

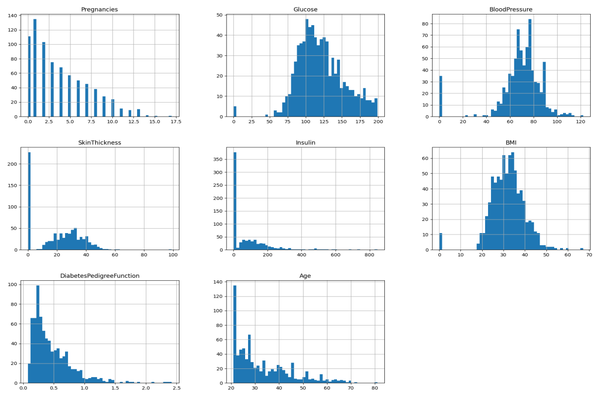
Descriptive Statistics

Variables	count	mean	std	min	25%	50%	75%	max
Pregnancies	768	4	3	0	1	3	6	17
Glucose	768	121	32	0	99	117	140	199
BloodPressure	768	69	19	0	62	72	80	122
SkinThickness	768	21	16	0	0	23	32	99
Insulin	768	80	115	0	0	30	127	846
BMI	768	32	8	0	27	32	37	67
DiabetesPedigreeFunction	768	0	0	0	0	0	1	2
Age	768	33	12	21	24	29	41	81
Outcome	768	0	0	0	0	0	1	1

Değişkenlerin ortalaması, standart sapması, minimum değeri, alt çeyreği, medyanı, üst çeyreği ve maksimum değerleri.

-HİSTOGRAM

Histogram grafiklerine göre 'BMI','BloodPressure' Değişkenlerinin normal dağıldığını, 'Glucose','Insulin', 'SkinThickness' değişkenlerinin normal dağılmadığı söylenebilir. Bu çalışmaya bağlı olarak veri düzenlemesi yapılırken normal dağılım gösteren değişkenlerin 0 değerlerinin yerine değişkenin ortalaması, normal Dağılım göstermeyen değişkenlerin 0 değerleri yerine değişkenin medyan değeri yazılmıştır.



-KORELASYON

Sonuç değişkeni (kişinin diyabet hastası olup olmaması) ile diğer tüm değişkenler arasındaki korelasyon katsayıları gözlemlenmektedir.
Sonuç değişkeni ile en yüksek korelasyonu gösteren değişken Glukozdur, sonrasında BMI gelmektedir.
Sonuç değişkeni ile en düşük korelasyonu gösteren değişken Kan Basıncıdır.

```
diabets_df3.corr()['Outcome'].sort_values(ascending=False)
```

```
Outcome                     1.000000
Glucose                     0.492782
BMI                         0.311924
Age                         0.238356
Pregnancies                 0.221898
SkinThickness               0.214873
Insulin                     0.203790
DiabetesPedigreeFunction    0.173844
BloodPressure               0.166074
Name: Outcome, dtype: float64
```

LOJİSTİK REGRESYON UYGULAMASI VE ANALİZİ

```
confusion_matrix(Y, Logreg.predict(X))
```

```
array([[441,  59],
       [120, 148]], dtype=int64)
```

```
print("Accuracy:" , metrics.accuracy_score(Y_test,Y_pred ))
```

```
Accuracy: 0.7708333333333334
```

```
print(classification_report(Y, Logreg.predict(X)))
```

```
              precision    recall  f1-score   support

           0       0.79      0.88      0.83       500
           1       0.71      0.55      0.62       268

    accuracy                           0.77       768
   macro avg       0.75      0.72      0.73       768
weighted avg       0.76      0.77      0.76       768
```

TN=441 FP=59 FN=120 TP=148.
0 sınıfı için precision: Diyabet olmama durumunda 0.79 oranında doğru pozitif tahmin yapılmıştır.
1 sınıfı için precision: Diyabet olma durumunda 0.71 oranında doğru pozitif tahmin yapılmıştır.
0 sınıfı için recall: Gerçekte diyabet olmayan kişilerin doğru sınıflandırma oranı 0.88dir.
1 sınıfı için recall: Gerçekte diyabet olan kişilerin doğru sınıflandırma oranı 0.55dir.
0 sınıfı için F1 skoru: Modelin sınıf 0'ı tahmin etme yeteneği 0.83'tür.
1 sınıfı için F1 skoru: Modelin sınıf 1'i tahmin etme yeteneği 0.62'dir.
Support: Veri setindeki toplam 500 sınıf 0, toplam 268 sınıf 1dir.
Modelin veri setindeki kişilerin yaklaşık %77 'sini doğru bir şekilde sınıflandırdığını gösterir.

KARAR AĞACI UYGULAMASI VE ANALİZİ

```
print("Accuracy: ", metrics.accuracy_score(Y_test, y_pred))
```

```
Accuracy:  0.7705627705627706
```

```
confusion_matrix(Y,clf.predict(X))
```

```
array([[430,  70],
       [110, 158]], dtype=int64)
```

```
              precision    recall  f1-score   support

           0       0.80      0.86      0.83       500
           1       0.69      0.59      0.64       268

    accuracy                           0.77       768
   macro avg       0.74      0.72      0.73       768
weighted avg       0.76      0.77      0.76       768
```

TN=430 FP=70 FN=110 TP=158.
0 sınıfı için precision: Diyabet olmama durumunda 0.80 oranında doğru pozitif tahmin yapılmıştır.
1 sınıfı için precision: Diyabet olma durumunda 0.69 oranında doğru pozitif tahmin yapılmıştır.
0 sınıfı için recall: Gerçekte diyabet olmayan kişilerin doğru sınıflandırma oranı 0.86dir.
1 sınıfı için recall: Gerçekte diyabet olan kişilerin doğru sınıflandırma oranı 0.59dir.
0 sınıfı için F1 skoru: Modelin sınıf 0'ı tahmin etme yeteneği 0.83'tür.
1 sınıfı için F1 skoru: Modelin sınıf 1'i tahmin etme yeteneği 0.64'dir.
Support: Veri setindeki toplam 500 sınıf 0, toplam 268 sınıf 1dir.
Modelin veri setindeki kişilerin yaklaşık %77 'sini doğru bir şekilde sınıflandırdığını gösterir.

MODELLERİN DEĞERLENDİRİLMESİ

MODEL	ACCURACY	PRECİSİON		RECALL		F1 SCORE	
		0	1	0	1	0	1
LOJİSTİK REGRESYON	0.7708	0.79	0.71	0.88	0.55	0.83	0.62
KARAR AĞACI	0.7705	0.80	0.69	0.86	0.59	0.83	0.64

Modellerin doğruluk değerleri arasında çok küçük bir fark bulunmaktadır.
0 sınıfı için Karar Ağacı 0.01 farkla pozitif tahminlerini daha iyi yapmıştır.
1 sınıfı için Lojistik Regresyon 0.02 farkla pozitif tahminlerini daha iyi yapmıştır.
0 sınıfı için Lojistik Regresyon 0.02 farkla gerçek pozitifleri daha iyi tahmin etmiştir
1 sınıfı için Karar Ağacı 0.04 farkla gerçek pozitifleri daha iyi tahmin etmiştir.
0 sınıfı için modellerin tahmin etme yeteneği aynıdır.
1 sınıfı için 0.002 farkla Karar Ağacının tahmin etme yeteneği daha iyidir.

KAYNAKÇA

https://erdincuzun.com/makine_ogrenmesi/model-degerlendirme-siniflandirma/
Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques. 2nd Edition, Morgan Kaufmann Publishers, San Francisco.
Veri Zarflama Analizi ve Veri Madenciliği ile Mevduat Bankalarında Etkinlik Ölçümü. Journal of BRSA Banking and Financial Markets, 4(2), 67-84.

İSTATİSTİK BÖLÜMÜ

TÜRKİYE'DE SİGORTA PRİMLERİNİN TAHMİNLENMESİ:

MAKİNE ÖĞRENMESİ VE ZAMAN SERİSİ MODELLERİNİN PERFORMANS KARŞILAŞTIRMASI

Ece AKÇA 18023025

Danışman: Doç. Dr. Selçuk ALP

ÖZET

Türkiye sigorta sektöründeki prim değerlerini tahmin eden en iyi yöntemi bularak diğer yıllarda da kullanılması amaçlanmıştır. Farklı zaman serisi yöntemleri ve yapay sinir ağları ile 2023 yılı test verisi olarak tahminlenmiş ve elde edilen sonuçlar MAE ve RMSE ile karşılaştırılarak karar verilmiştir.

Sigorta Nedir?

Sigorta, belirli riskler karşısında maddi ve manevi kayıpları güvence altına almak amacıyla sigorta şirketi ile imzalanan yasal sözleşmelerdir. Bu hizmet, kâr amacı güden özel sigorta şirketleri tarafından sağlanır ve iki ana kategoriye ayrılır: Hayat Sigortaları ve Hayat Dışı Sigortalar.
Hayat Sigortası: Sigortalının ölüm, riskli hastalık, maluliyet veya iş kaybı gibi durumlar için prim karşılığında güvence aldığı sigortalardır.
Hayat Dışı Sigortalar: Elementer sigorta olarak da bilinen bu sigorta türü, mal varlığının güvence altına alındığı sigortalardır.
Reasürans: Sigorta şirketlerinin, kendi risklerini azaltmak için başka sigorta şirketlerine primlerin bir kısmını devretmesidir. Bu, sigorta şirketlerinin de sigortalanması anlamına gelir ve bu işi yapan şirketlere reasürans denir.

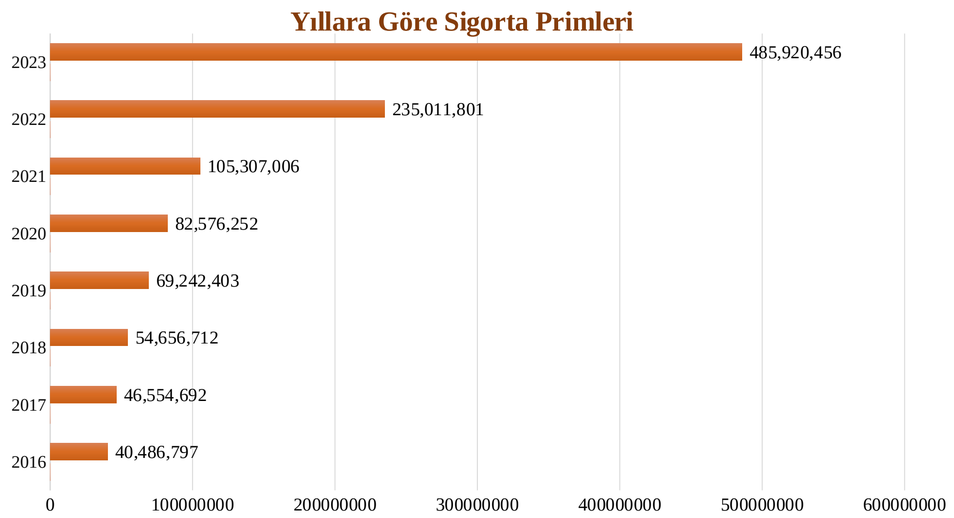
Türkiye'de Sigorta Sektörü

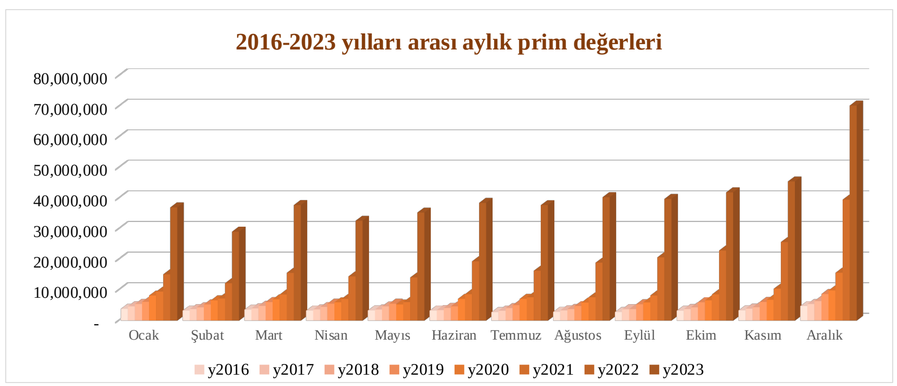
Türkiye'de sigorta sektörü, 2023 yılının ilk 9 ayında %128 büyüyerek 328 milyar liraya ulaşmış ve penetrasyon oranı 2 seviyesine çıkmıştır. Ülkede 71 sigorta şirketi faaliyet göstermekte ve sektörde rekabet artmaktadır. Karşılıklı güven ve toplumun olumlu tutumu, sektörün büyümesine katkı sağlamaktadır. Bu gelişmelerle, Türkiye'de sigorta sektörünün pazar payı artmaktadır.

Şirket Sayıları|Yıl	2019	2020	2021	2022	2023
Hayat ve Emeklilik Şirketi	22	21	21	21	18
Hayat Dışı Şirketi	38	39	41	45	49
Reasürans Şirketi	3	3	3	4	4
Toplam	63	63	65	70	71

Türkiye Sigorta Birliği ve Prim

Türkiye'deki sigorta verilerini toplayan ve koruyan kuruluş, sigorta şirketlerinin genel haklarını savunmak ve temsil etmek amacıyla kurulmuştur. 1900 yılından bu yana isim değiştirerek günümüze kadar gelmiş olan bu kuruluş, sigorta sektörünün köklü bir geçmişe sahip olduğunu gösterir. Ülkede sigorta bilincinin zamanla artması, düşük gelir seviyeleri ve sabit fiyat politikaları nedeniyle uzun süre düşük kalan poliçe satışları, kanunlardaki değişiklikler ve rekabetin artmasıyla yükselmiştir. Deprem ve trafik sigortalarının zorunlu hale gelmesi, hayat dışı sigortacılığının büyümesini sağlamıştır.





Box Jenkins Modeli Uygulamaları

Model	MAE	RMSE
ARIMA(0,2,3)[12]	12439543	14795097
ARIMA(2,1,1)(1,0,1)[12]	5028141	6490847
ARIMA(2,1,1)(1,0,0)[12]	6138889	7941656
ARIMA(2,1,2)(1,0,1)[12]	4940738	6387705

En küçük RMSE VE MAE değerine sahip ARIMA(2,1,2)(1,0,1)[12] modeli daha sonra değerlendirilmek için seçilmiştir.

Holt Winters Teknikleri Uygulamaları

Model	MAE	RMSE
Holt	22819311	24821403
Holt Damped	23856207	25861564
HW	9001911.5	10427990
HW Damped	2000913.9	2365616.2

En küçük RMSE VE MAE değerine sahip Düzeltmelli Holt Winters modeli daha sonra değerlendirilmek için seçilmiştir.

Yapay Sinir Ağları Uygulamaları

Model	MAE	RMSE
NNAR(2,1,2)[12]	6600507	7356030
NNAR(12,1,2)[12]	5149185	7272613
NNAR(1,2,10)[12]	10067851	11361826
NNAR(12,2,2)[12]	4267324	6298186
NNAR(12,2,3)[12]	3595129	6861309

En küçük RMSE değerine sahip NNAR(12,2,2)[12] modeli daha sonra değerlendirilmek için seçilmiştir.

Sonuçlar

Model	MAE	RMSE
ARIMA(2,1,2)(1,0,1)[12]	4940738	6387705
HW Damped	2000913.9	2365616.2
NNAR(12,2,3)[12]	3595129	6861309

Yukarıdaki çıktılardan seçmiş olduğumuz üç öngörü modelini karşılaştırdığımızda en küçük Ortalama Mutlak Hata ve Ortalama Karekök Sapması değerlerine sahip Holt Winters Damped modeli en yakın tahmini yapmıştır.
Makine öğrenmesi ile daha yakın tahminler yapılırken bu verimizde klasik ekonometrik modeller daha iyi tahmin yapmıştır. Farklı makine öğrenmesi ve ekonometrik modeller ile geliştirilmeye açıktır.
Gelecek yıllda da 2024 yılı test verisi olarak kabul edilip sonuçlar karşılaştırılabilir. Her yer bu durum tekrar edebilir. Öncesinde diğer yılların da tahmini sağlanıp en iyi model seçilirse sigorta şirketleri tahmin için bu modeli kullanabilir.

KAYNAKÇA

[1] Hakkimizda. (n.d.). https://tsb.org.tr/tr/hakkimizda
[2] USLU, M. (2016, March 5). Yapay Sinir Ağları (YSA) Nedir – Bölüm 2 ?. Veri Bilimci | Data Scientist. http://veribilimci.org/yapay-sinir-aglari-ysa-nedir-bolum-2/
[3] Şen, Z. (2004). Yapay Sinir ağları ilkeleri. Su Vakfı.

Kurum İçi

İSTATİSTİK BÖLÜMÜ

KALP YETMEZLİĞİ HASTALARININ SAĞ KALIMLARININ MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE TAHMİNLEMESİ

Elif Yaren ÖZYİĞİT 19023032

Danışman: Prof. Dr. Fatma NOYAN TEKELİ

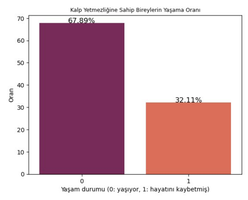
ÖZET

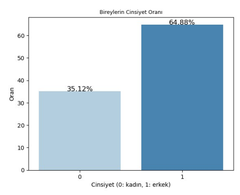
Kalp yetmezliği, kalbin yeterli miktarda kan pompalayamaması durumudur ve tedavi edilmediğinde ciddi sağlık sorunlarına neden olabilir. Bu çalışmada, UCI veri tabanından alınan kalp yetmezliğine ait sağlık verileri kullanılarak makine öğrenmesi algoritmaları üzerinden modeller kurulmuştur. Veri üzerinde karar ağacı algoritması ve rastgele orman algoritması kullanılarak kalp yetmezliği olan hastaların ölüm/yaşam durumları tahmin edilmiştir. Veri incelendiğinde eksik veri ile karşılaşılmamakla birlikte aykırı değerler gözlemlenmiştir. Bu değerler min-max standardizasyonu ile standardize edilmiştir. Modeller 5 çapraz doğrulama yöntemi ile veri seti üzerinde kurulmuştur. Kurulan modeller doğruluk, duyarlılık, kesinlik, özgüllük ve f1 skorları ile karşılaştırılmıştır. Kullanılan algoritmalar arasında en başarılı sonuçları veren algoritma rastgele orman algoritması olmuştur.

VERİ HAKKINDA

Bu çalışmada Faisalabad Kardiyoloji Enstitüsü ve Pakistan'da bulunan Müttefik hastanesinden toplanmış, kalp yetmezliği yaşayan 299 hastanın verisi kullanılmıştır. Veri University of California Irvine tarafından kurulan UCI Machine Learning Repository (UCI, 2020) sitesinden alınmıştır.

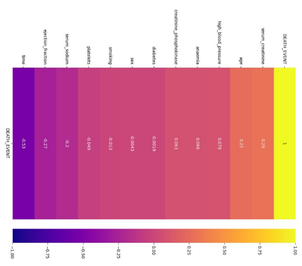
Değişkenler	Tanım	Veri Türü
Age	Hasta Yaşı (Yıl)	Nümerik
Anemia	Kırmızı kan hücrelerinin/hemoglobin seviyesinin azalması, kansızlık durumu (0: yok, 1: var)	Kategorik
Creatinine_phosphokinase	Kandaki CPK enziminin seviyesi (mcg/L)	Nümerik
Diabetes	Hasta diyabete durumu (0: yok, 1: var)	Kategorik
Ejection_fraction	Her bir kasılma sırasında kalpten çıkan kan yüzdesi	Nümerik
High_blood_pressure	Hasta hipertansiyon durumu (0: yok, 1: var)	Kategorik
Platelets	Kandaki tranbosit seviyesi (kiloplatelets/mL)	Nümerik
Serum_creatinine	Kandaki serum kreatinin seviyesi (mg/dL)	Nümerik
Serum_sodium	Kandaki serum sodyum seviyesi (mEq/L)	Nümerik
Sex	Cinsiyet (0: kadın, 1: erkek)	Kategorik
Smoking	Hasta sigara kullanma durumu (0: kullanmıyor, 1: kullanıyor)	Kategorik
Time	Hasta izlenme süresi (Gün)	Nümerik
DEATH_EVENT	Hastanın izlenme süresince ölme durumu (0: yaşıyor, 1: hayatını kaybetmiş)	Kategorik





Veri 299 gözlemden oluşup hastaya yönelik 13 klinik veri içermektedir. Hastaların 105'i kadın, 194'ü erkek olmak üzere, izlenilen süre içerisinde hayatını kaybetme durumu ayrıca bu veri seti içerisinde incelenmiştir.

Değişkenler arasında ilişki incelendiğinde diyabet durumu ejeksiyon fraksiyonu (ejection_fraction), kandaki tranbosit seviyesi (platelets), kandaki serum sodyum seviyesi (serum_sodium), cinsiyet (sex), sigara içme durumu (smoking), gözlemlenen zaman (time) değişkenleri ölüm olayı ile negatif yönde ilişkili olduğu tespit edilmiştir Korelasyon matrisi incelendiğinde yaşam durumu ve yaş değişkenleri arasında pozitif bir korelasyon görülmektedir. (0.25) Bu analizden yaşın artması ile ölüm riskinin de artması sonucu çıkartılabilir. Ejeksiyon frekasyonu ile yaşam durumu arasında güçlü sayılabilecek bir negatif korelasyon olduğunu görülmüştür.



VERİ ÖN İŞLEME

Veri incelendiğinde gözlemler arası kayıp veriye rastlanılmamıştır. Creatinine phosphokinase, ejection fraction, serum creatinine ve serum sodium değerlerinde aykırı değerler gözlemlenmiştir. Bu durumun önüne geçmek adına min-max scaler işlemi uygulanmıştır. Makine öğrenmesi uygulamaları gerçekleştirilirken veri setinin test serisi ve train serisi olarak ikiye ayrılması gerekmektedir. Bu işlem kurulan modellerin doğruluğunu ve genelleme yeteneğini değerlendirmek için önemlidir. Bu çalışmada verinin %80'i eğitim verisi %20'si test verisi olarak ayrılmıştır.

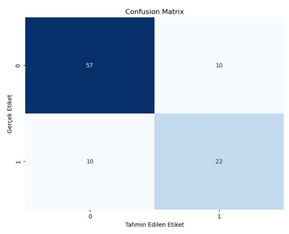
KARAR AĞACI ALGORİTMASI UYGULAMASI

Bu çalışmada karar ağacı modeli Python programı üzerinden sklearn kütüphanesi kullanılarak kurulmuştur. Model ilk olarak hiperparametre optimizasyonu yapılmadan default olarak kurulmuştur.

Bu uygulamada, kalp yetmezliğine sahip hastaların yaşama durumunun tahmin edilmesi amacıyla karar ağacı algoritması kullanılarak model kurulmuştur. Standartize edilmiş veri %33 test %67 eğitim seti olarak ayrılmıştır. Model hiperparametrelerinden olan ayırma kriteri gini index olarak seçilmiş, minimum yaprak sayısı 1, minimum ayrım 2 seçilmiştir.

Eğitim Seti Doğruluk Oranı:	1.0000
Test Seti Doğruluk Oranı:	0.7979

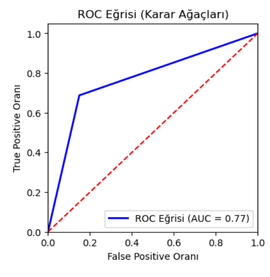
Modelin doğruluk değeri incelendiğinde modelin genelleme yeteneğinin çok iyi olmasa da iyi olduğu söylenebilmektedir. Model hakkında daha fazla bilgi edinmek için karmaşıklık matrisinin incelenmesinde fayda bulunmaktadır.



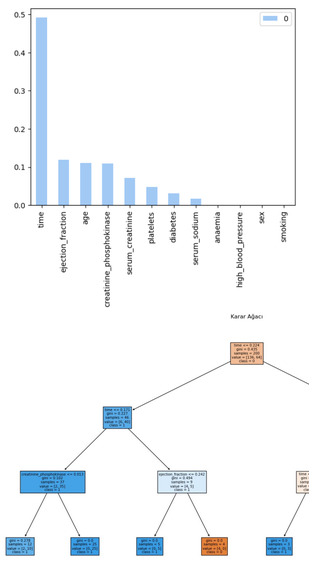
Karmaşıklık matrisine göre gerçekten yaşayan 57 kişinin "yaşıyor" olarak sınıflandırılarak doğru olarak tahmin edilmiştir. Gerçekten yaşayan 10 kişi ise "yaşamıyor" olarak sınıflandırılarak yanlış tahmin edilmiştir. Gerçekten yaşamayan 10 kişi "yaşıyor" olarak sınıflandırılarak yanlış tahmin edilmiştir. Gerçekten yaşamayan 22 kişi "yaşamıyor" olarak sınıflandırılarak doğru tahmin edilmiştir. Karmaşıklık matrisi ile modele dair performans değerleri hesaplanır.

Modelin tüm tahminlerinin doğru tahmin yapıldığını gösteren doğruluk değeri %78.9 olarak bulunmuştur. Bu oran tek başına modelin orta-iyi arası bir doğruluk oranına sahip olduğunu gösterse de sınıflar arası dengesizlik olduğu durumda yeterli ve doğru bir yorumlama metriği olmayacaktır. Pozitif sınıfa atanan örneklerin gerçekten pozitif olduğunu gösteren keskinlik değeri %68.8 olarak bulunmuştur. Bu sonuç modelin pozitif tahminlerinin orta seviyede güvenilir olduğunu gösterir yorumu yapılabilir. Gerçek pozitif verilerin yüzde kaçının doğru tahmin edildiğini gösteren duyarlılık değeri %68.8 bulunmuştur. Negatif tahmin edilen değerlerin gerçek negatif olma oranını gösteren özgüllük değeri %68.8 olarak bulunmuştur. F1 skoru ise %68.8 olarak bulunmuştur. Bu değer ile modelin genel olarak ortalama bir performans gösterdiği söylenebilir.

Metrikler	Değerler	Yüzdeler
Accuracy (Doğruluk)	0.789	%78.9
Precision (Kesinlik)	0.688	%68.8
Recall (Duyarlılık)	0.688	%68.8
Specificity (Özgüllük)	0.688	%
F1 Skoru	0.688	%68.8



ROC eğrisinin eğrisi grafiği incelendiğinde AUC değerinin 0.77 olduğu görülmüştür. Modelin genel performansının negatif ve pozitif sınırları ayırmada orta seviyede başarılı olduğu söylenebilir. ROC eğrisinin sol üst köşeye kısmen yakın olması modelin doğru pozitifleri orta-iyi bir oranda tahmin ettiğini, yanlış pozitifleri ise düşük oranda tuttuğu söylenebilir.



Karar ağacı algoritması ile ağaç oluşturulduktan sonra değişkenlerin önem sırası incelenmiştir. Modeli etkileyen en önemli özelliğin time değişkeni olduğu görülmüştür. Grafik üzerinde görüldüğü üzere karar ağacı bölünme noktalarını belirlemede ve modelin doğruluk oranını arttırma konusunda en çok kullanılan 3 özellik sırasıyla; gözlem zamanı, serum keratinin seviyesi, ejeksiyon frekasyonu değeridir. Grafik genel olarak incelendiğinde anemi, yüksek kan basıncı, cinsiyet, sigara içme durumu ve diyabet durumu modeli neredeyse modeli hiç etkilemediği gözlemlenmiştir.

```
K-fold Çapraz Doğrulama Sonuçları: [0.66666667
0.83333333 0.81666667 0.78333333 0.89830508]
Ortalama Doğruluk Oranı: 0.7997
```

Modele k = 5 olacak şekilde k-fold doğrulama uygulanmıştır. Modelin doğruluk oranı 0.66 ile 0.89 arasında olduğu gözlemlenmiştir. Ortalama doğruluk oranın ise 0.79 olduğu görülmüştür. İlk katmanın 0.66 olarak gözlemlenmesi modelin bazı veri kümelerinde zayıf performans gösterebileceğine işaret etmektedir. Sonuç olarak modelin genel olarak orta – iyi bir performans gösterdiği söylenebilir.

Oluşturulan ağacın kök düğümünün en önemli özellik olan time değişkeni olduğu görülmüştür. İkinci seviyedeki özellikler ise tekrar zaman değişkeni ve ejeksiyon frekasyonu olarak görülmüştür. Modelin daha net görülmesi için maksimum derinlik 3 olarak seçilmiştir.

RASTGELE ORMAN ALGORİTMASI UYGULAMASI

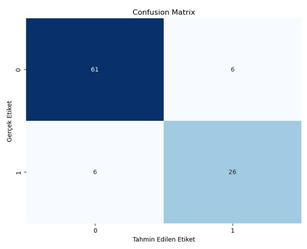
Bu uygulamada, kalp yetmezliğine sahip hastaların yaşama durumunun tahmin edilmesi amacıyla Random forest algoritması kullanılarak modeller kurulmuştur. Random forest daha önce de anlatıldığı üzere bir çok karar ağacının bir araya gelmesiyle oluşmaktadır. Ağaç oluşturulurken gini kriteri kullanılmıştır. Oluşturulan ağaç sayısı 1000 adettir.

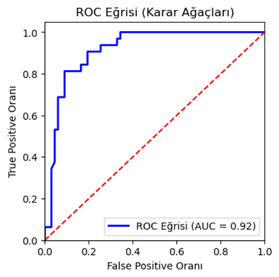
Eğitim Seti Doğruluk Oranı:	1.0000
Test Seti Doğruluk Oranı:	0.8788

Kurulan modelde test seti ile yapılan doğrulamada doğruluk(accuricy) 0.8788 olarak bulunmuştur. Bu sonuç test setinde %87.88 oranında doğru tahmin yapıldığını belirtmektedir. Modelin doğruluk değeri incelendiğinde modelin genelleme yeteneğinin iyi olduğu söylenebilmektedir. Model hakkında daha fazla bilgi edinmek için karmaşıklık matrisinin incelenmesinde fayda bulunmaktadır.

Modelin tüm tahminlerinin doğru tahmin yapıldığını gösteren doğruluk değeri %87.88 olarak bulunmuştur. Bu oran tek başına modelin yüksek doğruluk oranına sahip olduğunu gösterse de sınıflar arası dengesizlik olduğu durumda yeterli ve doğru bir yorumlama metriği olmayacaktır. Pozitif sınıfa atanan örneklerin gerçekten pozitif olduğunu gösteren keskinlik değeri %91.04 olarak bulunmuştur. Bu sonuç modelin pozitif tahminlerinin güvenilir olduğunu gösterir yorumu yapılabilir. Gerçek pozitif verilerin yüzde kaçının doğru tahmin edildiğini gösteren duyarlılık değeri %91.04 bulunmuştur. Negatif tahmin edilen değerlerin gerçek negatif olma oranını gösteren özgüllük değeri %81.25 olarak bulunmuştur. F1 skoru ise %91.04 olarak bulunmuştur. Bu değer ile modelin genel olarak güçlü ve dengeli bir performans gösterdiği söylenebilir. Bu metrikler incelendiğinde model pozitif ve negatif sınıfları için dengeli bir performans gösterse de pozitif sınıflarda başarısının daha fazla olduğu görülmüştür.

Metrikler	Değerler	Yüzdeler
Accuracy (Doğruluk)	0.8788	%87.88
Precision (Kesinlik)	0.9104	%91.04
Recall (Duyarlılık)	0.9104	%91.04
Specificity (Özgüllük)	0.8125	%81.25
F1 Skoru	0.9104	%91.04





Modele k = 5 olacak şekilde k-fold doğrulama uygulanmıştır. Modelin doğruluk oranı 0.73 ile 0.90 arasında olduğu gözlemlenmiştir. Ortalama doğruluk oranın ise 0.85 olduğu görülmüştür. İlk katmanın 0.73 olarak gözlemlenmesi modelin bazı veri kümelerinde zayıf performans gösterebileceğine işaret etmektedir. Sonuç olarak modelin genel olarak yüksek bir performans gösterdiği söylenebilir.

SONUÇ

Bu çalışmada makine öğrenmesi yöntemlerinden karar ağacı ve rastgele orman algoritmaları kullanılarak UCI Machine Learning Repository (UCI, 2020) sağlık verileri üzerinde ölüm yaşam tahminlemesi yapan modeller kurulmuştur. Karar ağacı algoritması kullanılarak yapılan çalışmada optimum başarıyı yakalamak adı çeşitli hiper parametrelerle ağaçlar oluşturulmuştur. Seçilen parametrelerle en iyi sonucu veren ağacın doğruluk değeri %78.9 olarak bulunmuştur. Ardından tahmini daha da geliştirmek ve overfitting sorunundan kaçınmak adına rastgele orman algoritması ile 1000 adet ağaç oluşturulmuştur. Rastgele orman algoritması ile kurulan model %87.88 doğruluk değeriyle daha başarılı bir model elde edilmiştir. Çalışma sonunda görülmüştür ki rastgele orman algoritmasının UCI verisi için tek bir ağaçla kurulan karar ağacı algoritmasından daha iyi performans göstermektedir.

KAYNAKÇA

[1] S. Bilekyiğit ve A. Eldem, "Kalp yetmezliği riskinin makine öğrenmesi yöntemleri ile analiz edilmesi," Yüksek lisans tezi, Bilgisayar Mühendisliği Programı, Karamanoğlu Mehmet Bey Üniversitesi, Karaman, 2022.

[2]] İ. Arslan, *Python ile Veri Bilimi*, 4. baskı, İstanbul: Pusula, 2010.

[3]] L. Breiman, *Machine Learning*, Springer, 2013.

İSTATİSTİK BÖLÜMÜ

HÜCUMA DÖNÜK FUTBOL OYUNCULARININ MARKET DEĞERLERİ İLE SINIFLANDIRILMASI VE ANALİZİ

EMİRCAN BAYKARA 19023073

Danışman: Dr. Öğr. Üyesi Erhan ÇENE

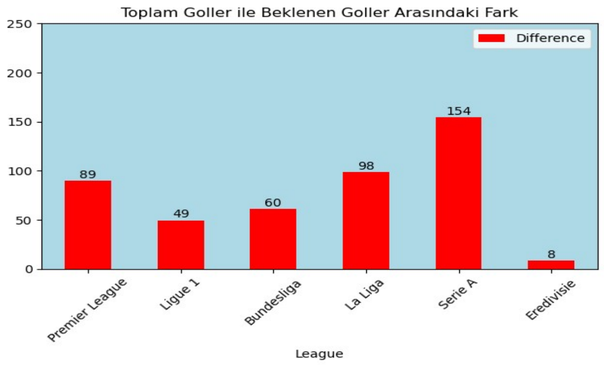
ÖZET

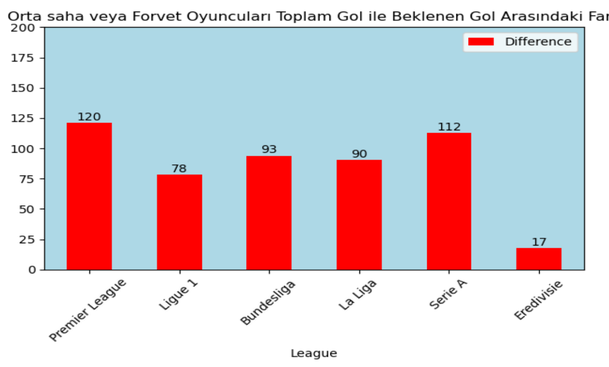
Futbol takımlarının hücum yeteneklerini detaylı bir şekilde inceleyerek, en az 10 maça çıkmış oyuncuların market değerlerine göre kümeleme sistemleri oluşturuldu. 6 büyük ligden 3411 futbol oyuncusuna ait geniş bir veri setini kullanarak, oyuncuları market değerlerine göre sınıflandırılıp ve gerçekleşen ile beklenen gol arasındaki farkı inceleyerek oyuncuların ve liglerin etkinliği değerlendirildi.

Bu kümeleme sistemine dayanarak, oyuncuların saha üzerindeki performanslarını daha derinlemesine inceleyerek, özellikle üçüncü bölgede (rakip ceza sahasına yakın alan) yaptıkları top sürmeleri, etkili pasları ve çalım yetenekleri ölçüldü. Bu analiz, takımların hücum stratejilerini geliştirmek ve en etkili oyuncuları belirlemek için önemli bir kılavuz sağlayacaktır.

AVRUPANIN ÖNDE GELEN FUTBOL LİGLERİNDE OYUNCU PERFONAMANS ANALİZLERİ

2022/2023 sezonunda Premier Lig, Eredivisie, Ligue 1, La Liga, Bundesliga ve Serie A gibi önde gelen futbol liglerindeki oyuncuların performansları incelendi. Bu analizde, her bir ligdeki oyuncuların istatistikleri üzerinden derinlemesine bir değerlendirme yapılarak farklı ligler arasındaki performans farklılıkları ve oyuncuların genel etkinliği anlaşılmaya çalışıldı.





Tablo 1'de, tüm mevkilerdeki oyuncuların beklenen ve gerçekleşen gol arasındaki farkı gösteren değerler sunulmuştur. Bu verilere göre, Serie A ligi diğer liglere kıyasla belirgin bir şekilde öne çıkmaktadır. Eredivisie ligi ise diğerlerine kıyasla daha düşük bir fark değerine sahiptir.

Tablo 2'de ise, sadece orta saha ve forvet oyuncularının beklenen ve gerçekleşen gol arasındaki farkı gösteren değerler sunulmuştur. Bu tabloda da Serie A ve Premier Lig öne çıkmaktadır, ancak diğer ligler arasındaki farklar daha belirgin değildir. Özellikle Eredivisie ligi, diğer liglere kıyasla daha düşük bir fark değerine sahiptir

OYUNCULARIN OYUN İÇİ KATKILARI

Rakip Sahada Top Sürme Değerlerine Göre İlk 5 Oyuncu

	Player	Nation	League	Team	Main_Pos	Age	Progression_PrgC	Market_Value
1	Vinicius Júnior	BRA	La Liga	Real Madrid CF	Forward	22	291	€150.00m
2	Jeremie Frimpong	NED	Bundesliga	Bayer 04 Leverkusen	Midfielder	21	219	€40.00m
3	Kylian Mbappé	FRA	Ligue 1	Paris Saint-Germain FC	Forward	23	210	€180.00m
4	Bukayo Saka	ENG	Premier League	Arsenal FC	Forward	20	201	€120.00m
5	Khvicha Kvaratskhelia	GEO	Serie A	SSC Napoli	Forward	21	191	€85.00m

Rakip Sahada Pas Atma Değerlerine Göre İlk 5 Oyuncu

	Player	Nation	League	Team	Main_Pos	Age	Progression.1_PrgP	Market_Value
1	Toni Kroos	GER	La Liga	Real Madrid CF	Midfielder	32	407	€15.00m
2	Joshua Kimmich	GER	Bundesliga	FC Bayern München	Midfielder	27	397	€75.00m
3	Orkun Kökçü	TUR	Eredivisie	Feyenoord	Midfielder	21	379	€35.00m
4	Lionel Messi	ARG	Ligue 1	Paris Saint-Germain FC	Forward	35	353	€35.00m
5	Bruno Fernandes	POR	Premier League	Manchester United FC	Midfielder	27	327	€75.00m

Rakip Sahada Topla Buluşma Değerlerine Göre İlk 5 Oyuncu

	Player	Nation	League	Team	Main_Pos	Age	Progression.2_PrgR	Market_Value
1	Vinicius Júnior	BRA	La Liga	Real Madrid CF	Forward	22	625	€150.00m
2	Dušan Tadić	SRB	Eredivisie	Ajax	Forward	33	585	€6.00m
3	Bukayo Saka	ENG	Premier League	Arsenal FC	Forward	20	582	€120.00m
4	Kylian Mbappé	FRA	Ligue 1	Paris Saint-Germain FC	Forward	23	561	€180.00m
5	Mohamed Salah	EGY	Premier League	Liverpool FC	Forward	30	508	€65.00m

Tablo 1'de top taşıma, Tablo 2'de üçüncü bölgede etkili pas sayıları, Tablo 3'te ise üçüncü bölgede topla buluşma değerleri yer almaktadır. Tablo 1'e baktığımızda, yüksek dribbling yeteneğine sahip genç oyuncuların ve piyasa değeri olarak yükselmeye daha yatkın futbolcuların öne çıktığı görülmektedir. Tablo 2'de ise, daha tecrübeli oyuncuların yanı sıra genç oyuncuların da bulunduğu ve milli futbolcumuz Orkun Kökçü'nün neden Benfica'ya rekor bir bedelle transfer olduğunu anlaşılabildiği gözlemlenmektedir. Tablo 3'te ise, ileri yaşına rağmen kalitesiyle dikkatleri üzerine çeken Dusan Tadic'in dikkat çektiği görülmektedir. Genel olarak, üç tablonun bir bütün olarak incelenmesi sonucunda, top taşıma ve çalım atma listesinde birinci olan Vinicius

Atılan Goller ile Beklenen Goller Arasındaki Fark

	Player	Nation	League	Team	Main_Pos	Age	Performance.8_G+A-PK	Expected.3_npxG+xAG	Perfonmance_Score-Expected_Score
1	Randal Kolo Muani	FRA	Bundesliga	Eintracht Frankfurt	Forward	23	34	19.8	14.2
2	Marcus Rashford	ENG	Premier League	Manchester United FC	Forward	24	38	24.9	13.1
3	Kylian Mbappé	FRA	Ligue 1	Paris Saint-Germain FC	Forward	23	46	33.4	12.6
4	Jamal Musiala	GER	Bundesliga	FC Bayern München	Midfielder	19	29	18.1	10.9
5	Eric Maxim Choupo-Moting	CMR	Bundesliga	FC Bayern München	Forward	33	21	10.2	10.8

Oyuncuların Kart Görme İstatistikleri

	Player	Nation	League	Team	Main_Pos	Age	Performance.9_CrdTotal
1	Érik Lamela	ARG	La Liga	Sevilla FC	Forward	30	18
2	Vinicius Júnior	BRA	La Liga	Real Madrid CF	Forward	22	17
3	Mattia Zaccagni	ITA	Serie A	SS Lazio	Forward	27	13
4	Valentín Castellanos	ARG	La Liga	Girona FC	Forward	23	12
5	Raphinha	BRA	La Liga	FC Barcelona	Forward	25	11

Bu tabloda, oyuncuların maçlarda gördüğü kart sayıları incelenmektedir. Vinicius Júnior, önceki tablolarda lider olarak öne çıkmışken, burada ise agresif tavırlarıyla dikkat çekiyor ve 17 kartla listenin üst sıralarında yer alıyor. Bu durum, onun hem yüksek performans gösterdiğini hem de maçlarda zaman zaman agresif davranışlar sergilediğini ortaya koyuyor.

Milletlerin Forvet ve Orta Saha Oyuncu Dağılımları

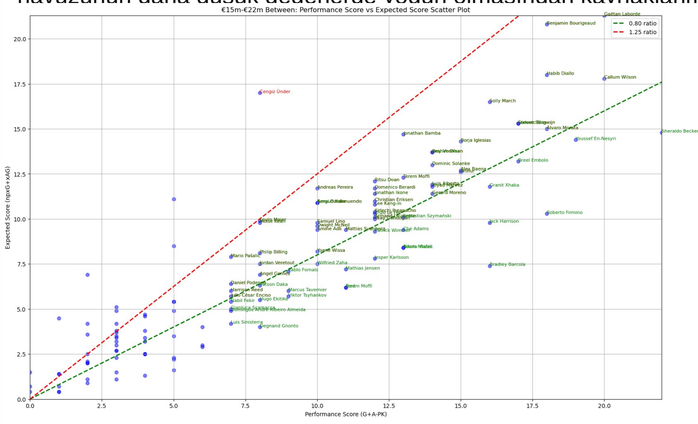
	Nation	Midfielders Count
1	FRA	107
2	ESP	106
3	NED	70
4	GER	63
5	ITA	50

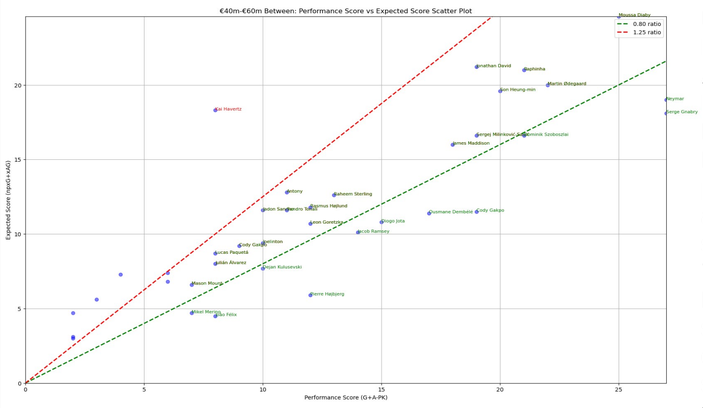
	Nation	Forwards Count
1	FRA	67
2	ESP	65
3	NED	62
4	GER	46
5	ENG	41

Fransa ve İspanya, futbol dünyasında orta saha ve forvet mevkilerinde zengin bir oyuncu havuzuna sahip olduğu görünüyor. Fransa ve İspanya'nın geçmişteki genişlemiş sömürge imparatorlukları, futbol yeteneklerinin çeşitliliğine ve gelişimine katkıda bulunmuş olabilir.

OYUNCULARIN DEĞERLERİNE GÖRE KÜMELEME SİSTEMİ

Veride toplam 3411 futbolcu bulunmaktadır. En az 10 maça çıkmış futbolcular bir grupta toplandı. Daha sonra, bu gruptan sadece forvet ve orta saha oyuncuları seçildi. Seçilen oyuncuların piyasa değerleri kullanılarak, 1 ile 10 arası bir puanlama sistemi oluşturuldu. Puanlama sistemi, oyuncuların değerlerinin yoğunluğuna göre belirlendi. Örneğin, 0-2 milyon euro arası değerler için 1 puan, 2-4 milyon euro arası değerler için 2 puan, 4-7 milyon euro arası değerler için 3 puan, 7-12 milyon euro arası değerler için 4 puan vb. şeklinde yapıldı. Bu değer aralıkları, elimizdeki oyuncu havuzunun daha düşük değerlerde yoğun olmasından kaynaklanm



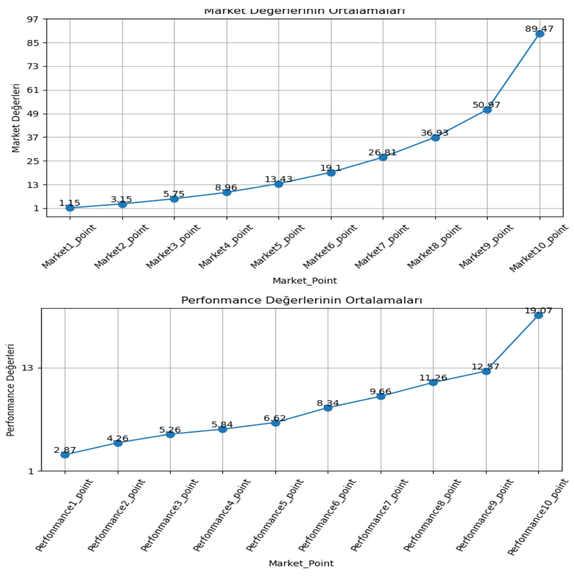


Bu iki tablo, oyuncuların gerçekleştirdikleri skor ile beklenen skor arasındaki oranı daha rahat gözlemlemek amacıyla hazırlanmıştır. Kırmızı çizginin üstündeki oyuncular, beklenen skorlarını karşılayamayan ve sezonu beklenenin oldukça altında tamamlayan oyuncuları göstermektedir. Yeşil çizginin altındaki oyuncular ise beklenenin üzerinde performans sergileyen oyuncular olarak gözlemlenmektedir. İki çizgi arasında kalan oyuncular ise beklenen skor katkısıyla gerçekleştirdikleri skorun ortalama düzeyde olduğunu göstermektedir.

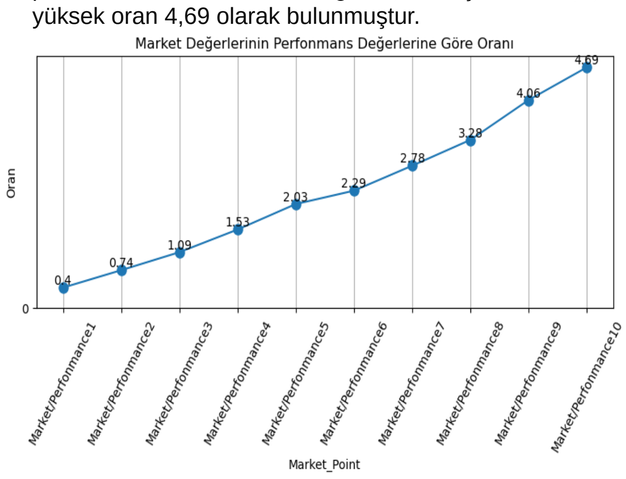
Tablo1'i incelediğimizde, bazı tanıdık isimlerin yanı sıra Avrupa piyasasında ismi çok duyulmamış oyuncuların da olduğunu görülmektedir. En tanınan isimlerden milli futbolcumuz Cengiz Ünder'in, bu sezon Fenerbahçe'ye büyük bir beklentiyle gelmesine rağmen, önceki sezon Avrupa'da beklenenin çok altında skor yaptığını ve bu transferin ne kadar riskli olduğu görülüyor. Dikkat çeken bir diğer oyuncu ise Bradeley Barcola, Lyon'da sergilediği üst düzey performansla parladı ve bu sezon PSG'de sahne alarak yeteneklerini daha geniş bir kitleye sergileme fırsatı buldu.

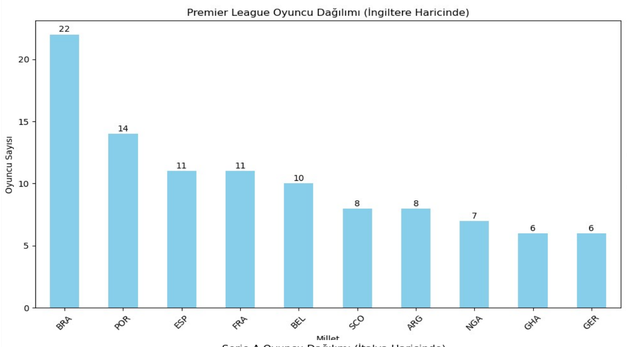
Tablo2'ye baktığımızda, Moussa Diaby'nin Leverkusen'de yakaladığı pozisyonları iyi bir şekilde değerlendirerek bu sezon Premier Lig ekiplerinden Aston Villa'ya transfer yaparak bu sezon daha üst düzey ligde boy gösterme fırsatı yakaladı. Kai Havertz'in Chelsea'deki düşük

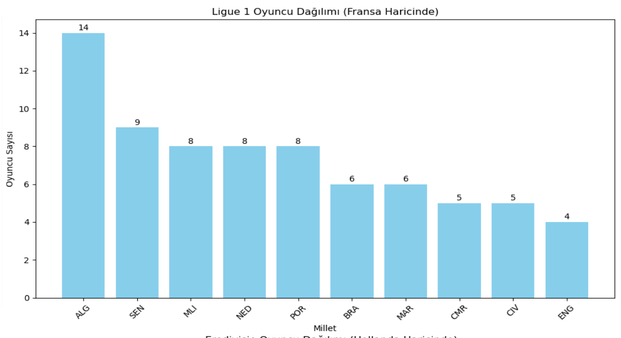
Futbolcuların Puanlama Sistemine Göre Piyasa Değerleri ve Performans Analizleri

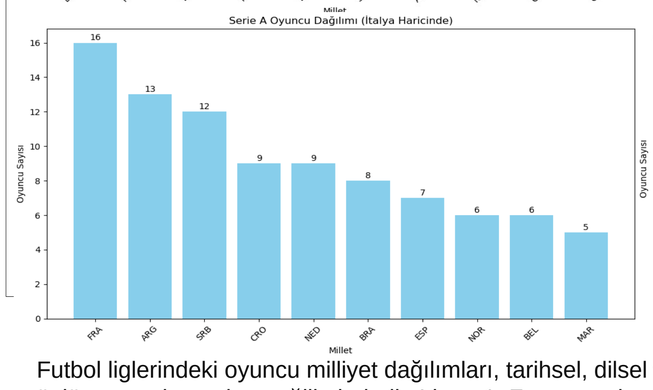


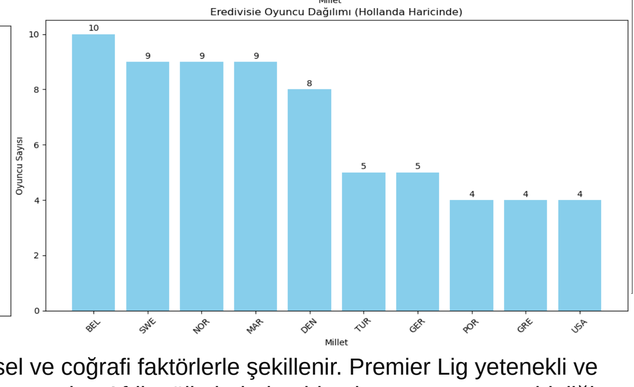
Transfer olduğunu göstermektedir. Ghabry ve Futbolcuların piyasa değerleri baz alınarak 1 ile 10 arasında puanlandığı bir sistem oluşturulmuştur. Piyasa değeri ile performans oranları incelendiğinde, en düşük oran 0,4, en yüksek oran 4,69 olarak bulunmuştur.











Futbol liglerindeki oyuncu milliyet dağılımları, tarihsel, dilsel ve coğrafi faktörlerle şekillenir. Premier Lig yetenekli ve ünlü oyuncuları çekme eğilimindedir. Ligue 1, Fransızca konuşulan Afrika ülkelerinden birçok oyuncuya ev sahipliği yapar, Fransa'nın sömürgecilik geçmişini yansıtır. Eredivisie, komşu ve az gelişmiş ülkelerden oyuncuları çekip geliştirir. Serie A, Bundesliga ve La Liga da benzer eğilimler gösterir. Dil ve kültür bağları, oyuncu tercihlerini etkiler. Bu dağılımlar, futbolun küresel yapısındaki tarihi, kültürel ve ekonomik bağlarla ilişkilidir.

KAYNAKÇA

https://fbref.com/en/
https://www.pythoncheatsheet.org/
https://matplotlib.org/cheatsheets/
https://www.transfermarkt.com.tr/
https://www.sofascore.com/tr/
https://www.premierleague.com/history/season-reviews/489
https://www.bundesliga.com/en/bundesliga/stats/players/2022-2023
https://www.legaseriea.it/en/serie-a/statistiche
https://www.ligue1.com/stats?seasonId=2022-2023&StatsActiveTab=1
https://eredivisie.eu/stats/
https://www.laliga.com/en-GB/stats/laliga-easports/scorers

İSTATİSTİK BÖLÜMÜ

RANDOM FOREST ALGORİTMASI İLE ÜRETKENLİK TAHMİNİ

Emre KARA 19023083

Danışman: Prof. Dr. Filiz Karaman

ÖZET

Fabrika işçilerinin üretkenliği, imalat ve hizmet sektörlerinde, işletmelerin rekabet gücünü artırmak ve maliyetlerini düşürmek için büyük önem taşıyan bir faktördür. Tekstil sektörü, yoğun iş gücü gerektiren bir endüstridir ve işçi üretkenliğinin optimize edilmesi sektörün gelişmesi açısından büyük önem arz etmektedir. İşçilerin üretkenliğinin doğru ve güvenilir tahmini, kaynakların verimli kullanılmasını sağlamakta ve üretim süreçlerinde meydana gelebilecek aksaklıkların önüne geçerek kar payını arttırmayı sağlamaktadır. Bu nedenle, işçi üretkenliğini etkileyen faktörlerin belirlenmesi ve bu faktörlere dayalı tahmin modellerinin geliştirilmesi, işletmelerin performansını ve rekabet gücünü artırmada önemli bir kriterdir. Bu çalışmada, bir tekstil fabrikasındaki işçilerin üretkenliği Random Forest algoritması ile tahmin edilmeye çalışılmıştır.

Tekstilde İşçi Üretkenliği

Pazardaki rekabetin her geçen gün arttığı tekstil sektöründe işletmeler, kısa zamanda kaliteli ürün üretmeyi dolayısıyla üretkenliklerini arttırmayı amaçlar. Üretkenlik, işçilerin belirli bir zaman dilimindeki çalışma verimliliklerini ve ürettikleri ürün veya hizmet miktarını ifade etmektedir. İşçilerin üretkenliğinin doğru ve güvenilir tahmini, kaynakların verimli kullanılmasını sağlamakta ve üretim süreçlerinde meydana gelebilecek aksaklıkların önüne geçerek kar payını arttırmayı sağlamaktadır. Bu nedenle, işçi üretkenliğini etkileyen faktörlerin belirlenmesi ve bu faktörlere dayalı tahmin modellerinin geliştirilmesi, işletmelerin performansını ve rekabet gücünü artırmada önemli bir kriterdir.

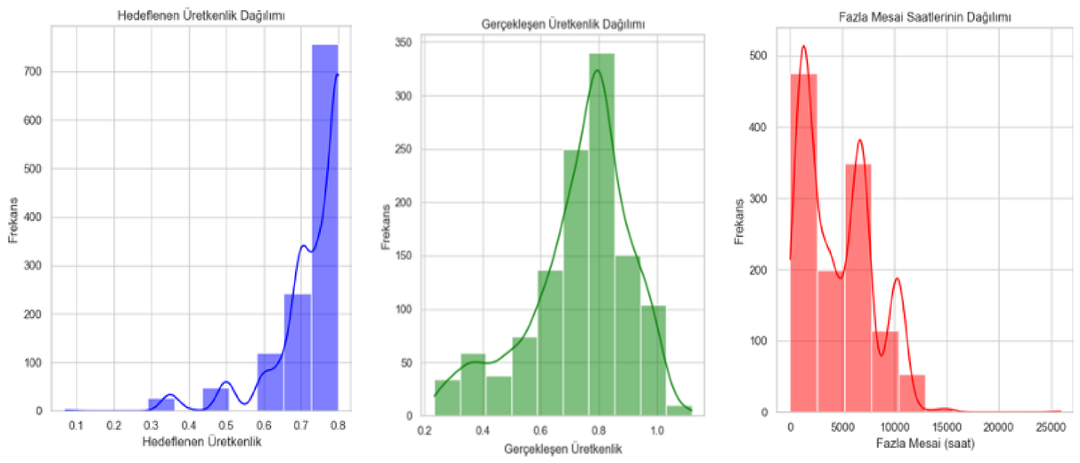
Random Forest Algoritması

Sınıflandırma analizi, gözlemlenen veri öğelerini önceden tanımlanmış kategorilere atamak için kullanılan bir makine öğrenmesi yöntemidir. Sınıflandırma, veri seti içindeki öznitelikler (features) ve etiketler (labels) kullanılarak model eğitimi yapılmasını kapsamaktadır. Eğitilen model, daha sonra etiketlenmemiş verileri sınıflandırmak için kullanılmaktadır.

Random Forest, birden fazla karar ağacını birleştirerek oluşturulan bir topluluk öğrenme tekniğidir. Her bir karar ağacı, veri setinden rastgele seçilen alt kümeler üzerinde bağımsız olarak eğitilir ve sınıflandırma görevinde çoğunluk oyu ile karar verilir. Bu yöntem, karar ağaçlarının karşılaştığı aşırı uyum (overfitting) sorununu azaltmakta ve genel tahmin başarısını arttırmaktadır.

İşçi Üretkenliği Verisi

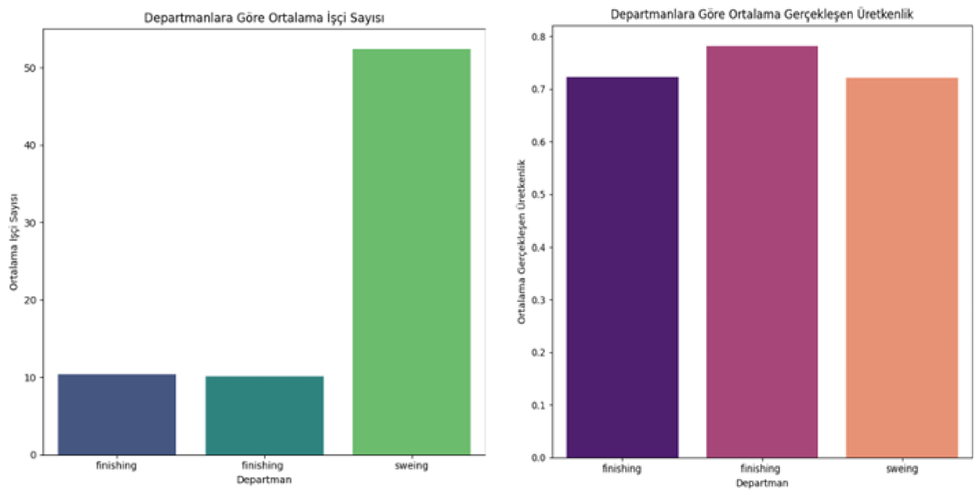
Veri seti, giyim üretimi yapan bir fabrikadan elde edilen metriklere göre işçi verimliliği verilerini içermektedir. Farklı tarihlerde, çeşitli departmanlarda ve takımlarda çalışan işçilerin üretkenlik ölçümlerini kapsamaktadır. Veri seti, 1197 gözlem ve 15 değişkenden oluşmaktadır. Veriler, fabrikanın farklı bölümlerinde, çeşitli zaman dilimlerinde ve takımlarda toplanmıştır.



İşçilerin hedeflenen üretkenlik oranlarının çoğunluğu 0.7 ve 0.8 arasında yoğunlaşmaktadır, işçiler için yüksek üretkenlik hedefi belirlemiştir ve işçilerin bu hedefe ulaşması beklenmektedir.

İşçinin üretkenlik oranları 0.6 ile 0.9 arasında yoğunlaşmaktadır. Bu dağılım, işçilerin büyük bir kısmının hedeflenen üretkenlik oranlarına yakın veya bu oranların üzerinde performans sergilediğini göstermektedir.

İşçilerin fazla mesai yapması üretkenlik üzerinde olumsuz etki yaratmaktadır. Fazla mesai saatlerinin artması, genellikle üretim hedeflerine ulaşmak için ekstra efor sarf edildiğini göstermektedir.



Grafiğie göre ortalama işçi sayısı grafiğinde dikiş (sewing) departmanında ortalama işçi sayısının diğer departmanlara göre oldukça yüksektir. Üretkenlik ortalaması grafiğinde bitirme (finishing) ve dikiş (sewing) departmanlarının ortalama üretkenlik oranlarının birbirine oldukça yakın olduğu görülmektedir.
Dikiş departmanının ortalama işçi sayısının daha yüksek olmasına rağmen üretkenlik oranının bitirme departmanı ile benzer olması, dikiş departmanında verimlilik artırma potansiyelinin olabileceğini göstermektedir.

Uygunluk Değerlendirmesi

1. Eksik veri analizi
2. Bağımlı ve bağımsız değişkenlerin kontrolü
3. Sürekli ve kategorik değişkenlerin bir arada bulunması
4. Random Forest Algoritması uygunluğu

İşçi Üretkenliğinde Random Forest Analizi

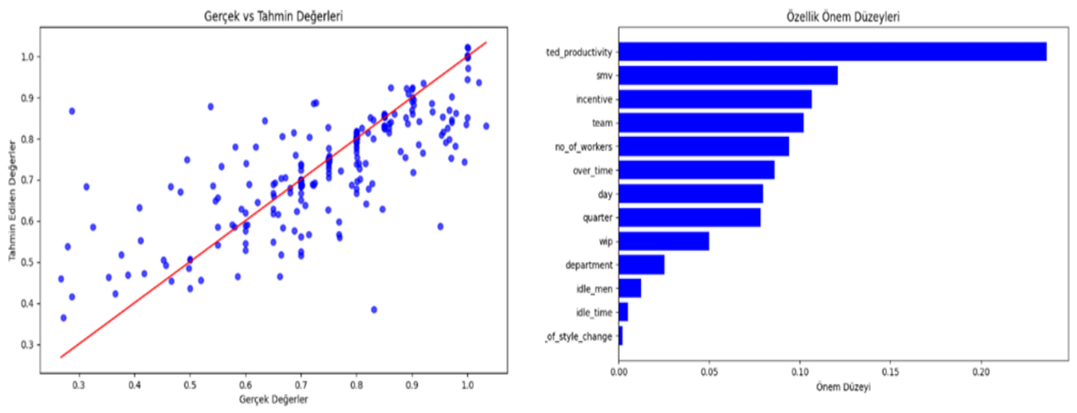
Bağımlı değişken
actual_productivity

%80 eğitim seti
% 20 test seti

Karar ağaç sayısı ve rastgelelelik parametreleri:
n_estimators 100 ağaç
random state 42

MSE	**0.015930**
RMSE	**0.126221**
R^2	**0.8985**

MSE değeri, modelin tahminlerindeki hata büyüklüğüdür. RMSE, MSE'nin karekökü alınarak hesaplanan ve hata büyüklüğünü gerçek birimlerde ifade etmektedir. Model değişkenliği %89.85 oranında açıklamaktadır. Model yüksek doğrulukta tahminler yapmakta ve tahminlerin büyük çoğunluğu gerçek değerlere yakınlık göstermektedir. Modelde “targeted_productivity”, “smv” ve “over_time" gibi değişkenleri üretkenlik tahmininde etkili faktör göstermiştir.



Sonuç

Bu çalışmada fabrika işçi verileriyle elde edilen Random Forest algoritması modeli, işçi üretkenliğini yüksek doğrulukla tahmin etmiştir. Modelin performans ölçümleri ve oluşturulan grafikler, modelin güvenilir ve doğru tahminler yaptığı sonucuna ulaştırmaktadır. Modelin doğru tahmin gücü, işletmelerin üretim ve işçi üretkenliklerini artırmada ve kaynaklarını daha verimli kullanmada fayda sağlamaktadır. Özellikle modele göre “targeted_productivity”, “smv” ve “over_time” gibi değişkenlerin üretkenlik üzerinde büyük etkisi olduğu belirlenmiştir.

KAYNAKÇA

[1] Pham, H. T., & Afify, A. (2010). Data mining techniques for manufacturing process control. Journal of Intelligent Manufacturing, 21(5), 613-624.
[2] Lee, H. W., & Chen, Y. T. (2015). The impact of training on employee productivity in a Taiwanese textile company. International Journal of Production Research, 53(1), 177-188.
[3] Nembhard, D. A., Shi, L., & Aktan, M. (2009). Workforce flexibility and manufacturing performance. Production Planning & Control, 20(5), 455-472.

YILDIZ TEKNİK ÜNİVERSİTESİ
Fen Edebiyat Fakültesi

İSTATİSTİK BÖLÜMÜ

Borsa İstanbul'da Hisse Senedi Fiyatlarının Yapay Zeka ile Tahmin Edilmesi

Enes Kaya
18023078

Tez Danışmanı: Prof. Dr. Filiz Karaman

Özet

Bu tez çalışmasında, Borsa İstanbul'daki hisse senetlerinin günlük kapanış fiyatları kullanılarak, gelecek fiyat hareketlerinin tahmin edilmesi amaçlanmıştır. Hisse senetlerinin geçmiş fiyat hareketlerinden yola çıkarak gelecekteki fiyatları tahmin edebilmek için, Python programlama dilinde çeşitli veri kütüphaneleri kullanılmış ve yapay zeka tekniklerinden Yinelemeli Sinir Ağı (Recurrent Neural Network, RNN) modeli uygulanmıştır.
Bu çalışmanın sonuçları, hisse senedi fiyatlarının tahmin edilmesinde RNN modelinin etkinliğini ve potansiyelini ortaya koymaktadır. Tez, finansal piyasalarda yapay zeka uygulamalarının önemini vurgulamakta ve bu alanda gelecekte yapılacak çalışmalara ışık tutmaktadır.

Borsa İstanbul Hisse Senedi ve Veri Seti Tanımı

Bu çalışmada, Borsa İstanbul'da işlem gören PETKM hisse senedinin günlük kapanış fiyatları kullanılmıştır. Veriler, 2015-2024 yılları arasındaki günlük kapanış fiyatlarını içermektedir. Bu veri seti, hisse senedi fiyat hareketlerini incelemek ve gelecekteki fiyatları tahmin etmek için kullanılmıştır.

Borsa İstanbul Tarihçesi



Borsa İstanbul (BIST), 1986 yılında İstanbul Menkul Kıymetler Borsası (İMKB) adıyla kuruldu. Türkiye'nin en büyük ve en önemli finansal piyasası olan İMKB, 2013 yılında İstanbul Altın Borsası ve Vadeli İşlem ve Opsiyon Borsası'nın birleşmesiyle Borsa İstanbul adını aldı. Borsa İstanbul, hisse senetleri, tahviller, vadeli işlemler ve diğer finansal araçların alım satımının yapıldığı bir platformdur. Türkiye'nin ekonomik büyümesine katkı sağlayarak yatırımcılar için çeşitli fırsatlar sunmaktadır.

Veri Seti Kaynağı:

```
from isyatirimhisse import StockData, Financials
stock_data = StockData()
hisse_senedi= 'PETKM'

df = stock_data.get_data(
    symbols=hisse_senedi,
    start_date='02-01-2018'
)
print(df)
```

Veriler, Python programlama dilinde kullanılan ''isyatirimhisse'' adlı bir Python paketi aracılığıyla elde edilmiştir. Bu paket, Borsa İstanbul'daki hisse senetlerinin geçmiş günlük fiyat değerlerini sağlamaktadır ve finansal analizler için güvenilir bir kaynaktır.

	DATE	CLOSING_TL	LOW_TL	HIGH_TL	VOLUME_TL	CLOSING_USD
0	2018-01-02	4.4864	4.357389	4.497589	2.264539e+08	1.1894
1	2018-01-03	4.4639	4.435900	4.536844	2.366647e+08	1.1845
2	2018-01-04	4.6097	4.469548	4.632179	2.963055e+08	1.2238
3	2018-01-05	4.6995	4.626571	4.738730	4.090684e+08	1.2524
4	2018-01-08	4.7163	4.710691	4.800418	3.418260e+08	1.2570
...	...	...	...	...	...	...
1591	2024-05-14	19.7800	19.410000	19.920000	1.259177e+09	0.6133
1592	2024-05-15	19.9000	19.650000	20.060000	8.895238e+08	0.6163
1593	2024-05-16	20.2600	19.950000	20.400000	2.032486e+09	0.6289
1594	2024-05-17	20.7600	20.320000	20.880000	2.127218e+09	0.6437
1595	2024-05-20	21.3600	20.760000	21.380000	2.900767e+09	0.6631

	LOW_USD	HIGH_USD	VOLUME_USD	XU100_TL	XU100_USD	CODE
0	1.1552	1.1924	6.003708e+07	1175.2420	311.571156	PETKM
1	1.1771	1.2039	6.280077e+07	1160.5215	307.945455	PETKM
2	1.1866	1.2297	7.866240e+07	1154.9119	306.610231	PETKM
3	1.2330	1.2629	1.090180e+08	1166.3794	310.835953	PETKM
4	1.2555	1.2794	9.110258e+07	1163.0451	309.977671	PETKM
...	...	...	...	...	...	...
1591	0.6018	0.6176	3.904110e+07	10173.4200	315.437739	PETKM
1592	0.6086	0.6213	2.754993e+07	10158.6300	314.611240	PETKM
1593	0.6193	0.6332	6.309089e+07	10319.9600	320.346636	PETKM
1594	0.6300	0.6474	6.595614e+07	10643.5800	330.022758	PETKM
1595	0.6444	0.6637	9.004707e+07	10739.5700	333.399292	PETKM

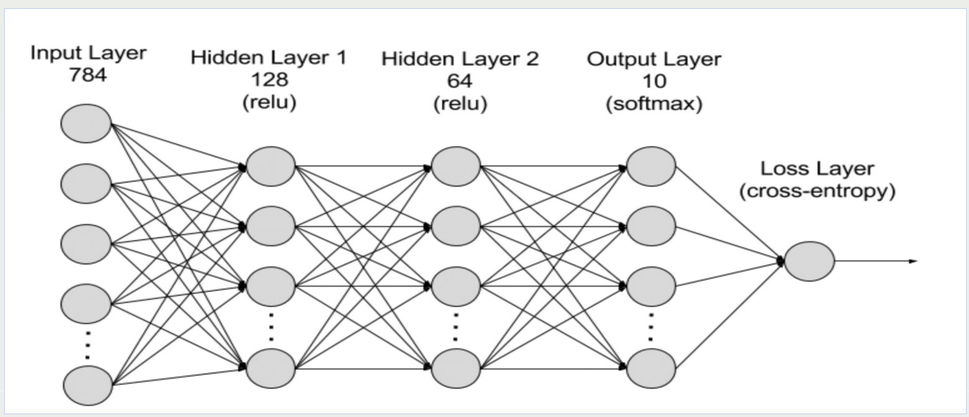
Veri Setinin Özellikleri:

Zaman Aralığı: 2015-2023 yılları arası
Frekans: Günlük kapanış fiyatları
Veri Tipleri: Tarih, kapanış fiyatı, işlem hacmi, açılış fiyatı, en yüksek ve en düşük fiyat,BIST100 endeks fiyatı,(***TL ve USD fiyatları ayrı olarak girilmiştir.)**

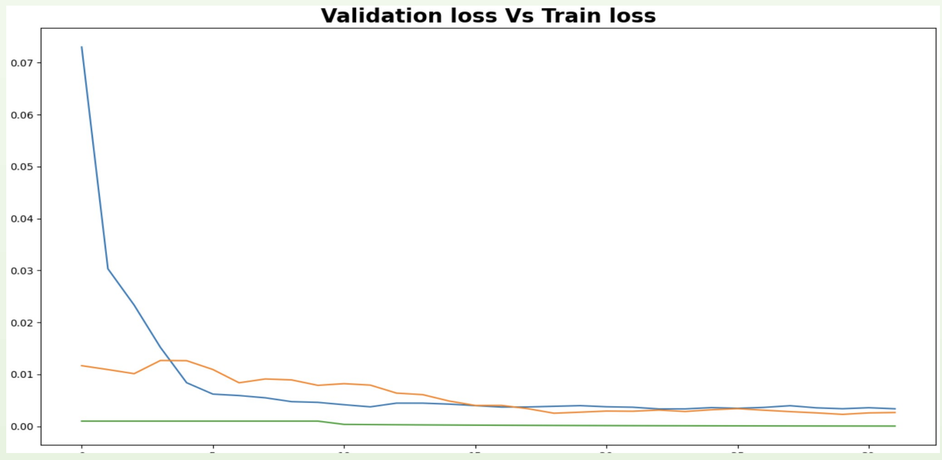
Kullanılan Algoritma (RNN)

Bu çalışmada, zaman serisi verisi olan hisse senedi fiyatlarının tahmini için Yinelemeli Sinir Ağı (Recurrent Neural Network, RNN) algoritması kullanılmıştır. RNN'ler, ardışık bağımlılıkları ve geçmiş verilerden öğrenerek gelecekteki fiyat hareketlerini tahmin etme yetenekleri nedeniyle ve verimizin de bir zaman serisi olması sebebiyle tercih edilmiştir. PETKM hisse senedinin günlük kapanış fiyatları kullanılarak yapılan analizde, RNN modelinin geçmiş fiyat verilerini başarılı bir şekilde öğrenip gelecekteki fiyatları tahmin edebildiği (accuracy) doğruluk puanımızın (0.79) yüksek olmasından görülmüştür. Çalışmanın sonuçları, RNN modelinin finansal piyasalarda etkin bir tahmin aracı olarak kullanılabileceğini ve yapay zeka tekniklerinin bu alandaki potansiyelini vurgulamaktadır.

RNN MODEL

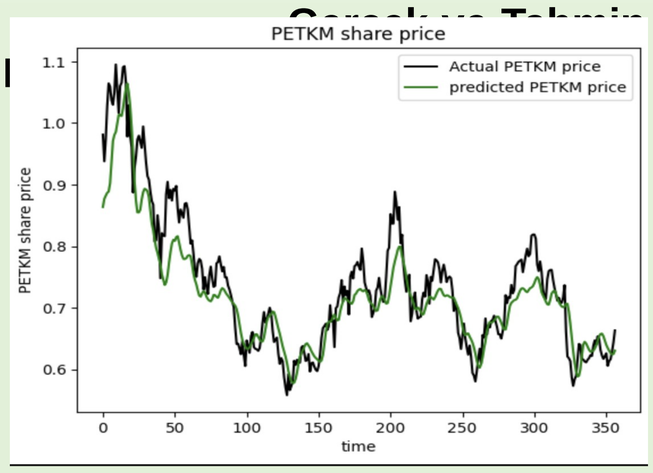


Model Eğitim ve Doğrulama Kayıpları



Bu grafikte modelimizin eğitim sürecini ve doğrulama hatalarını görmekteyiz. Eğitim süreci boyunca modelin eğitim (train) ve doğrulama (validation) kayıplarını görmekteyiz.
Eğitim kaybının (mavi) ve doğrulama kaybının (turuncu) zamanla azaldığı ve stabil hale geldiği görülmektedir, bu da modelin iyi bir şekilde genelleme yapabildiğini ve düzgün çalıştığını göstermektedir.

Gerçek ve Tahmin Edilen PETKM Hisse Senedi



Bu grafik, modelimizin tahmin ettiği hisse senedi fiyatı ile gerçek fiyatını karşılaştırmaktadır.
PETKM hisse senedinin gerçek fiyatları (siyah) ile , modelimiz tarafından tahmin edilen fiyatları (yeşil) olarak görmekteyiz. Modelimizin tahminlerinin gerçek fiyat hareketlerine yakın olduğu görülmektedir, bu da modelimizin doğruluğunu vurgulamaktadır.

```
# predicting next day
real_data = [model_inputs[len(model_inputs+1) - sequence_length:len(model_inputs+1),0]]
real_data = np.array(real_data)
real_data = np.reshape(real_data, (real_data.shape[0], real_data.shape[1], 1))
✓ 0.0s

print(real_data.shape)
prediction = regressor.predict(real_data)
prediction = scaler.inverse_transform(prediction)
print(prediction)
✓ 0.2s
(1, 50, 1)
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 197ms/step
[[0.6431367]]
```

```
from sklearn.metrics import r2_score

print("Accuracy score of the predictions for {0}: {1}".format(hisse_senedi, r2_score(actual_prices, predicted_prices)))
✓ 0.0s
Accuracy score of the predictions for PETKM: 0.7977515891669311
```

SONUÇ

Modelimiz PETKM hisse senedinin ertesi günkü fiyatını 0.6431 USD (yaklaşık 20.75 TL) olarak tahmin etmiştir. PETKM hissesi gün içinde 20.75 TL'yi görüp, kapanışa doğru 21.36 TL'ye yükselmiştir. Bu ani artış, son çeyrek bilanço etkisi ve şirket temettü haberi gibi temel analiz unsurlarının etkisiyle gerçekleşmiştir. Bu durum, hisse senedi fiyatlarının teknik analiz kadar temel analizlerden de etkilendiğini göstermektedir.
Modelimiz, piyasa koşulları normal seyrettiğinde yüksek doğrulukla tahminler yapmaktadır. Modelin doğruluk oranı %79.77 olarak hesaplanmıştır, bu da modelin performansının yüksek ve başarılı olduğunu göstermektedir.
Temel analiz unsurları fiyatlarda değişimlere yol açabilir. Yine de, modelimizin güçlü tahmin yetenekleri yatırımcılara değerli bilgiler sunarak doğru kararlar alınmasına ve hissenin gelecekteki fiyatı hakkında fikir sahibi olmasına yardımcı olacaktır.

KAYNAKÇA

- https://stackoverflow.com/questions/54116693/modulenotfounderror-no-module-named-seaborn-in-python-ide
- https://medium.com/deep-learning-turkiye/rnn-nedir-nas%C4%B1l-%C3%A7al%C4%B1%C5%9F%C4%B1r-9e5d572689e1
- https://www.investaz.com.tr/hisse-senetleri/hisse-senedi-tarihcesi
- https://www.kaggle.com/code/volkanmazlum/notebookb3749a21c5

İSTATİSTİK BÖLÜMÜ

1996-2021 YILLARI ARASINDAKİ BASKETBOLCULARIN İSTATİSTİKLERİNE GÖRE ANALİZLERİ

Eren Boz 19023002

Danışman: Doç. Dr. Elif TUNA

ÖZET

Bu çalışmamda sizler ile 1996-2021 yılları arasında oynamış olan basketbolcuların oynadığı dönemleri, o dönemde attığı sayıları, aldığı ribaundları, asist sayılarını ve bazı kişisel özelliklerini(boy,kilo,yaş) içinde barındıran bir veri seti üzerinden farklı analizler ve bu analizleri tablolar haline getirerek kendi yorumlamalarımla birlikte bir sunum hazırlamak istedim. Bu sunumda bazen tek bir basketbolcu üzerinden bazen ise bu veri setindeki bütün basketbolcuların bulunduğu analizler üzerinden ilerlemeyi tercih ettim ve bunun üzerine yorumlar yaptım.

1996-2021 YILLARI ARASINDAKİ BASKETBOLCULARIN İSTATİSTİKLERİNE GÖRE ANALİZLERİ

Basketbolda İstatistik Nedir ve Neden Gereklidir?

Basketbolda İstatistik Nedir?

Basketbolda istatistik, oyuncu performanslarını, takım dinamiklerini ve oyun içi stratejileri değerlendirmek için kullanılan sayısal verilerdir. Bu veriler, oyun sırasında toplanan çeşitli parametrelerden elde edilir. Basketbolda yaygın olarak kullanılan bazı istatistikler şunlardır:

- Puan (Points - PTS): Oyuncunun bir maçta attığı toplam sayı.
- Ribaund (Rebounds - REB): Oyuncunun bir maçta aldığı toplam ribaund sayısı (savunma ve hücum ribaundları dahil).
- Asist (Assists - AST): Oyuncunun takım arkadaşlarına yaptığı ve onların sayı yapmasını sağlayan pas sayısı.
- Blok (Blocks - BLK): Oyuncunun rakip takımın atışlarını engelleme sayısı.
- Top Çalma (Steals - STL): Oyuncunun rakip takımın topunu çalma sayısı.
- Top Kaybı (Turnovers - TO): Oyuncunun yaptığı top kaybı sayısı.
- Şut Yüzdesi (Field Goal Percentage - FG%): Oyuncunun yaptığı başarılı şutların toplam şut girişimlerine oranı.
- Serbest Atış Yüzdesi (Free Throw Percentage - FT%): Oyuncunun yaptığı başarılı serbest atışların toplam serbest atış girişimlerine oranı.

Neden Gereklidir?

- Performans Değerlendirme
- Karar Verme Süreci
- Stratejik Planlama
- Gelişim ve Eğitim
- Taraftarlar ve Medya

Sonuç

Basketbolda istatistikler, oyunun çeşitli yönlerini anlamak ve değerlendirmek için kritik bir araçtır. Performans değerlendirmesinden stratejik planlamaya, eğitim ve gelişimden taraftar deneyimine kadar birçok alanda önemli rol oynar. Bu nedenle, basketbol istatistikleri hem oyuncular hem de takımlar için vazgeçilmez bir araçtır.

Oyuncu Bazında İstatistik Değerlendirmesi

	name	team	age	draft_year	year	total_pts	total_ast	total_reb
7298	Kobe Bryant	LAL	34.0	1996	2012	31603	5894	6575
7651	Kobe Bryant	LAL	35.0	1996	2013	31685	5931	6600
8184	Kobe Bryant	LAL	36.0	1996	2014	32465	6127	6799
8709	Kobe Bryant	LAL	37.0	1996	2015	33626	6311	7043
0	Kobe Bryant	LAL	37.0	1996	2016	33626	6311	7043
1	Kobe Bryant	LAL	37.0	1996	2017	33626	6311	7043

Bu tabloya baktığımız zaman ‘Kobe Bryant’ isimli basketbolcunun istatistiklerini görüyoruz. Tabloyu değerlendirecek olursak Kobe Bryant’ın 34 yaşından 37 yaşının sonuna kadar olan grafikleri karşımıza çıkıyor fakat 37 yaşından sonra basketbolu bıraktığı için istatistikleri zaman geçse de değişmemiş oluyor ve sabit bir şekilde kalıyor. Bu tabloları aynı kodu kullanıp sadece koddaki isimleri değiştirerek bütün oyuncular için oluşturabiliriz ve bunu basketboldaki son 4 senesi olarak değil son 2 senesi olarak ya da son 10 senesi olarak da değerlendirmeye alabilmemiz mümkün. Aktif basketbol hayatı boyunca her sene istatistiklerinin değiştiğini gözlemleyeceğiz.

Top 10 oyuncu sıralaması

```
stats.sort_values('pts_total',ascending=False).head(10)
```

Bu kodu kullanarak sağ tarafta gördüğümüz en çok sayı atmış ilk 10 oyuncuyu sırasıyla görüntüleyebiliriz.
bu sıralamaları verimizdeki bütün değişkenlerimiz için yapabiliriz ve bu alanlarda en üstte bulunan ilk 10 oyuncuyu görebiliriz.

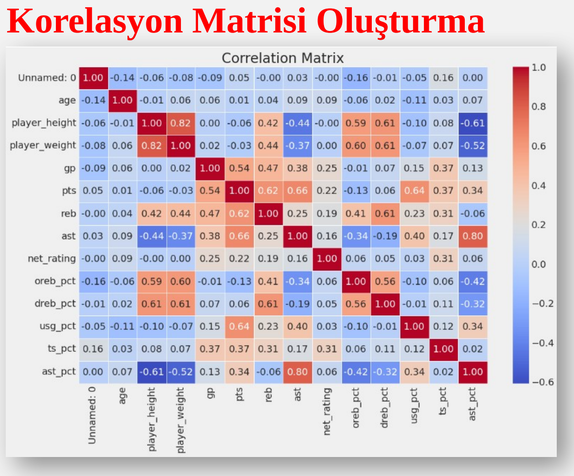
player_name	gp	pts	pts_total	reb	reb_total	ast	ast_total
LeBron James	1310	26.933333	35283.0	7.461111	9774.0	7.433333	9738.0
Kobe Bryant	1346	24.200000	32573.0	5.200000	6999.0	4.760000	6407.0
Dirk Nowitzki	1522	20.109524	30607.0	7.347619	11183.0	2.338095	3559.0
Carmelo Anthony	1191	22.500000	26798.0	6.272222	7470.0	2.727778	3249.0
Tim Duncan	1392	18.889474	26294.0	10.768421	14990.0	3.010526	4191.0
Paul Pierce	1343	18.931579	25425.0	5.436842	7302.0	3.357895	4510.0
Vince Carter	1541	16.345455	25188.0	4.227273	6514.0	2.972727	4581.0
Ray Allen	1300	18.883333	24548.0	4.050000	5265.0	3.377778	4391.0
Allen Iverson	914	26.064286	23823.0	3.692857	3375.0	6.000000	5484.0
Kevin Durant	884	26.884615	23766.0	7.076923	6256.0	4.238462	3747.0

```
stats.sort_values('reb_total',ascending=False).head(5)
```

Bir diğer örnekte de en çok ribaund alan ilk 5 oyuncuyu bu şekilde görüntülemiş olabiliriz.

player_name	gp	pts	pts_total	reb	reb_total	ast	ast_total
Tim Duncan	1392	18.889474	26294.0	10.768421	14990.0	3.010526	4191.0
Dwight Howard	1182	15.929412	18829.0	11.858824	14017.0	1.329412	1571.0
Dirk Nowitzki	1522	20.109524	30607.0	7.347619	11183.0	2.338095	3559.0
Pau Gasol	1226	16.561111	20304.0	9.044444	11088.0	3.155556	3869.0
Tyson Chandler	1160	7.821053	9072.0	8.721053	10116.0	0.794737	922.0

Korelasyon Matrisi Oluşturma



Veri setimizde 14 tane değişkenimiz olduğundan burada 14 satır 14 sütun olacak şekilde bir korelasyon matrisi oluşturabiliriz. Bu korelasyon matrisi, iki değişken arasındaki doğrusal ilişkiyi gösterir. Korelasyon katsayısı -1 ile 1 arasında değişir. 1'e yakın pozitif değerler güçlü pozitif korelasyonu, -1'e yakın negatif değerler güçlü negatif korelasyonu ve 0'a yakın değerler zayıf veya hiç korelasyon olmadığını gösterir.

Korelasyon Matrisi Analizi:

Player Height ve Player Weight:
Korelasyon: 0.82
Yorum: Bu iki değişken arasında güçlü bir pozitif korelasyon var. Yani, oyuncu boyu arttıkça oyuncu ağırlığı da artma eğilimindedir.

Ast ve Pts:
Korelasyon: 0.66
Yorum: Asist ve puan arasında pozitif bir korelasyon var. Bu, daha fazla asist yapan oyuncuların genellikle daha fazla puan aldığını gösterir.

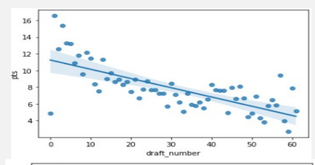
Oreb_pct ve Player Weight:
Korelasyon: 0.60
Yorum: Hücum ribaund yüzdesi ve oyuncu ağırlığı arasında pozitif bir korelasyon var. Daha ağır oyuncuların hücum ribaundlarında daha etkili olduğunu gösterir.

Ast_pct ve Ast:
Korelasyon: 0.80
Yorum: Asist yüzdesi ve asist sayısı arasında güçlü bir pozitif korelasyon var. Bu, daha fazla asist yapan oyuncuların genellikle daha yüksek asist yüzdesine sahip olduğunu gösterir.

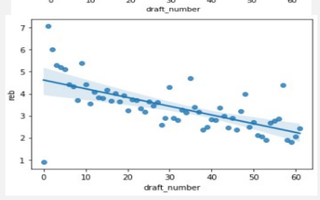
Dreb_pct ve Player Weight:
Korelasyon: 0.61
Yorum: Savunma ribaund yüzdesi ve oyuncu ağırlığı arasında pozitif bir korelasyon var. Daha ağır oyuncuların savunma ribaundlarında daha etkili olduğunu gösterir.

Regresyon Grafikleri

```
sns.regplot(x='draft_number',y='pts',data=avg_by_pick)
plt.show()
```




```
sns.regplot(x='draft_number',y='reb',data=avg_by_pick)
plt.show()
```



Regresyon grafiklerine bakarak da oyuncuların draft edilme sıralarıyla sezon içerisindeki sayı ve ribaund istatistikleri arasındaki bağlantıyı görebiliriz.

SONUÇ

Bu çalışmada, basketbol verilerini analiz ederek korelasyon matrisleri,oyuncu grafik tabloları ve en iyi 10 oyuncudan oluşan tablolar gibi grafik ve tablolar oluşturduk ve yorumladık. Korelasyon matrisi, oyuncu performansları arasındaki ilişkileri anlamamıza yardımcı oldu. Özellikle oyuncu boyu ile ağırlığı arasında güçlü pozitif bir korelasyon, asist ve puan arasında anlamlı bir ilişki ve bazı performans metrikleri arasında dikkat çeken negatif korelasyonlar tespit ettik. Bu analizler, basketbol takımlarının oyuncu seçiminde ve stratejik kararlarında önemli bir rol oynayabilir.

Bu tür analizlerin, basketbol oyuncularının ve takımlarının performansını değerlendirirken güçlü bir araç olduğunu gördük. Ancak, korelasyonun nedensellik anlamına gelmediğini ve bu tür analizlerin daha derinlemesine araştırmalar için bir başlangıç noktası olarak kullanılması gerektiğini unutmamak önemlidir. Bu çalışma, basketbolda istatistiksel analizlerin önemini ve bu analizlerin nasıl yapılabileceğini göstermiştir.

KAYNAKÇA

[1] https://tr.nbafantasy.nba.com/statistics
[2] https://www.kaggle.com/
[3] https://www.eurohoops.net/

İSTATİSTİK BÖLÜMÜ

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KALP YETMEZLİĞİNİN TAHMİN EDİLMESİ

20023009 Eren Enöz

Danışman: Prof. Dr.Gülhayat Gölbaşı Şimşek

ÖZET

Bu tez çalışması , çeşitli ülkelerden toplanan hasta verilerinin incelenerek gelecekte gerçekleşmesi olası kalp yetmezliğinin tahmin edilmesini amaçlamaktadır. Verilerimiz betimsel istatistiklerle özetlenmiş ve analiz öncesinde gerekli ön işleme yöntemleri uygulanmıştır. Ardından korelasyon yardımıyla değişkenlerin kalp rahatsızlığına olan yönelimleri incelenmiştir. Son olarak denetimli makine öğrenmesi yöntemlerinden lojistik regresyon ,destek vektör makinesi ve en yakın komşu algoritması kullanılarak kalp yetmezliği riski tahmin edilmiştir.

HASTALARIN DEMOGRAFİK, KLİNİK VE EGZERSİZ TESTİ SONUÇLARI

İndeks	Age	Sex	ChestPain Type	RestingBP	Cholesterol	FastingBS	Resting ECG	MaxHR	ExerciseAngina	Old peak	ST Slope	Heart Disease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
5	39	M	NAP	120	339	0	Normal	170	N	0.0	Up	0
6	45	F	ATA	130	237	0	Normal	170	N	0.0	Up	0
7	54	M	ATA	110	208	0	Normal	142	N	0.0	Up	0
8	37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
9	48	F	ATA	120	284	0	Normal	120	N	0.0	Up	0
10	37	F	NAP	130	211	0	Normal	142	N	0.0	Up	0

918 hastaya ait bilgiler Amerika , İsviçre , Macaristan gibi ülkelerden toplanmıştır ve açık kaynak olarak paylaşılmaktadır.. 11 farklı değişken ve hastaların hali hazırda bir kalp rahatsızlığına sahip olma durumuna göre verilerimiz listelenmiştir. Dinlenme sırasındaki kan basıncı [RestingBP] değerleri 120 mg/dl'nin altında olanlar 0 olarak kayda geçmiştir. Verilerimizde boş değer bulunmamaktadır.

BETİMSEL İSTATİSTİKLER

	Sayı	Ortalama	Standart sapma	Minimum Değer	25%	50%	75%	Maksimum Değer
Age	918.00	53.51	18.51	28.00	47.00	54.00	60.00	77.00
RestingBP	918.00	132.40	18.51	0.00	120.00	130.00	140.00	200.00
Cholesterol	918.00	198.80	109.38	0.00	173.25	223.00	267.00	603.00
FastingBS	918.00	0.23	0.42	0.00	0.00	0.00	0.00	1.00
MaxHR	918.00	136.81	25.46	60.00	120.00	138.00	156.00	202.00
Oldpeak	918.00	0.89	1.07	-2.60	0.00	0.60	1.50	6.20
HeartDisease	918.00	0.55	0.50	0.00	0.00	1.00	1.00	1.00

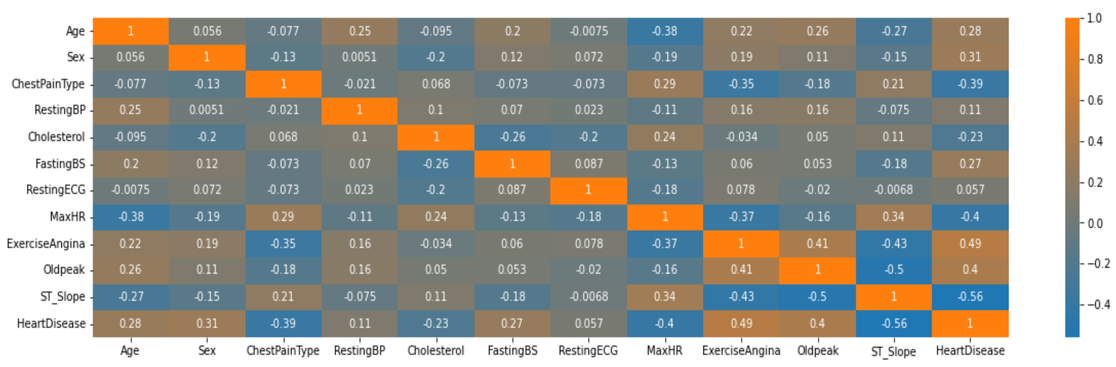
Sayısal değerlerimizin temel istatistikleri tablodaki gibidir. Hastalarımızın yaş ortalaması yaklaşık olarak 54 olup en genç hastamız 28, en yaşlı hastamız ise 77 yaşındadır. Hastalarımızın belirli bir kalp rahatsızlığına sahip olma durumu yaklaşık olarak yarı yarıyadır. Dinlenme sırasındaki kan basıncı [RestingBP] değerleri yalnızca 120 mg/dl'nin altında olan değerleri içerdiğinden minimun değer 0 gelmektedir maksimum kan basıncı ise 200 mg/dl'dir.

VERİ ÖN İŞLEME

	Age	Sex	ChestPain Type	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	Exercise Angina	Oldpeak	ST Slope	Heart Disease
0	40	1	1	140	289	0	1	172	0	0.0	2	0
1	49	0	2	160	180	0	1	156	0	1.0	1	1
2	37	1	1	130	283	0	2	98	0	0.0	2	0
3	48	0	0	138	214	0	1	108	1	1.5	1	1
4	54	1	2	150	195	0	1	122	0	0.0	2	0

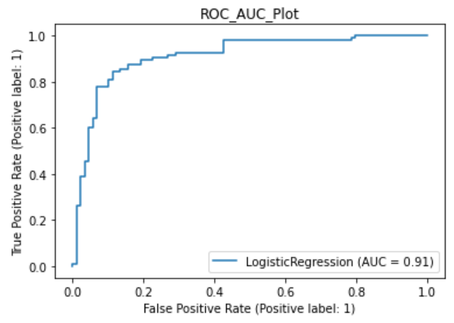
Makine Öğrenmesi algoritmalarında kategorik veriler problem oluşturabilmektedir bunun olmaması adına bazı değişiklikler yapılmıştır. Veri tipi değişen sütunlarımız : Sex., ChestPain Type, RestingECG, ExerciseAngina ve ST Slope. Ayrıca Oldpeak, Age, RestingBP, Cholesterol, MaxHR sütunlarına normalizasyon ve standartlaştıma işlemleri uygulanmıştır.

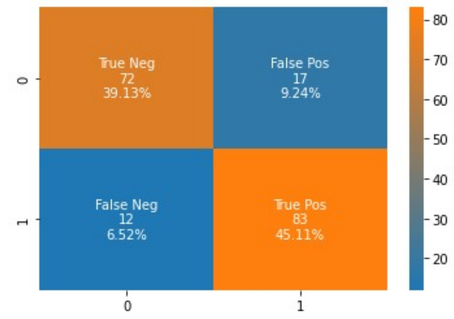
Korelasyon Matrisi



Araştımanın özü diğer değişkenler ve kalp rahatsızlığı arasındaki etkileşimi incelemek olduğundan korelasyon grafiğinde bu değişkenler ve kalp rahatsızlığı arasındaki değerlere odaklanmak mantıklı olacaktır. Bu durumda RestingBP ve RestingECG değişkenlerinin kalp rahatsızlığıyla olan etkileşimi diğer değişkenlere daha düşük olduğunu hatta RestingBP ve RestingECG'nin korelasyonlarının 0'a yakınlığı sebebiyle kalp rahatsızlığı üzerine pozitif veya negatif bir etkisi olmadığı gözükmektedir.

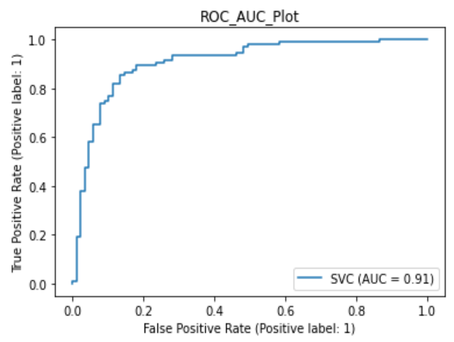
LOJİSTİK REGRESYON





Lojistik Regresyon modelinin test sonuçlarına göre elde edilen hata matrisi kalp rahatsızlığına sahip olanları %45,11 doğru ve %6,52 yanlış tahminle kalp rahasızlığına sahip olanları tespit etme konusunda başarılı olduğu gözlenmektedir. Hasta olmayanları tespit etme konusunda ise %39,13 doğru %9,24 hatalı tahmine sahip olduğu görülmektedir. Bu sonuçlara göre modelin hem hasta olanları hem de hasta olmayanları tahmin etme konusuna başarılı olduğu gözlenmektedir.

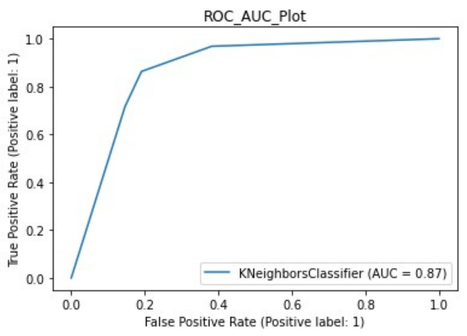
DESTEK VEKTÖR MAKİNESİ





Destek Vektör Makinesi yöntemiyle elde edilen hata matrisi ve grafiği yukarıdaki şekildeki gibidir. Modelimize göre kalp rahatsızlığına sahip olanları %46,20 doğru ve %5,43 yanlış tahminle kalp rahatsızlığına sahip olanları tespit etme konusunda başarılı olduğu gözlenmektedir. Hasta olmayanları tespit etme konusunda ise %39,67 doğru %8,70 hatalı tahmine sahip olduğu görülmektedir. Bu sonuçlara göre modelin hem hasta olanları hem de hasta olmayanları tahmin etme konusuna başarılı olduğu gözlenmektedir.

EN YAKIN KOMŞU ALGORİTMASI





En Yakın Komşu Algoritması yöntemiyle elde edilen hata matrisi ve grafiği yukarıdaki şekildeki gibidir. Modelimize göre kalp rahatsızlığına sahip olanları %41,30 doğru ve %10,33 yanlış tahminle kalp rahatsızlığına sahip olanları tespit etme konusunda başarılı olduğu gözlenmektedir ancak iyileştirelebilicek durum olduğu söylenebilir. Hasta olmayanları tespit etme konusunda ise %33,15 doğru %15 hatalı tahmine sahip olduğu görülmektedir ancak hatalı tahmin yüzdesinin yüksekliği dikkat edilmesi gerekilen bir detaydır. Bu sonuçlara göre modelin hem hasta olanları hem de hasta olmayanları tahmin etme konusuna başarılı olduğu fakat geliştirilebilir olduğu gözlenmektedir.

Sonuç

Model	Doğruluk	Çapraz geçerlilik Skoru	ROC_AUC Skoru
Lojistik Regresyon	84.24%	91.08%	84.13%
Destek Vektör	85.87%	90.72%	85.75%
En Yakın Komşu Algoritması	83.70%	89.58%	83.61%

Destek Vektör Makinesi (SVC), hem doğruluk hem de ROC_AUC skoru açısından en iyi performansı gösteren modeldir. Bu model, sınıfları daha doğru ve genel olarak ayırt etmede daha başarılıdır.
Lojistik Regresyon, çapraz doğrulama skorunda en yüksek değeri alarak, genel performansta tutarlılık göstermektedir. Bu model, özellikle eğitim ve test setleri arasında tutarlı performans gerektiren durumlar için uygun olabilir.
K-En Yakın Komşu (K-NN), diğer iki modelin biraz gerisinde kalmaktadır, ancak yine de makul bir performans göstermektedir. K-NN, parametre ayarlamalarına (örneğin, k değeri) oldukça duyarlıdır ve bu, performansını etkileyebilir

KAYNAKÇA

- Kachuee M, Fazeli S, Sarrafzadeh M. ECG heartbeat classifcation: a deep transferable representation. In: 2018 IEEE international conference on healthcare informatics (ICHI). New York: IEEE; 2018. p. 443–4
- Machine learning and statistical methods for predicting mortality in heart failure Published: 09 November 2020
- Moody GB, Mark RG. The impact of the mit-bih arrhythmia database. IEEE Eng Med Biol Mag. 2001;20(3):45–50

İSTATİSTİK BÖLÜMÜ

KARAR AĞAÇLARI YARDIMIYLA SİGORTA SEKTÖRÜNDE KASKO RİSK ANALİZİ

Faden Alaca 19023064

Danışman: Doç. Dr. Elif TUNA

ÖZET

Kasko risk analizi uygulamasında, sigortacılık sektörünün kasko branşında bazı poliçe bilgileri üzerinden risk analizi yapılmıştır. Uygulamada kullanılan veri setinde 10.000 veri bulunmaktadır. Bu çalışmada, veri setindeki değişkenleri kullanarak net prim üzerindeki etkilerini tahmin edilmiş ve bu şekilde risk analizi yapılmıştır. Analizlerle, sigorta şirketlerine yeni kayıt yaptıran kişilerin poliçedeki bilgilerine göre net prim üzerindeki etkilerini öngörülebilecektir. Bu sayede sigorta şirketleri, yeni müşterilere daha doğru primler belirleyerek risk yönetiminde daha etkili olabilirler. Bu çalışmada modelleme tekniği olarak karar ağaçları kullanılmıştır. Bu analiz, Python programı ile gerçekleştirilmiş olup, veri temizleme, model oluşturma ve sonuçların değerlendirilmesi aşamalarını kapsamaktadır.

AMAÇ

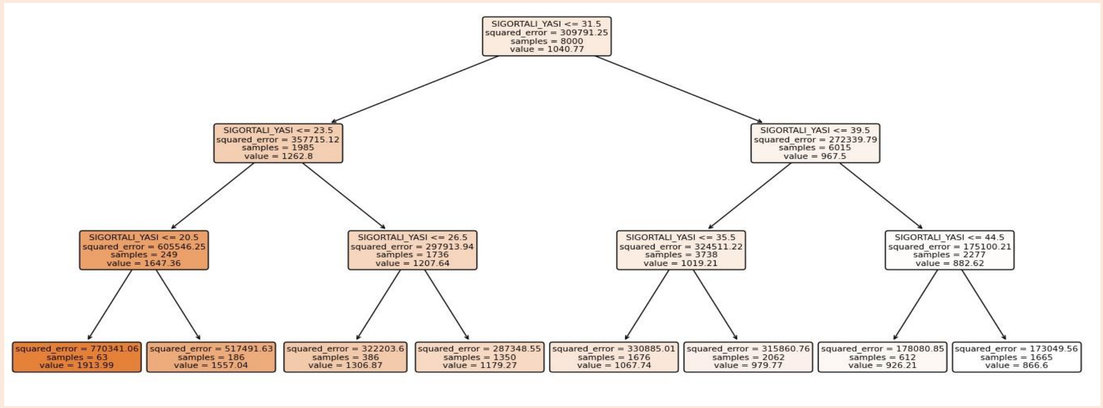
Yapılan araştırmanın amacı, kasko risk analizinde sigortacılık sektörünün kasko branşında poliçe bilgileri üzerinden karar ağaçları yardımıyla risk analizi yapılmasıdır. Yapılan araştırmada sigortalının yaşı, aracın model yılı, aracın yaşı, aracın bedeli ve hasarsızlık kademesi değişkenleri yer almaktadır. Bu bağlamda, veri setindeki değişkenleri kullanarak net prim üzerindeki etkilerini tahmin etmek ve karar ağaçlarıyla risk analizi yapmak amaçlanmaktadır.

YÖNTEM

Tez kapsamında, karar ağaçları yardımıyla Kasko risk analizi uygulamasında, sigortacılık sektörünün kasko branşında poliçe bilgileri üzerinden risk analizi yapılmıştır. Veriler Python programı ile analiz edilmiş olup, veri temizleme, model oluşturma ve sonuçların değerlendirilmesi aşamalarını kapsamaktadır.

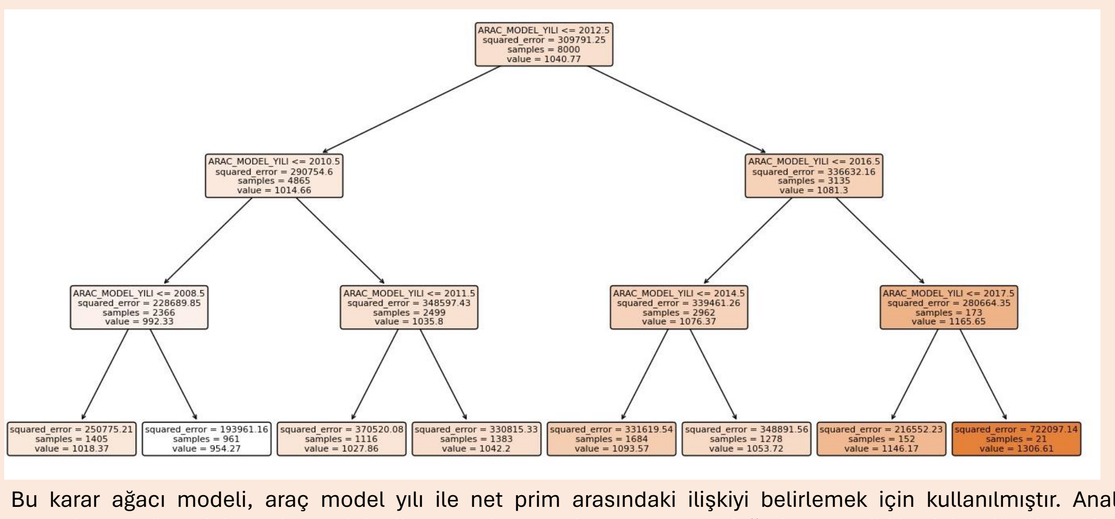
ANALİZLER

• SİGORTALI YAŞI DEĞİŞKENİYLE NET PRİM ANALİZİ



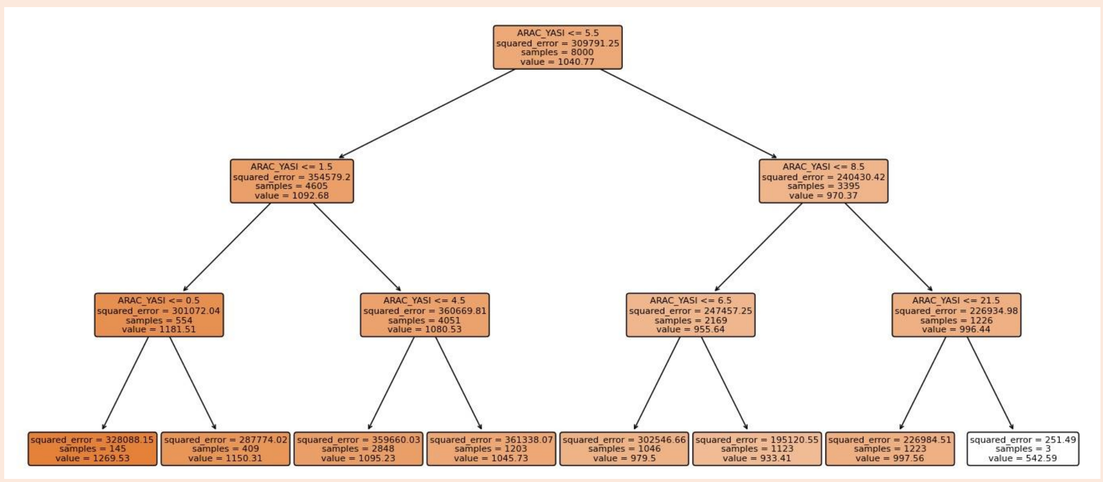
Bu karar ağacı modeli, sigortalı yaşı ile net prim arasındaki ilişkiyi belirlemek için kullanılmıştır. Analiz sonuçlarına göre, genç sigortalı yaş gruplarında net prim değerleri daha yüksek seviyelerde bulunmaktadır. Yani sigortalı yaşı arttıkça net prim değerlerinde genel bir düşüş gözlemlenmiştir. Her bir düğümdeki örnek sayıları ve ortalama değerler, bu ilişkinin detaylarını ortaya koymaktadır. Bu çalışma, sigortalı yaşı ile net prim arasındaki ilişkiyi derinlemesine inceleyerek, sigorta sektöründeki stratejik kararlar için önemli bir veri kaynağı sunmaktadır.

• ARAÇ MODEL YILI DEĞİŞKENİYLE NET PRİM ANALİZİ



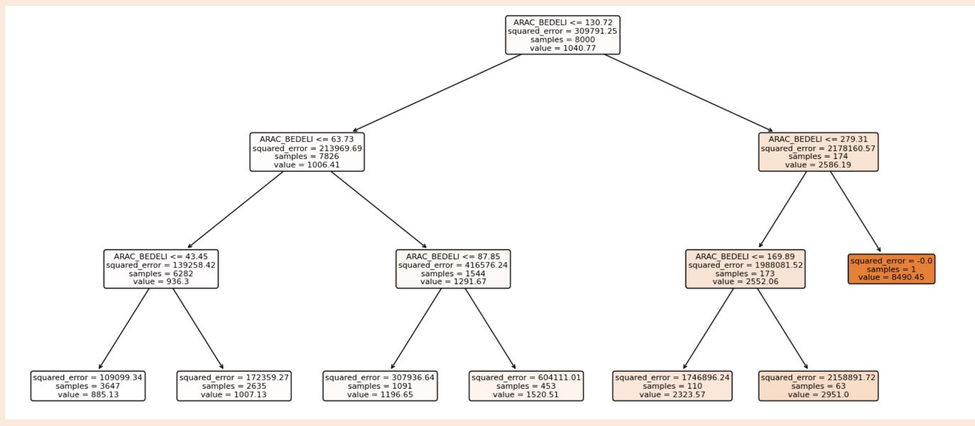
Bu karar ağacı modeli, araç model yılı ile net prim arasındaki ilişkiyi belirlemek için kullanılmıştır. Analiz sonuçlarına göre, daha eski model yıllarına sahip araçların net prim değerleri genellikle daha düşüktür. Araç model yılı arttıkça, net prim değerlerinde genel bir artış gözlemlenmektedir. Özellikle 2016 yılı ve sonrası model yıllarındaki araçlar için prim değerleri, genel ortalamaya göre daha yüksek seviyelerdedir. Bu analiz sonucunda, araç model yılı ve net prim arasındaki ilişkiyi derinlemesine inceleyerek, sigorta sektöründeki stratejik kararlar için önemli bir veri kaynağı sunmaktadır.

• ARAÇ YAŞI DEĞİŞKENİYLE NET PRİM ANALİZİ



Bu karar ağacı modeli, araç yaşı ile net prim arasındaki ilişkiyi belirlemek için kullanılmıştır. Analiz sonuçlarına göre, daha genç yaşa sahip araçların net prim değerleri genellikle daha yüksektir Araç yaşı arttıkça, net prim değerlerinde genel bir düşüş gözlemlenmektedir. Bu analiz sonucunda, araç yaşı ve net prim arasındaki ilişkiyi derinlemesine inceleyerek, sigorta sektöründeki stratejik kararlar için önemli bir veri kaynağı sunmaktadır.

• ARAÇ BEDELİ DEĞİŞKENİYLE NET PRİM ANALİZİ



Bu karar ağacı modeli, araç bedeli ile net prim arasındaki ilişkiyi belirlemek için kullanılmıştır. Analiz sonuçlarına göre, daha düşük bedelli araçların net prim değerleri genellikle daha düşük seviyelerdedir. Araç bedeli arttıkça, net prim değerlerinde genel bir artış gözlemlenmektedir. Özellikle, araç bedeli 169.89'un üzerinde olan araçlar için prim değerleri genel ortalamaya göre oldukça yüksek seviyelerdedir. Sigorta şirketlerinin araç bedeli ve net prim arasındaki ilişkiyi derinlemesine inceleyerek, sigorta sektöründeki stratejik kararlar için önemli bir veri kaynağı sunmaktadır.

• HASARSIZLIK KADEMESİ DEĞİŞKENİYLE NET PRİM ANALİZİ

Bu karar ağacı modeli, hasarsızlık kademesi ile net prim arasındaki ilişkiyi belirlemek için kullanılmıştır. düşük hasarsızlık kademesine sahip sigortalıların net prim değerleri genellikle daha yüksek seviyelerdedir Hasarsızlık kademesi düştükçe, net prim değerlerinde genel bir azalma gözlemlenmektedir Özellikle, hasarsızlık kademesi 4'ün üzerinde olan sigortalılar için prim değerleri genel ortalamaya göre oldukça yüksek seviyelerdedir. Sigorta şirketlerinin, hasarsızlık kademesi ve net prim arasındaki ilişkiyi derinlemesine inceleyerek, sigorta sektöründeki stratejik kararlar için önemli bir veri kaynağı sunmaktadır.

• SONUÇLAR

Bu çalışma, sigorta sektöründe stratejik kararlar için beş farklı değişkenin (sigortalı yaşı, araç model yılı, araç yaşı, araç bedeli ve hasarsızlık kademesi) net prim ile olan ilişkisini karar ağacı modelleri kullanarak incelemiştir. Genç sigortalıların net prim değerlerinin daha yüksek olduğu, yaş arttıkça ise net prim değerlerinde genel bir düşüş gözlemlenmiştir. Araç model yılı arttıkça net prim değerlerinde genel bir artış gözlemlenmiş, daha eski model yıllarına sahip araçların net prim değerlerinin daha düşük olduğu belirlenmiştir. Daha genç yaşa sahip araçların net prim değerlerinin daha yüksek olduğu, araç yaşı arttıkça ise net prim değerlerinde genel bir düşüş gözlemlenmiştir. Araç bedeli arttıkça net prim değerlerinde de artış gözlemlenmiş, daha düşük bedelli araçların net prim değerlerinin daha düşük olduğu belirlenmiştir. Düşük hasarsızlık kademesine sahip sigortalıların net prim değerlerinin daha yüksek olduğu, hasarsızlık kademesi arttıkça net prim değerlerinde genel bir düşüş gözlemlenmiştir. Bu bulgular, sigorta şirketlerinin stratejik kararlarında dikkate alması gereken önemli verileri sunmaktadır.

KAYNAKÇA

[1] https://miracozturk.com/python-ile-siniflandirma-analizleri-karar-agaclari/
[2] https://medium.com/yaz%C4%B1l%C4%B1m-ve-bili%C5%9Fim-kul%C3%BCb%C3%BC/decision-trees-karar-a%C4%9Fa%C3%A7lar%C4%B1-python-d57056105578
...

İSTATİSTİK BÖLÜMÜ

TEMIZ SU TÜKETİMİ TAHMİNİNDE ZAMAN SERİSİ YÖNTEMLERİNİN KIYASLANMASI

Fatma ERDEM - 19023017

Danışman: Doç. Dr. Serpil KILIÇ DEPREN

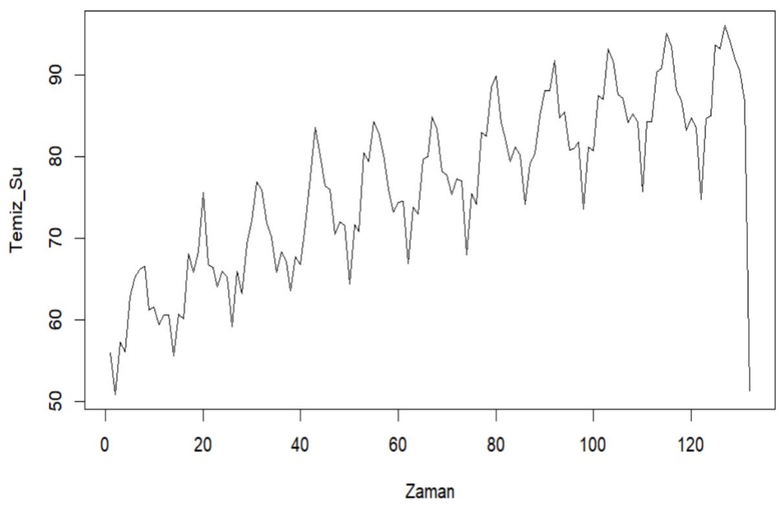
ÖZET

Bu çalışma, İstanbul Su ve Kanalizasyon İdaresi'nin (İSKİ) 2009-2019 yılları arasındaki İstanbul'a sağladığı temiz su miktarlarının aylara göre verisini kullanarak gelecek aylar için kullanılacak olan en iyi zaman serisi modelini bulmayı amaçlamaktadır. Çalışmada, zaman serilerindeki trend ve mevsimsellik incelenmiş olup, zaman serisi grafiği, ACF grafiği ve periyodun belirlenmesi gibi analizler yapılmıştır. Verinin altında yatan uzun vadeli eğilimleri ve belirli dönemlerdeki dalgalanmaları ortaya çıkarmak amacıyla toplamsal ve çarpımsal ayrıştırma yöntemleri kullanılmıştır. Regresyon analizinde toplamsal ve çarpımsal regresyon modelleri ile gelecekteki değerleri tahmin etmeyi amaçlar. Üstel Düzleştirme yöntemi ile Toplamsal ve Çarpımsal Winters yöntemleri ile mevsimsel bileşeni de içeren düzleştirme yöntemi uygulanmıştır. Ayrıca, Box-Jenkins modeli ile çeşitli denemeler yapılmış, modellerin anlamlılığı test edilmiş ve tahmin serileriyle modelin güvenilirliği analiz edilmiştir. Sonuç olarak, İstanbul'un temiz su tüketimini etkileyen trend ve mevsimsel bileşenler başarılı bir şekilde ayrıştırılmış, çeşitli yöntemlerin performans metrikleri değerlendirilerek bu veri setinde en iyi sonucu veren yöntemlerin Çarpımsal Winters ve ARIMA olduğuna karar verilmiştir.

GİRİŞ

Su, insanlık tarihi boyunca hayatın sürdürülebilmesi ve ekonomik kalkınmanın sağlanabilmesi için temel bir kaynaktır. Büyük şehirlerde temiz suyun düzenli ve yeterli bir şekilde temin edilmesi son yılların ciddi sorunları arasında yer almaktadır. Halk sağlığı ve yaşam kalitesi açısından büyük bir öneme sahiptir.

İstanbul, yoğun ve hızlı artan nüfusu ile artan su tüketimine sahiptir. Su kaynaklarının yönetimi konusunda kritik zorluklarla karşılaşmaktadır. Bu nedenle, su tüketiminin doğru bir şekilde tahmin edilmesi, gelecekteki ihtiyacın öngörülmesi, su yönetimi stratejilerinin etkin bir şekilde planlanması açısından büyük önem taşır.



Zaman serisi grafiğinde 2009-2019 yılı arasında pozitif yönlü bir artış ve belirgin dalgalanmalar görülmektedir. Dolayısıyla, temiz su miktarına ilişkin serinin artan bir trende ve mevsimselliğe sahip olduğu söylenebilir. Artan trendle İstanbul'un su ihtiyacının arttığını, İSKİ'nin temiz su sağlamak için gelecek yıllarda önlem alması gerektiğinin önemli bir göstergesi olduğu söylenebilir.

YÖNTEMLER

AYRIŞTIRMA YÖNTEMİ: Zaman serisi verilerini bileşenlerine ayırarak analiz etmeyi amaçlar. Genellikle zaman serisinin trend, mevsimsellik ve rastgele bileşenleri ayırt etmede kullanılır.

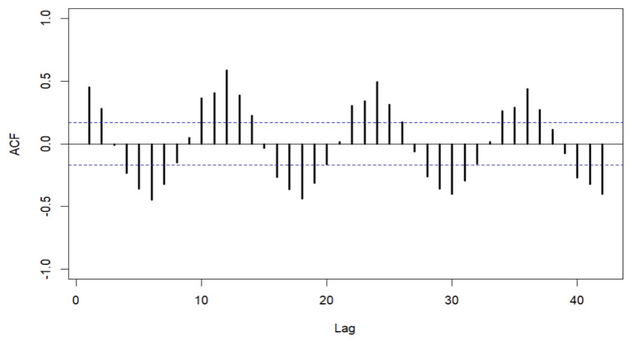
REGRESYON ANALİZİ: Toplamsal regresyon modeli, bağımlı değişkenin bağımsız değişkenlerin doğrusal kombinasyonu olarak ifade edildiği bir regresyon modelidir. Çarpımsal regresyon modeli, bağımlı değişkenin bağımsız değişkenlerin çarpımı şeklinde ifade edildiği bir modeldir.

ÜSTEL DÜZLEŞTİRME YÖNTEMİ: Trend ve mevsimselliği dikkate alan bir düzleştirme yöntemidir. Verilerin geçmiş değerlerine ağırlıklı ortalamalar vererek gelecekteki değerlerini tahmin eder. Bu yöntemde, daha yakın geçmişteki verilere daha fazla ağırlık verilirken, daha uzak geçmişteki verilere daha az ağırlık verilir.

ARIMA: ARIMA modeli, zaman serisi verisini durağan hale getirdikten sonra otoregresif (AR) ve hareketli ortalama (MA) bileşenlerini kullanarak tahmin yapar.

VERİ SETİ VE BULGULAR

Veri trende sahip olduğundan, periyodun belirlenebilmesi için öncelikle periyot serisi oluşturulur. Bu serinin otokorelasyon değerleri ve ACF grafiği incelendiğinde, ilk pozitif en büyük değer ile ikinci pozitif en büyük değer arasındaki farkın 12 olduğu görülür. Bu durum, periyodun 12 olduğuna işaret eder.



Ayrıştırma Yöntemi

ACF grafiğinde veride belirgin bir mevsimsellik olduğu gözlemlenir. Bu mevsimselliğin giderilmesi için 1. dereceden mevsimsel fark işlemi uygulanmış ve veri durağanlaştırılmıştır. Seride hem trend hem de mevsimsellik olması sebebiyle ayrıştırma yöntemleriyle analize başlanır.

Toplamsal ayrıştırma yöntemi ile merkezi hareketli ortalama ile mevsimsel bileşen bulunur. Endeks değerleri oluşturularak saflaştırılmıştır. Regresyon analizi sonucunda modelin anlamlı olduğu %95 güvenle belirlenmiş ve trend bileşeni hatadan arındırılmıştır. Hatadan arındırılan mevsimsel ve trend bileşenleri birleştirilerek tahmin serisi elde edilmiştir. ACF grafiği ve Box-Ljung testi sonuçlarına göre hataların arasında ilişki olmadığı tespit edilmiştir, bu nedenle toplamsal ayrıştırma yöntemi uygun bulunmuştur. Aynı süreçler çarpımsal ayrıştırma yöntemi için de ilerletilmiştir. Çarpımsal ayrıştırma yöntemi de uygun bulunmuştur.

	Toplamsal Ayrıştırma Yöntemi	Çarpımsal Ayrıştırma Yöntemi
p-value	2.20E-16	2.20E-16
Box-Ljung Testi	0.489	0.364
Hata Kareler Ortalaması	3.804	3.791

Kullanılacak yönteme karar verilebilmesi için hata kareler ortalamasına bakılır. Çarpımsal ayrıştırma yönteminin hata kareler ortalamasının daha küçük olması sebebiyle analizler yapılırken çarpımsal ayrıştırma yönteminin kullanılması tercih edilir.

Regresyon Analizi

Analizi yapılan zaman serisinde hem mevsimsellik hem de trend olduğundan, toplamsal model ve çarpımsal model kullanılması uygundur. Her iki modelde de zaman serisinin ikinci harmoniğinin regresyon analizinin çıktısına bakıldığında sin2 ve cos2 değişkenlerine ait p-value değerlerinin 0,05'ten büyük olduğu ve bu sebeple değişkenlerin anlamsız olduğu söylenebilir. Bu durumda s2 ve c2 değişkenleri modelden çıkarılmalıdır. Her iki modelde de Durbin Watson Test İstatistiği 1,5 ile 2,5 değerleri arasında olmadığı için hatalar arası ilişki olabileceği ipucunu verir. Box-Ljung testi yapıldığında p-value < a = 0,05 değerinden küçük olduğu için Her iki modelde de hatalar arasında ilişki vardır. Bu sebeple regresyon modellerine gecikmeli değerleri eklenerek var olan otokorelasyon problem ortadan kaldırılır. Model performans metriklerine bakıldığında çarpımsal regresyon modelinin toplamsal regresyon modeline göre daha iyi sonuç verdiği söylenebilir.

Toplamsal Regresyon Modeli					
1.Harmonik		2. Harmonik			
sin1	cos1	sin1	cos1	sin2	cos2
2.23E-13	2.49E-15	1.52E-13	1.63E-15	0.135	0.119

Toplamsal Regresyon Modeli	
Durbin Watson Test	1.69E-05
Box-Ljung Testi	0.04147

MAE	4.690
RMSE	5.631
MAPE	6.318
R2	0.841

Çarpımsal Regresyon Modeli					
1.Harmonik		2. Harmonik			
sin1	cos1	sin1	cos1	sin2	cos2
7.98E-10	3.39E-14	8.62E-10	5.47E-14	0.428	0.500

Çarpımsal Regresyon Modeli	
Durbin Watson Test	1.86E-07
Box-Ljung Testi	4.73E-05

MAE	2.669
RMSE	4.040
MAPE	0.037
R2	0.831

Üstel Düzleştirme Yöntemi

Üstel düzleştirme yöntemi, zaman serisi verilerinin tahmin edilmesi ve düzeltilmesi için güçlü bir tekniktir. Bu yöntem, geçmiş verilerin üstel ağırlıklarla (önceki verilere giderek azalan ağırlıklar) ortalamasını alarak gelecekteki değerleri tahmin eder. Veri setinde hem trend hem de mevsimsellik varsa Winters Düzleştirme

	Toplamsal Winters Yöntemi	Çarpımsal Winters Yöntemi
l	59.077	57.837
b	0.355	0.526
s	12	12
Box-Ljung Testi:	1	1
RMSE	3.367	3.322

Toplamsal ve çarpımsal Winters yöntemine ait çıktıya bakıldığında ortalama düzeyin başlangıç değerinin "l", eğimin başlangıç değerinin "b" ve mevsimsel terimin başlangıç değerlerinin periyodunun "s" 12 tane değeri gösterdiği görülmektedir. Her iki yöntemde de Box Ljung test istatistiği sonucunda p-value= 1 > a=0,05 olduğundan modelin anlamlı olduğu %95 güvenle söylenebilir. Hatalar arası ilişki yoktur. Hem toplamsal hem çarpımsal modelde hatalar arası ilişki olmadığı için RMSE değerlerine bakılır. Toplamsal model için RMSE= 3.366579, çarpımsal model için RMSE= 3.321892 olarak bulunmuştur. Bu durumda RMSE değeri daha küçük olduğunda öngörülerin çarpımsal model ile yapılması daha uygundur.

Box - Jenkins Modeli

Box – Jenkins modelinin uygulanabilmesi için verinin durağan hale getirilmesi gerekmektedir. 1. dereceden mevsimsel fark alındıktan sonra PACF grafiği daha hızlı azalırken ACF grafiğinin daha yavaş azaldığı söylenebilir. Modelin otokorelasyon modeli olduğu görülmektedir. ACF ve PACF grafiklerine bakıldığında en uygun model ARIMA(0,1,0)(1,1,0)12'dir denebilir. Ancak yine de birkaç modelin daha test edilmesi daha güvenilir sonuçlar elde etmemizi sağlar. ARIMA(0,1,0)(1,1,0)12, ARIMA(0,1,0)(2,1,0)12, ARIMA(0,1,0)(0,1,1)12, ARIMA(0,1,0)(0,1,2)12, ARIMA(0,1,0)(1,1,1)12 modelleri denenmiştir.

Model	p-value	RMSE	BIC
ARIMA(0,1,0)(1,1,0)12	0.002	3.780	678.310
ARIMA(0,1,0)(2,1,0)12	0.181	-	-
ARIMA(0,1,0)(0,1,1)12	0.000	3.715	676.020
ARIMA(0,1,0)(0,1,2)12	0.847	-	-
ARIMA(0,1,0)(1,1,1)12	0.854	-	-

P-value değerleri kontrol edildiğinde ARIMA(0,1,0)(1,1,0)12 ve ARIMA(0,1,0)(0,1,1)12 modellerinin anlamlı olduğu görülmektedir. Sonuç olarak RMSE ve BIC değerleri daha küçük olduğundan ARIMA(0,1,0)(0,1,1)12 modelinin öngörülerin yapılmasında kullanılması daha uygundur.

SONUÇ

Çarpımsal regresyon, çarpımsal ayrıştırma, çarpımsal Winters, ARIMA(0,1,0)(1,1,0)12 ve ARIMA(0,1,0)(0,1,1)12 ve modelleri performans metriklerine göre kıyaslanırsa, bu veri üzerinde en iyi tahmin modelini oluşturan yöntem Çarpımsal Winters Yöntemi olarak belirlenir. Daha sonra ise en iyi tahmin modeli ARIMA(0,1,0)(0,1,1)12 olarak belirlenir.

	Belirlenen Modellerin Performans Metrikleri				
Metrik	Çarpımsal Regresyon	Çarpımsal Ayrıştırma	Çarpımsal Winters Yöntemi	ARIMA(0,1,0)(1,1,0)12	ARIMA(0,1,0)(0,1,1)12
MAE	2.669	1.855187	1.385	1.74	1.665
RMSE	4.040	3.791443	3.321	3.78	3.715
MAPE	0.037	2.675925	1.966	2.423	2.328

Gelecek çalışmalarda, özellikle Çarpımsal Winters ve ARIMA modellerinin performansını iyileştirmek için daha fazla model geliştirilebilir. Ayrıca, farklı mevsimsel desenlerin ve trendlerin daha derinlemesine incelenerek tahminlerin daha hassas hale getirilmesi sağlanabilir.

KAYNAKÇA

[1] https://medium.com/@nrmnbabalik/zaman-serileri-2-b%C3%B6l%C3%BCm-smoothingy%C3%B6ntemleri-89078ad7e71b

[2] Temuçin, T., & Temiz, İ. (2016). Türkiye Dış Ticaret İhracat Hacminin Projeksiyonu: Holt-Winters ve Box-Jenkins Modellerinin Bir Kıyaslaması. Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 21(3), 937-960.

[3] https://www.kaggle.com/code/prashant111/arima-model-for-time-series-forecasting

İSTATİSTİK BÖLÜMÜ

UYUM ANALİZİ VE DİYABET VERİSİ ÜZERİNE BİR UYGULAMA

Gizem IŞIK 21023601

Danışman: Dr. Öğr. Üyesi Doğan YILDIZ

ÖZET

Diyabet, günümüzün önde gelen sağlık sorunlarından biridir. Ulusal Diyabet Federasyonu'na göre 2021 yılında 537 milyon olan yetişkin diyabetli sayısının 2030'da 643 milyona, 2045'te ise 783 milyona ulaşması beklenmektedir. Federasyon ayrıca 2021 yılında her 5 saniyede 1 olmak üzere 6,7 milyon ölümün sorumlusunun diyabet olduğu bildirmiştir. Yüksek diyabet riski olan bireylerde yaşam tarzı değişiklikleri ile diyabet önemli ölçüde önlenebilir ya da başlaması geciktirilebilir. Bu çalışmada, diyabet ile ilişkili olan bazı faktörlerin Basit Uyum Analizi ile incelenmesi ve bu faktörler arasında varolan ilişkilerin gerçek bir veri seti üzerinde incelenmesi amaçlanmıştır. Çalışmada, UC Irvine Machine Learning Repository veri tabanında bulunan "National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset" veri kümesi kullanılmıştır. Verilerin analizi ve grafiklerin çiziminde SPSS 27 programı kullanılmıştır.

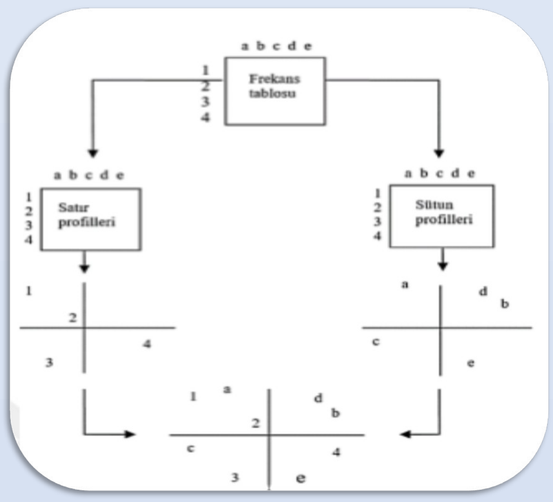
DİYABET



Diyabet, insülin eksikliği, insülin direnci veya her ikisinin bir kombinasyonu sonucunda kandaki glukoz seviyesinin yükselmesiyle ortaya çıkan, kronik bir hastalıktır. Diyabet; tip 1, tip 2, spesifik nedenlere bağlı diyabet ve gebelik diyabeti olmak üzere başlıca dört gruba ayrılır. Tip 2 diyabet, obezite ve fiziksel inaktiviteye bağlı olarak genellikle daha sık görülmekte ve tüm diyabet olgularının %90'dan fazlasını oluşturmaktadır. Temelinde genetik olarak yatkın kişilerde yaşam tarzı ile tetiklenen ve giderek artan insülin direnci ve zamanla azalan insülin salınımı söz konusudur. Türkiye, dünyada diyabetin hızla yayıldığı ülkelerin başında gelmektedir.

UYUM ANALİZİ

Uyum analizi, çok değişkenli veri setlerinde değişkenler arasındaki ilişkileri görselleştirmek ve analiz etmek için kullanılan istatistiksel bir yöntemdir. Uyum analizi ile iki yada daha fazla kategorik değişken arasındaki ilişki çözümlenirken;değişkenlerin kategorileri indirgenmiş boyutta bir uzayda noktalar olarak tanımlanır. Birbirine benzer kategoriler uzayda birbirine daha yakın noktalar olarak konumlanır.



Uyum Analizi üç aşamadan oluşmaktadır:

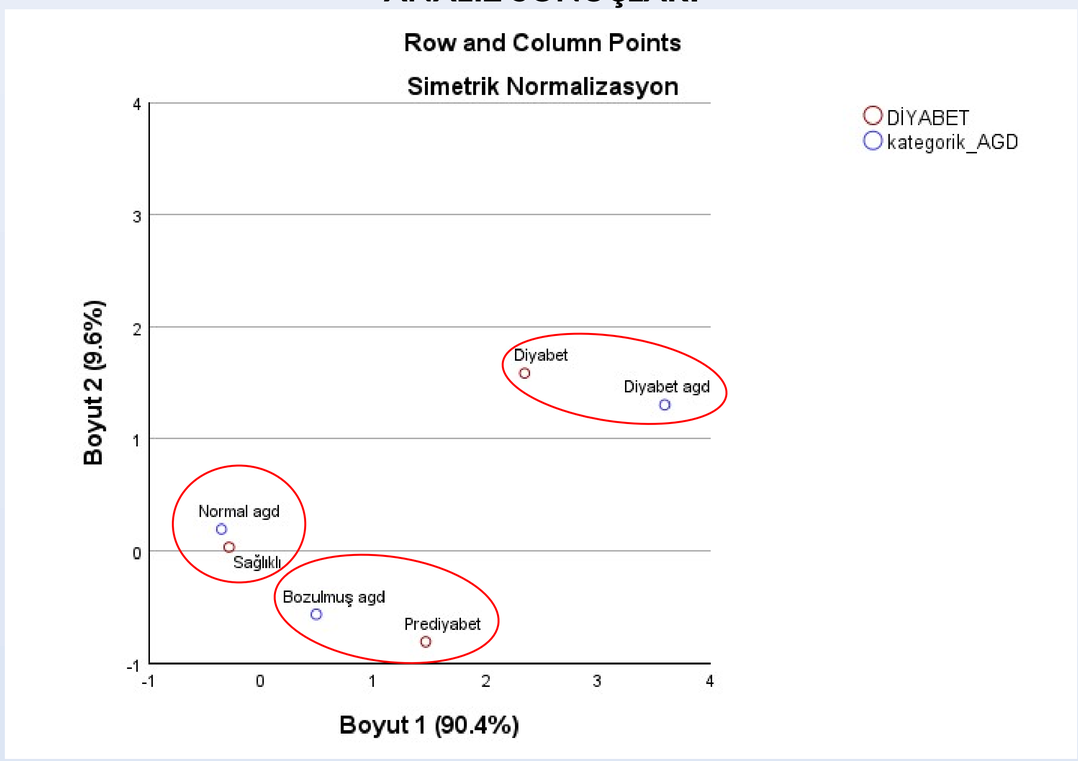
- Satır ve sütun profilleri hesaplanır.
- Satır ve sütun profilleri iki-boyutlu uzayda ayrı ayrı resmedilir.
- Satır ve sütun profilleri iki boyutlu ortak bir grafik (harita) üzerinde gösterilir.

UYGULAMA

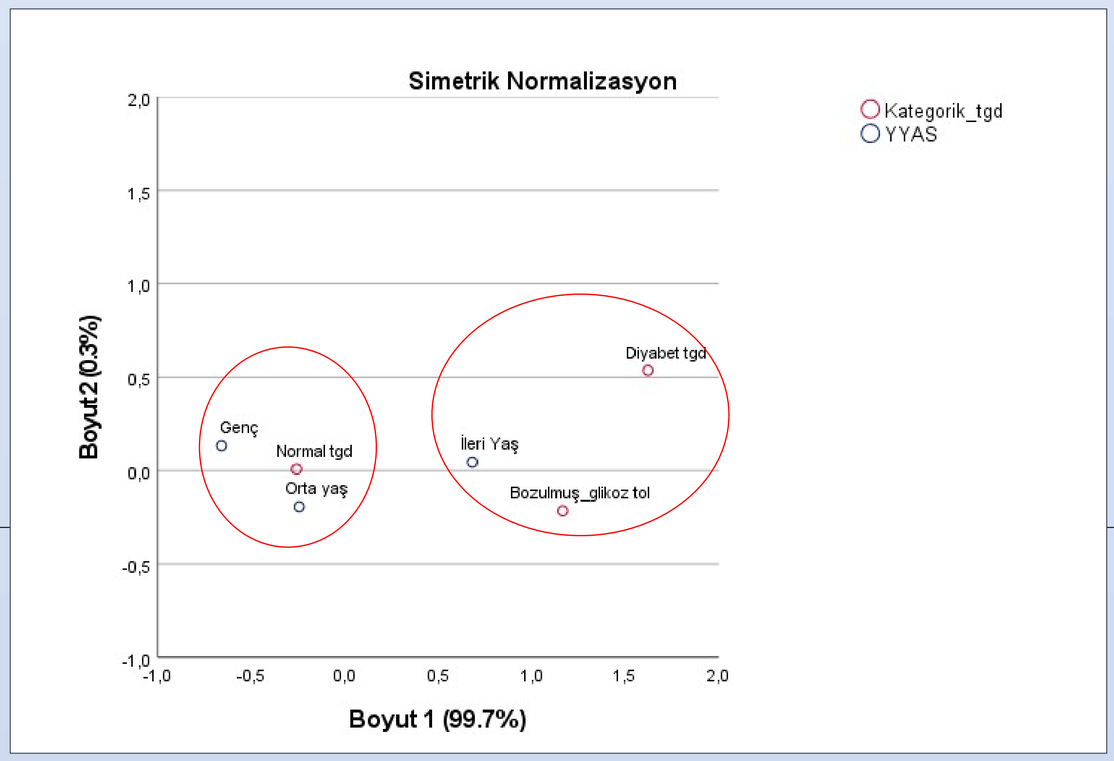
Değişken Adı	Değişkenin Tanımı
Yaş	Katılımcının yaşı
Cinsiyet	Katılımcının cinsiyeti
Egzersiz	Katılımcı sıradan bir hafta içinde orta veya kuvvetli yoğunlukta egzersiz yapma(ma) durumu
Beden Kitle İndeksi	Kilogram cinsinden ağırlığın, metre cinsinden boyun karesine bölünmesi ve ardından bir ondalık basamağa yuvarlanmasıyla hesaplanan değer.
Açlık Glikoz Değeri	En az 9 saatlik açlık sonrası ölçülen açlık glikoz değeri
Tokluk Glikoz Değeri	75 gram glikoz aldıktan 2 saat sonra ölçülen tokluk glikoz değeri
Diyabet	Katılımcıların "Hamilelik dışında, bir doktor veya sağlık profesyoneli tarafından size diyabetiniz olduğu söylendi mi?" sorusuna verilen yanıtlardan oluşmaktadır. (1=Sağlıklı, 2= Prediyabet, 3= Diyabet)
HOMA-IR	İnsülin direnci değeri Açlık glikozu (mg/dL)*İnsülin (µU/mL)/405 değeri kullanılarak hesaplanmıştır.
İnsülin (µU/mL)	İnsülin değeri

Ulusal Sağlık ve Beslenme İnceleme Araştırması (NHANES), ve ABD nüfusunun sağlık ve beslenme durumunu değerlendirmek için uygulanan bir programdır. Orijinal veri seti 2277 gözlemden ve 8 değişkenden oluşmaktadır. Veri seti Diyabet değişkenin kategorilerine göre 3'e ayrılmıştır. Ardından her kategorinin kendi içindeki aşırı gözlemleri ayıklanmıştır. Diyabet değişkeninin kategorileri arasındaki sınıf dengesizliğini azaltmak amacıyla "Prediyabet" ve "Diyabet" kategorilerine göre daha fazla gözlem içeren "Sağlıklı" kategorisinin %25'i herhangi bir koşul belirtmeksizin rassal olarak alınmıştır. Bunların yanı sıra veri setinde bulunan açlık glikoz değeri ve insülin değişkenleri kullanılarak HOMA-IR değişkeni oluşturulmuştur. Sonuçta 509 gözlem ve 9 değişkenden oluşan veri seti elde edilmiştir. Uyum analizi yapabilmek için değişkenin en az 3 kategoriden oluşması gerektiğinden veri setinde bulunan sürekli ya da kesikli sayısal değişkenler kategorik hale getirilmiştir.

ANALİZ SONUÇLARI



Diyabet-Açlık Glikoz Değeri değişkenleri ile yapılan uyum analizinde ,"Diyabet" ve "Diyabet agd" kategorilerinin yakın olduğu, benzer şekilde "Normal agd ile "Sağlıklı" kategorilerinin, "Bozulmuş agd" ile de "Prediyabet" kategorilerinin yakın oldukları görülmektedir. Diyabet değişkenin seviyelerinin açlık glikoz seviyelerindeki artışa paralel olarak artış gösterdiği söylenebilir.



Tokluk Glikoz Değeri değişkeni ile Yaş değişkeni arasında yapılan Basit Uyum Analizi sonucunda elde edilen iki boyutlu grafikte "Normal tgd" kategorisi ile "Genç" ve "Orta yaş" kategorilerinin profillerinin benzer olduğu görülmektedir. "Diyabet tgd" ve "Bozulmuş glikoz toleransı" kategorileri ile "İleri yaş" kategorisinin yakın olarak konumlandıkları görülmektedir.

KAYNAKÇA

[1] T.C. Sağlık Bakanlığı, Türkiye Halk Sağlığı Kurumu Ankara. (2014). Yayın No:816.
[2] Alpar, R. (2021). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, 6 .Baskı. Ankara: Detay Yayıncılık.
[3] International Diabetes Federation. (2021). IDF Diabetes Atlas (10th edition). Erişim tarihi: 10.04.2024, Erişim Adresi: IDF Atlas 10th Edition 2021.pdf

İSTATİSTİK BÖLÜMÜ

Sanata İlgi ile Mutluluk arasındaki ilişkinin incelenmesi.

Göktuğ Yavuzer - 19023026

Danışman: Prof. Dr. Ali Hakan BÜYÜKLÜ

ÖZET

Sanat insanların duygu ve düşüncelerini aktarabilmeleri için en etkili yollardan biridir. Sanat, hayatın merkezinde yer alan ancak kelimelere dökmesi zor olanları fark etmemizi ve kendimizi tanımamızı da sağlar. Aynı zamanda da farklı kültürlerle ilgili fikirlerimizi genişletebilmek ve dünyanın sesini duymak için etkili bir yoldur. Ayrıca sanat toplumun bir aynası olarak kabul edilebilir. Bu araştırmada da sanatın toplumda gördüğü ilgi ve bu ilginin mutluluk ile olan ilişkisi araştırılmıştır.

SANAT NEDİR

Türk Dil Kurumu'ndaki anlamıyla sanat Duygu ve düşünceleri göze ve gönle hitap edecek şekilde ifade etme konusundaki yaratıcılık olarak tanımlanmıştır. Sanat insanlık tarihi kadar eskidir ve insanlarla birlikte gelişmektedir.

SANATIN TARİHİ VE GELİŞİMİ

Sanatın ilk belirtileri milattan önce 100.000 yılına kadar uzanmaktadır. Zaman içerisinde gelişen sanat giderek çeşitlenmiş ve gelişmiştir. Mağara duvarlarında yapılmaya başlanan resimler tuvallere geçmiştir. Bu gelişim sanatın tanımı ve kapsamı hakkında tartışmalara yol açmış birçok filozof sanatın anlamını sorgulamıştır, bugün sanatın 10'dan fazla çeşidi olmakla beraber hâlâ tartışmalar devam etmektedir.



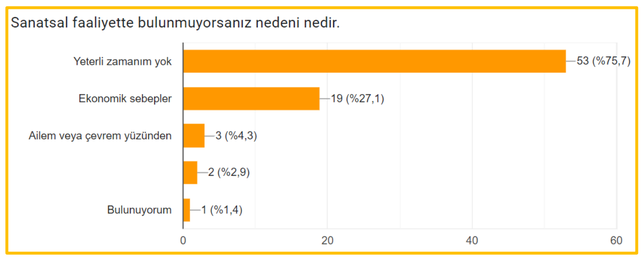
KATILIMCILARIN ÖZELLİKLERİ

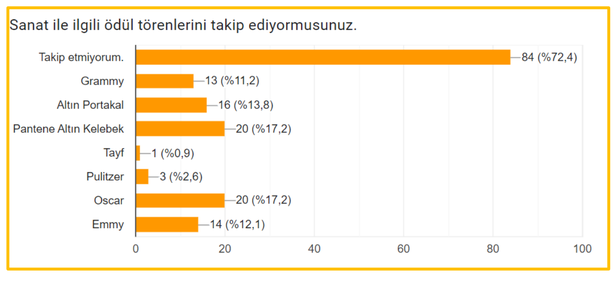
Soru	Seçenekler	N	%
Cinsiyet	Erkek	60	51,7%
	Kadın	56	48,3%
Yaş	18'den küçük	54	46,6%
	18-25	4	3,4%
	26-35	35	30,2%
	36-45	10	8,6%
	46-55	13	11,2%
Eğitim	İlkokul	2	1,7%
	Ortaokul	7	6,0%
	Lise	39	33,6%
	Önlisans	36	31,0%
	Lisans	12	10,3%
	Yüksek Lisans	17	14,7%
	Doktora	3	2,6%

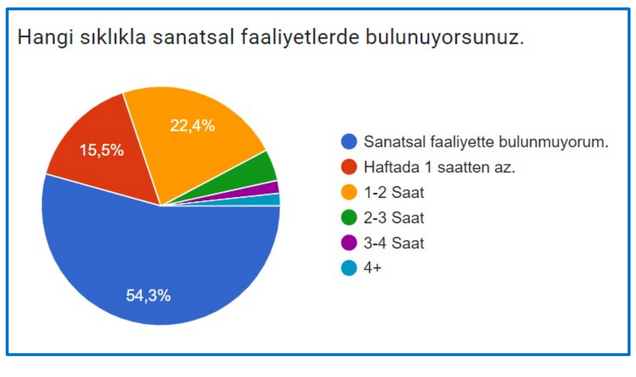
Gelir	Düzenli gelirim yok	44	37,9%
	10 Bin - 18 Bin	42	36,2%
	18 Bin - 50 Bin	28	24,1%
	50 Bin - 100 Bin	2	1,7%
İş Grubu	Banka - Finans - Sigorta	9	7,8%
	Çalışmıyorum	49	42,2%
	Genel İşler	31	26,7%
	Gıda Sanayi	5	4,3%
	Güvenlik	4	3,4%
	İnşaat	2	1,7%
	Sağlık	4	3,4%
	Taşımacılık	4	3,4%
	Ticaret	8	6,9%
Sanatsal Faaliyette Bulunma Sıklığı	Sanatsal faaliyette bulunmuyorum.	63	54,3%
	Haftada 1 saatten az.	26	22,4%
	1-2 Saat	5	4,3%
	2-3 Saat	2	1,7%
	3-4 Saat	2	1,7%
	4+	18	15,5%

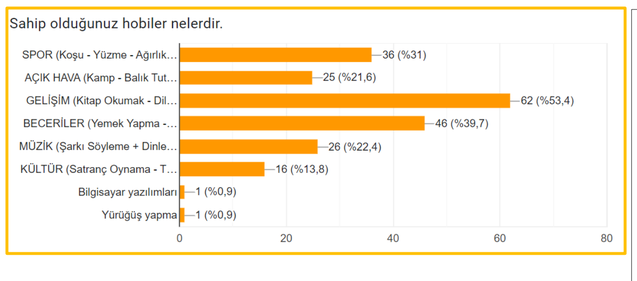
VERİ

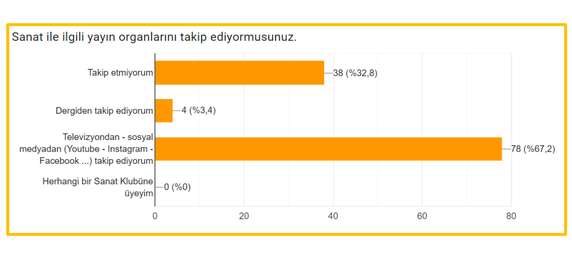
Araştırmada veriler anket yapma yöntemiyle toplanmıştır. Ankete 116 kişi katılmıştır. Ankete katılımda herhangi bir kısıt veya şart bulunmamaktadır. Katılımcıların dağılımları aşağıda verilerle ve grafiklerle gösterilmiştir.

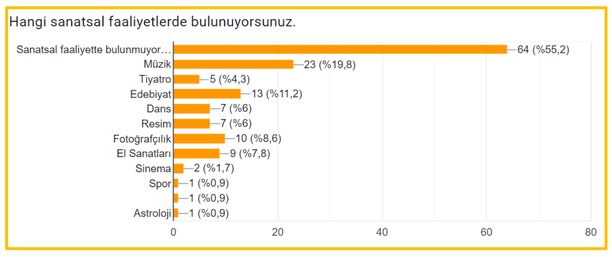












SANATA İLGİ ÖLÇEĞİ

Taşkesen (2014) tarafından geliştirilen bireylerin sanata olan ilgilerini düşünce, duygu ve davranış boyutunda ölçmeyi amaçlayan bir ölçek niteliğinde olup 20 soru ve 2 boyuttan oluşmaktadır. 3, 5, 9, 13 ve 16 numaralı sorular ters puanlanmaktadır. 5li ölçeğe sahiptir, cevapları (1) Katılmıyorum ile (5) Katılıyorum arasındadır.

MUTLULUK ÖLÇEĞİ

Demirci, İ. & Ekşi. H. (2018) tarafından geliştirilen Mutluluk Ölçeği kişilerin mutluluk düzeylerini ölçmek amacıyla geliştirilmiştir. Tek faktörlü olmakla beraber 6 sorudan oluşmaktadır ve 5'li likert ölçeğine sahiptir. Cevaplar (1) Bana Uygun Değil ile (5) Bana Tamamen Uygun arasında değer almaktadır.

SONUÇLAR

Araştırma sonucunda kişilerin sanata ilgileri ve mutlulukları ölçülerek demografik özellikleri ile analizler gerçekleştirilmiştir ve aşağıdaki sonuçlara ulaşılmıştır;

- **Katılımcıların yarısının sanatsal faaliyette bulunmadığı ve bunun sebebinin büyük çoğunluğunun (%75) yeterli zaman bulamadığı ve ekonomik sebepler (%27) olduğu,**
- **Katılımcıların üçte ikisinin sanat ile ilgili yayın organlarını takip ettiğini, %72'sinin ödül törenlerini takip etmediğini,**
- **Katılımcıların yarısının Gelişim (Kitap Okuma, Dil Öğrenme, Yazma), üçte ikisinin Beceriler (Yemek Yapma - El Sanatları - Fotoğrafçılık) ve Spor (Koşu - Yüzme - Ağırlık Kaldırma - Bisiklet Sürme) , beşte birinin Müzik (Şarkı Söyleme + Dinleme - Enstrüman Çalma - Beste Yapma) ve Açık Hava (Kamp - Balık Tutma - Bahçe İşleri) , onda birinin ise Kültür (Satranç Oynama - Tarihi Yerler ve Müzeleri Gezme) hobilerine sahip olduğu,**
- **Sanata ilgi duyan kadın ve erkeklerin davranışlarının farklılık gösterdiğini,**
- **Sanatsal faaliyette bulunma sıklığının Davranış bakımından sıklığa göre farklılık gösterdiği,**
- **Sanata İlginin ve Mutluluğun yaş, gelir, iş türü ve eğitim seviyesine göre bir farklılık göstermediğini,**
- **Sanata ilgi arttıkça mutluluğun da arttığını,**
- **Kadınların erkeklerden daha çok sanat ile ilgili yayın organlarını takip ettiğini,**
- **Kadınların erkeklerden daha çok hobiye sahip oldukları çıkarımları yapılabilir.**

KAYNAKÇA

[1] Üstün A. (2023) Üniversite öğrencilerinin boş zaman ve sanata ilgi tutumları arasındaki ilişkinin incelenmesi. Yüksek Lisans Tezi. Selçuk Üniversitesi / Sağlık Bilimleri Enstitüsü / Rekreasyon Ana Bilim Dalı

[2] Taşkesen S. (2014) Lise öğrencilerinin kişilik tipleri ile sanata olan ilgileri arasındaki ilişkinin incelenmesi. Yüksek Lisans Tezi. Gazi Üniversitesi / Eğitim Bilimleri Enstitüsü / Güzel Sanatlar Eğitimi Ana Bilim Dalı / Mesleki Resim Eğitimi Bilim Dalı

[3] https://www.sanatsanat.com/sanat-felsefesi-nedir/

İSTATİSTİK BÖLÜMÜ

Küresel Barış ve Dünya Mutluluk Endeksleri İlişkisi Temelinde

Korelasyon ve Kümeleme Analizleri

Gözde İÇER 16023069

Danışman: Prof. Dr. Ersoy ÖZ

Dünya Mutluluk Endeksi ve Küresel Barış Endeksi'ni temel alan bu çalışmada, barışın mutluluk ile olan ilişkisi bazı istatistiksel yöntemlerle incelenmiş ve analiz edilmiştir. Endekslerdeki değişkenlerin kendi aralarındaki ilişkiye bakabilmek adına korelasyon analizi uygulanmıştır. Barış skorunun neredeyse her değişkenle ilişkisi olduğu görülmüştür. Endeksleri, ülkeler bazında karşılaştırabilmek için öncelikle hiyerarşik kümeleme ile küme sayısı belirlenmiş, daha sonra k-means kümeleme kullanılarak ülkeler benzerliklerine göre kümelere ayrılmıştır.

Mutluluk ve Barış İlişkisi

Başlangıç olarak, içimizdeki gücü artırmak, şevkat gibi pozitif duygulara sahip olmak bizi hayatın zorluklarına karşı donatır. Zihnimiz, mutluyken daha sakin bir hal alır. Sakinliğimiz gündelik aktivitelerimize yansır ve bize zor gelebilecek şeylerle başa çıkmamızı sağlar. Aynı zamanda pozitif duygular başkalarıyla iletişimizi güçlendirir. Kavgalardan, kötü duygulardan, iletişimsizlikten kaçınmış oluruz. Çevreyle ve insanlarla uyum içinde yaşarız. Bu da barışı beraberinde getirmektedir.

Dünya Mutluluk Raporu

Dünya Mutluluk Raporu verilerini "Gallup World Poll" isimli bir şirket araştırmaktadır. Her yıl 150'den fazla ülkede 15 yaşından büyük 1000 kişiye anket yapılmaktadır. Raporda mutluluğu ölçmek için bazı değişkenler kullanılmıştır. Bunlardan biri "Cantril Ladder" denilen "Cantril Merdiveni"dir. Cantril merdiveni ölçeğinden raporda şöyle bahsedilmiştir; "Lütfen basamakları altta 0'dan, üstte 10'a kadar numaralandırılmış bir merdiven hayal edin. Merdivenin tepesi sizin için mümkün olan en iyi hayatı, alt kısmı ise sizin için mümkün olan en kötü hayatı temsil eder. Şu anda kişisel olarak merdivenin hangi basamağında olduğunuzu hissettiğinizi söylersiniz?" (WHR 2017, Helliwell ve diğerleri, s.9).

Diğer açıklayıcı değişkenler ise;

- Kişi başına Gayri Safi Yurt İçi Hasıla (GDP per capita, GDP),
- Sosyal Destek (social support, SS),
- Sağlıklı Yaşam Beklentisi (healthy life expectancy, HLE),
- Yaşam Seçimlerini Yapma Özgürlüğü (freedom to make life choices, FRE),
- Cömertlik (generosity, GEN),
- Yolsuzluk algısı (perceived (absence of) corruption, PER).
- Negatif etki (negative affect)
- Pozitif etki (positive affect)

Küresel Barış Endeksi

Dünyadaki şiddet eğilimine dikkat çekmek ve küresel barış gelişmelerini raporlamak amacıyla kurulan Ekonomi ve Barış enstitüsü "Küresel Barış Endeksi(Global Peace Index, GPI)"ni yayımlamaktadır. Bu endeks 25 farklı değişken ve 163 ülke ile dünyadaki barış oranını araştırmaktadır. Bu değişkenlerle elde edilen bulguları GPI skoru adı altında toplamışlardır.

- GPI skoruna göre, daha barışçıl olduğu düşünülen ülkelerin endeks puanları daha düşüktür.

Analizlerde Küresel Barış Endeksi'nden GPI skorun Dünya Mutluluk Endeksi'ndeki değişkenler ile ilişkisi araştırılmaktadır. Veri setinde GPI skoru, diğer değişkenler ile doğru orantıyı sağlayacak biçimde dönüşüm sağlanmıştır.

Değişken Adı	Min	Max	Ort ± SS	Eksik Gözlem
Life Ladder	1,44	7,69	5,62 ± 1,13	-
GPI Score	-3,44	-1,12	-2,02 ± 0,47	-
GDP per capita	7,07	11,67	9,51 ± 1,13	8
Social Support	0,36	0,97	0,79 ± 0,12	
Healthy Life Expectancy	52,2	74,6	65,13 ± 5,52	3
Freedom to make life choices	0,22	0,96	0,79 ± 0,11	1
Generosity	-0,26	0,59	0,03 ± 0,16	8
Perceptions of Corruption	0,15	0,94	0,72 ± 0,17	7
Positive Affect	0,26	0,84	0,65 ± 0,1	-
Negative Affect	0,11	0,51	0,29 ± 0,08	-

- Ardından eksik gözlemler ortalama değer ile doldurulmuştur

Normallik Kontrolü

Veri sayısı 50'den büyük olduğu için Kolmogorov-Smirnov sonuçlarına bakılmıştır. Sadece generosity değişkeni normal dağılıma sahip çıkmıştır. (p-value>0,05)

Korelasyon Analizi

Korelasyon analizi, değişkenler arasında bir ilişkinin var olup olmadığını tespit etmemizi sağlar. Eğer bir ilişki söz konusu ise bu ilişkinin yönü ve büyüklüğü hakkında bilgi verir. Korelasyon katsayısı $-1 \leq r \leq 1$ arasındadır. Korelasyon katsayısının 0'a yaklaşması değişkenler arasındaki ilişkinin zayıfladığını ifade eder. – yönlü ilişki negatif bir ilişki var demektir. + yönlü ilişki pozitif bir ilişki olduğunu belirtir.

Korelasyon Analizi Sonuçları (Spearman Korelasyon)

Korelasyon analizine göre, Küresel Barış Endeksi skoru ile Mutluluk Endeksi değişkenleri arasında ilişki mevcuttur. Cömertlik ve yolsuzluk algısı değişkenlerinin barış skoruyla ilişkisi yoktur(p-value>0,05). Pozitif etki değişkeniyle olan ilişkisi de zayıftır. Barış skorunun en ilişkili olduğu değişken sağlıklı yaşam beklentisi olup, yüksek dereceye çok yakındır(r=0,595). Kalan diğer 5 değişkenle ilişkisi orta derecededir. Barış skorunun sadece negatif etki değişkeniyle negatif yönde ilişkisi bulunmaktadır. Barış skoru ile pozitif yönde ilişkisi olan değişkenler ise boyutlarına göre(çoktan aza) şu şekildedir; Sağlıklı yaşam beklentisi, kişi başına Gayri Hasi Yurtiçi Hasıla, sosyal destek, yaşam seçimlerini yapma özgürlüğü, pozitif etki. Barış skoru ile mutluluk endeksi değişkenlerinin korelasyon analizi sonuçlarına göre, barış skoru, diğer değişkenlerin çoğunluğuyla pozitif yönde ilişkiye sahiptir. Analizin tümüne baktığımda dikkatimi çeken diğer bir nokta ise, sağlıklı yaşam beklentisi ve kişi başına GHYH değişkenleri arasında ilişkinin pozitif yönde ve çok yüksek olduğudur. Ayrıca cömertlik değişkenin diğer tüm değişkenlerle ilişkisi neredeyse yoktur.

Kümeleme Analizi

Kümeleme analizi, bir veri setinde çeşitli gruplar olup olmadığını belirlemek ve eğer varsa bu grupları tanımlamak için kullanılır. Çok değişkenli bir istatistiksel yöntemdir. Kümeleme yöntemleri izledikleri yaklaşımlara göre iki temel gruba ayrılırlar. Bunlar, aşamalı (hiyerarşik) kümeleme yöntemleri ve aşamalı olmayan kümeleme yöntemleridir. Analizlerde, hiyerarşik(Ward metodu) ve hiyerarşik olmayan(k-means) kümeleme yöntemleri kullanılmıştır.

Kümeleme Analizi Sonuçları

Öncelikle analizi yapabilmek için bazı varsayımlar kontrol edilmiştir.

- Değişkenlerde uç değerler olmamalı.

Analizde her gözlemin etkisini görebilmek için uç değerler çıkarılmamıştır.

- Çoklu doğrusal bağlantı olmamalı.
- Örneklem evreni temsil edebilmeli.

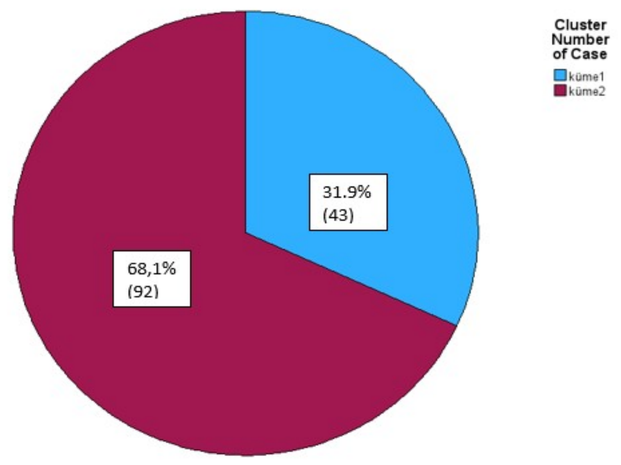
KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,832
Bartlett's Test of Sphericity	Approx. Chi-Square	702,532
	Df	45
	Sig.	<,001

Çoklu doğrusal bağlantı için Barlett testi;
Sig.<0,05 olduğundan H_0 reddedilebilir.
H_0 : Varyans-Kovaryans matrisi bir birim matristir.
H_1 : Varyans-Kovaryans matrisi bir birim matris değildir.
Kaiser-Meyer-Olkin testi analiz edilen değişkenlerin örnek sayısının yeterli olup olmadığını test etmek için kullanılır. Test oranının 0,50'den yüksek olması beklenir ve 0,83 değeri iyi bir örnek sayısına sahip olduğunu göstermektedir.

Hiyerarşik Kümeleme

Bu yöntemleri karşılaştıran literatürdeki çalışmalarda, genellikle Ward yöntemi en iyi sonuç veren yöntem olarak görülmektedir. Ward yöntemi ile bir kümedeki varyansı en aza indirecek gruplar oluşturulmaktadır. Dendrograma bakılıp, 2 kümede toplanabilecekleri görülmüştür.

K-Means Kümeleme



Küme1	Küme2
Afganistan, Benin, Botsvana, Burkina Faso, Kamerun, Çad, Kongo(Brazavil), Kongo(Kinşasa), Etiyopya, Gabon, Gambiya, Gana, Guatemala, Gine, Hindistan, Fildişi Sahili, Kenya, Laos, Liberya, Madagaskar, Malavi, Mali, Moritanya, Moğolistan, Mozambik, Myanmar, Namibya, Nepal, Nijer, Nijerya, Pakistan, Filipinler, Senegal, Sierra Leone, Güney Afrika, Tacikistan, Tanzanya, Togo, Uganda, Yemen, Zambiya, Zimbabve	Arnavutluk, Arjantin, Ermenistan, Avustralya, Avusturya, Azerbaycan, Bahreyn, Bangladeş, Belçika, Bolivya, Bosna Hersek, Brezilya, Bulgaristan, Kanada, Şili, Çin, Kolombiya, Kosta Rika, Hırvatistan, Kıbrıs, Çekya, Danimarka, Dominik Cumhuriyeti, Ekvador, Mısır, El Salvador, Estonya, Finlandiya, Fransa, Gürcistan, Almanya, Yunanistan, Honduras, Macaristan, İzlanda, Endonezya, İran, Irak, İrlanda, İsrail, İtalya, Japonya, Ürdün, Kazakistan, Kosova, Kuveyt, Kırgızistan, Letonya, Lübnan, Libya, Litvanya, Malezya, Mauritius, Meksika, Moldova, Karadağ, Fas, Hollanda, Yeni Zelanda, Nikaragua, Kuzey Makedonya, Norveç, Panama, Paraguay, Peru, Polonya, Portekiz, Romanya, Rusya, Suudi Arabistan, Sırbistan, Singapur, Slovakya, Slovenya, Güney Kore, İspanya, Sri Lanka, Filistin Devleti, İsveç, İsviçre, Tayvan, Tayland, Tunus, Türkiye, Ukrayna, Birleşik Arap Emirlikleri, Özbekistan, Venezuela, Vietnam, Birleşik Krallık, Amerika Birleşik Devletleri, Uruguay

Kümeleme sonuçlarına bakarsak, Küme 2'de yığılma söz konusu olduğunu görebiliriz. Küme1'deki ülkelerin ortak özelliği çoğunluğunun Asya ve Afrika-Karayip-Pasifik ülkeleri olmasıdır. Küresel Barış Skoru ve Mutluluk Endeksi değişkenlerinden gelen verilerle 2 kümeye ayrılan bu tabloda coğrafi konumun etkin olduğunu söyleyebiliriz.

KAYNAKÇA

[1] https://avys.omu.edu.tr/storage/app/public/yukselt/118305/k%C3%BCme%20analizi.pdf
[2] https://worldhappiness.report/ed/2024/
[3] Murtagh ve Legendre, 2014:275

İSTATİSTİK BÖLÜMÜ

İKİ DEĞİŞKENLİ NORMAL DAĞILIM TESTİNDEN SİMÜLASYON OLUŞTURMA

Hakan AKKAYA 20023013

Danışman: Doç. Dr. Atıf Ahmet EVREN

ÖZET

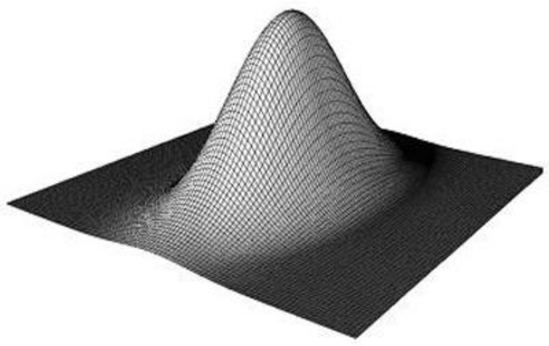
Simülasyonlar, bir sistem veya sürecin davranışını modellemek için oluşturulan yapay ortamlardır. İstatistiksel simülasyonlar özellikle belirli bir olasılık dağılımına dayanan verilerin veya belirsizliklerin etkilerini anlamak için kullanılır. Bu çalışmada iki değişkenli normal dağılım testi özellikleri anlatılmakta ve bu testin formülleri kullanılarak simülasyon oluşturulup grafiklerle incelenmektedir.

İKİ BOYUTLU NORMAL DAĞILIM

İki boyutlu rasgele değişkenin bileşenleri X ve Y olsun. $\mu_X, \mu_Y \in R, \sigma_x > 0, \sigma_y > 0$ ve $-1 < \rho < 1$ olmak üzere, iki değişkenli normal dağılımın olasılık yoğunluk fonksiyonu $x, y \in R$ için,

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]}$$

şeklindedir.



İKİ DEĞİŞKENLİ NORMAL DAĞILIMIN ÖZELLİKLERİ

a) X ve Y değişkenlerinin marjinalleri normaldir. Yani $X \sim N(\mu_x, \sigma_x^2)$ ve $Y \sim N(\mu_y, \sigma_y^2)$ dir.

b) $\rho_{x,y} = \rho$ dur. Yani X ve Y arasındaki korelasyon ρ dur.

c) a, b ϵ R için, $aX + bY \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\rho\sigma_x\sigma_y)$ dir.

d) Koşullu dağılımlar normaldir. Yani,

$\mu_{y|x} = \mu_y + \rho(\frac{\sigma_y}{\sigma_x})(x - \mu_x)$, $\mu_{x|y} = \mu_x + \rho(\frac{\sigma_x}{\sigma_y})(y - \mu_y)$,

$\sigma_{y|x}^2 = \sigma_y^2(1-\rho^2)$, $\sigma_{x|y}^2 = \sigma_x^2(1-\rho^2)$, olmak üzere,

$Y|X = x \sim N(\mu_{y|x}, \sigma_{y|x}^2)$ ve $X|Y = y \sim N(\mu_{x|y}, \sigma_{x|y}^2)$ dir.

İstatiksel Modelleme ve Simülasyon

İstatistiksel modelleme ve simülasyon, bir sistemin parçalarının birbirleriyle ve bütün sistemin kendisiyle etkileşimini araştırmak için kullanılan önemli yöntemlerden biridir. Genellikle betimsel istatistikler ve keşifçi veri analizi ardından kullanılır ve değişkenler arasındaki ilişkileri kurma veya geleceğe yönelik tahminler yapma gibi istatistiksel amaçlar taşır.

Ancak, bu yöntemlerin kullanımı belirli istatistiksel varsayımların sağlanmasına bağlıdır. Her modelleme yöntemi farklı varsayımlara dayandığı için, bu yöntemlerin dikkatli bir şekilde uygulanması önemlidir. Varsayımların doğruluğu incelenmelidir, çünkü bu, modelin doğruluğunu ve geçerliliğini etkileyebilir. İstatistiksel modelleme ve simülasyonun birlikte kullanılması, sistemlerin daha iyi anlaşılmasını sağlar ve gelecekteki olayların daha doğru tahmin edilmesine yardımcı olabilir.

Excel de simülasyon oluşturma

Excelde farklı istatistiksel ters fonksiyon yöntemleri kullanılarak simülasyon oluşturulabilir. Bunlardan bazıları şunlardır;

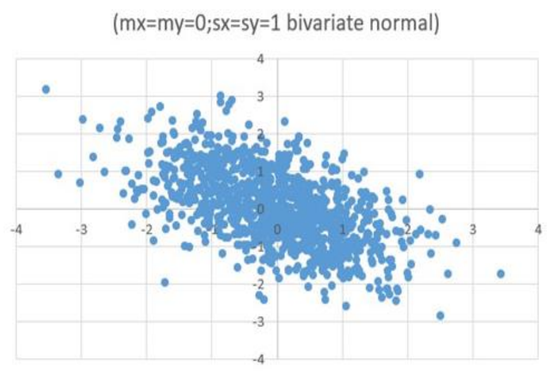
BİNOM.TERS	LOGNORM.TERS
F.TERS	NORM.S.TERS
FISHERTERS	NORM.TERS
GAMA.TERS	

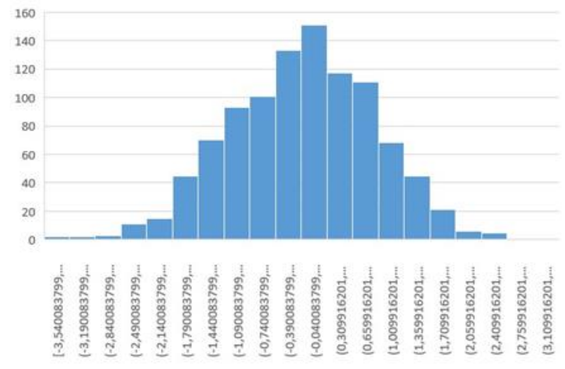
İKİ BOYUTLU NORMAL DAĞILIM İLE EXCELDE SİMÜLASYON OLUŞTURMA

$\mu_x, \mu_y = 0$, $\sigma_x, \sigma_y = 1$ için iki boyutlu normal dağılım simülasyonu;

- Öncelikle rasgele değerler atanmalıdır. Bunun için =RASTGELEARADA(alt_sınır, üst_sınır) formülünü kullanarak, belirtilen aralıkta bir rastgele sayı üretilir. Bu simülasyonda =RASTGELEARADA(0, 10000)/10000 formülü ile virgülden sonra dört basamak olacak şekilde 0 ile 1 arasında rasgele 1000 olasılık değeri üretiliyor.
- =NORM.S.TERS() fonksiyonu standart normal kümülatif dağılımın tersini verir. Dağılımın ortalaması 0, standart sapması 1'dir. Her bir satırda rasgele atanılan sayıları, bu fonksiyonu kullanarak x değişkeni olarak atıyoruz.
- Daha sonra x bilindiğinde y değerini bulabilmek için öncelikle simülasyonda =RASTGELEARADA(0, 10000)/10000 ile rasgele değerlerimizi atıyoruz.
- y değerlerini bulabilmek için NORM.TERS fonksiyonunu kullanılıyor. Bu fonksiyon belirtilen ortalama ve standart sapma için normal kümülatif dağılımın tersini verir.
- Ortalama ve standart sapmaları bulabilmek içi -1 ile 1 arasında rasgele ρ değeri atanıyor.
- $\mu_y = \rho.x$, $\sigma_y = 1 - \rho^2$ formülleri ile y değerleri için ortalama ve standart sapmaları hesaplanır.
- NORM.TERS fonksiyonunu uygulamak için, NORM.TERS(olasılık,ortalama,standart_sapma) olacak şekilde sırasıyla atanılan rasgele olasılık değeri daha sonra ortalamaları ve standart sapmaları ekleyerek X bilindiğinde Y değeri hesaplanır.

Örnek) $\rho = -0{,}56$ için dağılım ve histogram grafikleri;





Eğer iki değişken arasında hiç ilişki yoksa korelasyon katsayısı sıfır ya da sıfıra yakın bulunur. Eğer iki değişken birbiriyle yüzde yüz oranında ilişkili ise korelasyon maksimum 1 değeri alır fakat tamamen ters ilişkili ise -1 değerine yakın bir değer alır.

$\mu_x, \mu_y, \sigma_x, \sigma_y$ bilindiğinde iki boyutlu normal dağılım simülasyonu;

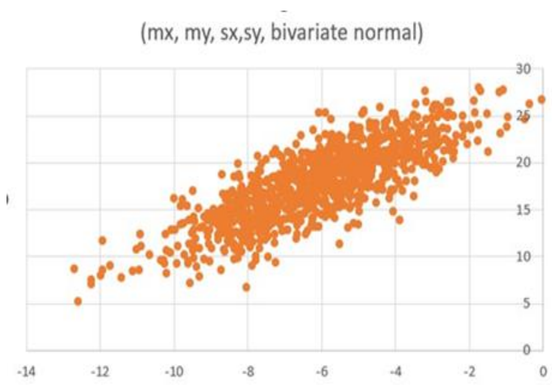
- Burada önceki aşamalarda bulduğumuz x ve y değerlerinden sonra $Z = \frac{T-\mu_t}{\sigma_t}$, $T \sim Normal(\mu_t, \sigma_t)$ formülü ile ortalamalar ve standart sapmalar bilindiğinde X ve Y değerlerini;

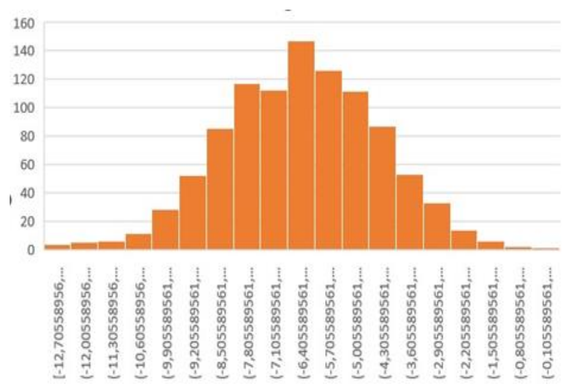
$$x = \frac{X-\mu_x}{\sigma_x} \rightarrow X = x.\sigma_x + \mu_x \quad , X \sim Normal(\mu_x, \sigma_x),$$

$$y = \frac{Y-\mu_y}{\sigma_y} \rightarrow Y = y.\sigma_y + \mu_y \quad , Y \sim Normal(\mu_y, \sigma_y)$$

olacak şekilde Excel'de hesapladıktan sonra atıyoruz.

Örnek) $\rho = 0{,}79,\ \mu_x = -6,\ \sigma_x = 2,\ \mu_y = 18,\ \sigma_y = 4$ için dağılım ve histogram grafikleri;





KAYNAKÇA

[1]http://www.isum.yildiz.edu.tr/sayfa/H%C4%B0ZMETLER%C4%B0M%C4%B0Z/%C4%B0STAT%C4%B0ST%C4%B0KSEL-MODELLEME/63

[2] https://www.probabilitycourse.com/chapter5/5_3_2_bivariate_normal_dist.php

[3] İbrahim Uysal ve Abdullah Kılıç, Normal Dağılım İkilemi, Yıl 2022

İSTATİSTİK BÖLÜMÜ

TÜRKİYE'DE İŞLENEN SUÇ İSTATİSTİKLERİ ÜZERİNE PANEL VERİ ANALİZİ

Harun YILDIRIM 18023032

Danışman: Dr. Öğr. Üyesi Erhan ÇENE

ÖZET

Çalışmada Türkiye'de bölgelere göre işsizlik, intihar oranı, boşanma hızı ve okuma yazma oranı değişkenlerinin suç üzerine etkisi 2013 - 2022 dönemi için panel veri analizi ile incelenmiştir. Panel veri analizinde kullanılan birimleri Türkiye'deki 81 il oluşturmaktadır (Türkiye İstatistik Kurumu'nun sınıflandırmasıdır). İşsizlik, intihar oranı, boşanma hızı ve okuma yazma oranı değişkenleri arasındaki etkiler statik panel veri analizinde en sık kullanılan sabit ve tesadüfi etki modelleri kullanılarak saptanmaya çalışılmıştır. Değişkenlerin birbirleri arasındaki ilişkiye bakmak için korelasyon analizi, tahminciler arasından uygun seçim için Hausman testi, birimler arası korelasyonu ölçmek için Paseran's CD testi, birimlerin durağanlığı sınamak için Harris – Tzavalis (HT) panel birim kök testi, ortalama karşılaştırması yapılabilmesi için Friedmans'R testi, gruplar arası varyansların homojen olup olmadığını sınamak için Levene, Brown ve Forsythe testi kullanılmıştır.

Çalışmada elde edilen sonuçlara göre, kaba boşanma hızındaki binde birlik artış kişi başı işlenen suç sayısını ortalama 1,24 arttırmakta; işsiz insan sayısındaki binde birlik artış suç sayısını ortalama 0,07 artırmaktadır. İntihar oranı yalnızca bir modelde anlamlıdır ve bu modele göre intihar oranındaki %1'lik artış, bin suç sayısını ortalama 3,52 azaltmaktadır. Bu araştırmada, okuma-yazma oranındaki 1 birim artışın suç sayısını ortalama 1.84 düşürdüğünü göstermektedir.

Panel Veri Analizi

Panel Veri, aynı birimlerin çeşitli karakteristiklerinin (yatay kesit değişkenlerinin) zaman boyunca gözlenmesi ile elde edilen veri setidir. Panel veri setinde hem yatay kesit hem de zaman serisi boyutu bulunmaktadır. Zaman serisi boyutu, aynı yatay kesit birimlerinin belirli bir dönem boyunca izlenmesi ile oluşturulmaktadır
Panel veri setinde yer alan birimlerin özelliklerine ait zaman serilerinde eksik gözlem bulunmadığında, bu tür panel veri setleri dengeli panel olarak adlandırılmaktadır. Bazı panel birimlerine ait gözlemlerin eksik olduğu panel veri setlerine ise dengesiz panel denmektedir.
Panel veri analizi, regresyon ve zaman serisi analizinin birleşimi gibi düşünülebilir (Frees, 2004). Bu analiz, kesitte yer alan birimlerin zaman boyunca tekrarlı gözlemleri söz konusu olduğundan, tekrarlı varyans ve varyans modellerine dayanmaktadır.

Veri

Bu çalışmada TÜİK tarafından sağlanan detaylı istatistikler kullanılmıştır. Veri seti 2013-2022 yılları arasında il bazlı Türkiye'de işlenmiş suç sayılarını, işsizlik sayılarını, intihar oranlarını, boşanma hızını ve okuma yazma bilen insan oranını içermektedir. Bağımlı değişken olarak Türkiye'de işlenmiş suç sayıları; bağımsız değişken olarak Türkiye'de işsizlik sayıları, intihar oranları, boşanma hızı ve okuma yazma bilen insan oranı kullanılmıştır.

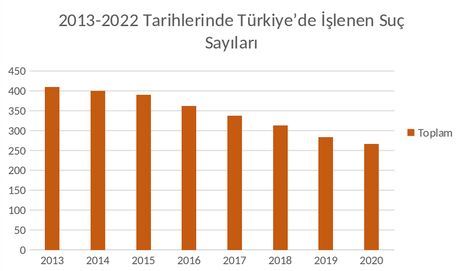
Kullanılan Yöntemler

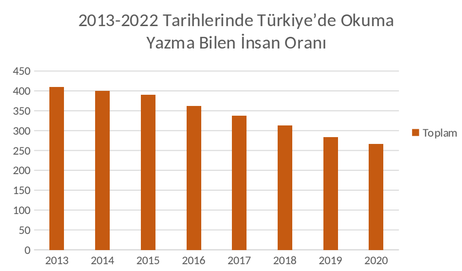
Bu çalışmada, panel veri analizi yöntemleri kullanılarak elde edilen sonuçları incelemek için bir dizi model ve test uygulanmıştır. Analiz sürecinde yararlanılan başlıca modeller ve testler şunlardır;

- Sabit Etkiler Modeli : Gruplar veya zaman dilimlerine ait etkilerin sabit olduğu varsayımı altında bağımsız değişkenlerin bağımlı değişken üzerindeki etkilerini doğru bir şekilde tahmin etmek için kullanılır.
- Tesadufi Etkiler Modeli : Bireysel veya grup düzeyindeki etkilerin rastgele varyasyonlar olarak kabul edildiği durumlarda, daha genelleştirilebilir ve verimli tahminler elde etmek için kullanılır.
- Harris – Tzavalis (HT) panel birim kök testi : Verinin durağanlığını sınamak için kullanılır.
- Hausmann Testi : Sabit Etkiler Modeli ile Tesadüfi Etkiler Modeli arasında seçim yapmak amacıyla, tesadüfi etkilerin modelde anlamlı olup olmadığını test etmek için kullanılır.
- Pesaran CD Testi : Birimler arası korelasyonu ölçmek için kullanılır.
- Levene, Brown ve Forsythe Testi : Gruplar arası varyansların homojen olup olmadığı sınamak için kullanılır
- Friedmans' R Testi : Gruplar arası ortalama karşılaştırması yapılabilmesi için kullanılır.

Veri Görselleştirme

Yapılan analiz ve kurulan model sonuçlarının daha anlaşılabilir olması açısından çeşitli yöntemler kullanılarak, elde edilen veriler görselleştirilmiştir. Veri görselleştirme aşaması, bu çalışma içerisinde önemli bir yere sahiptir. Aşağıda yer alan sütun grafikleri, Türkiye'de işlenen suç sayılarını ve okuma yazma oranı 2013– 2022 yıllarına göre gösterilmektedir.





Türkiye haritası üzerinde ise 2013 – 2022 yıllarında Türkiye'de işlenen suç sayıları il bazlı olarak gösterilmektedir. İşlenen en çok suçun olduğu ilimiz ilk olarak Van sonrasında sırasıyla Amasya ve Ankaradır.

Çalışmanın Amacı

Bu çalışmanın amacı, Türkiye'de 2013-2022 yılları arasında il bazlı işlenen suç sayıları ile işsizlik oranları, intihar oranları, boşanma hızı ve okuma yazma bilen insan oranı gibi çeşitli sosyo-ekonomik faktörler arasındaki ilişkileri analiz etmektir. TÜİK tarafından sağlanan detaylı istatistikler kullanılarak, bağımlı değişken olarak işlenen suç sayıları ve bağımsız değişkenler olarak işsizlik oranları, intihar oranları, boşanma hızı ve okuma yazma bilen insan oranı ele alınmaktadır. Bu çalışma, bu değişkenlerin işlenen suç sayıları üzerindeki etkilerini inceleyerek, suç oranlarını etkileyen temel sosyo-ekonomik faktörleri belirlemeyi hedeflemektedir. Analiz sonuçları, suç oranlarını azaltmaya yönelik politika önerileri geliştirilmesine katkıda bulunabilir.

Analiz Sonuçları

Tablo 2.1'de tesadüfi etkiler tahminleyeni sonuçları, varsayım ihlallerine dirençli tahminleyenlerle birlikte verilmektedir. Sütunlarda gösterilen modeller sırasıyla tesadüfi etkiler modeli (RE), değişen varyans ve ardışık korelasyona (otokorelasyon) dirençli model (RE_RB), AR1 korelasyonu dikkate alan model (RE_AR1), değişen varyans ve eşanlı korelasyona dirençli model (PCSE), değişen varyans, otokorelasyon ve eşanlı korelasyona dirençli modeldir (PCSE_AR1).

Değişken	RE	RE_RB	RE_AR1	PCSE	PCSE_AR1
issizliksayi	.0105709	.0105709	.0073469	.0102781	.072641
	.0037926	.0094051	.0023733	.0009033	.0016469
	0.005	0.261	0.002	0.0000	0.0000
intiharoraniyuzbin	-10.3357	-10.3357	-8.782342	-352.6739	-29.30763
	51.89727	19.75908	22.21527	97.02536	22.13653
	0.842	0.601	0.693	0.0000	0.186
bosanmahizibin	993.5512	993.5512	787.7233	2638.862	1204.819
	389.6071	269.7072	192.1899	302.7233	321.0877
	0.011	0.000	0.000	0.0000	0.000
okumayazmaorani	-383.0178	-383.0178	-374.9229	-184.1779	-446.452
	80.41649	82.70660	82..89659	53.58073	102.1857
	0.000	0.000	0.000	0.001	0.000
_cons	2571.474				
	885.9835	2571.474	2897.514	964.456	2724.872
	0.004	555.5867	633.9844	408.677	660.259
		0.000	0.000	0.018	0.0000
chi2	65.75	31.69	61.95	322.12	67.62
P	0.000	0.000	0.000	0.000	0.000

Modellerin tümünde Wald X^2 test istatistiği anlamlıdır. Katsayıların testleri incelendiğinde kaba boşanma hızının her modelde anlamlı (p=0,0000), intihar oranı RE modelinde %84; RE_AR1 modelinde %69 ve diğer dört modelde en çok %0,5 anlamlılık seviyelerinde anlamlı olduğu anlaşılmaktadır. İşsiz insan sayısı değişkeni ise yalnızca RE_RB modelinde (p=0,261) anlamsız, diğer modellerde anlamlıdır. Katsayıların değerlerinde çok büyük değişiklikler olmadığı gözlenmektedir.

Varsayımlar için yapılan testlerde otokorelasyon ve eşanlı korelasyon için sıfır hipotezlerinin kabul edildiği gibi reddedildiği de gözlendiği için, elde edilen modellerden PCSE ile tahmin edilen katsayılar incelendiğinde, suç sayısı üzerinde kaba boşanma hızının ve işsiz insan sayısının pozitif, intihar oranı ve okuma yazma bilen insanların oranının negatif etkisi olduğu görülmektedir. Bu modele göre işsiz insan sayısı oranının suç sayısı üzerinde anlamlı bir etkisi yoktur.

Sonuç ve Öneriler

Analiz sonuçlarına göre, işsiz insan sayısındaki %1'lik artış suç sayısını ortalama 0.07 arttırmakta. İşsiz insan sayısı, belirli bir dönemde çalışmak isteyen ancak iş bulamayan bireylerin sayısını ifade eder. İşsiz insanların artışı ile birlikte bireyler hem maddi hem de manevi sorunlar yaşayabilmektedir. Üstelik işsizlik süreci doğal haliyle insanların psikolojilerinde sorunlara yol açabilirken, işsizlik sürecinde insanlar arasında yaşanan büyük çaplı sorunlar ve kavgalar bireylerin ruhsal sağlıklarında ciddi bozulmalara neden olabilmektedir. Yaşadıkları sorunlar ile baş edemeyen kişiler de daha sabırsız, öfkeli ve şiddete yönelik hale gelerek suça yönelebilmektedir.

İntihar oranındaki %1'lik artış, suç sayısını ortalama 3,52 azaltmaktadır.İntihar oranının suç sayısı üzerinde negatif etkisi olması beklentiye ters bir durumdur. Değişkenler arası ters yönlü ve beklentiye aykırı bu ilişki, çoklu doğrusal bağlantıdan ve/veya düşük serbestlik derecesinden kaynaklanıyor olabilir.

Toplumların sosyal ve ekonomik gelişiminde okuma-yazma oranının kritik bir rol oynadığı genel kabul görüyor. Bu araştırmada, okuma-yazma oranındaki %1 artışın suç sayısını ortalama 1.84 düşürdüğünü gösteriyor. Bu durum, eğitimli bireylerin daha iyi istihdam edilmesi ve yasadışı faaliyetlere olan ihtiyacın azalmasıyla ilişkilendiriliyor. Ayrıca, okuryazar bireylerin daha bilinçli olmaları ve yasalara daha fazla uymaları, toplumsal sorunları daha etkili bir şekilde çözmelerine yardımcı oluyor. Bu nedenle, okuma-yazma becerilerini geliştirmeye yönelik çabaların, suçla mücadelede önemli bir strateji olabileceği belirtiliyor.

Boşanma hızındaki bir artışın suç sayısını anlamlı derecede artırdığı görülmektedir. Boşanma hızının binde bir birim artmasının (yani boşanma oranının binde 1 artması) suç sayısını yaklaşık 1204.819 birim artıracağını gösterir.Bu, boşanmaların toplumsal yapı üzerindeki geniş etkilerini ve potansiyel olarak suç oranlarındaki artışı yansıtabilir.

KAYNAKÇA

- BALTAGI, Badi H. (2010), Econometric Analysis of Panel Data, Fourth Edition, John Wiley & Sons Ltd.
- https://biruni.tuik.gov.tr/medas/?kn=73&locale=tr
- HSIAO, Cheng (2003), *Analysis of Panel Data*, 2nd press, Cambridge University Press, New York.

İSTATİSTİK BÖLÜMÜ

MAKİNE ÖĞRENİMİ İLE PETROL KUYULARINDA ANOMALİ TESPİTİ VE HATA TAHMİNİ

Helin Tuğba YAŞASIN 20023052

Danışman: Arş. Gör. Dr. Coşkun PARİM

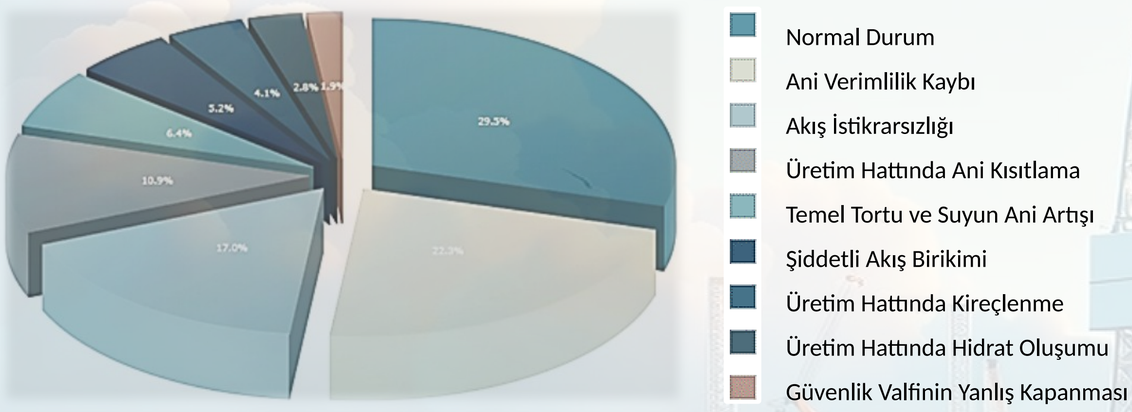
ÖZET

Bu çalışma, petrol kuyularında meydana gelen anomalilerin tespit edilmesi ve hataların tahmin edilmesi amacıyla gerçekleştirilmiştir. Endüstriyel sensör verileri kullanılarak anormalliklerin erken tespiti ve operasyonel verimliliğin artırılması hedeflenmiştir.

PETROL KUYULARI

Petrol kuyuları, yer altındaki petrol rezervuarlarına ulaşmak ve bu rezervuarları ekonomik olarak üretmek amacıyla açılan kuyulardır. Petrol kuyuları, petrolün çıkarılması ve taşınması süreçlerinde kritik bir rol oynar ve petrol endüstrisinin temel yapı taşlarından biridir. Deniz üstü petrol kuyuları, deniz tabanında bulunan petrol rezervuarlarına ulaşmak amacıyla inşa edilen kuyulardır. Bu kuyular, özellikle derin denizlerde ve uzak deniz alanlarında bulunan büyük petrol rezervlerini çıkarmak için kullanılır. Deniz üstü petrol kuyularının işletilmesi, karada bulunan kuyulara göre daha karmaşık ve maliyetlidir ancak büyük rezervuarlar nedeniyle ekonomik olarak önemlidir.

PETROL KUYULARI VERİ SETİNDEKİ OLAYLAR



MAKİNE ÖĞRENMESİ ALGORİTMALARI

Yapay Sinir Ağı: İnsan beyninin yapısı ve işlevine göre modellenen bir tür makine öğrenimi algoritmasıdır. Çok katmanlı yapısı ile veriler arasındaki karmaşık ilişkileri öğrenebilir.

Rastgele Orman: Birden çok karar ağacını bir araya getirerek sınıflandırma ve regresyon problemlerinde makine tahmin doğruluğunu artıran bir makine öğrenmesi algoritmasıdır.

Karar Ağacı: Veriyi dallara ayırarak her bir dalın sonunda bir karar veya tahminle sonuçlanan hiyerarşik bir modeldir.

AdaBoost: Zayıf öğrenicileri ardışık olarak eğiten ve her adımda hatalara daha fazla ağırlık vererek bu hataları düzelten bir topluluk öğrenme algoritmasıdır. Her yeni model, önceki modellerin zayıf yönlerini hedef alarak daha güçlü ve doğru bir sınıflandırıcı oluşturur.

İzolasyon Ormanı: Veri noktalarının izolasyon sürelerini kullanarak anormallikleri tespit eden ve özellikle anomali algılama için etkili olan bir makine öğrenmesi yöntemidir.

Gradient Boosting: Hataları düzelterek, her aşamada yeni ağaçlar ekleyen ve tahmin performansını artıran bir makine öğrenmesi algoritmasıdır.

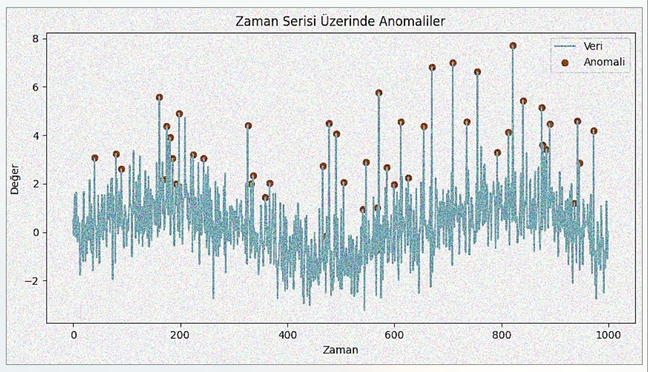
ANOMALİ TESPİTİ

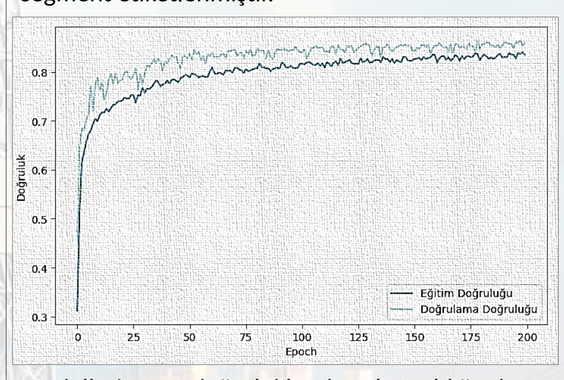
Petrol endüstrisinde, doğal akışlı deniz üstü petrol kuyularında istenmeyen durumların erken tespiti büyük bir önem taşımaktadır. Bu kuyularda meydana gelen anormallikler, üretim verimliliğini, kuyunun genel sağlığını olumsuz etkileyebilir ve yüksek maliyetlere yol açabilir. Bu nedenle, anormallik tespiti, operasyonel riskleri azaltmak, üretim kayıplarını önlemek ve maliyetleri azaltmak için kritik bir öneme sahiptir. Makine öğrenimi tekniklerinin kullanımıyla, anormalliklerin otomatik olarak tespit edilmesi ve hatta bazı durumlarda önceden tahmin edilmesi mümkün hale gelir. Anormallik tespiti için makine öğrenimi modelleri, çeşitli girdi değişkenlerini kullanabilir. Bu değişkenler arasında basınç, sıcaklık, akış hızı, su ve petrol oranları gibi kuyudan elde edilen sensör verileri bulunmaktadır.

Değişken	Değişken Türü	Açıklama
Petrol Kuyularındaki Anormal Olaylar	Bağımlı	Normal Durum, Temel Tortu ve Suyun Ani Artışı, Güvenlik Valfinin Yanlış Kapanması, Şiddetli Akış Birikimi, Akış İstikrarsızlığı, Ani Verimlilik Kaybı, Üretim Hattında Ani Kısıtlama, Üretim Hattında Kireçlenme, Üretim Hattında Hidrat Oluşumu
B-KBS	Bağımsız	Kalıcı Kuyubaşı Basınç Sensöründeki Basınç
S-KBS	Bağımsız	Kalıcı Kuyubaşı Basınç Sensöründeki Sıcaklık
B-ÜKS	Bağımsız	Üst Düzey Üretim Kısma Sensöründeki Basınç
S-ÜKS	Bağımsız	Üst Düzey Üretim Kısma Sensöründeki Sıcaklık
B-GLS	Bağımsız	Gaz Lift Enjeksiyon Sensöründeki Basınç
A-GLS	Bağımsız	Gaz Lift Enjeksiyon Sensöründeki Akış Hızı

HATA TAHMİNİ

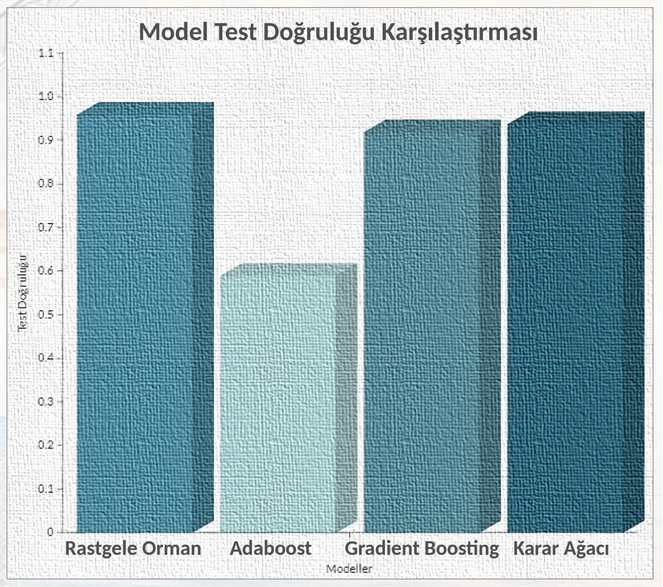
Petrol kuyularından elde edilen sensör verileri kullanılarak anomali tespiti için bir veri seti oluşturulmuştur. Bu veri seti, deniz üstü petrol kuyularından elde edilen gerçek sensör ölçümlerini, yazılım tarafından simüle edilmiş veya elle çizilmiş veriler içermektedir. Veri seti, sekiz farklı türde anormal olayı ve bir normal durumu içermektedir. Veri seti, her bir olay türü için ayrı klasörlerde düzenlenmiş ve her veri toplama oturumu için ayrı dosyalarda saklanmıştır. Her bir dosya, belirli bir zaman penceresi kullanılarak işlenmiş ve her bir segment etiketlenmiştir.

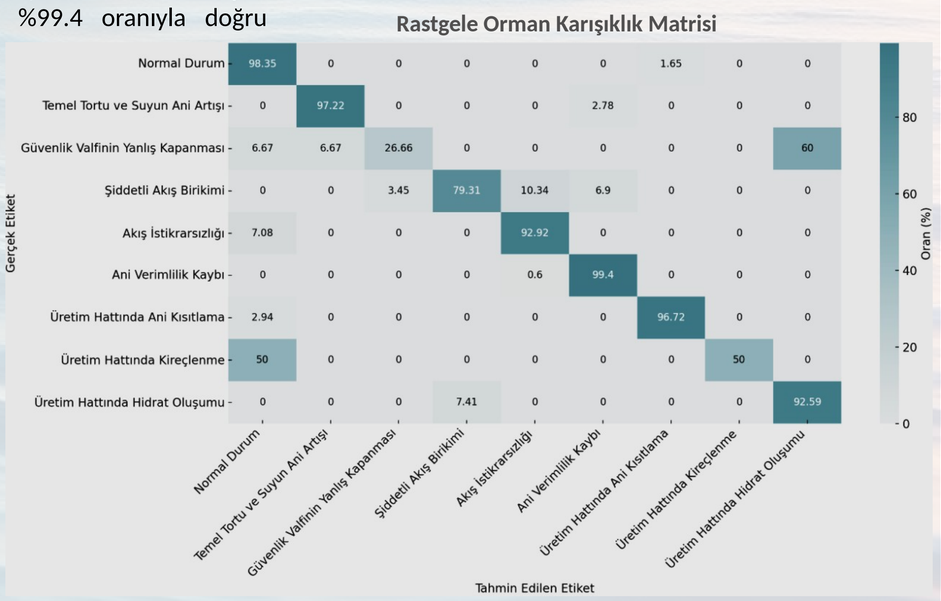




Anomali tespiti ve hata tahmini için çeşitli makine öğrenimi modelleri kullanılmıştır. Rastgele orman, karar ağacı, adaboost, gradient boosting, izolasyon ormanı ve yapay sinir ağı modelleri bu amaçla eğitilmiştir. Zaman serisinin ortalarında ve sonlarına doğru daha sık anomali tespitleri bulunmaktadır. Bu, sistemin belirli zaman dilimlerinde daha sık sorun yaşadığını veya bu dönemlerde ölçümlerde daha fazla sapma olduğunu gösterir. Modelin eğitim ve doğrulama grafiğinde görüldüğü üzere, eğriler benzer bir artış göstermektedir. Modelin öğrenme kapasitesinin iyi olduğu sonucuna varılabilir.

Modellerin test doğrulukları karşılaştırıldığında en yüksek doğruluk değerinin %95.97 ile Rastgele Orman modeli olduğu görülmektedir. Karar ağacı %93.7, Gradient Boosting %92 ve Adaboost %60 doğruluk oranına sahiptir. En yüksek doğruluk değerine rastgele orman modeli ile ulaşıldığı için, rastgele orman modelinin karışıklık matrisi incelenmiştir. Normal durum %98.35 doğruluk oranıyla doğru tahmin edilmiştir. Temel tortu ve suyun ani artışı %97.22 doğruluk oranıyla tahmin edilmiştir. Güvenlik valfinin yanlış kapanması %26.6 doğruluk oranıyla doğru tahmin edilmiş, %60 oranında yanlış tahmin edilmiştir. Bunun sebebi güvenlik valfinin kapandığı durumda ani basınç değişikliklerine ve akış durmalarına neden olduğu için hidrat oluşumu ile benzer sensör verileri göndermesidir. Ayrıca güvenlik valfinin yanlış kapanması nadir bir olay olduğundan model bu olayı tanımakta zorlanmaktadır. Şiddetli akış birikimi %79.31 doğru tahmin edilmiştir. Akış istikrarsızlığı %92.92 doğruluk oranıyla tahmin edilmiştir. Ani verimlilik kaybı, %99.4 oranıyla doğru tahmin edilmiştir. Üretim hattında ani kısıtlama %96.72 doğruluk oranıyla tahmin edilmiştir. Üretim hattında kireçlenme %50 doğruluk oranıyla doğru tahmin edilmiştir. Modelin tanıması için yeterli veri yoktur. Üretim hattında hidrat oluşumu ise %92.59 doğruluk oranıyla doğru tahmin edilmiştir.





KAYNAKÇA

[1]Zhang, Y., Li, G., & Wang, X. (2018). A Support Vector Machine-Based Approach for Detecting Anomalies in Oil Production Data. Energy Exploration & Exploitation, 36(6), 1367-1380.

[2] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

[3] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

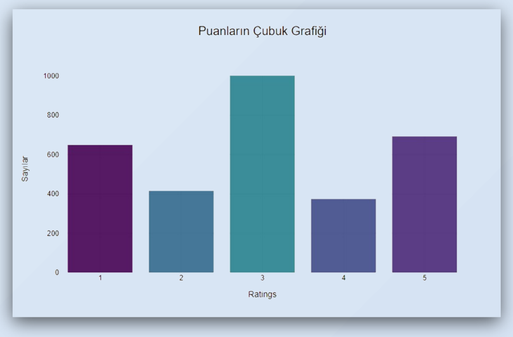
TÜRKİYE'DEKİ SEYAHAT VE KONAKLAMA İNCELEMELERİNİN DUYGUSAL ANALİZİ: BERT MODELİ İLE DUYGU ANALİZİ YAKLAŞIMI

İrem Nur AKCAN 19023056

Danışman : Doç. Dr. Gülder KEMALBAY

Bu proje, Türkiye'deki seyahat ve konaklama yerlerine yönelik dünya genelindeki kullanıcı değerlendirme yorumlarının duygu analizini yapmayı amaçlamaktadır. Günümüzde, seyahat eden bireyler, ziyaret edecekleri yerler hakkında bilgi almak için genellikle çevrimiçi yorumlara başvurmaktadır. Bu nedenle, bu yorumların analiz edilmesi, hem işletmelerin hizmet kalitesini arttırması hem de potansiyel ziyaretçilerin daha bilinçli kararlar vermesi açısından büyük önem taşımaktadır. Öncelikle, veri seti çevrimiçi bir platformdan toplanmış ve veriler ön işleme tabi tutulmuştur. Yorumların içindeki anlamsız karakterler ve durak kelimeler temizlenmiş, ardından BERT modelinin sınıflandırma görevleri için tasarlanmış ve özelleştirilmiş bir modeli olan BertForSequenceClassification ile eğitilmiştir. Modelin eğitimi ve ince ayarı tamamlandıktan sonra, yorumlar üzerinden duygu analizi gerçekleştirilmiştir. Analiz sonuçları, pozitif, negatif ve nötr duygu dağılımlarını içeren çeşitli görsellerle desteklenmiştir.Bu çalışmanın bulguları, Türkiye'deki seyahat ve konaklama yerlerinin genel müşteri memnuniyeti düzeylerini ve belirli konulardaki güçlü ve zayıf yönlerini ortaya koymuştur.

BETİMSEL İSTATİSTİKLER

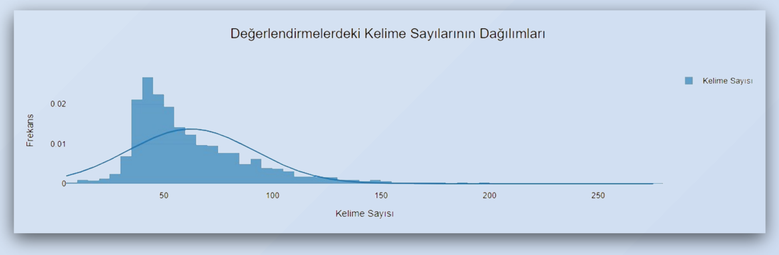


	Ratings	Sayılar
0	3	1001
1	5	692
2	1	649
3	2	415
4	4	374

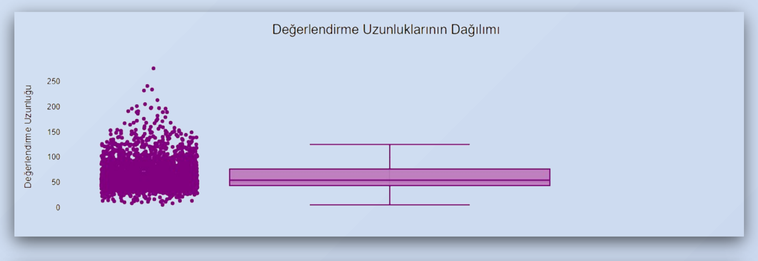
	Sütun	Boş Değer
0	ratings	0
1	reviews	0

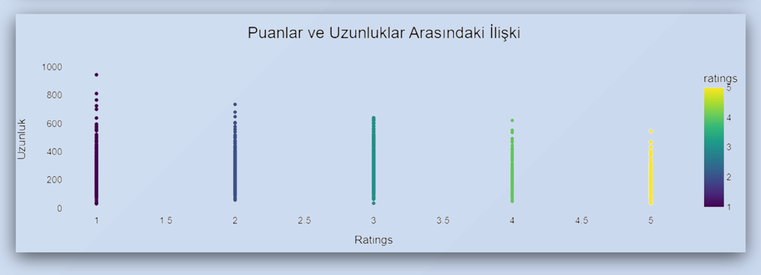
	İstatistik	Değer
0	count	3030
1	mean	3.03
2	std	1.4
3	min	1
4	25%	2
5	50%	3
6	75%	4
7	max	5

Puan dağılımını gösteren çubuk grafikte, en yüksek frekans 1001 kez ile 3 puanda, en düşük frekans ise 374 kez ile 4 puanda gözlemlenmiştir. Her bir puanın frekansları sırasıyla; 1 puan 692, 2 puan 649, 3 puan 1001, 4 puan 415 ve 5 puan 374 kez verilmiştir. Veri setinde eksik veri bulunmamaktadır. İstatistiksel özet bilgilerde, veri sayısı 3030, ortalama 3.03, standart sapma 1.4, minimum değer 1, maksimum değer 5 olarak belirlenmiştir. Bu analiz, puanların genel olarak ortalama civarında yoğunlaştığını ve veri setinin eksiksiz olduğunu göstermektedir.



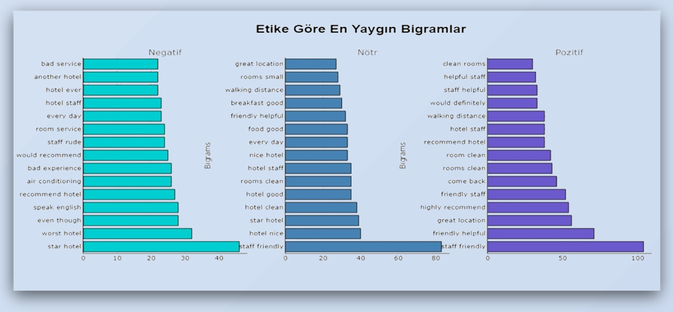
Değerlendirme puanları ve bu değerlendirmelerin uzunlukları analiz edilmiştir. İlk grafik, değerlendirmelerdeki kelime sayılarının dağılımını göstermektedir. Kelime sayıları genellikle 50 civarında yoğunlaşmakta, ardından düşüş eğilimi göstermektedir. İkinci grafik, değerlendirme uzunluklarının kutu grafiğini sunmakta ve değerlendirmelerin büyük çoğunluğunun 50 ile 100 kelime arasında olduğunu, bazı uç değerlerin ise 200 kelimenin üzerinde olduğunu göstermektedir. Üçüncü grafik, puanlar ile değerlendirme uzunlukları arasındaki ilişkiyi göstermektedir. Bu grafikte, yüksek puanlar (4 ve 5) genellikle daha kısa değerlendirmelerle ilişkilendirilirken, düşük puanlar (1 ve 2) daha uzun değerlendirmelerle ilişkilidir. Bu bulgular, kullanıcıların değerlendirme puanlarına göre yazı uzunluklarının farklılaştığını ortaya koymaktadır.





NGRAMLAR

Doğal Dil İşleme (NLP) alanında, "ngram" terimi, ardışık kelime Dizilimlerini ifade eder. Ngramlar, dil modelleri ve metin analizi gibi uygulamalarda sıkça kullanılır. Tek bir kelimeden oluşan ngramlara "unigram", iki kelimeden oluşanlara "bigram", üç kelimeden oluşanlara ise "trigram" denir. Ngramlar, metin içerisindeki kelime ilişkilerini ve yapılarını anlamamıza yardımcı olur. Görselde, otel yorumlarıyla ilgili bigramların dağılımı görülmektedir. Görsel, bigramları üç farklı duygu kategorisine ayırmaktadır: negatif, nötr ve pozitif. Bu ayrım, kullanıcıların otel yorumlarında en sık kullandıkları iki kelimelik ifadeleri, duygu durumlarına göre analiz ederek ortaya koymaktadır.

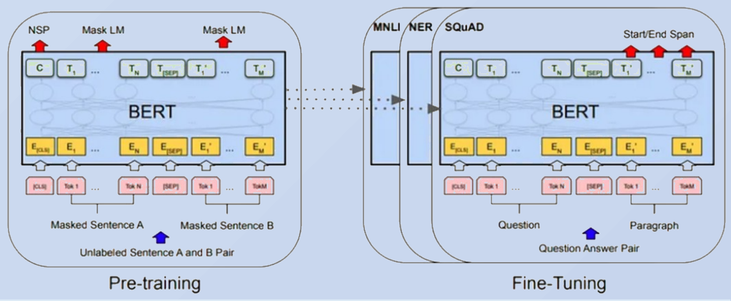


BERT ALGORİTMASI

BERT (Bidirectional Encoder Representations from Transformers), dil işleme görevlerinde çığır açan bir derin öğrenme modelidir. 2018 yılında Google tarafından geliştirilmiştir ve birçok doğal dil işleme görevinde, örneğin metin sınıflandırma, adlandırılmış varlık tanıma ve soru-cevap sistemlerinde, üstün performans sergilemektedir.

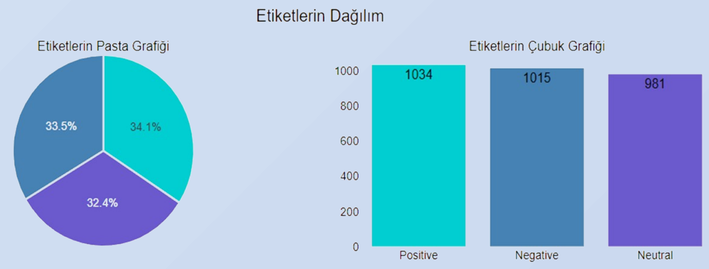


BERT modeli iki aşamada eğitilir: ön eğitim (pre-training) ve ince ayar (fine-tuning). Ön eğitim aşamasında, BERT büyük miktarda metin verisi üzerinde dilin yapısını ve anlamını öğrenir. Bu süreçte Maskelenmiş Dil Modeli (MLM) ve Cümle Çifti Tahmini (NSP) görevleri kullanılır. MLM'de rastgele seçilen kelimeler maskelenir ve model bunları bağlamına göre tahmin eder. NSP'de ise model, iki cümlenin ilişkili olup olmadığını tahmin eder. Bu sayede BERT, hem kelime hem de cümle düzeyinde bağlamsal anlayış geliştirir.

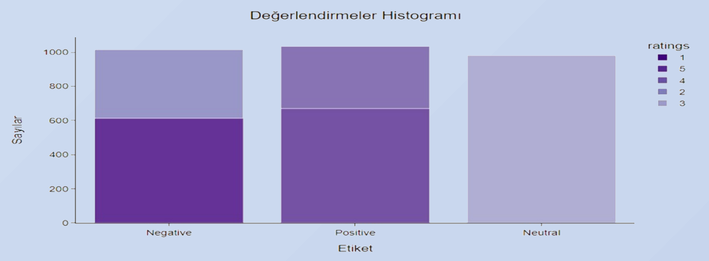


İnce ayar aşamasında, önceden eğitilmiş BERT modeli belirli bir NLP görevi için özelleştirilir ve göreve özgü etiketli verilerle yeniden eğitilir. Bu süreç, modelin göreve özel bilgileri öğrenmesini ve performansını optimize etmesini sağlar. Metin sınıflandırma, duygu analizi ve soru-cevap gibi görevler için ince ayar yapılır. Modelin tüm parametreleri yeniden eğitildiği için, BERT görevin gerektirdiği örüntüleri ve ilişkileri öğrenir. Bu, genellikle daha kısa sürelidir ve daha az veri gerektirir.

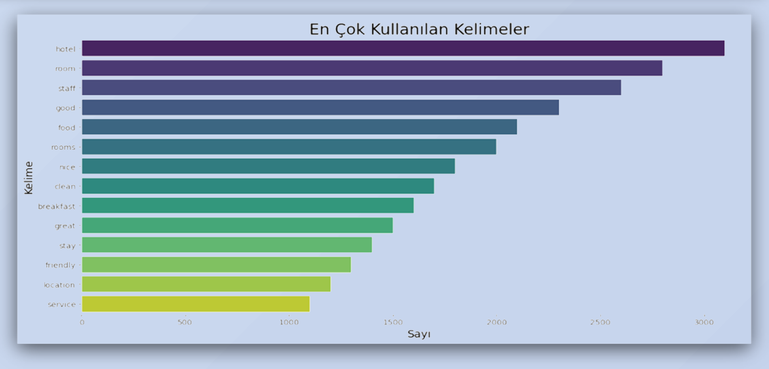
LABEL ENCODING



Puanlar üzerinde Label Encoding işlemi gerçekleştirilmiş ve sonuçlar çeşitli grafiklerle görselleştirilmiştir. Histogramda, negatif, pozitif ve nötr etiketlerin puan dağılımı incelenmiş; negatif ve pozitif etiketler yaklaşık 1000, nötr etiketler ise yaklaşık 900 değerlendirme almıştır. Pasta grafiğinde, pozitif etiketlerin %34.1, negatif etiketlerin %32.4 ve nötr etiketlerin %33.5 oranında dağıldığı görülmektedir. Label Encoding işlemi, kategorik verilerin sayısal değerlere dönüştürülmesini sağlayarak analizlerin daha anlamlı ve karşılaştırılabilir olmasını mümkün kılmıştır. Ayrıca, etiket sınıflarının dağılımlarının oldukça dengeli olduğu gözlemlenmiştir.

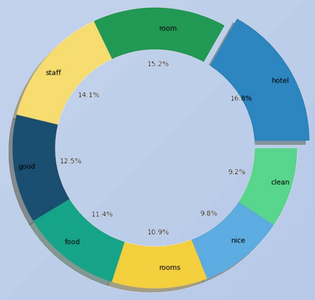


EN ÇOK KULLANILAN KELİMELER

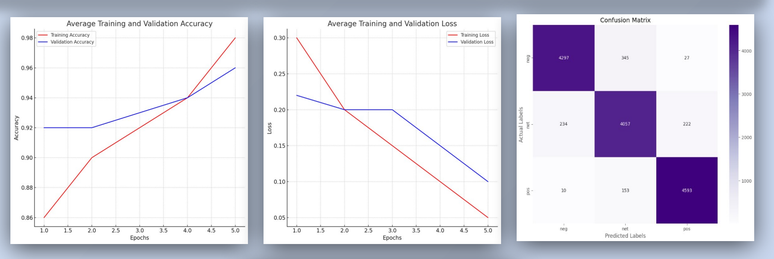


Bu iki grafik, otel değerlendirmelerinde en sık kullanılan kelimelerin analizini sunmaktadır. Yatay çubuk grafikte "hotel", "room", "staff", "good", "food", "rooms", "nice", "clean", "breakfast", "great", "stay", "friendly", "location" ve "service" kelimeleri yer almaktadır. En sık kullanılan kelime "hotel" olup, yaklaşık 3000 kez kullanılmıştır. İkinci grafik ise aynı verileri yüzdesel dağılım olarak sunan bir donut grafiktir. Bu grafikte de "hotel" kelimesi %16.8 ile en büyük dilimi oluşturmakta olup, "room" %15.2 ve "staff" %14.1 gibi diğer kelimeler izlemektedir. Bu grafikler, otel değerlendirmelerinin hangi konulara odaklandığını görsel olarak açıkça ortaya koymaktadır.

	word	count
0	hotel	3119
1	room	1557
2	staff	1557
3	good	1312
4	food	1019



ÇALIŞMA SONUÇLARI




```
Classification Report:
              precision    recall  f1-score   support

           0     0.9463    0.9203    0.9331      4669
           1     0.8907    0.8990    0.8948      4513
           2     0.9486    0.9657    0.9571      4756

    accuracy                         0.9289     13938
   macro avg     0.9285    0.9283    0.9283     13938
weighted avg     0.9291    0.9289    0.9289     13938
```

Modelin performansı, eğitim ve doğrulama kayıplarının epoch'lar boyunca azaldığını ve doğruluklarının arttığını gösteren grafiklerle değerlendirilmiştir. Bu da modelin doğrulama verisi üzerinde de iyileştiğini göstermektedir. Eğitim doğruluğu hızlı bir artış gösterirken, doğrulama doğruluğu daha sabit bir artış sergilemiştir. Sınıflandırma raporu ise metriklere göre modelin performansını %92.89 genel doğrulukla değerlendirmiştir. Precision, recall ve f1-score değerleri tüm sınıflar için yüksektir. Karışıklık matrisi, modelin negatif, nötr ve pozitif sınıflarını ne kadar doğru tahmin ettiğini ortaya koymuştur; model pozitif sınıfta en doğru tahminleri yaparken, nötr sınıfta görece karışıklık yaşamıştır. Bu bulgular, BertForSequenceClassification modelinin otel yorumlarının duygu analizinde etkili bir şekilde kullanılabileceğini ve yüksek doğruluk oranıyla başarılı sonuçlar verdiğini kanıtlamaktadır.

KAYNAKÇA

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171-4186

[2] Foundations of Statistical Natural Language Processing by Christopher D. Manning and Hinrich Schütze

[3] "Natural Language Processing with Classification and Vector Spaces" by Coursera (offered by deeplearning.ai)

İSTATİSTİK BÖLÜMÜ

D SEGMENTİ ARAÇLARIN İSTATİKSEL ANALİZİ VE GÖRSELLEŞTİRİLMESİ

Mehmet Ali Devran

19023042

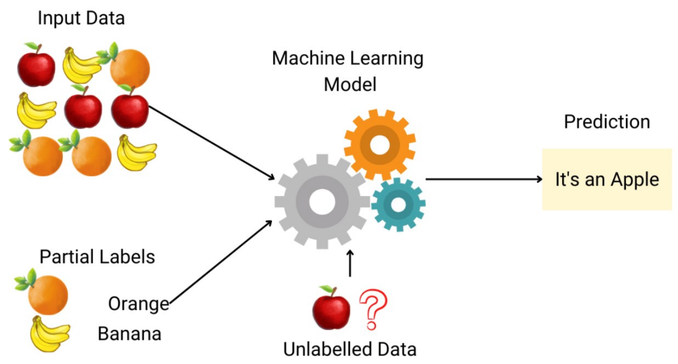
Danışman: Arş. Gör. Dr. Coşkun PARİM

ÖZET

Bu çalışmanın amacı, D segmenti araç pazarındaki rekabeti ve satışları etkileyen faktörlerin istatiksel analizini yapmaktır. Bunu ise, bu segmentte bulunan araçların fiyat, azami hız, şehir içi ortalama yakıttüketimi, bagaj hacmi, azami tork, ve beygir gücü gibi faktörlerin etkileri araştırılarak yapılmıştır. Analiz süresince Chernoff yüzleri ve radar grafiği gibi görsel araçlar, bu faktörler arasındaki ilişkiyi dahaaçık bir şekilde görselleştirmek için kullanılmıştır. Ayrıca, makine öğrenimi yöntemlerinden k-ortalamalar kümeleme yöntemi, araçların özelliklerine göre kümelenmesi ve optimal sayıda kaç tane küme oluşturulması gerektiğini gözlemlemek için kullanılmıştır. Bu çalışmanın bulguları, D segmenti araçların pazarlama dinamikleri üzerine yapılan istatistiksel araştırmalara ve otomotiv endüstrisindeki satış / pazarlama alanlarına da katkı sağlamaktır.

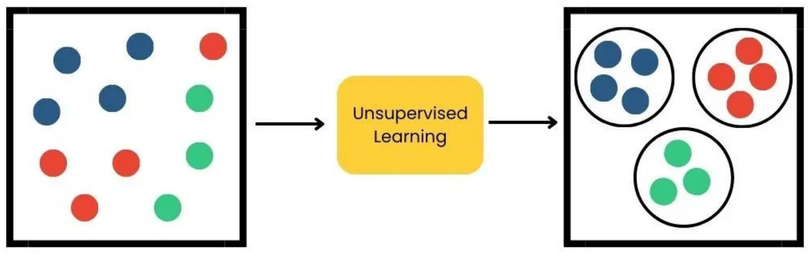
Gözetimli Makine Öğrenimi (Supervised Learning)

Gözetimli öğrenme, önceden gözlenerek sonuçları bilinen (etiketlenmiş) verilerden yararlanarak bu verilerin ve sonuçların tümünü kapsayan bir fonksiyon elde etmeyi amaçlayan makine öğrenmesi yöntemidir. Kısaca gözetimli öğrenme (Supervised Learning) eğitim verileri üzerinden bir fonksiyon üreten tekniktir. Bu öğrenme tekniğinin algoritması girdilerle (etiketlenmemiş veri) çıktılar (etiketlenmiş veri) arasında en iyi eşleme yapan fonksiyonu üretmek üzere çalışmaktadır. Eğitim verisi, girdi nesneleri ve sonuçlanması istenilen çıktılardan meydana gelmektedir. Sistem eğitilirken veri setinde bulunan her bir örneğe ait giriş ve çıkışlar verilir.



Gözetimsiz Makine Öğrenimi (Unsupervised Learning)

Gözetimsiz öğrenme, etiketlenmemiş ham veri kümesindeki veriler arasında var olan ancak açıkça görülmeyen bağıntının açığa çıkarılması işlemidir . Gözetimsiz öğrenme (Unsupervised Learning) gözlemlere bağlı bir makine öğrenmesi yöntemidir. Diğer bir deyişle çıktı verilerini kullanmadan sadece girdiler üzerinde öğrenme işlemini gerçekleştirmeye çalışır. Genellikle kümeleme, olasılık 32 yoğunluk tahmini, öznitelikler arasındaki ilişkilerin bulunması ve boyut indirgeme gibi amaçlar için kullanılmaktadır. Gözetimsiz öğrenme tekniği özellikle olasılık teorisine dayanan kümeleme işlemlerinde kullanılmaktadır.

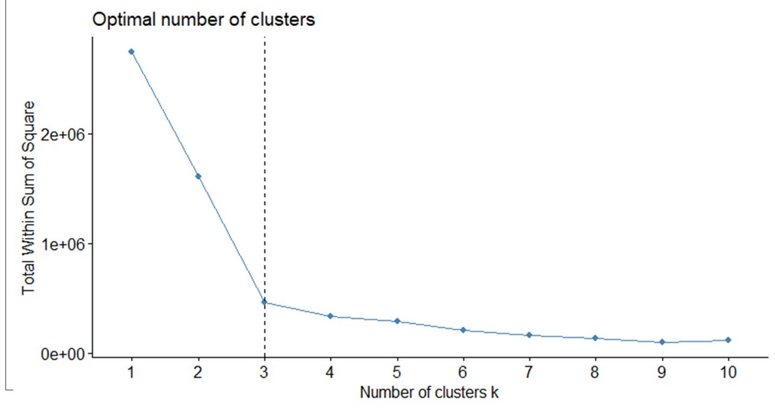


Veri Setinin Tanımlanması

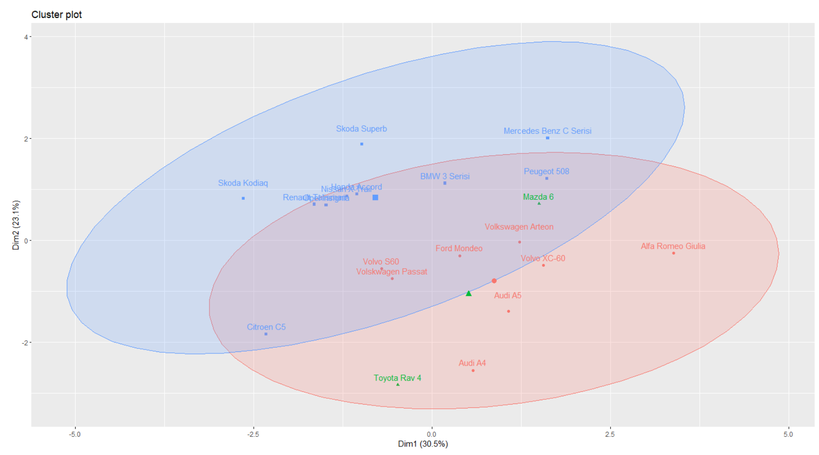
Model	Beygir Gücü (hp)	Azami Tork (nm)	Azami Hız (km/s)	Silindir Hacmi (cc)	Yakıt Deposu (lt)	Bagaj Hacmi (lt)	Uzunluk (mm)	Şehir içi Ortalama Tüketim (lt)
BMW 3 Serisi	170	250	232	1597	59	480	4709	7.7
Mercedes Benz C Serisi	201	300	246	1496	66	455	4751	8.7
Volskwagen Passat	150	340	220	1968	59	586	4775	4.6
Volvo S60	161	265	180	1969	60	442	4761	6.2
Skoda Superb	150	250	219	1498	66	595	4869	6.8
Audi A4	204	400	210	1968	40	460	4721	5.5
Peugeot 508	225	300	248	1598	62	487	4750	7.6
Volkswagen Arteon	197	400	237	1968	66	563	4862	5.2
Opel Insignia	122	285	200	1496	62	490	4897	5.3
Citroen C5	130	300	188	1499	53	580	4500	4.4
Skoda Kodiaq	150	250	198	1498	58	835	4697	7.2
Mazda 6	187	252	240	2498	62	490	4800	8.0
Honda Accord	190	260	190	1498	56	570	4901	7.3
Ford Mondeo	150	370	215	1995	60	458	4871	5.8
Renault Talisman	160	270	207	1333	50	608	4853	6.7
Nissan X-Trail	158	270	198	1332	60	565	4690	7.6
Alfa Romeo Giulia	280	400	240	1995	58	480	4643	9.0
Audi A5	204	400	210	1968	58	465	4733	5.6
Toyota Rav 4	222	270	180	2487	55	580	4600	4.4
Volvo XC-60	200	420	205	1969	71	483	4688	6.8

K-Ortalamalar Kümeleme Yöntemi

K-ortalama algoritması, veri kümesindeki n adet nesnenin k tane kümeye bölümlenmesini sağlamaktadır. Kümeleme sonucunda kümelerin içinde yer alan nesneler arasındaki benzerliklerin çok olması, diğer kümelerin nesneleri ile 43 arasındaki (inter-cluster) benzerliklerin çok düşük olması beklenmektedir alama algoritmasının iki temel bileşeni vardır. Birinci bileşen için k merkez değerleri rastgele biçimde belirlenmeli yani, k merkez değeri girdi olarak sunulmuş olmalıdır. Diğer bileşen için ise veri kümesindeki her nesne kendisine en yakın merkez noktanın olduğu demete dâhil olmalıdır.



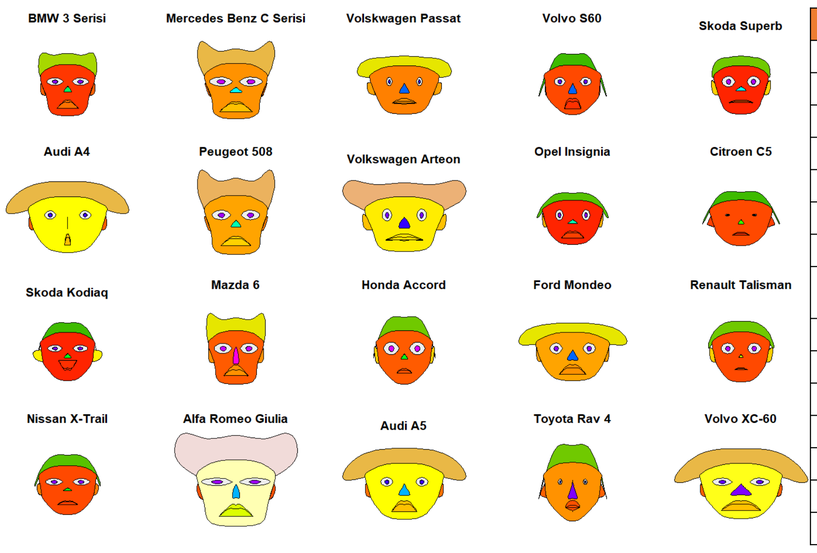
Grafikte, küme sayısı arttıkça toplam kare değerlerinin azaldığı ve bu düşüş hızının belirgin bir şekilde yavaşladığı bir eğilim görüldü. Grafikteki kırılma noktası, bu düşüş hızının önemli ölçüde yavaşladığı ve optimal küme sayısının belirlendiği noktayı temsil eder. Bu dirsek noktası, analizimizde kullanılacak optimal küme sayısını belirlememize yardımcı oldu. Sonuç olarak, veri setimiz için en uygun küme sayısının 3 olduğunu belirledik. Bu küme sayısının ötesinde, varyansın azalma hızının azaldığı ve artık belirgin bir iyileşme sağlanmadığı gözlemlendi.



Veri setindeki araba modelleri, k-means kümeleme algoritması kullanılarak farklı gruplara ayrıldı. Bu grupların oluşumu Şekil incelendiğinde, benzer özelliklere sahip araçların aynı kümeye yerleştirildiği görülüyor.

Mavi kümede, Peugeot 508, BMW 3 Serisi, Citroen C5 gibi orta sınıf sedan araçlar, Mercedes Benz C Serisi ve Nissan X-Trail gibi farklı tiplerdeki araçlarla birlikte yer alıyor. Kırmızı kümede, daha büyük boyutlu ve lüks sedan modelleri bulunurken, yeşil kümede ise daha kompakt ve pratik araçlar bulunuyor. Bu kümeler, araçların boyutu, performansı ve segmenti gibi özelliklere dayanarak oluşturulmuş gibi görünüyor.

Chernoff Yüzleri

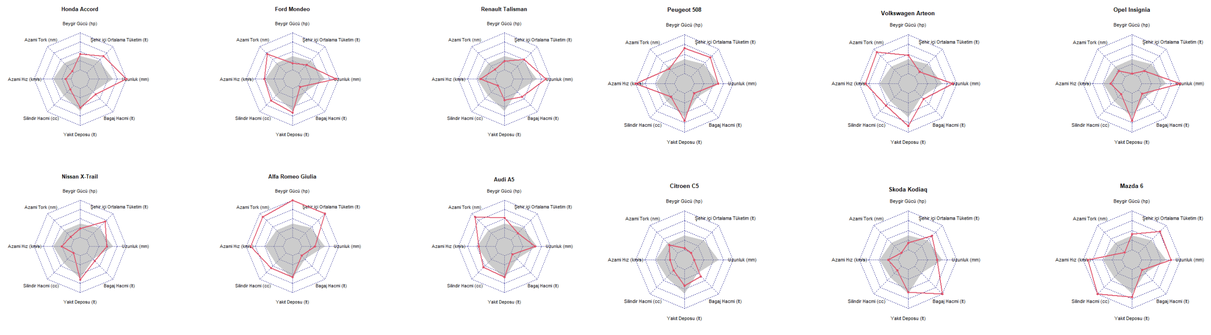


Parametre	Özellik
Yüz Uzunluğu	Beygir Gücü (hp)
Yüz Genişliği	Azami Tork (nm)
Yüz Şekli	Azami Hız (km/s)
Ağız Uzunluğu	Silindir Hacmi (cc)
Ağız Genişliği	Yakıt Deposu (lt)
Gülümseme	Bagaj Hacmi (lt)
Göz Uzunluğu	Uzunluk (mm)
Göz Genişliği	Şehir İçi Ortalama Tüketim (lt)
Saç Uzunluğu	Beygir Gücü (hp)
Saç Genişliği	Azami Tork (nm)
Saç Şekli	Azami Hız (km/s)
Burun Uzunluğu	Silindir Hacmi (cc)
Burun Genişşiği	Yakıt Deposu (lt)
Kulak Uzunluğu	Uzunluk (mm)
Kulak Genişliği	Bagaj Hacmi (lt)

Rstudio kullanarak oluşturduğumuz Chernoff yüzleri yukarıdaki gibi görselleştirilmiştir. Parametre sayısı özellik sayısından fazla olduğu için bazı özellikler birden fazla parametreye atanmıştır, bu yüzden yüzleri daha dikkatli bir şekilde incelememiz gerekir.

Yüzleri incelediğimizde araçların kendi içinde belli özelliklere göre ayrıldığını açıkça görebilmekteyiz. Örneğin, “BMW 3 Serisi”, “Mercedes Benz C Serisi”, “Peugeot 508”, “Mazda 6” ve “Alfa Romeo Giulia” birbirlerine yakın göz genişliğine sahiptir. Bu bize, bu araçların en çok şehir içi yakıt tüketimine sahip olduklarını gösterir. Eğer, en çok beygir gücüne sahip arabayı bulmak istiyorsak yüz uzunluğu ve saç uzunluğu parametrelerine bakmamız gerekir en uzun saça ve yüze “Alfa Romeo Giulia” modeli sahip olduğu için en çok beygir gücüne “Alfa Romeo Giulia” modeli sahiptir diyebiliriz.

Radar Grafiği



Radar grafiklerindeki her eksen, araba modellerinin farklı özelliklerini temsil eder. İncelemek, farklı araba modellerinin özelliklerindeki farklılıkları ve benzerlikleri hızla farkedilmesini sağlar. Yukaridaki şekillerde görülebildiği üzere güçlü araba modelleri Alfa Romeo Giulia ve Audi A5, yüksek beygir gücü (hp) ve azami tork (nm) değerleriyle öne çıkarken, daha ekonomik modeller Citroen C5 ve Renault Talisman daha düşük güç ve tork seviyelerine sahiptir.
Diğer özellikler, yakıt deposu (lt), bagaj hacmi (lt), uzunluk (mm) ve şehir içi ortalama tüketim (lt) gibi, her araba modelinin farklı kullanım avantajlarını ve dezavantajlarını yansıtır. Bu analiz, kullanıcıların ihtiyaçlarını ve tercihlerini belirleyerek en uygun araçları seçmelerine yardımcı olabilir. Özellikle, performans ve pratiklik kriterlerine önem veren kullanıcılar için radar grafikleri değerli bir kaynak olabilir.

KAYNAKÇA

[1] Uzun, E. (2007). İnternet Tabanlı Bilgi Erişimi Destekli Bir Otomatik Öğrenme Sistemi. Edirne: Trakya Üniversitesi Fen Bilimleri Enstitüsü, Doktora Tezi.

[2] Nizam, H. ve ark. (2014). Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması. XIX. Türkiye'de İnternet Konferansı.

[3] Hinton G. ve Sejnowski T.J. (1999). Hinton G. ve Sejnowski T.J. (editorler), 1999, "Unsupervised Learning and Map Formation: Foundations of Neural Computation", MIT Press, ISBN 0-262-58168-X. MIT Press, ISBN 0-262- 58168-X.

İSTATİSTİK BÖLÜMÜ

ARAÇ ÖZELLİKLERİ VE FİYAT TAHMİN ANALİZİ

Muhammed Yasir ÖZTÜRK 19023501

Danışman: Doç. Dr. Öyküm Esra YİĞİT

ÖZET

Bu çalışmada, otomotiv sektöründe araç özelliklerinin fiyatlandırma üzerindeki etkisi Random Forest ve Lineer Regresyon modelleri kullanılarak analiz edilmiştir. Araç özellikleri arasında motor gücü, yakıt tüketimi, motor boyutu gibi faktörler bulunmaktadır. Random Forest modeli, daha düşük RMSE değerleri ile Lineer Regresyon modeline kıyasla daha iyi performans göstermiştir. Bu sonuçlar, Random Forest modelinin araç fiyatlarını tahmin etmede daha doğru ve güvenilir olduğunu göstermektedir.

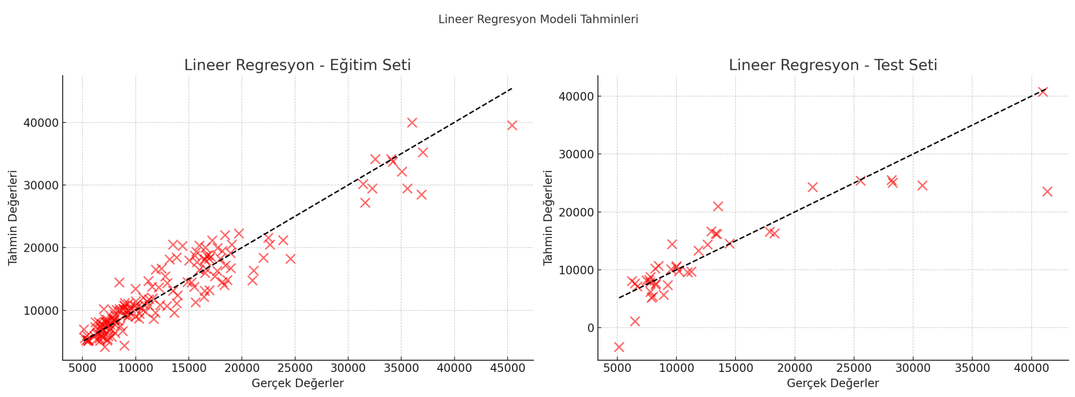
GİRİŞ

Otomotiv sektörü, araç fiyatlandırması için birçok değişkenin dikkate alındığı karmaşık bir alandır. Araçların fiyatlarını etkileyen çeşitli faktörler arasında motor gücü, yakıt tüketimi, motor boyutu gibi teknik özellikler bulunmaktadır. Bu çalışmada, araçların bu özelliklerinin fiyat üzerindeki etkisini incelemeyi amaçlamaktadır. Amacımız, otomobil alıcılarına ve satıcılarına araç özelliklerinin fiyat üzerindeki etkilerini daha iyi anlamalarına yardımcı olacak bir analiz sunmaktır. Ayrıca, bu çalışma, otomotiv sektöründe fiyatlandırma stratejilerini geliştirmeye ve piyasa dinamiklerini daha iyi kavramaya katkı sağlayabilir. Veri setimiz, çeşitli marka ve modeldeki araçların özelliklerini ve fiyatlarını içermektedir. Bu kapsamlı veri seti kullanılarak, araç özellikleri ile fiyatlar arasındaki ilişkileri belirlemek ve bu bilgileri kullanarak araç fiyatlarını tahmin edebilecek modeller geliştirmek hedeflenmiştir. Bu çalışmanın sonucunda, otomotiv sektöründe daha bilinçli kararlar alınmasına ve fiyatlandırma stratejilerinin optimize edilmesine katkıda bulunmasını amaçlamaktadır.

YÖNTEM

İki farklı model kullanılarak araç fiyat tahminleri yapılmıştır: Random Forest ve Lineer Regresyon. Veri seti eğitim ve test setlerine ayrılarak modellerin performansları değerlendirilmiştir. Performansların karşılaştırılma metrikleri olarak ise RMSE kullanılmıştır.

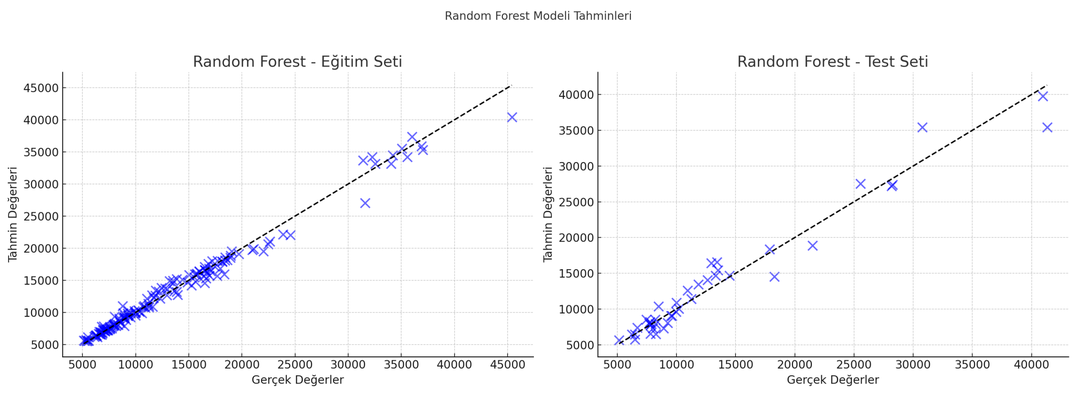
Lineer Regresyon Modeli Tahminleri



Eğitim Seti (Training Set):
Gerçek ve tahmin edilen fiyatlar karşılaştırılmıştır.
Siyah kesik çizgi mükemmel tahmin performansını (y=x) temsil eder.
Eğitim setinde model, genellikle doğru tahminler yapmaktadır.

Test Seti (Test Set):
Test setindeki tahmin performansı değerlendirilmiştir.
Test setinde, modelin eğitim setine göre daha fazla hata yaptığı görülmektedir.

Random Forest Modeli Tahminleri

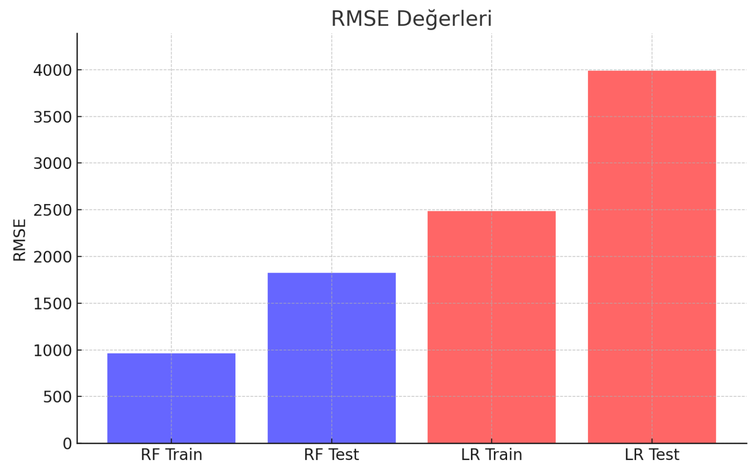


Eğitim Seti (Training Set):
Gerçek ve tahmin edilen fiyatlar karşılaştırılmıştır.
Siyah kesik çizgi mükemmel tahmin performansını (y=x) temsil eder.
Eğitim setinde model, doğru tahminler yapmaktadır.

Test Seti (Test Set):
Test setindeki tahmin performansı değerlendirilmiştir.
Test setinde, model genel olarak doğru tahminler yapmaktadır ve sapmalar minimaldir.

Model Performans Karşılaştırması

RMSE Değerleri (Kök Ortalama Kare Hatası):



Random Forest Modeli:
Eğitim Seti RMSE: 961.02
Test Seti RMSE: 1823.91

Lineer Regresyon Modeli:
Eğitim Seti RMSE: 2484.14
Test Seti RMSE: 3989.54

Karşılaştırma

Eğitim Seti Performansı:
Random Forest modeli, eğitim setinde Lineer Regresyon modeline kıyasla daha düşük bir RMSE değerine sahiptir (961.02 vs. 2484.14). Bu, Random Forest modelinin eğitim verileri üzerinde daha iyi tahminler yaptığını gösterir.

Test Seti Performansı:
Random Forest modeli, test setinde de Lineer Regresyon modeline kıyasla daha düşük bir RMSE değerine sahiptir (1823.91 vs. 3989.54). Bu, Random Forest modelinin yeni ve görülmemiş veriler üzerinde de daha iyi tahminler yaptığını ve daha genel geçer bir model olduğunu gösterir.

SONUÇ

Bu çalışmada, araç özelliklerinin fiyatlandırma üzerindeki etkisini değerlendirmek amacıyla Random Forest ve Lineer Regresyon modelleri kullanılmıştır. Elde edilen sonuçlar, Random Forest modelinin hem eğitim hem de test setlerinde Lineer Regresyon modeline kıyasla daha düşük RMSE (Kök Ortalama Kare Hatası) değerlerine sahip olduğunu göstermiştir. Bu bulgu, Random Forest modelinin araç fiyatlarını tahmin etmede daha doğru ve güvenilir olduğunu ortaya koymaktadır.

Bulguların Önemi

Bu çalışmanın bulguları, otomobil alıcıları ve satıcıları için önemli bilgiler sunmaktadır. Özellikle araç fiyatlandırmasında kullanılabilecek daha güvenilir ve doğru modellerin geliştirilmesi, piyasa dinamiklerini daha iyi anlamaya ve optimize etmeye yardımcı olabilir. Random Forest modelinin üstün performansı, bu modelin araç fiyat tahminlerinde tercih edilebilir bir yöntem olduğunu göstermektedir.

Uygulama Alanları

Otomotiv sektöründe fiyatlandırma stratejilerinin geliştirilmesi ve piyasa dinamiklerinin anlaşılması için bu çalışmanın sonuçları önemli bir katkı sağlamaktadır. Otomobil üreticileri, satıcılar ve alıcılar, bu bulgular doğrultusunda daha bilinçli kararlar alabilir ve fiyatlandırma stratejilerini optimize edebilirler.

KAYNAKÇA

[1] Kaggle. (2021). *Car Price Prediction Multiple Linear Regression Dataset*. Retrieved from https://www.kaggle.com/datasets/hellbuoy/car-price-prediction.
[2] Erdem Erzurumlu. (2021). Lineer Regresyon Nedir?. Retrieved from https://medium.com/@erdemerzurumlu/lineer-regresyon-560233d9c9e1.
[3] Data Science TR. (2021). *Makine Öğrenmesi Dersleri: Random Forest Regresyon*. Retrieved from https://medium.com/data-science-tr/makine-%C3%B6%C4%9Frenmesi-dersleri-5a-random-forest-regresyon-2a91715a8b66
.

Istatistik BÖILÜMÜ

BİTİRME ÇALIŞMASININ BAŞLIĞI

Noah Onur Yuce 19023020

Danışman: Dr. Selçuk ALP

ÖZET

Bu çalışma, Amazon'daki startup şirketlerin yeteneklerini artırmak için AI tahmin modellerini kullanmaktadır. Tanınmış markaların aksine, startup'lar hemen tanınabilir bir marka bilinirliğine sahip değildir ve Amazon'un anahtar kelime odaklı platformunda stratejik olarak hareket etmek zorundadır. Satışlar ve anahtar kelime özellikleri arasındaki ilişkileri analiz ederek, AI modellerimiz etkili anahtar kelimeleri tahmin etmelerini, seçimlerini optimize etmelerini, satışlarını artırmalarını ve Reklam Maliyetlerini (ACOS) azaltmalarını sağlayacak pratik bilgiler sunmaktadır. Bu bilgiler sayesinde, bütçelerini stratejik ve kârlı bir şekilde kullanmalarına yardımcı olmaktadır.

Leveraging Prediction Models to Optimize Amazon Keyword Efficiency for Startups"

Keyword Usage on Amazon:

On Amazon, a platform central to e-commerce, keywords are vital for product visibility and sales. These search terms directly connect customer searches to the relevant product listings. Startups, unlike established brands, cannot rely solely on brand recognition and must instead strategically use keywords to ensure their products are seen by potential customers.

Keyword Attributes: This research focuses on several key attributes of keywords:

1. Impressions: The number of times a keyword brings a product listing into view on a customer's screen.
2. Conversion Rate: The percentage of clicks that result in a purchase, reflecting the keyword's ability to drive sales.
3. Clicks: The total number of times customers have clicked on the product listing via the keyword.

These attributes are crucial as they directly influence the ability of a startup to engage with potential customers and drive sales on Amazon.

Importance of Keyword Attributes: Understanding which attributes most significantly affect sales and visibility can drastically improve a startup's strategic planning and budgeting for advertising. Effective keyword selection can lead to higher sales, better customer reach, and lower advertising costs, crucial for startups operating with limited marketing budgets.

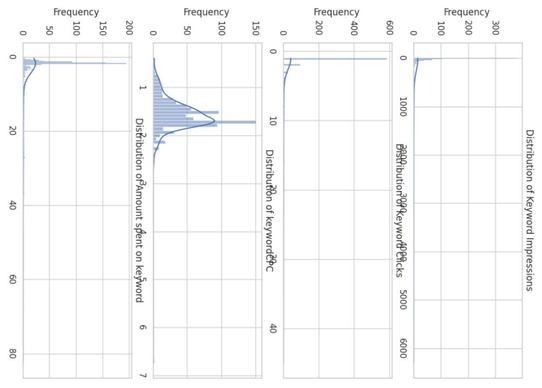
Prediction Model Implementation: To enhance the strategic selection of keywords, we employ prediction models capable of analyzing vast datasets to predict the impact of different keyword attributes on sales. The model identifies the importance of each attribute, allowing startups to prioritize keywords that offer the best return on investment. By integrating AI into keyword strategy, startups can optimize their advertising spend, lower their Advertising Cost of Sales (ACOS), and achieve better market penetration.

Data Source and Initial Checks:

This study utilized real transactional data from 'Fairway Finds,' a brand that experiences wide variability in keyword performance. An initial assessment of the data revealed a highly skewed distribution, which is characteristic of genuine keyword performance on Amazon.

Unlike conventional studies that often rely on normally distributed datasets, this research required embracing the natural skewness — where some keywords might receive over 100,000 clicks while others barely reach 100.

Similarly, conversion rates varied dramatically, with some keywords achieving a 100% conversion rate despite low impressions, and others showing high impressions but no conversions.



Selection of Prediction Models:

Given the skewed nature of the data, models traditionally used in normally distributed scenarios were unsuitable. Instead, this research focused on models that excel in handling skewed and non-linear data without assumptions of normality:

- ***Random Forest***: An ensemble learning method that builds multiple decision trees and merges them together to get a more accurate and stable prediction. It's robust against overfitting and is known for its high accuracy.

- ***XGBoost (Extreme Gradient Boosting)***: A decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework. It is particularly well-suited for large datasets and known for its execution speed and model performance.

- ***GBM (Gradient Boosting Machines)***: This model builds trees one at a time, where each new tree helps to correct errors made by previously trained trees. GBM has a strong predictive power but requires careful tuning of parameters.

Why These Models Are Optimal:

These models were chosen due to their ability to handle large variations in data effectively. Each is capable of managing outliers and skewed data, making them particularly advantageous for this dataset.

Furthermore, these models do not require the transformation of data to fit a normal distribution, thus preserving the originality and interpretability of the data. They also provide feature importance scores, which are crucial for understanding which attributes most significantly impact keyword performance.

Summary of Model Results

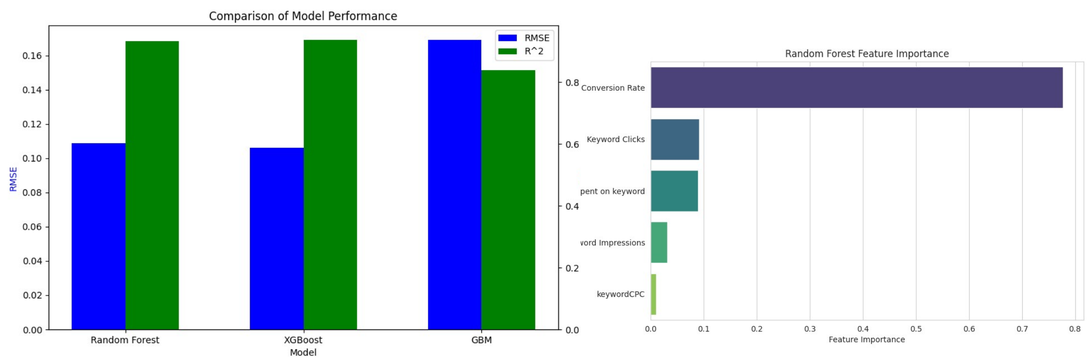
The analysis conducted using Random Forest, XGBoost, and GBM models has provided significant insights into keyword optimization on Amazon.

The Random Forest model demonstrated an RMSE of 0.1088 and an R^2 of 0.9329, indicating high accuracy and reliability. XGBoost followed closely with an RMSE of 0.1061 and an R^2 of 0.9362, offering slightly better efficiency in handling data variability. The GBM model, while effective, showed a higher RMSE of 0.1692 and a lower R^2 of 0.8378, suggesting some limitations in capturing the full complexity of the data compared to the other models.

Random Forest Feature Importance

The feature importance analysis from the Random Forest model highlighted several key factors that most significantly influence sales outcomes on Amazon:

- ***"Conversion Rate"*** emerged as the most crucial feature, underscoring its pivotal role in predicting successful sales transactions.
- "***Keyword Clicks***" and Amount ***Spent on Keyword***" also showed substantial influence, indicating their importance in driving visibility and engagement.
- Lesser but still notable impacts were observed from "***Keyword Impressions***" and "***Keyword CPC***", which help in understanding broader keyword effectiveness.



Practical Application of Insights

The insights gained from the feature importance can guide startups in allocating their advertising budgets more effectively:

- ***Utilizing Importance as Coefficients:***

The importance values derived from the Random Forest model can be used as coefficients to weigh the influence of each keyword attribute in a predictive model. This approach allows for a quantitatively informed method to select keywords that are likely to maximize sales and minimize costs.

- ***Optimizing Keyword Selection:***

By prioritizing keywords with higher conversion rates and appropriate cost-per-click values, startups can enhance their product visibility while keeping advertising spend efficient. This strategy ensures that each dollar spent on ads contributes positively to overall sales performance.

Conclusion

This research provides a data-driven pathway for startups on Amazon to optimize their keyword strategies, thereby enhancing their competitive edge in a crowded marketplace. By applying sophisticated Prediction modeling techniques, we can decode complex patterns and offer actionable strategies that drive business success.

KAYNAKÇA

[1]https://www.nvidia.com/en-us/glossary/xgboost/
[2]https://www.ibm.com/topics/random-forest
[3]https://rexven.com/courses/amazon-course

İSTATİSTİK BÖLÜMÜ

DİKKAT EKSİKLİĞİ HİPERAKTİVİTE BOZUKLUĞUNUN DEPRESYON, ANKSİYETE VE STRES ÖLÇEĞİNİN KULLANILMASIYLA YORDANMASI

Okan GÜNEŞ 17023051

Danışman: Prof.Dr Gülhayat GÖLBAŞI

ÖZET

Bu çalışma Dikkat Eksikliği Hiperaktivite Bozukluğunun (DEHB) Depresyon, Anksiyete, Stres Ölçeğinin kullanılmasıyla yordanmasını amaçlamaktadır. Çalışmada Stres, Depresyon ve Anksiyete Ölçeğinin kullanılmasıyla DEHB'li bireylerde cinsiyet değişkenine, yaş değişkenine, Anksiyete puanına, Stres puanına ve depresyon puanına göre anlamlı bir ilişkisinin bulunma durumunun istatistiksel olarak tespit edilmesini hedeflemektedir.

SPSS

MATERYAL VE METOT: Bu çalışmada yapılan regresyon analizi kapsamında iki yöntem kullanılmıştır. Bu analizler, Çoklu regresyon ve lojistik regresyon analizleridir. Bu analizler SPSS 29 programı kullanılarak gerçekleştirilmiştir.

BULGULAR

ADHD Total Değerine İlişkin Regresyon Analiz Bulguları

Değişkenler	B	Standart Hata	β	t	p	%95 Güven Aralığı (B)		VIF
						Alt	Üst	
Sabit	20,482	1,502		13,635	<0,001	17,533	23,431	
Cinsiyet	-0,532	0,748	-0,022	-0,712	0,477	-2,000	0,936	1,066
Yaş	-0,096	0,048	-0,061	-1,995	**0,046**	-0,191	-0,002	1,027
Depresyon	0,200	0,094	0,088	2,134	**0,033**	0,016	0,384	1,899
Anksiyete	0,710	0,128	0,241	5,562	**<0,001**	0,459	0,961	2,084
Stres	0,963	0,137	0,319	7,011	**<0,001**	0,693	1,233	2,304
R: 0,584	**R^2**: 0,341	**Düzeltilmiş R^2**: 0,337		**F**: 75,811		**p**: **<0,001**		
Bağımlı Değişken: ADHD Total								

- Regresyon modelinde yer alan değişkenlerin ADHD Total değerindeki varyansın %33,7'sini açıkladığı ve regresyon modelinin istatistiksel olarak anlamlı olduğu (F=75,811; p=<0,001),
- Yaşın ADHD Total değeri üzerinde negatif yönde anlamlı etkisinin olduğu (β=-0,061),
- Depresyon (β=0,088), Anksiyete (β=0,241) ve Stres (β=0,319) değişkenlerinin ADHD Total değeri üzerinde pozitif yönde anlamlı etkilerinin olduğu,
- Varyans şişme değerlerinin (Variance Inflation Factor-VIF) <10 olduğu ve bağımsız değişkenler arasında çoklu bağlantı sorunun olmadığı belirlenmiştir.

ADHD Total Gruplarına İlişkin Lojistik Regresyon Analiz Bulguları

Değişkenler	B	Standart Hata	p	Exp(B)	%95 Güven Aralığı (Exp(B))	
					Alt	Üst
Cinsiyet*	0,125	0,172	0,469	1,133	0,808	1,587
Yaş	-0,024	0,012	**0,036**	0,976	0,954	0,998
Depresyon	0,036	0,022	0,098	1,037	0,993	1,082
Anksiyete	0,124	0,030	**<0,001**	1,132	1,067	1,202
Stres	0,163	0,033	**<0,001**	1,177	1,104	1,256
Sabit	-1,565	0,362	<0,001	0,209	-	-
Nagelkerke R^2=0,306						
Bağımlı Değişken: ADHD Total (0 Grubu < Medyan; 1 Grubu ≥ Medyan; Medyan=30,00)						

*Referans grup = Kadın

- Regresyon modelinde yer alan değişkenlerin ADHD değerindeki varyansın %30,6'sını açıkladığı,
- ADHD değerinin >30 olma durumunu yaştaki 1 birimlik artışın 0,97 kat azalttığı,
- ADHD değerinin >30 olma durumunu 1 birimlik artışta Anksiyetenin 1,13 kat, Stresin 1,17 kat artırdığı belirlenmiştir.

RANDOM FOREST ALGORİTMASI VE SUPPORT VECTOR MACHİNE ALGORİTMASI

MATERYAL VE METOT: Keşifçi veri analizi ve sonrasında 2 adet makine öğrenimi algoritması uygulanmıştır. Keşifçi veri analizile aynı işleve inde, hedef değişkenin kendisinden üretildiği "ADHDTotal" değişkeni ve ayrıca veri setinin index`i sahip olan "SıraNo" değişkeni veri setinden çıkarılmıştır. Makine öğrenimi algoritmaları da çalışmanın bir sınıflandırma problemi olmasından dolayı Random Forest Classifier ve Support Vector Machine olmak üzere 2 adet makine öğrenimi algoritması kullanılıp, sonuçları paylaşılmıştır.

Random Forest Algoritması: Denetimli sınıflandırma algoritmalarından biridir. Hem regresyon hem de sınıflandırma problemlerinde kullanılmaktadır. Algoritma, birden fazla karar ağacı üreterek sınıflandırma işlemi esnasında sınıflandırma değerini yükseltmeyi hedefler.

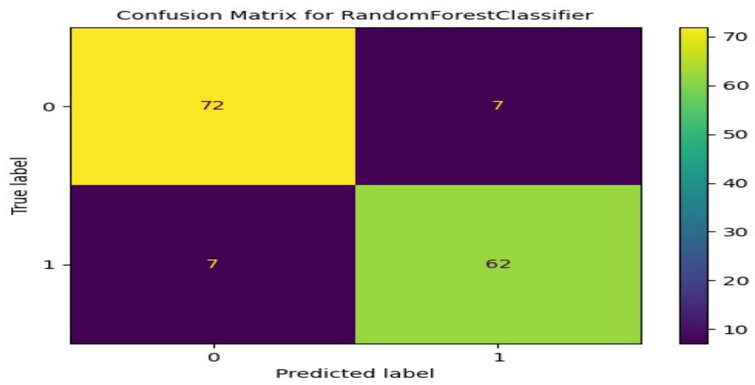
Random Forest Classifier

Classification Report:

RANDOMFORESTCLASSIFIER

	precision	recall	f1-score	support
0	0.91	0.91	0.91	79
1	0.90	0.90	0.90	69
accuracy			0.91	148
macro avg	0.90	0.90	0.90	148
weighted avg	0.91	0.91	0.91	148

Confusion Matrix:



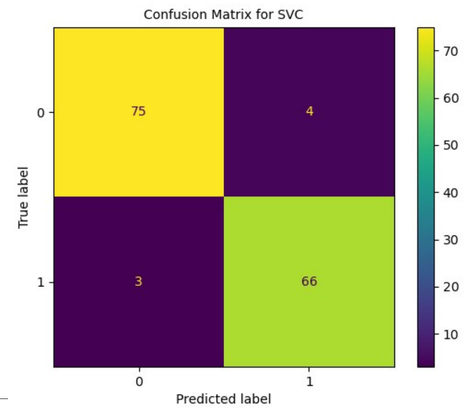
Support Vector Machine

Classification Report:

SVC

	precision	recall	f1-score	support
0	0.95	0.96	0.96	78
1	0.96	0.94	0.95	70
accuracy			0.95	148
macro avg	0.95	0.95	0.95	148
weighted avg	0.95	0.95	0.95	148

Confusion Matrix:



KAYNAKÇA

[1] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[3] https://scikit-learn.org/stable/modules/svm.html

İSTATİSTİK BÖLÜMÜ

TÜRKİYE'DE KONUTLARDAKİ ELEKTRİK TÜKETİMİNİN İLLER BAZINDA MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TAHMİN EDİLMESİ

Osman Nuri AKYOL 18023002

Prof. Dr. Ersoy ÖZ

ÖZET

Gelişen teknoloji, sanayileşme ve ekonomik dinamikler, dünya genelinde enerji talebini artırmaktadır. Bu artışın büyük bir kısmı elektrik enerjisi talebinden kaynaklanmaktadır. Elektriğe olan bağımlılık, günlük yaşam ve endüstriyel faaliyetler için kritik öneme sahiptir. Bu çalışmada, Türkiye'deki farklı illerdeki konutların gelecek aylardaki elektrik tüketimini makine öğrenmesi yöntemleriyle tahmin etmek amaçlanmaktadır. 2016-2023 yılları arasındaki il bazlı aylık tüketim verileri, meteorolojik, demografik ve ekonomik faktörler gibi değişkenlerle analiz edilecektir. Öznitelik mühendisliği teknikleri kullanılarak her ilin özelliklerini yansıtan tahmin modelleri oluşturulacaktır. SARIMA, Random Forest, LightGBM ve XGBoost gibi algoritmaların performansları karşılaştırılarak en uygun model belirlenecektir. Bu tahmin modeli, Türkiye'nin enerji üretim planlaması ve dağıtım şirketlerinin operasyonlarına katkı sağlayarak enerji kaynaklarının daha etkin kullanılmasına yardımcı olacaktır.

1.Kullanılan Modeller

1.1 SARIMA SARIMA, zaman serisi verilerinde trend, mevsimsellik ve otokorelasyon gibi özellikleri modellemek için kullanılan bir yöntemdir. Bu model, zaman serisinin bir sonraki adımını tahmin etmek için kullanılır ve performansı parametrelere ve veri setine bağlıdır.

1.2 Random Forest Random Forest, sınıflandırma ve regresyon problemlerinde kullanılan güçlü bir makine öğrenmesi algoritmasıdır. Çok sayıda karar ağacı oluşturup birleştirerek çalışır, aşırı uyumu azaltır ve yüksek doğruluk sağlar. Büyük ve karmaşık veri setlerinde etkilidir.

1.3 LightGBM LightGBM, hızlı ve yüksek performanslı bir makine öğrenmesi algoritmasıdır. Büyük ve yüksek boyutlu veri setleri için optimize edilmiştir. Karar ağaçları kullanarak çalışır ve hafızayı verimli kullanır. Hızlı hesaplama kapasitesiyle dikkat çeker.

1.4 XGBoost XGBoost, yüksek performanslı ve güçlü bir makine öğrenmesi algoritmasıdır. Karar ağaçlarını optimize ederek yüksek doğruluk ve genelleme yeteneği sağlar. Hız, performans ve paralel işlem yetenekleriyle büyük veri setlerinde etkilidir.

2. Veri Analizi

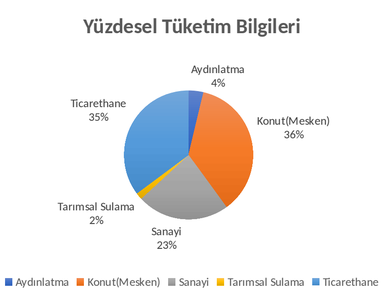
2.1Veri Kapsamı Bu çalışma, Türkiye'nin elektrik tüketiminin en yüksek olduğu İstanbul, Ankara, İzmir, Bursa ve Antalya illerini analiz etmektedir. Bu şehirler, Türkiye'nin toplam elektrik tüketiminin %43'ünü oluşturmaktadır ve ekonomik faaliyetlerin yoğun olduğu bölgelerdir.

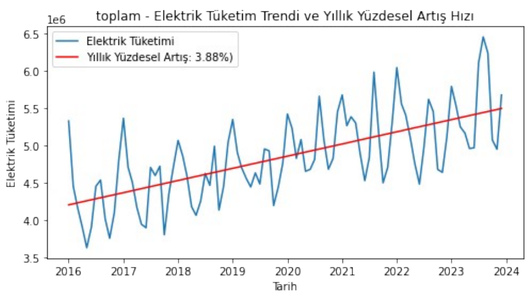
2.2.1 Değişkenler

1.MWh: İlin aylık elektrik tüketimi (MWh).
2.MWhTotal: Türkiye genelinde aylık toplam elektrik tüketimi.
3.HouseSales: İlin aylık konut satış sayısı.
4.Sıcaklık: İlin aylık ortalama sıcaklık değeri.
5.Gün: İlin aylık ortalama güneşlenme süresi.
6.Yağış: İlin aylık yağışlı gün sayısı.
7.Yüzde: İlin Türkiye genelindeki toplam elektrik tüketimi içindeki yüzdelik payı.

2.3 Veri Analizi Veri analizi Python programlama dili kullanılarak yapılmıştır. Veriler, her il için ayrı DataFrame'ler (df) olarak tanımlanmıştır. Pandas, numpy, matplotlib, seaborn ve sklearn kütüphaneleri kullanılarak veriler analiz edilmiş ve görselleştirilmiştir.

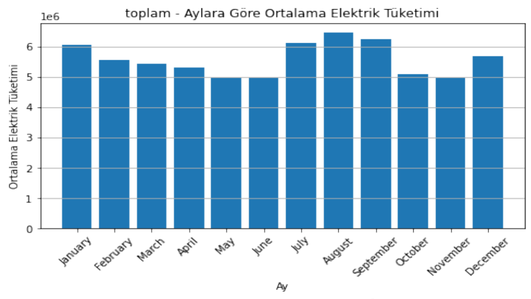
2.3.1 Türkiye'de Elektrik Tüketim Verileri 2023 Aralık ayı itibariyle Türkiye'deki toplam elektrik tüketimine profil abone grubu bazında bakıldığında, konutlardaki tüketimin %36 ile en yüksek paya sahip olduğu görülmektedir.

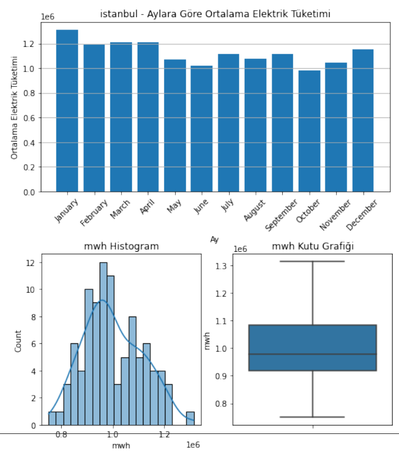




Yıllık ortalama artış hızına yüzdesel olarak bakıldığında ise yıllık yüzdesel artış değeri %3,88 olarak hesaplanmıştır.

Türkiye'nin aylık ortalama elektrik tüketimi, yıl boyunca bazı değişiklikler göstermektedir. Özellikle yaz aylarında (Haziran, Temmuz, Ağustos) ve kış aylarında (Aralık, Ocak, Şubat) tüketim artmaktadır. Bu artışlar, mevsimsel değişikliklerin ve hava koşullarının elektrik tüketimi üzerindeki etkisini yansıtmaktadır.





2.3.2 İller' de Elektrik Tüketim Verileri (İstanbul)

İstanbul'daki konutların elektrik tüketimi, Türkiye genelindeki ortalama konut tüketiminin yaklaşık %20'sine denk gelmektedir. Yıllık ortalama artış hızı %2,20 olarak hesaplanmıştır. İstanbul'da 2016-2023 yılları arasında aylık ortalama elektrik tüketimi 1 milyon megavat saat (MWh) civarındadır. Tüketim genel olarak sabit bir seyir izlerken, kış aylarında artış gözlemlenmektedir.



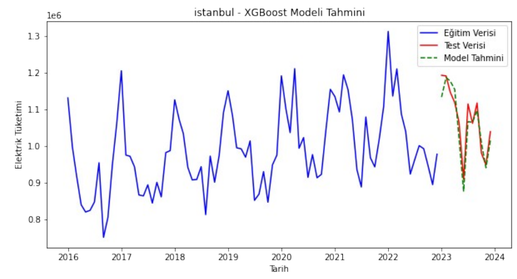
3. Uygulama

Bu çalışmada, İstanbul, Ankara, İzmir, Bursa ve Antalya şehirlerindeki konutların elektrik tüketimini tahmin etmek için SARIMA, Random Forest, LightGBM ve XGBoost modelleri kullanılmıştır. SARIMA modeli mevsimsel etkileri yakalamak, temel istatistik modellerinden olduğu ve yenilikçi makine öğrenmesi algoritmaları ile karşılaştırmak için tercih edilmiştir. Random Forest, LightGBM ve XGBoost ise yüksek doğruluk oranları ve yenilikçi algoritmalar olduğu için seçilmiştir.

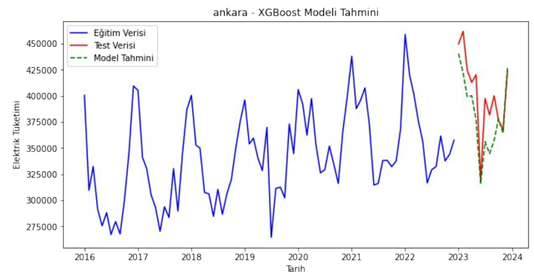
Bu modellerin performansı R^2, MAE ve RMSE metrikleri ile incelenmiş, tahmin edilen ve gerçek değerler arasındaki farklar ve yüzdesel farklar hesaplanmıştır. Elektrik tüketimi tahminleri için 84 aylık veri eğitim seti, son 12 aylık veri ise test seti olarak kullanılmıştır. Modellerin her ay yaptığı tahminler, gerçek değerlerle karşılaştırılmış ve GridSearchCV ile en iyi parametreler belirlenmiştir. Performans metrikleri ve yüzdesel farklar analiz edilerek modellerin etkinliği değerlendirilmiştir.

3.1 İllerde En İyi Performans Gösteren Modeller ve Grafikleri

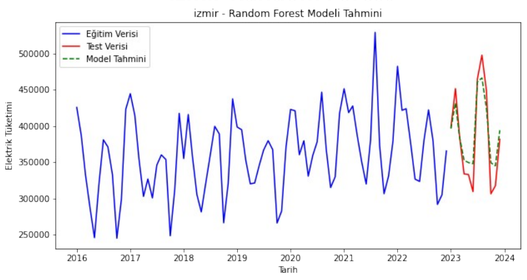
İstanbul-XGBoost



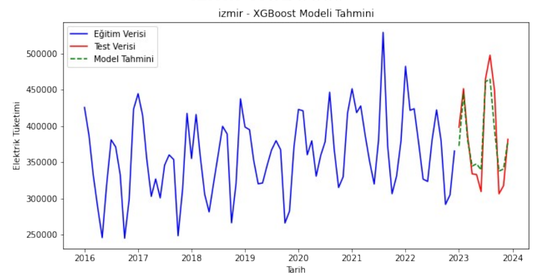
Ankara-XGBoost



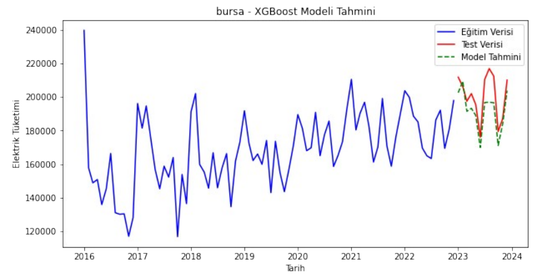
İzmir-Random Forest



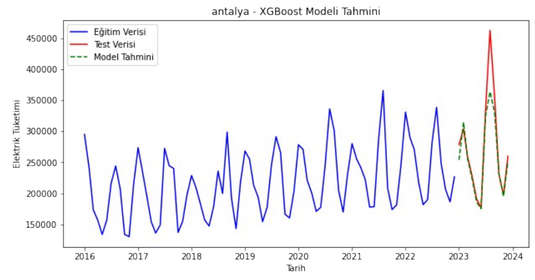
İzmir-XGBoost



Bursa-XGBoost



Antalya-XGBoost



4. Sonuç: İller Bazında Çıktıların Değerlendirilmesi İller için konutlardaki elektrik tüketim tahminlerinde kullanılan dört modelin performans metrikleri karşılaştırılmıştır. Bu modellerin performanslarını değerlendirmek için R^2 ,MAE (Ortalama Mutlak Hata), RMSE (Kök Ortalama Kare Hatası), toplam tahmin ve gerçek değerler arasındaki farklar gibi çeşitli performans ölçütleri incelenmiştir. Aşağıdaki tabloda ise modellerin illerde gösterdiği performans metrikleri sunulmuştur.

İl	Performans Metrikleri	SARIMA	Random Forest	LightGBM	XGBoost
İstanbul	R^2	0,32	0,67	0,58	0,86
	Toplam fark (%)	-7,43%	-2,47%	-1,98%	-1,12%
	Aylık Ortalama Fark (%)	7,42%	3,40%	4,07%	2,60%
Ankara	R^2	0,33	0,42	0,38	0,52
	Toplam fark (%)	-8,92%	-6,99%	-7,17%	-5,20%
	Aylık Ortalama Fark (%)	9,29%	7,47%	7,75%	5,33%
İzmir	R^2	0,46	0,86	0,6	0,85
	Toplam fark (%)	-4,02%	1,60%	-1,85%	-0,48%
	Aylık Ortalama Fark (%)	8,46%	5,57%	8,39%	5,39%
Bursa	R^2	0,32	0,38	0,4	0,54
	Toplam fark (%)	-0,36%	-4,64%	-4,09%	-4,19%
	Aylık Ortalama Fark (%)	4,48%	4,56%	5,95%	4,38%
Antalya	R^2	0,57	0,81	0,6	0,85
	Toplam fark (%)	-8,16%	-6,12%	-6,39%	-0,48%
	Aylık Ortalama Fark (%)	10,24%	5,00%	9,17%	5,39%

Genel Değerlendirme Analizler sonucu, beş ilde konutlardaki elektrik tüketim tahminlerinde en iyi performansı genel olarak XGBoost modelinin gösterdiği belirlenmiştir. XGBoost, yüksek R^2, düşük MAE ve RMSE değerleri, düşük toplam ve aylık fark yüzdeleri ile tutarlı ve doğru tahminler sunarak diğer modellere göre üstün performans sergilemiştir. İstanbul, Ankara ve Antalya'da XGBoost açık ara en başarılı model olmuştur. İzmir'de ise Random Forest ve XGBoost benzer performans göstermiştir. Bursa'da SARIMA modeli toplam tüketime en yakın sonucu vermiş olsa da, XGBoost genel performansta daha iyi gözükmektedir. Bu çalışma, XGBoost gibi yenilikçi modellerin elektrik tüketimi tahmininde başarılı olduğunu göstermiştir. Doğru tahminler, enerji kaynaklarının verimli kullanılmasına katkı sağlayacaktır. Farklı iller, zaman aralıkları ve değişkenlerle bu modellerin performansının daha geniş kapsamlı incelenmesi, daha iyi model performanslarının elde edilmesine katkı sağlayabilir.

KAYNAKÇA

[1] Karaca, Coşkun ve Hacer KARACAN. "Çoklu Regresyon Metoduyla Elektrik Tüketim Talebini Etkileyen Faktörlerin İncelenmesi." Selçuk Üniversitesi Mühendislik, Bilim ve Teknoloji Dergisi 4.3 (2016): 182-195.
[2] Bayar, Ecem. "Diferansiyel Polinom Sinir Ağı ile Elektrik Tüketim Tahmini". MS thesis. Sakarya Üniversitesi, 2019.
[3] Aydın, Beytiye Derya. "Meteorolojik Değişkenlerin Elektrik Tüketimi Üzerindeki Etkisi". Fen Bilimleri Enstitüsü, 2017.

İSTATİSTİK BÖLÜMÜ

NASA TURBO FAN JET MOTORLARI ARIZA TAHMİNİ İLE UYGUN İSTATİSTİKSEL MODEL SEÇİMİ

Öykü ARDA 19023072

Danışman: DOÇ. DR. Öyküm Esra YİĞİT

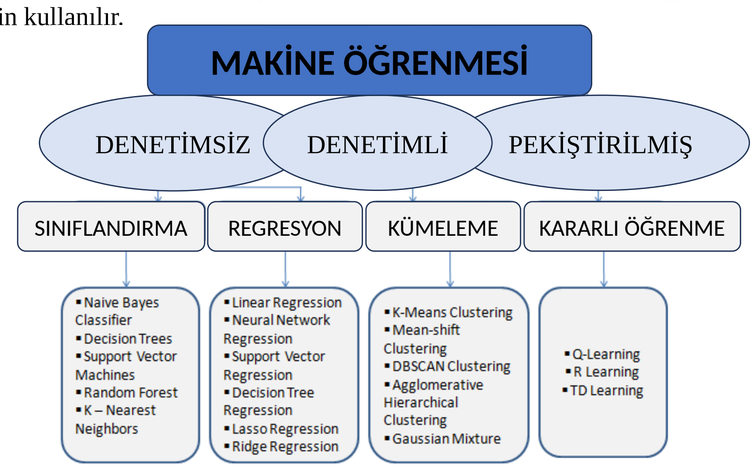
ÖZET

NASA Turbo fan Jet Motorları birçok farklı yöntem ile arıza tahminine konu olmuştur. Makine öğrenmesi, farklı bilgisayar yazılım programları ile istatistiksel modeller kurulmuş ve araştırmalar yapılmıştır. Bu çalışmanın temel amacı, NASA'nın Turbo Fan Jet motorlarındaki arızaları tahmin etmek için uygun istatistiksel modellerin seçilmesi uygulamasıdır. Python; Jüpiter Notebook üzerinde yapılan bu çalışma motor bozulmaları hakkında önceden elde edilen tahminlerin hangi model ile en iyi açıklandığı gösterilmiştir.

MAKİNE ÖĞRENMESİ

Makine öğrenimi, bilgisayarların belirli bir alanda (veri analizi, görüntü tanıma, doğal dil işleme vb.) insan benzeri yetenekler kazanmasını sağlar. Makine öğrenimi, çeşitli endüstrilerde (sağlık, finans, ulaşım, pazarlama vb.) veri analizi, tahmin, sınıflandırma ve öneri sistemleri gibi birçok uygulama alanında kullanılır. Bu teknoloji, büyük miktarda veriyi anlamak ve değerlendirmek için güçlü bir araç sağlar, böylece bilgiye dayalı kararlar alınabilir ve yeni fırsatlar keşfedilebilir.

Makine öğrenmesi, veriyi kullanarak modeller oluşturarak tahminler veya sınıflandırmalar yapmayı amaçlayan bir alandır. Bu alanda iki ana yöntem olan denetimli ve denetimsiz öğrenme, farklı veri yapıları ve problemler için kullanılır.



Bu çalışmada Makine öğrenimi metotlarından Denetimli Makine Öğrenmesi ; Lineer Regresyon, Lasso Regresyon, Ridge Regresyon, Karar Ağaçları, Rastgele Orman, Destek Vektör Regresyon, Gradient Boosting kullanılmıştır.

Denetimli Makine Öğrenmesi

Denetimli makine öğrenmesi, etiketli veri kullanarak model eğitme sürecidir. Bu yöntemde, her eğitim verisi örneği, bir girdi (bağımsız değişken) ve bir çıktı (bağımlı değişken) içerir. Amaç girdiler ve çıktılar arasındaki ilişkiyi öğrenerek, yeni girdiler için doğru çıktıları tahmin edebilen bir model oluşturmaktır.

Regresyon

- Doğrusal Regresyon: Sürekli hedef değişkenlerin tahmini için kullanılır.
- Polinomsal Regresyon: Doğrusal olmayan veri ilişkilerini modellemek için kullanılır.
- Destek Vektör Regresyonu (SVR): Veriyi daha karmaşık şekillerde ayırarak doğrusal olmayan regresyon yapar.
- Ridge ve Lasso Regresyonu: Doğrusal regresyonun özel türleridir ve aşırı uyumlamayı (overfitting) önlemek için düzenleme (regularization) kullanır.

Sınıflandırma (Classification)

- Lojistik Regresyon: İkili veya çok sınıflı sınıflandırma problemlerinde kullanılır.
- K-En Yakın Komşu (KNN): Verilerin benzerliklerine dayalı sınıflandırma yapar.
- Destek Vektör Makineleri (SVM): Veriyi en iyi ayıran düzlemi bulur.
- Karar Ağaçları: Veriyi karar düğümleri üzerinden sınıflandırır.
- Rassal Ormanlar: Birden fazla karar ağacının ortalamasını alarak daha doğru tahminler yapar.
- Yapay Sinir Ağları: Özellikle karmaşık ve büyük veri setlerinde kullanılır**.**

ÖZELLİK SEÇİMİ

Özellik seçimi (Feature selection), makine öğrenmesi modellerinde veri setindeki gereksiz veya alakasız özellikleri (değişkenleri) çıkararak modelin performansını artırma sürecidir. Bu süreç, modelin daha hızlı ve daha etkili olmasını sağlar, aşırı uyum (overfitting) riskini azaltır ve daha iyi genelleştirme yeteneklerine sahip modeller oluşturur.

Bu çalışmada Sarmalayıcı Özellik Seçimi yöntemlerinde Biri olan RFE(Yinelemeli Özellik Eleme) yöntemi ile eğitim ve test verilerinde kullanılacak değişkenler belirlenmiştir.



NASA TURBO FAN VERİ SETİ(CMAPPS)

Makinelerin veya teknolojik aletlerin sağlık yönetimi, kesinti ve arızalardan kaçınmak amacıyla varlıkların durumunu tahmin etmek endüstride önemli bir konudur. Bu veri seti, NASA'nın varlık bozulma modellemesi için çok iyi bilinen halka açık veri setidir. Turbo fan jet motorlarında arızaya kadar çalıştırılmış ve simüle edilmiş verileri içerir. Motor bozulma simülasyonu C-MAPSS kullanılarak gerçekleştirilmiştir. Dört farklı set, farklı çalışma koşulları ve arıza modları kombinasyonları altında simüle edilmiştir. Arıza gelişimini karakterize etmek için çeşitli sensör kanallarını kaydeder. FD001, FD002, FD003, FD004 diye adlandırılan veri setleri ayrı kaydedilmiştir. Dört veri setinde motorlara ait bilgi ve çalışma koşullarına göre değişen farklı belli parametreler vardır.

	unit_nr	time_cycles	setting_1	setting_2	setting_3	s_1	s_2	s_3	s_4	s_5	...	s_12	s_13	s_14	s_15	s_16	s_17	s_18	s_19	s_20	s_21
0	1	1	-0.0007	-0.0004	100.0	518.67	641.82	1589.70	1400.60	14.62	...	521.66	2388.02	8138.62	8.4195	0.03	392	2388	100.0	39.06	23.4190
1	1	2	0.0019	-0.0003	100.0	518.67	642.15	1591.82	1403.14	14.62	...	522.28	2388.07	8131.49	8.4318	0.03	392	2388	100.0	39.00	23.4236
2	1	3	-0.0043	0.0003	100.0	518.67	642.35	1587.99	1404.20	14.62	...	522.42	2388.03	8133.23	8.4178	0.03	390	2388	100.0	38.95	23.3442
3	1	4	0.0007	0.0000	100.0	518.67	642.35	1582.79	1401.87	14.62	...	522.86	2388.08	8133.83	8.3682	0.03	392	2388	100.0	38.88	23.3739
4	1	5	-0.0019	-0.0002	100.0	518.67	642.37	1582.85	1406.22	14.62	...	522.19	2388.04	8133.80	8.4294	0.03	393	2388	100.0	38.90	23.4044

5 rows × 26 columns

FD001 adlı ilk veri setinin ilk beş satırında görüldüğü üzere;

- **unit_nr**: Jet motoru ünitesinin numarası.
- **time_cycles**: Zaman döngüsü numarası, yani belirli bir motorda verinin alındığı döngü sayısı.
- **setting_1, setting_2, setting_3**: Motordaki üç farklı ayarın değerleri.
- **s_1, s_2, ..., s_21**: Motorun çeşitli sensörlerinden alınan 21 farklı sensör verisi olarak tanımlanmıştır.

Numara	Sembol	Tanım	Birim
1	T2	Fan girişindeki toplam ısı	ºR
2	T24	Düşük basınçlı ek motorun çıkışındaki toplam ısı	ºR
3	T30	Yüksek basınçlı ek motorun çıkışındaki toplam ısı	ºR
4	T50	Düşük basınçlı türbin çıkışındaki toplam ısı	ºR
5	P2	Fan girişindeki basınç	psia
6	P15	Harici baypas borusunun toplam basıncı	psia
7	P30	Yüksek basınçlı ek motorun çıkışındaki toplam basınç	psia
8	Nf	Fanın fiziksel hızı	rpm
9	Nc	Çekirdek motorun fiziksel hızı	rpm
10	epr	Motor basınç oranı (P50/P2)	-
11	Ps30	Yüksek basınç ek motorundaki statik basınç	psia
12	Phi	Yakıt akış oranı ve P30	pps/psi
13	NRf	Fanın düzeltilmiş hızı	ppm
14	NRc	Çekirdek motorun düzeltilmiş hızı	rpm
15	BPR	Baypas oranı	-
16	farB	Ateşleme odasının gaz-hava oranı	-
17	htBleed	Hava tahliye dağılımı	-
18	Nf_dmd	Sabitlenmiş fan hızı	rpm
19	PCNfR_dmd	Çekirdek motorun sabit düzeltilmiş hızı	rpm
20	W31	Yüksek basınçlı türbinin soğutucu hava tahliye akışı	lbm/s
21	W32	Düşük basınçlı türbinin soğutucu hava tahliye akışı	lbm/s

Farklı çalışma koşullarını gösteren sensörler bu sensörlerin her biri sıcaklık, fan, basınç gibi değişkenleri gösterir. Bazı değişkenler motor bozulmasına etkisi olmadığından temizlenmiştir. Özellik seçimi olan RFE yöntemi ile her veri seti için farklı değişkenler seçilmiştir. Eğitim ve test verilerine ayrılan her veri seti için gerekli hiper parametre optimizasyonu ile Doğrusal Regresyon, Lasso Regresyon Ridge Regresyon, Karar Ağaçları, Rastgele Orman, Destek Vektör Regresyonu, Gradient Boosting modellerinin sonuçlarını bulunmuştur.

SONUÇ

Sonuç olarak FD001, FD002, FD003, FD004 veri setlerine yapılan analizler ve kurulan model sonuçlarına göre RMSE aşağıdaki tablodaki değerler olarak bulunmuştur.

FD001 için MODEL	**RMSE**
Doğrusal Regresyon	76.05
Lasso Regresyon	75.35
Ridge Regresyon	76.05
Karar Ağaçları	73.97
Random Forest	75.99
Destek Vektör Regresyon	**72.28**
Gradien Boosting	75.89

FD002 için MODEL	**RMSE**
Doğrusal Regresyon	61.91
Lasso Regresyon	60.75
Ridge Regresyon	61.91
Karar Ağaçları	70.45
Random Forest	62.62
Destek Vektör Regresyon	**60.00**
Gradien Boosting	62.61

FD003 için MODEL	**RMSE**
Doğrusal Regresyon	113.41
Lasso Regresyon	112.66
Ridge Regresyon	113.41
Karar Ağaçları	**92.74**
Random Forest	117.07
Destek Vektör Regresyon	109.86
Gradien Boosting	114.81

FD004 için MODEL	**RMSE**
Doğrusal Regresyon	79.18
Lasso Regresyon	78.50
Ridge Regresyon	79.18
Karar Ağaçları	**70.27**
Random Forest	86.19
Destek Vektör Regresyon	73.98
Gradien Boosting	83.14

Uygulamanın temel amacı, NASA Turbo Fan jet motorlarında arıza tahmini için uygun istatistiksel modellerin seçilmesidir. Makine öğrenimi çeşitleri, öznitelik seçimi ve hiper parametre optimizasyonu gibi temel kavramlar detaylı bir şekilde ele alınmış ve uygulanmıştır.

- FD001 için arıza tahmini problemini çözmek ve motor arızalarını önceden tespit etmek için uygulanabilecek en uygun yöntem, diğerlerinden düşük hata oranına sahip yüzde **72,28** ile **Destek Vektör Regresyonu** modelidir.
- Diğer veri seti FD002 için en uygun seçilip arıza tahmininde birçok sorunu önleyemeye yardımcı olan ve tahminlemeleri diğerlerine göre daha isabetli yapan model yüzde **60,00** ile **Destek Vektör Regresyon** olarak bulunmuştur.
- FD003 veri setinin arıza tahmini başarısı en düşük RMSE değeri ve diğerlerine göre en uygun olan model yüzde **92,74** ile **Karar Ağaçları** modeli olmuştur.
- Son veri seti FD004 veri seti için tahminleme başarısı diğer modellerle karşılaştırıldığında en uygun istatistiksel model yüzde **70,27** ile **Karar Ağaçları** olarak sonuçlanmıştır.

KAYNAKÇA

[1] https://www.turhost.com/blog/makine-ogrenmesi-machine-learning-nedir/
[2]https://www.linkedin.com/pulse/makine-%C3%B6%C4%9Frenimi-y%C3%B6ntemleri-ve-dikkat-edilmesi-gereken-orhan-cansu/
[3] https://www.kaggle.com/datasets/behrad3d/nasa-cmaps
[4] https://www.analyticsvidhya.com/blog/2023/05/recursive-feature-elimination/

Amazon Çok Satan 100 Kitabın Veri Kazıma Yöntemiyle Elde Edilmesi ve Python ile Analizi

Rana ÇÖZEL 21023610

Danışman: Dr. Öğr. Üyesi Doğan Yıldız

ÖZET

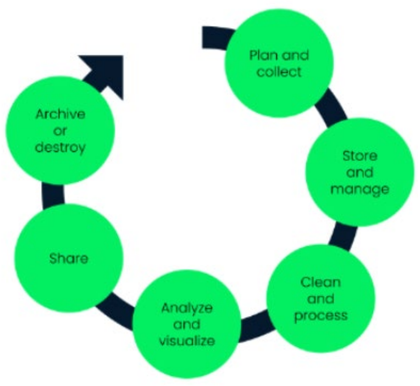
Bu çalışmada, Amazon'un çok satan kitaplar sekmesinde bulunan kitapları veri kazıma yöntemi ile elde edip, bu verileri Python kullanarak detaylı analizler yapılıp tüketicilerin tercih ettiği kitaplar hakkında analizler gerçekleştirmeyi amaçlamaktadır. Çalışmanın odak noktası, çeşitli faktörlerin (örneğin, kitap türü, yazarın popülaritesi, yayın tarihi, kullanıcı puanı vb.) kitap satışları üzerindeki etkisini anlamaktır. Veri analizi yardımı ile kitapların satış performansını etkileyen nedenler araştırılır. Öncelikle veri toplarken kullanılan Python kütüphaneleri hakkında ve yazılan koda hakkında bilgi verilir. Ardından çekilen veri setlerini ve yapılan analizler detaylıca açıklanır. Analizler, bu değişkenlerin kitap satışları üzerindeki etkisini istatistiksel olarak değerlendirir ve önemli bulguları ortaya koyar. Bu çalışmanın sonuçları kitap endüstrisindeki trendleri anlamak için değerli bir bakış sunar ve yayıncılar, yazarlar ve pazarlamacılar için stratejik kararlar alırken kullanılabilecek pratik bilgiler sağlar.

Veri nedir?

Veri, işlenmemiş bilgiye verilen addır. Araştırmalar, deneyler, gözlemler ve internet gibi kaynaklardan elde edilen her türlü değer veri olarak tanımlanabilir. Veriler tek başına hiçbir işlevi veya anlamı yoktur. Genel olarak, veri nicel ve nitel kategorilere ayrılır ve birincil veya ikincil olabilir. Nitel veriler metin tabanlıyken, nicel veriler sayısaldır.



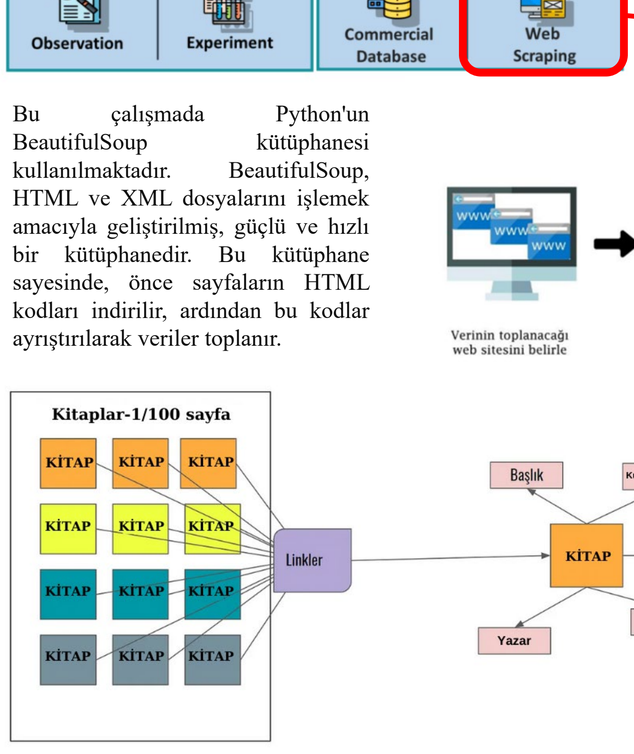




Veri İşleme ve Kullanımı

Veriler, her zaman ham haliyle anlamlı olmayabilir bu nedenle bilgiye dönüştürülmek üzere işlenir. Bu işlem çeşitli adımlar içerebilir:

1) Veri Toplama: İlgili verilerin toplanması.
2) Veri Temizleme: Eksik veya hatalı veriler düzeltilir.
3) Veri Analizi: Çeşitli yöntemlerle verilerin incelenmesi
4) Bilgi Elde Etme**:** Analiz sonuçlarının yorumlanması yoluyla anlamlı bilgi elde etmek.



Veri kazıma (web scraping), web sitelerinden otomatik yöntemlerle veri toplama işlemidir. Bu yöntem, kullanıcıların manuel olarak toplaması zor veya zaman alıcı bilgileri otomatik olarak elde etmelerini sağlar. Veri kazıma, farklı yazılım araçları ve programlama dilleri kullanılarak yapılır.

Bu çalışmada Python'un BeautifulSoup kütüphanesi kullanılmaktadır. BeautifulSoup, HTML ve XML dosyalarını işlemek amacıyla geliştirilmiş, güçlü ve hızlı bir kütüphanedir. Bu kütüphane sayesinde, önce sayfaların HTML kodları indirilir, ardından bu kodlar ayrıştırılarak veriler toplanır.



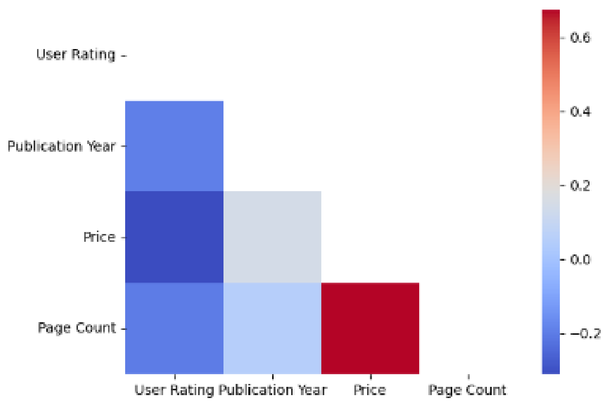
Çalışmada görüleceği üzere, öncelikle tüm kategorilerdeki linkler toplanır ve ardından bu linklere tek tek girilerek istenen veriler elde edilir. Stratejimizin şeması yandaki görselde gösterilmiştir.

KAYNAKÇA

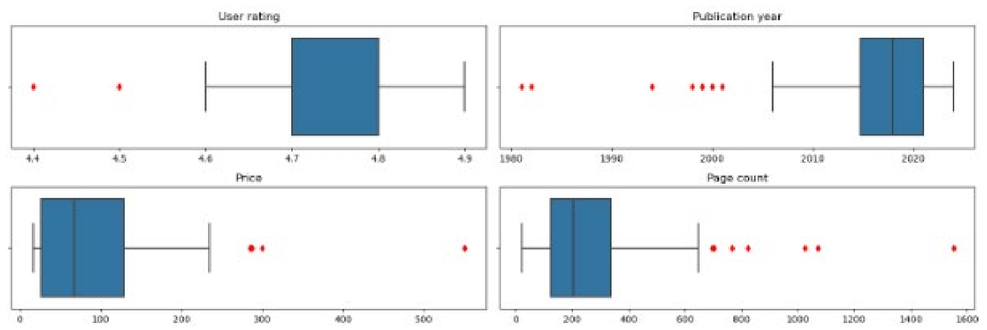
[1] FDEZRİVEROLA, F. (2014). Web scraping technologies in an API world, Briefings in Bioinformatics, Volume 15, Issue 5, September, Pages 788–797
[2] GLEZ-PEÑA, D., LOURENÇO, A., LÓPEZ-FERNÁNDEZ, H., REBOİRO-JATO, M.,
[3] Gürsakal, N. (2014). Büyük veri. Dora.

VERİ ANALİZİ

Bu kısımda veri kazıma yolu ile elde edilen veriler analiz edilerek bir takım çıktılar elde edilir. Bu araştırma sonucunda elde edilen bazı çıktılar aşağıdaki gibidir.

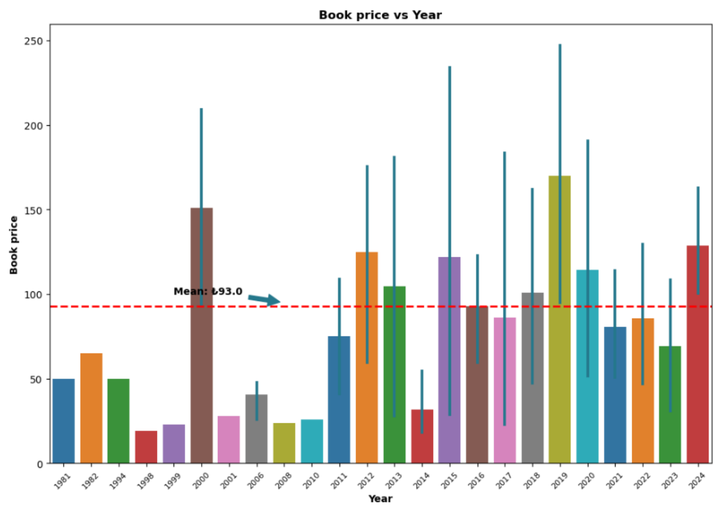


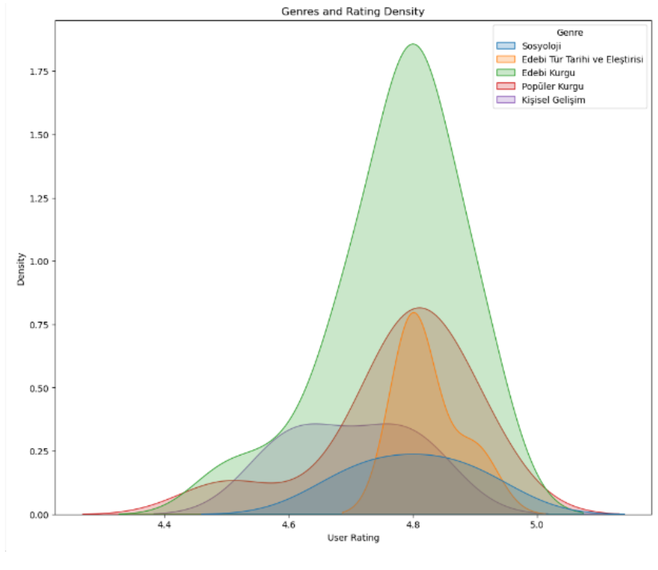
Yandaki çıktı bir ısı haritasıdır. Bu ısı haritası; Kullanıcı Puanı, Yayın Yılı, Fiyat ve Sayfa Sayısı olmak üzere dört değişken arasındaki korelasyonları görselleştirmektedir. Her bir hücre, iki değişken arasındaki korelasyon katsayısını temsil ederken, renk skalası korelasyonun pozitif mi yoksa negatif mi olduğunu göstermektedir. Renkler kırmızıdan maviye doğru bir geçiş sergiler; kırmızı tonlar pozitif korelasyonu, mavi tonlar negatif korelasyonu ve beyaz tonlar ise zayıf veya nötr korelasyonu belirtir.



Yukarıdaki çıktılar kutu grafiklerini temsil etmektedir. Kutu grafikleri (box plots), dört farklı değişkenin dağılımını ve uç değerlerini göstermektedir: Kullanıcı Puanı, Yayın Yılı, Fiyat ve Sayfa Sayısı. Her bir grafik, değişkenlerin merkezi eğilimlerini, dağılımlarını ve uç değerlerini görselleştirir.

Bu grafikte, farklı yıllar için kitap fiyatları gösterilmektedir. Yatay eksen yılları, dikey eksen ise kitap fiyatlarını temsil etmektedir. Ayrıca her bir yılın ortalama kitap fiyatını belirten bir kırmızı çizgi de eklenmiştir. Bu, tüm yıllar için ortalama kitap fiyatının 93 olduğunu gösterir. grafik kitap fiyatlarının yıllar boyunca tutarsız bir şekilde değiştiğini ve bazı yıllarda önemli dalgalanmalar gösterdiğini ortaya koymaktadır.





Bu yoğunluk grafiği, farklı türlerdeki kitapların kullanıcı puanlarının dağılımlarını göstermektedir. Edebi Kurgu türünün yoğunluk grafiği en yüksek tepeye sahip, yani bu türdeki kitaplar yüksek kullanıcı puanları almaya daha meyillidir. Sosyoloji türünün ise nispeten daha düşük puanlar aldığı gözlemlenebilir.

İSTATİSTİK BÖLÜMÜ

AMAZON SATICILARININ EN İYİ SATICI SIRALAMASI TAHMİNİ

Salih Arda MERMER - 19023012

Danışman: Doç. Dr. Selçuk ALP

ÖZET

Bu tezde e-ticaret sektörünün en popular platformlarından biri olan Amazon üzerinden, ürünlerin verileri, sektörde oldukça yaygın olarak kullanılan Helium10 uygulaması ile rastgele bir şekilde seçilmiştir. Seçilen bu veriler python programlama dili kullanılarak analiz edilmiş ve analiz sürecinde çeşitli regresyon modelleri uygulanarak satıcıların, ürün kategorisi içerisindeki başarısını gösteren "en iyi satıcı sıralaması" verisinin basit bir tahmini yapılmıştır.

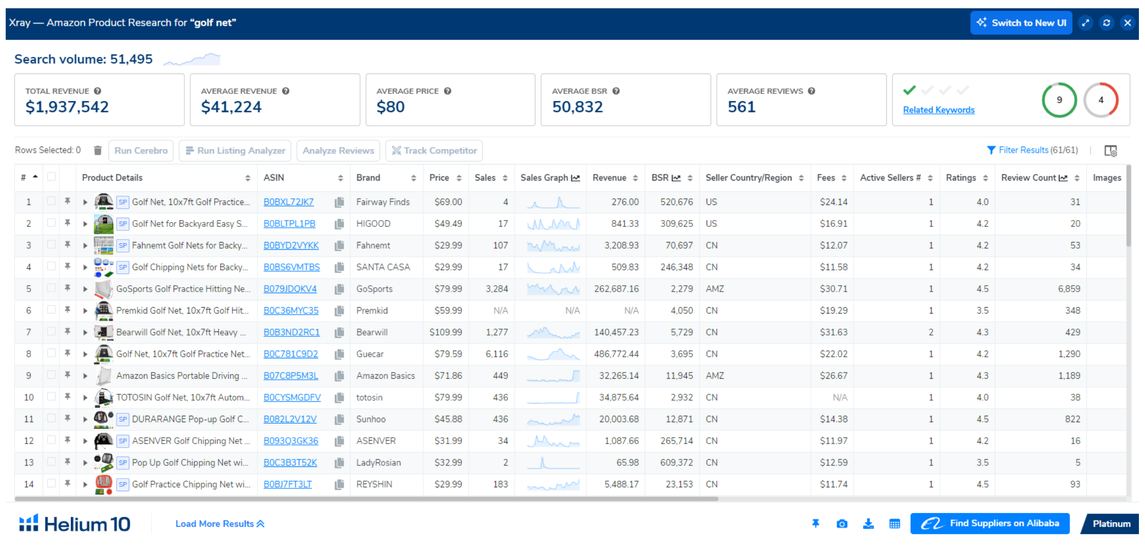
AMAZON EN İYİ SATICI SIRALAMASI NEDİR?

Amazon En İyi Satıcı Sıralaması (Best Seller Rank, kısaca BSR), Amazon'da satılan ürünlere atanan ve aynı kategori veya alt kategorideki diğer ürünlere göre popülerlik ve satış performansını gösteren sayısal bir sıralamadır. BSR, bir ürünün son dönem ve geçmişteki satış performansının bir ölçüsüdür. Daha düşük bir BSR numarası, daha yüksek bir satış sıralaması anlamına gelir ve bu da daha iyi satışları gösterir. Sıralama, ürünün ait olduğu kategori veya alt kategoriye özgüdür ve bir ürün birden fazla kategoride listeleniyorsa farklı BSR'lere sahip olabilir. Amazon, BSR'yi genellikle saatlik olarak günceller, bu da sıralamanın gün içinde satışlara bağlı olarak dalgalanabileceği anlamına gelir. BSR'nin hesaplanmasında kullanılan kesin algoritma gizlidir, ancak son dönemdeki satış sayısını ve bu satışların hızını dikkate aldığı genel olarak kabul edilmektedir. Ani satış artışları, bir ürünün BSR'sini önemli ölçüde iyileştirebilir. Yüksek bir BSR (düşük sayısal değer), bir ürünün Amazon'da görünürlüğünü artırabilir, çünkü en çok satanlar listelerinde ve arama sonuçlarında daha belirgin hale gelebilir. BSR, satılan birim sayısına dayandığından, daha düşük fiyatlı ve yüksek satış hacmine sahip ürünler, daha yüksek fiyatlı ve daha az satan ürünlere göre daha iyi BSR'lere sahip olabilir. BSR'yi anlamak ve izlemek, satıcıların ürünlerinin performansını ve pazarlama stratejilerinin etkinliğini değerlendirmelerine yardımcı olabilir. Ancak, bu metrik birçok faktörden sadece biridir ve müşteri yorumları, fiyat ve genel piyasa trendleri gibi diğer faktörlerle birlikte değerlendirilmelidir.

HELİUM10 NEDİR?

Helium 10, Amazon satıcılarına yönelik kapsamlı bir yazılım araçları setidir. Bu araçlar, ürün araştırmasından anahtar kelime optimizasyonuna, listeleme analitiğinden envanter yönetimine kadar geniş bir yelpazede hizmetler sunar. Helium 10, satıcıların rekabetçi pazarda başarılı olmalarına yardımcı olmak amacıyla tasarlanmıştır. Araçlar arasında Black Box, Xray, Cerebro ve Magnet gibi popüler özellikler bulunur. Black Box, satıcılara karlı ürün fikirleri bulmalarında yardımcı olurken, Xray aranan ürünlerin verilerini çeker. Cerebro ve Magnet ise anahtar kelime araştırması ve optimizasyonu için kullanılır. Ayrıca, Helium 10, Amazon satıcılarının satışlarını artırmalarına ve işlerini daha verimli bir şekilde yönetmelerine olanak tanıyan detaylı analiz ve raporlama yetenekleri sunar. Bu sayede, Helium 10, Amazon'da satış yapan kişilere rekabet avantajı sağlamayı amaçlayan kapsamlı bir çözüm olarak öne çıkar.

Bu tezde Helium10'un "Xray" aracını kullanarak rastgele belirlenen kelimeler üzerinden aranan ürünlerin verilerini çekilmiş ve bir araya toplanmıştır.



KULLANILAN VERİLER?

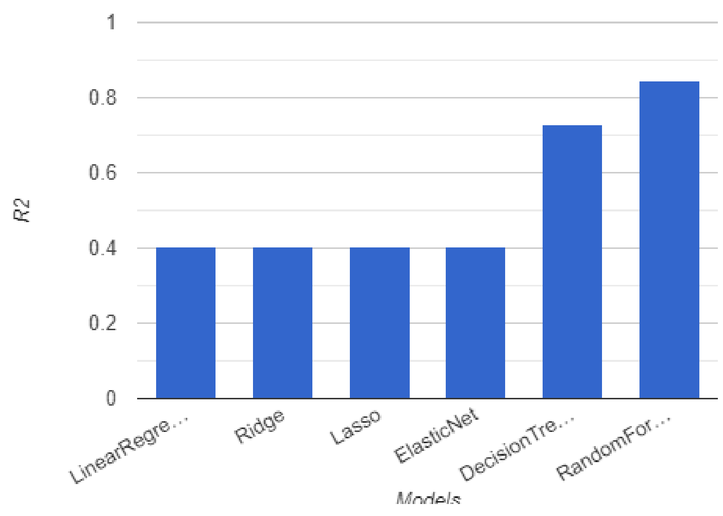
Çekilen veri 28 sütundan oluşan oldukça büyük bir veri seti. Bu veriler arasında kullandığımız veriler:
BSR: Tahmin etmek istediğimiz, en iyi satıcı sıralaması, bağımlı değişken.
Price: Ürünün fiyatı.
Sales: Ürünün satış rakamları
Active Sellers: Aynı ürünü satan satıcı sayısı.
Ratings: Ürünün puanı.
Review Count: Ürüne yapılan yorum sayısı.
Images: Ürünün resim sayısı.
Review Velocity: Ürünün yorumlarının artış trendi.
Weight: Ürünün ağırlığı.
Volume: Ürünün hacmi.
Brand: Ürünün markası.
Category: Ürünün ait olduğu kategorinin ismi.
Size Tier: Ürünün, amazon tarafından belirlenen boyut kategorisi.
Fullfilment: Ürünün satıcı mı yoksa amazon tarafından mı kargolandığını belirten değer.
Search Keyword: Ürünün arandığı ana kelime.

Örnek Veri:

	Product Details	ASIN	Brand	Price $	Sales	Revenue	BSR	Seller Country/Region	Fees $	Active Sellers #	Ratings	Review Count
0	Afrin No Drip Severe Congestion Maximum Strength Nasal Spray - 12 Hour Nasal Spray Relief for Nose Congestion, Nasal Swelling, and Allergies - 2 x 0.5 Fl Oz Bottles - Pack of 1	B07HG5JMBR	Afrin	16.46	13960	229781.60	2107	AMZ	3.86	9	4.7	8195
1	Afrin No Drip Extra Moisturizing 12 Hour Nasal Congestion Relief Pump Mist, Sinus Spray Decongestant - 15 mL	B000GCJSYG	Afrin	7.12	9893	70438.16	2602	AMZ	3.81	16	4.7	7789
2	Afrin No Drip Severe Congestion 12 Hours Relief Nasal Decongestant Bottle, 0.67 Fl Oz (Pack of 2)	B00E5PTOJO	Afrin	17.59	5608	98644.72	5122	US	6.50	27	4.8	3259
3	Amazon Basic Care No Drip Nasal Spray, 12 Hour Nasal Decongestant, Pump Mist, Oxymetazoline HCl, Cold and Allergy Nasal Spray, 1 Fl Oz (Pack of 1)	B07MLTLN75	Amazon Basic Care	5.99	0	0.00	895	AMZ	3.68	2	4.5	12505
4	Afrin No Drip Extra Moisturizing Pump Nasal Mist Twin Pack, - Nasal Relief Spray For Nose Congestion and Nasal Relief, 1 Fluid Ounce Total (0.5 FL OZ Bottles) (Packaging May Vary)	B08H9CKYKL	Afrin	14.24	0	0.00	3741	AMZ	7.25	12	4.8	1811
5	Afrin Allergy Sinus Nasal Spray - Fast and Powerful Nasal Decongestion From Allergies, For Adults and Children 6 Years and Older, 0.50 Fl Oz (Pack of 1)	B019NX6JCM	Afrin	7.56	2752	20805.12	9476	AMZ	3.65	9	4.5	1094

BEKLENTİLER VE SONUÇLAR?

- Beklentiler:
Bu tez araştırmasında, sektördeki deneyimimden; ürünün fiyatı, satış rakamı, ürünün listelendiği kategori, yorum sayısı, puanı ve kargolayan firma'nın modele en çok etki edeceğini bekliyordum.
- Kullanılan Model:
Verileri analiz ederken Lineer Regresyon, Ridge, Lasso, Elastic Net, Karar Ağaçları ve Random Forest modelleri test edildi ve %84'lük bir başarı oranı ile Random Forest modelinden en iyi sonuç alındı.



Bu model bize bir satıcının belirli bir kategorideki BSR değerini tahmin etmemizi sağlayarak, satıcının o kategoride ne kadar başarılı olduğunu anlamamızı ve rekabet edeceğimiz satıcıları belirlememize yardımcı olacaktır.

KAYNAKÇA

[1] https://towardsdatascience.com/random-forest-regression-5f605132d19d
[2] https://revenuegeeks.com/helium10-xray/
[3] https://stackoverflow.com/

İSTATİSTİK BÖLÜMÜ

HAVAYOLU YOLCU TAŞIMACILIĞI ALANINIDA MÜŞTERİ MEMNUNİYETİNİ ETKİLEYEN FAKTÖRLER: İSTANBUL HAVAALANI

Seren TÜRKMENGİL 19023043

Danışman: Prof. Dr. Ali Hakan BÜYÜKLÜ

ÖZET

Müşteri memnuniyetinin ön plana çıktığı sektörlerden biri de havayolu yolcu taşımacılığı sektörüdür. Günümüzde yolcu talepleri açısından büyük bir artış yaşanan hava yolu taşımacılığında müşteri memnuniyetinin doğru şekilde sağlanabilmesi, havayolları işletmeciliğindeki gelişimi daha da hızlandıracak bir olgu olarak ele alınmaktadır. Yanı sıra müşteri memnuniyeti, önemi, müşteri memnuniyeti konusunda belirleyici unsurlar, memnuniyet unsurunun getirileri gibi hususlar işlenmiştir. Çalışmamıza, İstanbul Havalimanı'nı kullanan 201 müşterinin gönüllü katılımıyla anket yapılıp memnuniyet durumları ile memnuniyetsizlikleri incelenmiştir. Havayolu müşterilerinin memnuniyet durumları ve bu durumlar üzerinde etkili olan faktörlerin, müşteri gözünden ele alınarak havayolu yolcu taşımacılığındaki memnuniyet unsurlarının geliştirilmesi, memnuniyetsizliğe sebep olan unsurların ortadan kaldırılması, tespit edilen hususların literatüre kazandırılarak işletmeler tarafından değerlendirmeye alınıp sektörün daha iyi bir hale getirilmesi amaçlanmıştır.

Havaalanı ve Havaalanı Sistemi

Havaalanları içerisinde; uçak iniş-kalkış trafiği ve yolcu trafiğinin düzenlenmesi için kullanılan binalar, uçuşlara hazırlık aşamasında gerekli bakım ve onarım yapıları, uçuş sonrası iniş veya kalkışların sağlanması için gerekli donanımı içeren yapılar yer alır. Bu çerçevede, havaalanları genellikle karada veya su üzerinde inşa edilen tesislerdir ve bu tesisler, hava araçlarının iniş-kalkış ve yer hareketleri sırasında kullanabilecekleri belirli alanları içerir. Havaalanı sistemi , mekanik , teknolojik altyapı, ve elektronik sistemlerin birleşimi ve alt bileşenden oluşur.

Havaalanı İşletmeciliğinde Yolcu Memnuniyeti

Son dönemlerde havayolu işletmeciliğini yaygınlaşması ve bu nedenle bu alanda yaşanan rekabet ortamı, havayolu işletmeleri açısından yolcuların beklenti ve isteklerine daha fazla yönelim olmasını, diğer bir ifade ile yolcu memnuniyetinin üst düzeyde sağlanması gerekliliğini ortaya çıkarmıştır. Havayolu işletmelerinin artan sayısı ile birlikte bu işletmeler belirli bir müşteri profili yakalamakla birlikte mevcut müşterilerini kaybetmeden yeni müşteriler kazanma yoluna gitmektedirler.

AMAÇ

Yapılan çalışmada, havayolu yolcu taşımacılığı alanında müşteri memnuniyetine ya da memnuniyetsizliğine neden olan unsurların tespit edilmesi, mevcut havayolu taşımacılığını kullanan müşterilerin elde tutulması ve bunlara ek olarak havayolu taşımacılığına yeni müşterilerin kazandırılması amaçlanmıştır. Bu bağlamda hava taşımacılığında İstanbul Havaalanı'nı kullanan müşteriler üzerinde araştırmalar yapılıp memnuniyet durumları ile memnuniyetsizlikleri incelenmiş ve bunların sebepleri tespit edilmeye çalışılmıştır.

YÖNTEM

Tez kapsamında, İstanbul Havaalanı'nı kullanan müşterilerin memnuniyet ve memnuniyetsizliklerini etkileyen faktörlerin analizi için anket çalışması yapılmıştır.Anket verilerine göre faktörler tespit edilmiştir.Bu analizler için SPSS uygulaması kullanılmıştır.

ANALİZLER

Faktör Analizi

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,758
Bartlett's Test of Sphericity	Approx. Chi-Square	230,837
	df	21
	Sig.	<,001

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,826
Bartlett's Test of Sphericity	Approx. Chi-Square	294,672
	df	21
	Sig.	<,001

Ankette yer alan 5'li likert ölçekli 14 soru için yapılan KMO ve Bartlett's testi iki kısımda incelenmiştir.

Çıktılara baktığımızda 0,758 ve 0,826 olasılık değerlerini 0,50'den büyük olduğu için sıfır hipotezi reddedilmektedir. Bu durumda, veri setinin faktör analizi için uygun olduğu söylenebilir.

	Faktörler	
	1 (Hizmet ve Kolaylık)	2 (Destekleyici Hizmetler)
İstanbul Havalimanına şehirden ulaşım kolaydır. (ulasım_kolaylıgı)	0,740	
İstanbul Havalimanında oturma yerleri yeterlidir. (oturma_alanı_yeterliligi)	0,535	
İstanbul Havalimanında uçaklar zamanında kalkmaktadır. (ucak_kalkıs_zamanı)	0,568	
İstanbul Havalimanında check-in bekleme süresi normaldir. (check_in_kolaylıgı)	0, 593	
İstanbul Havalimanında içinde kafe ve restaurant sayısı yeterlidir. (cafe_yeterliligi		0,649
İstanbul Havalimanında gümrük işleri hızlı ve sorunsuzdur. (gumruk_kolaylıgı)		0,507

Faktör analizi sonucunda, ulaşım kolaylığı (0,740), oturma alanı yeterliliği (0,535), uçak kalkış zamanı (0,568), ve check-in kolaylığı (0,593) değişkenleri 1. faktörde yüksek faktör yüklerine sahiptir. Cafe yeterliliği (0,649) ve gümrük kolaylığı (0,507) değişkenleri ise 2. faktörde yüksek faktör yüklerine sahiptir. Bu sonuçlar, ilgili değişkenlerin belirlenen faktörlerle güçlü bir ilişkiye sahip olduğunu göstermektedir.

Değişken	İlk Öz Değer	İndirgenmiş Değer
İstanbul Havalimanına şehirden ulaşım kolaydır. (ulasım_kolaylıgı)	1,000	0,740
İstanbul Havalimanında içinde kafe ve restaurant sayısı yeterlidir. (cafe_yeterliligi)	1,000	0,649
İstanbul Havalimanında fiyatlar uygundur . (fiyat_uygunlugu)	1,000	0,213
İstanbul Havalimanında oturma yerleri yeterlidir. (oturma_alanı_yeterliligi)	1,000	0,535
İstanbul Havalimanında uçaklar zamanında kalkmaktadır. (ucak_kalkıs_zamanı)	1,000	0,568
İstanbul Havalimanında check-in bekleme süresi normaldir. (check_in_kolaylıgı)	1,000	0,593
İstanbul Havalimanında bagaj teslimi hızlı ve kolaydır. (bagaj_alımı_kolaylıgı)	1,000	0,453
İstanbul Havalimanında bagaj alımı hızlı ve rahattır. (bagaj_teslim_kolaylıgı)	1,000	0,423
İstanbul Havalimanında gümrük işleri hızlı ve sorunsuzdur. (gumruk_kolaylıgı)	1,000	0,507
İstanbul havalimanında uçak bileti satın alabileceğim yerler yeterlidir. (ucak_bileti)	1,000	0,454
İstanbul Havalimanında personel sorun çözümünde samimidir. (personal_davranısı)	1,000	0,496
İstanbul Havalimanının temizlik ve hijyen konusunda yeterlidir. (temizlik_hijyen)	1,000	0,421
İstanbul Havalimanında yönlendirici levhalar yeterli ve açıklayıcıdır. (levha_yeterliligi)	1,000	0,338
İstanbul Havalimanının araç park yerleri ve yolcu indirme yerleri yeterlidir. (arac_park_yeri_yeterliligi)	1,000	0,374

Faktör analizi sonucunda, başlangıçta her değişkenin değeri 1 iken, analiz sonrası en düşük indirgenmiş değere sahip olan değişken "İstanbul Havalimanı'nda fiyatlar uygundur" (0,213) olmuştur. Bilgi kaybı 0,50'nin üzerinde olan değişkenler; yani fiyat uygunluğu, bagaj alımı kolaylığı, bagaj teslim kolaylığı, uçak bileti, personel davranışı, temizlik ve hijyen, levha yeterliliği ve araç park yeri yeterliliği, veri setinden çıkarılmalıdır.

Kruskal-Wallis Testi Sonuçları

Independent-Samples Kruskal-Wallis Test Summary

Total N	201
Test Statistic	16,321[a]
Degree Of Freedom	4
Asymptotic Sig.(2-sided test)	,003

a. The test statistic is adjusted for ties.

H0: İstanbul Havalimanı'nda müşteri memnuniyetini etkileyen hizmet ve kolaylık faktörü ile müşterilerin eğitim durumlarına göre anlamlı bir farklılık göstermemektedir.

H1: İstanbul Havalimanı'nda müşteri memnuniyetini etkileyen hizmet ve kolaylık faktörü ile müşterilerin eğitim durumlarına göre anlamlı bir farklılık göstermektedir.

incelendiğinde İstanbul Havaalanı'nda müşteri memnuniyetini etkileyen hizmet ve kolaylık faktörü ile müşterilerin eğitim durumları arasında anlamlı farklılıklar olduğu sonucuna ulaşılmıştır. (p<0.05).

Independent-Samples Kruskal-Wallis Test Summary

Total N	201
Test Statistic	4,972[a]
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	,174

a. The test statistic is adjusted for ties.

H0: İstanbul Havalimanı'nda müşteri memnuniyetini etkileyen hizmet ve kolaylık faktörü ile müşterilerin seyahat tercihlerine göre anlamlı bir farklılık göstermemektedir.

H1: İstanbul Havalimanı'nda müşteri memnuniyetini etkileyen hizmet ve kolaylık faktörü ile müşterilerin seyahat tercihlerine göre anlamlı bir farklılık göstermektedir.

İstanbul Havaalanı'nda müşteri memnuniyetini etkileyen hizmet ve kolaylık faktörü ile müşterilerin seyahat tercihleri arasında anlamlı farklılıklar olmadığı sonucuna ulaşılmıştır (p>0.05).

Independent-Samples Kruskal-Wallis Test Summary

Total N	201
Test Statistic	1,797[a]
Degree Of Freedom	4
Asymptotic Sig.(2-sided test)	,773

a. The test statistic is adjusted for ties.

H0: İstanbul Havalimanı'nda müşteri memnuniyetini etkileyen destekleyici hizmetler faktörleri ile müşterilerin gelir durumlarına göre anlamlı bir farklılık göstermemektedir.

H1: İstanbul Havalimanı'nda müşteri memnuniyetini etkileyen destekleyici hizmetler faktörleri ile müşterilerin gelir durumlarına göre anlamlı bir farklılık göstermektedir.

İstanbul Havaalanı'nda müşteri memnuniyetini etkileyen destekleyici hizmetler faktörü ile müşterilerin gelir durumları arasında anlamlı farklılıklar olmadığı sonucuna ulaşılmıştır(p>0.05).

Independent-Samples Kruskal-Wallis Test Summary

Total N	201
Test Statistic	3,741[a]
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	,291

a. The test statistic is adjusted for ties.

H0: İstanbul Havalimanı'nda müşteri memnuniyetini etkileyen destekleyici hizmetler faktörleri ile müşterilerin seyahat tercihlerine göre anlamlı bir farklılık göstermemektedir.

H1: İstanbul Havalimanı'nda müşteri memnuniyetini etkileyen destekleyici hizmetler faktörleri ile müşterilerin seyahat tercihlerine göre anlamlı bir farklılık göstermektedir.

İstanbul Havaalanı'nda müşteri memnuniyetini etkileyen destekleyici hizmetler faktörü ile müşterilerin seyahat tercihleri arasında anlamlı farklılıklar olmadığı sonucuna ulaşılmıştır(p>0.05).

H0: İstanbul Havalimanı'nda müşteri memnuniyetini etkileyen hizmet ve kolaylık faktörü ile müşterilerin gelir durumlarına göre anlamlı bir farklılık göstermemektedir.

H1: İstanbul Havalimanı'nda müşteri memnuniyetini etkileyen hizmet ve kolaylık faktörü ile müşterilerin gelir durumlarına göre anlamlı bir farklılık göstermektedir.

İstanbul Havaalanı'nda müşteri memnuniyetini etkileyen hizmet ve kolaylık faktörü ile müşterilerin gelir durumları arasında anlamlı farklılıklar olmadığı sonucuna ulaşılmıştır (p>0.05).

Independent-Samples Kruskal-Wallis Test Summary

Total N	201
Test Statistic	8,367[a]
Degree Of Freedom	4
Asymptotic Sig.(2-sided test)	,079

a. The test statistic is adjusted for ties.

H0: İstanbul Havalimanı'nda müşteri memnuniyetini etkileyen destekleyici hizmetler faktörleri ile müşterilerin eğitim durumlarına göre anlamlı bir farklılık göstermemektedir.

H1: İstanbul Havalimanı'nda müşteri memnuniyetini etkileyen destekleyici hizmetler faktörleri ile müşterilerin eğitim durumlarına göre anlamlı bir farklılık göstermektedir.

İstanbul Havaalanı'nda müşteri memnuniyetini etkileyen destekleyici hizmetler faktörü ile müşterilerin eğitim durumları arasında anlamlı farklılıklar olmadığı sonucuna ulaşılmıştır (p>0.05).

Independent-Samples Kruskal-Wallis Test Summary

Total N	201
Test Statistic	2,911[a]
Degree Of Freedom	4
Asymptotic Sig.(2-sided test)	,573

a. The test statistic is adjusted for ties.

Sonuçlar

İstanbul Havaalanı'nda müşteri memnuniyetinin sağlanması konusunu ele aldığımız çalışmamızda yaptığımız araştırmalar sonucunda müşteri memnuniyeti konusunda yalnızca havalimanında sunulan hizmetlerin değil, müşterilerin belirli birtakım özelliklerinin de etkili olduğu gibi genel bir sonuca ulaştığımızdan öncelikli olarak bahsetmemiz gerekmektedir. Tüm bu sonuçlar altında İstanbul Havaalanı'nı kullanan müşterilerin memnuniyet seviyelerini etkileyen hizmet ve Kolaylık faktörü ile eğitim durumu, gelir durumu ve seyahat tercihleri arasındaki ilişkiler incelenmiştir. Analiz sonucunda, Hizmet ve Kolaylık faktörü ile eğitim durumu arasında anlamlı farklılıklar bulunmuştur. Ancak, Hizmet ve Kolaylık faktörü ile gelir durumu ve seyahat tercihleri arasında anlamlı bir fark tespit edilememiştir. Aynı şekilde, Destekleyici Hizmetler faktörü ile eğitim durumu, gelir durumu ve seyahat tercihleri arasındaki ilişkiler de incelenmiştir. Destekleyici Hizmetler faktörü ile bu üç değişken arasında anlamlı bir fark bulunamamıştır.

KAYNAKÇA

[1] Doganis, R. (2010). *Flying Off Course: Airline Economics and Marketing* (4th ed.). Routledge.
[2] https://mustafaotrar.net/istatistik/kruskal-wallis-h-testi/
[3] https://dergipark.org.tr/tr/download/article-file/108451

İSTATİSTİK BÖLÜMÜ

ÖĞRENCİLERİN HARCAMA ALIŞKANLIKLARININ İNCELENMESİ

SÜMEYYE SERT 18023068

Danışman: Doç. Dr. Reşit ÇELİK

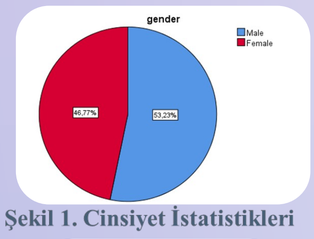
ÖZET

Bu tez çalışmasının amacı; öğrencilerin bireysel ve toplam birim öğrenci maliyetlerini belirlemek ve bu maliyetleri oluşturan bileşenleri ortaya koymaktır. Bu çalışmada ele alınan veri seti, çeşitli demografik gruplar ve akademik geçmişlerden 1000 öğrencinin harcama alışkanlıklarını temsil eden kurgusal verilerin gerekli düzenlemeler sonucu elde edilen veri setinden; öğrencilerin aylık gelir ve harcamalarını analiz etmek amaçlanmıştır. Veri setinde yaş, cinsiyet, okul yılı, branş, aylık gelir, alınan maddi yardım ve farklı harcama kategorilerindeki harcamalar gibi bilgiler yer alıyor. Harcama kategorileri arasında ise eğitim, barınma, yiyecek, ulaşım, kitap ve malzeme, eğlence, kişisel bakım, teknoloji, sağlık ve çeşitli harcamalar yer alıyor. Ek olarak veri seti her öğrenci için harcamalarına ilişkin tercih edilen ödeme yöntemini içerir. Verimizin analizinde; Özet istatistikleri, Ki-Kare testi, Regresyon analizi ve Korelasyon incelemeleri yapılmıştır.

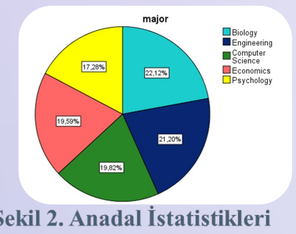
EĞİTİM

Eğitim her dönemde ve her şartta dikkat edilmesi ve tartışılması gereken bir olgudur. Özellikle modern dönemlerde, her insan örgün ya da yaygın bir şekilde eğitim gördüğü için bu olgunun önemi kendiliğinden ortaya çıkmaktadır. Zira eğitim hemen fark edilebileceği gibi çok karmaşık ve aynı zamanda birçok unsur ile birlikte düşünülmesi gereken bir olgudur. Bireyde var olan potansiyeli tespit edip, bu potansiyeli en üst düzeye çıkarabilme süreci olarak tanımlanabilecek eğitim olgusu toplumdan topluma, zamandan zamana farklı biçimlerde şekillenmiş ve her yeni şekli ile de toplumları etkilemiştir.

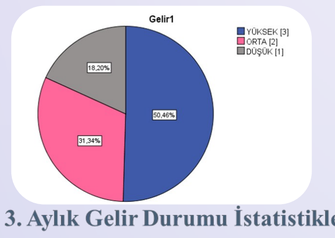
ÖĞRENCİ HARCAMA ALIŞKANLIKLARI VERİSİNİN ANALİZİ



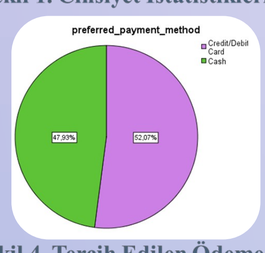
Şekil 1. Cinsiyet İstatistikleri



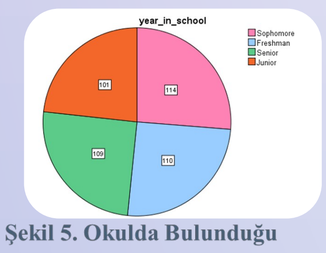
Şekil 2. Anadal İstatistikleri



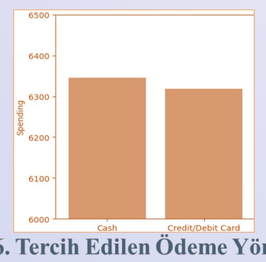
Şekil 3. Aylık Gelir Durumu İstatistikleri



Şekil 4. Tercih Edilen Ödeme Yöntemi İstatistikleri



Şekil 5. Okulda Bulunduğu Yılının İstatistikleri



Şekil 6. Tercih Edilen Ödeme Yöntemi İstatistikleri Toplam Harcamalar

Kİ – KARE TESTİ

Ki-Kare bağımsızlık testi, kategorik veriler arasındaki bağımsızlık ilişkisini değerlendirmek için kullanılan bir istatistiksel testtir. Bu test, örnekleme dağılımı olarak ki-kare dağılımını kullanır ve parametrik olmayan bir test olarak kabul edilir.

Hipotezlerin Belirlenmesi:

- H_0 (Sıfır Hipotezi): Değişkenler arasında bir ilişki yoktur (bağımsızdır).
- H_1 (Alternatif Hipotezi): Değişkenler arasında bir ilişki vardır (bağımlıdır).

$$\chi^2_{test} = \sum \frac{(Oi - Ei)\text{^}2}{Ei}$$

Cinsiyet ile Tercih Edilen Ödeme Yöntemi İlişkisi

Hipotez:

H_0 : Cinsiyet ile tercih edilen ödeme yöntemi arasında bir ilişki yoktur.

H_1 : Cinsiyet ile tercih edilen ödeme yöntemi arasında bir ilişki vardır.

gender * preferred_payment_method Crosstabulation

			preferred_payment_method		Total
			Cash	Credit/Debit Card	
gender	Female	Count	97	106	203
		Expected Count	97,3	105,7	203,0
	Male	Count	111	120	231
		Expected Count	110,7	120,3	231,0
Total		Count	208	226	434
		Expected Count	208,0	226,0	434,0

Ödeme Yöntemi Kategorisi Arasındaki İlişki Tablosu

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,003[a]	1	,955		
Continuity Correction[b]	,000	1	1,000		
Likelihood Ratio	,003	1	,955		
Fisher's Exact Test				1,000	,516
N of Valid Cases	434				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 97,29.

b. Computed only for a 2x2 table

Yöntemi Kategorisinin Ki-Kare Testi Tablosu

Kadınların sayısı ile erkeklerin sayısı arasında yaklaşık 30 kişi bir fark bulunurken, her ikisinin de harcamalarında kart kullanımı daha fazla olduğu yorumunu yapabiliriz.

Ki-Kare testi sonuçlarına göre “Cinsiyet” bağımsız değişkeni ile “Tercih Edilen Ödeme Yöntemi” bağımlı değişkeni arasında **p** değeri **0,955**'dir. Bu iki değişken arasında elde edilen **p** değeri **0,05** olan **α** değerinden büyük olduğundan dolayı modelimizin anlamsız olduğunu ve iki değişken arasında anlamlı bir ilişki olmadığını söyleyebiliriz. Bu yüzden H_0 hipotezinin kabulüne ve H_1 hipotezinin reddine karar verilir.

Yaş ve Öğrencilerin Öğrenim Giderleri Arasındaki İlişkisi

Hipotez:

H_0 : Yaş ile öğrencilerin öğrenim giderleri arasında bir ilişki yoktur.

H_1 : Yaş ile öğrencilerin öğrenim giderleri arasında bir ilişki vardır.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	2830,339[a]	2814	,411
Likelihood Ratio	1726,464	2814	1,000
Linear-by-Linear Association	,709	1	,400
N of Valid Cases	434		

a. 3224 cells (100,0%) have expected count less than 5. The minimum expected count is ,11.

Tablo 3. Yaş Kategorisi ile Öğrencilerin Öğrenim Giderleri Kategorisinin Ki-Kare Testi Tablosu

Ki-Kare testi sonuçlarına göre “Yaş” bağımsız değişkeni ile “Öğrenim Giderleri” bağımlı değişkeninin Pearson Chi-Square’a karşılık gelen significance değeri **0,411** olarak görülmektedir. Bu değere göre **0,411>0,05** olduğu için H_0 hipotezi reddedilemez. Bu da demek oluyor ki yaş ile öğrencilerin öğrenim giderleri arasında bir ilişki yoktur, diğer bir deyişle anlamlı bir fark yoktur.

REGRESYON ANALİZİ

Regresyon analizi, iki veya daha fazla nicel değişken arasındaki ilişkiyi ölçmek için kullanılan istatistiksel bir yöntemdir. Bu yöntem, bağımlı değişken olarak adlandırılan bir değişkenin, bağımsız değişken olarak adlandırılan bir veya birden fazla değişken tarafından nasıl etkilendiğini analiz etmeye yardımcı olur.

Öğrencilerin Burs Desteği ile Yemek Harcamaları Arasındaki İlişkisi

Hipotez:

H_0: Öğrencilerin burs desteği ile yemek harcamaları arasında bir ilişki yoktur.

H_1: Öğrencilerin burs desteği ile yemek h

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	643,717	1	643,717	,086	,769[b]
	Residual	3232064,746	432	7481,631		
	Total	3232708,463	433			

a. Dependent Variable: food

b. Predictors: (Constant), financial_aid

Tablo 4. Öğrencilerin Burs Desteği ile Yemek Harcamaları Arasındaki İlişkisi Tablosu

ANOVA tablomuzda bulunan anlamlılık değeri için **F** değeri etkili olup modeldeki anlamlılık değerimiz **0,76**’dır. Regresyon analizi testimizin sonuçlarına göre “Yemek Harcamaları” bağımlı değişken olup “Burs Desteği” bağımsız değişkeni arasında elde edilen **p** değeri **0,05** olan **α** değerinden büyük olduğundan dolayı modelimizin anlamsız olduğunu ve iki değişken arasında anlamlı bir ilişki olmadığını söyleyebiliriz. Bu yüzden H_0 hipotezinin kabulüne ve H_1 hipotezinin reddine karar verilir.

KORELASYON İNCELEMESİ

Maddeler Arası Korelasyon İncelemesi

Correlations

			age	monthly_income	financial_aid	tuition	housing	food	transportation	books_supplies	entertainment	personal_care	technology	health_wellness	miscellaneous
Spearman's rho	age	Correlation Coefficient	1,000	-,039	-,099*	,043	,047	,002	-,077	,002	-,002	-,055	,065	-,050	,011
		Sig. (2-tailed)	.	,414	,039	,370	,333	,969	,108	,963	,975	,253	,173	,299	,817
		N	434	434	434	434	434	434	434	434	434	434	434	434	434
	monthly_income	Correlation Coefficient	-,039	1,000	,078	-,030	-,004	,048	,033	,005	-,062	-,014	,040	-,029	-,053
		Sig. (2-tailed)	,414	.	,106	,533	,929	,317	,495	,925	,195	,776	,403	,546	,274
		N	434	434	434	434	434	434	434	434	434	434	434	434	434
	financial_aid	Correlation Coefficient	-,099*	,078	1,000	-,027	,007	-,008	,014	-,086	,027	,043	-,033	,027	,033
		Sig. (2-tailed)	,039	,106	.	,581	,877	,868	,768	,075	,570	,374	,487	,575	,499
		N	434	434	434	434	434	434	434	434	434	434	434	434	434
	tuition	Correlation Coefficient	,043	-,030	-,027	1,000	-,050	,005	-,019	,036	-,043	-,006	-,016	,068	-,027
		Sig. (2-tailed)	,370	,533	,581	.	,295	,910	,695	,454	,374	,898	,747	,159	,572
		N	434	434	434	434	434	434	434	434	434	434	434	434	434
	housing	Correlation Coefficient	,047	-,004	,007	-,050	1,000	-,043	-,024	-,027	,022	,062	,048	-,016	,068
		Sig. (2-tailed)	,333	,929	,877	,295	.	,376	,625	,581	,641	,199	,316	,737	,160
		N	434	434	434	434	434	434	434	434	434	434	434	434	434
	food	Correlation Coefficient	,002	,048	-,008	,005	-,043	1,000	,038	,134**	,023	,091	,065	-,018	,094
		Sig. (2-tailed)	,969	,317	,868	,910	,376	.	,432	,005	,634	,059	,174	,703	,051
		N	434	434	434	434	434	434	434	434	434	434	434	434	434
	transportation	Correlation Coefficient	-,077	,033	,014	-,019	-,024	,038	1,000	,036	,000	,115*	-,047	,046	,014
		Sig. (2-tailed)	,108	,495	,768	,695	,625	,432	.	,459	,994	,017	,326	,335	,770
		N	434	434	434	434	434	434	434	434	434	434	434	434	434
	books_supplies	Correlation Coefficient	,002	,005	-,086	,036	-,027	,134**	,036	1,000	-,021	-,007	,065	-,009	-,017
		Sig. (2-tailed)	,963	,925	,075	,454	,581	,005	,459	.	,666	,890	,179	,849	,720
		N	434	434	434	434	434	434	434	434	434	434	434	434	434
	entertainment	Correlation Coefficient	-,002	-,062	,027	-,043	,022	,023	,000	-,021	1,000	,005	,089	-,040	,031
		Sig. (2-tailed)	,975	,195	,570	,374	,641	,634	,994	,666	.	,911	,063	,404	,515
		N	434	434	434	434	434	434	434	434	434	434	434	434	434
	personal_care	Correlation Coefficient	-,055	-,014	,043	-,006	,062	,091	,115*	-,007	,005	1,000	,065	,014	,050
		Sig. (2-tailed)	,253	,776	,374	,898	,199	,059	,017	,890	,911	.	,179	,767	,300
		N	434	434	434	434	434	434	434	434	434	434	434	434	434
	technology	Correlation Coefficient	,065	,040	-,033	-,016	,048	,065	-,047	,065	,089	,065	1,000	,002	,061
		Sig. (2-tailed)	,173	,403	,497	,747	,316	,174	,326	,179	,063	,179	.	,966	,204
		N	434	434	434	434	434	434	434	434	434	434	434	434	434
	health_wellness	Correlation Coefficient	-,050	-,029	,027	,068	-,016	-,018	,046	-,009	-,040	,014	,002	1,000	,006
		Sig. (2-tailed)	,299	,546	,575	,159	,737	,703	,335	,849	,404	,767	,966	.	,906
		N	434	434	434	434	434	434	434	434	434	434	434	434	434
	miscellaneous	Correlation Coefficient	,011	-,053	,033	-,027	,068	,094	,014	-,017	,031	,050	,061	,006	1,000
		Sig. (2-tailed)	,817	,274	,499	,572	,160	,051	,770	,720	,515	,300	,204	,906	.
		N	434	434	434	434	434	434	434	434	434	434	434	434	434

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

- Katılımcıların aylık geliri ile aldığı finansal yardımı arasında pozitif yönde zayıf bir ilişki vardır.
- Katılımcıların eğitim harcamaları ile eğlence harcamaları arasında negatif yönde zayıf bir ilişki vardır.
- Katılımcıların yemek harcamaları ve sağlık ile harcamaları arasında negatif yönde zayıf bir ilişki vardır.
- Katılımcıların yaşı ile sağlık harcamaları arasında negatif yönde zayıf bir ilişki vardır.

SONUÇ

Bu tez çalışmasında yapılan analizler öğrencilerin gelirlerini ve giderlerini açıklamaktadır .Aynı zamanda öğrencilerin aylık gelir ve harcamalarını analiz ederek, finansal refahlarını ve geleceklerini iyileştirmek için sunulacak önerilere yol açmaktadır.

KAYNAKÇA

[1] Veri Seti- https://www.kaggle.com/datasets/sumanthnimmagadda/student-spending-dataset

[2]https://openaccess.hku.edu.tr/

[3]https://search.trdizin.gov.tr/tr/yayin/detay/95343/turkiyede-yuksekogretimde-ogrenci-harcama-ve-maliyetleri

İSTATİSTİK BÖLÜMÜ

Makine Öğrenmesi ve Açıklanabilir Yapay Zeka Yöntemleri ile Sigortacılık Sektöründe Sahtecilik Tespitinin Analizi

Ulaş TABAR 20023044

Danışman: Prof. Dr. Ersoy ÖZ

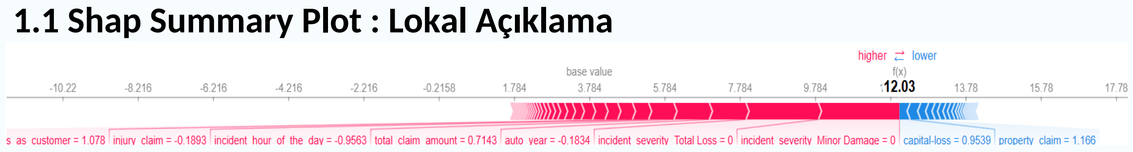
ÖZET

Bu çalışmada sigortacılık sektöründe önemli bir yeri olan araç sigortası hasar tespitinde sahtecilik saptama problemini makine öğrenmesi teknikleri kullanarak tahmin etmek ve bu tahminleri Açıklanabilir Yapay Zeka(XAI) yöntemleriyle anlamlandırmak amaçlanmıştır. Çalışma kapsamında ilgili veri seti 1000 gözlemden, kategorik ve nümerik 36 değişkenden oluşmaktadır.Veri önişleme, aykırı değer analizi ve eksik veri analizi uygulanarak veri seti son haline getirilmiş, makine öğrenmesi modellerinden en iyi sonucu veren Lightgbm algoritması baz model alınarak geliştirilmiştir. Hiperparametre optimizasyonları ve k-katlı çapraz doğrulama işlemleri sonucunda model en optimum sonucu verecek hale getirilmiş, sonuçlar accuracy, f1 score ve roc auc metrikleri ile ölçülmüştür.Daha sonra lokal ve global sonuçları olan 5 adet XAI yöntemi belirlenmiş ve her birinin modeli açıklaması yorumlanmıştır. Lightgbm algoritması XAI yöntemlerinin çıktılarının daha gerçekçi yorumlanabilmesi için hiperparametreleri kullanılarak kontrollü bir şekilde karmaşıklaştırılmıştır.

1.SHAP (Shapley Additive Explanations) Yöntemi

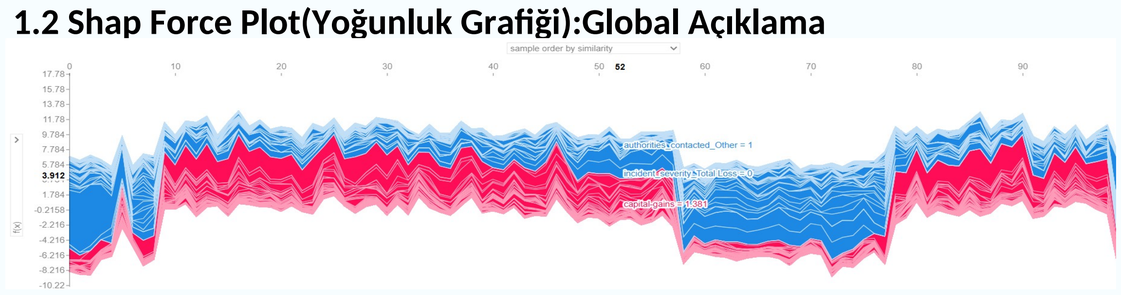
SHAP her bir değişkenin, modelin çıktısına olan katkısını belirlemek için oyun teorisinden türetilmiş Shapley değelerini kullanarak model kararlarını açıklayan bir yöntemdir.

1.1 Shap Summary Plot : Lokal Açıklama



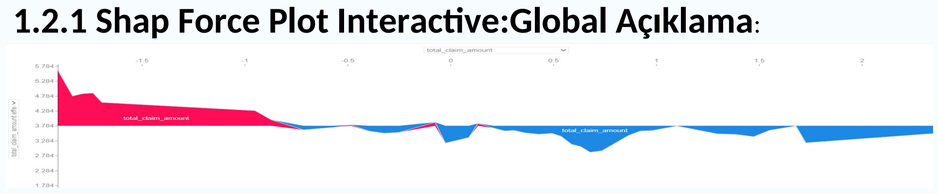
Yukarıdaki grafik veri setindeki tek satır veriye uygulanan bir shap grafik türüdür.Tek satır veride tahmin sonucuna pozitif/negatif etki eden gözlemler ve değerleri ayrıca tahmin sonucunun 1 sınıfına ait olma olasılığı gözükmektedir. Diğer SHAP grafiklerinde de görülecek olan 1 sınıfına ait olma olasılığı log-odds cinsdindedir. Bunun sebebi SHAP yöntemindeki Shapley değerleri hesaplanırken log-odds(logaritmik ihtimaller) cinsinden hesaplanmaktadır. Tek gözlem biriminde tahmin sonucuna pozitif etki yapan değişkenler ve değerleri kırmızı renkte, negatif etki yapanlar ise mavi renkte gözükmektedir.12.03 log-odds değeri sigmoid fonksiyonuna girildiğinde 1 sınıfına ait olma olasılığı %99 çıkmaktadır.

1.2 Shap Force Plot(Yoğunluk Grafiği):Global Açıklama



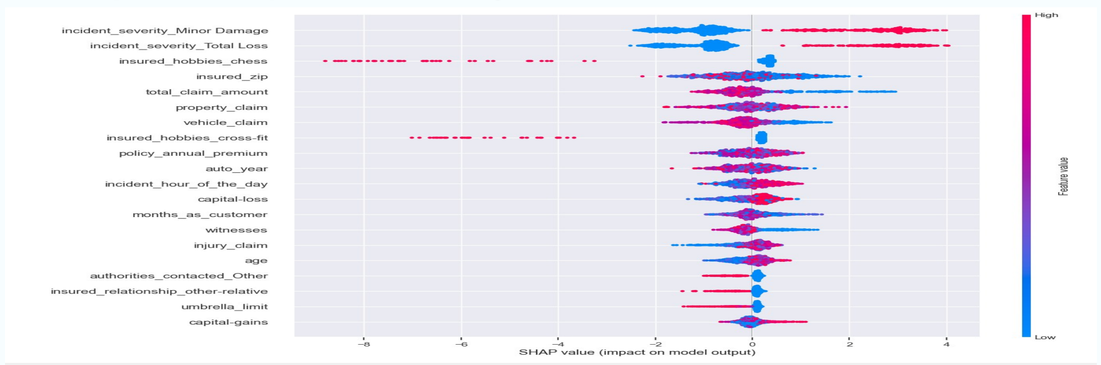
Shap Yoğunluk Grafiği 1.1'deki Shap Summary Plot'ın dikey hale getirilip tüm veriye uygulanmış halidir.Pyhtonda oluşturulan html uzantılı çıktısında tahminin pozitif(kırmızı) ve negatif(mavi) yönde olduğu alanların üstüne gelindiğinde ilgili noktaya etki eden değişkenler ve değerleri görülmektedir.Bu grafiğin kullanışlı olmasının asıl sebebi ise interaktif olarak istenilen değişkenin shap değerlerinin bütün veri için hedef değişkeni üzerindeki etkisinin görülebilmesidir

1.2.1 Shap Force Plot Interactive:Global Açıklama:



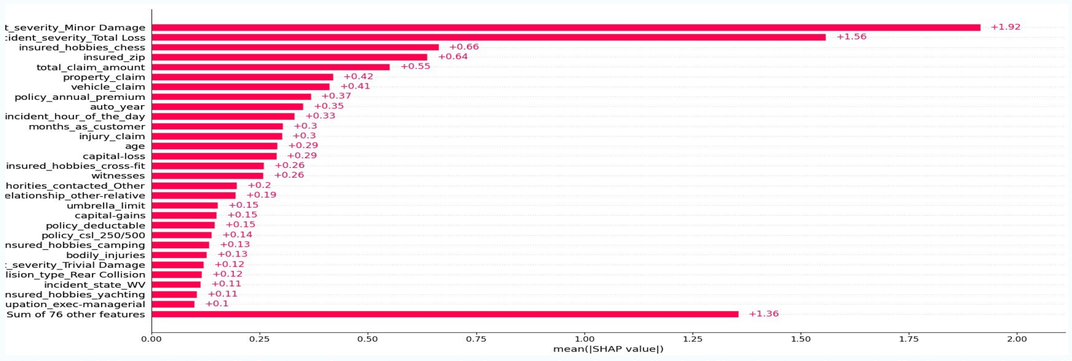
Trafik sigortası bulunan sigortalının müşterisi olduğu şirketten kaza sonrası istediği tazminat tutarının yüksek olması çoğunlukla beklenen bir senaryodur.Ancak Shap Force Plot sayesinde durumun veri setinde tam tersi olduğu açıkça gözükmektedir.Sigortalının talep ettiği 'toplam tazminat tutarı' değeri arttıkça hedef değişkenin pozitif olma olasılığına etki eden shapley değerleri azalmakta dolayısıyla hedef değişkene etki azalmaktadır.

1.3 Shap Summary Plot:Global Açıklama



Yukarıdaki global açıklama grafiğindeki değişkenler shapley değerlerinin önem düzeylerine göre sıralanmıştır.Sahteciliğe en çok etkisi olan değişkenin 'düşük şiddetli ihbar' olduğu gözükmektedir.Grafik ayrıca değişkenlerin shapley değerleriyle etkilerini detaylandırmaktadır.Örneğin müşterinin sigorta paketindeki ek koruma limiti(unbrella limit) değerlerinin artması hedef değişkene olumsuz etki yapmaktadır.

1.4. Shap Bar Plot: Global Açıklama



Yukarıdaki bar plot klasik feature importance çıktılarının shapley değerleri ile oluşturulmuş hali olarak anlaşılabilir. Her bir değişkenin shapley değerleri ve hedef değişkeni etkilemeleri yatay eksende gösterilmiştir. Grafiğe göre minimum hasar değişkeni en yüksek etkiye sahip değişkendir. Grafik için eşik değer(0.40) belirlenerek belirli bir değer üzerindeki değişkenlere ayrı ayrı odaklanılabilir. Bu değişkenlere yapılacak her bir düzeltme işlemi sonucunda yenide kurulacak model ile değişkenin önem düzeyindeki değişim bu grafik yeniden oluşturularak karşılaştırılabilir ve iyileştirmeler takip edilebilir.

2.LIME (Local Interpretable Model-Agnostic Explanations) Yöntemi

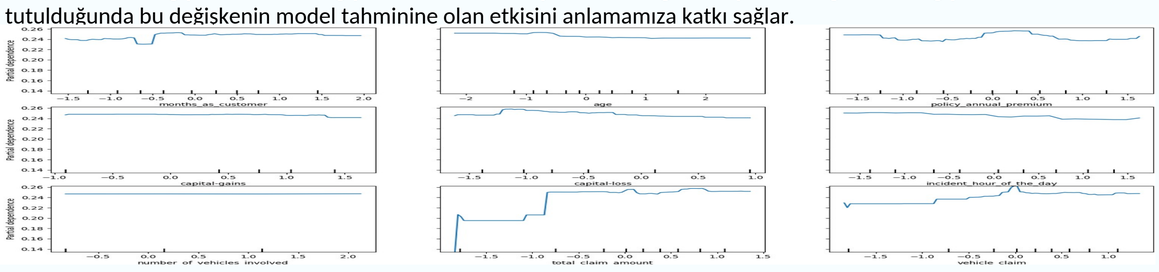
Lime, modelin belirli bir tahminini açıklamak için,orijinal modelin yakınında yerel ve anlaşılabilir model oluşturur ve bu yerel modelin kararlarını kullanarak orijinal modelin tahminini açıklar.



Yukarıdaki grafik tanımından da anlaşılacağı üzere yerel tahminler yapan bir yöntemdir.Tek bir gözlem birimindeki değişken değerlerine ve sonuçlarına bakıldığında ilgili satırın tahmin olasılıklarına göre model çıktısı 0 yani sahtekarlık yok çıkmıştır.Bu olasılık değerlerine en çok etki eden 5 değişken ve değerleri en sağ tarafta belirtilmiştir.Ortadaki çıktıda ise yorum şu şekilde yapılır:sigortalının yaşadığı hasarın minimum olması kategorisinin 1 sınıfının olma olasılığı olan 0.30 değerine etkisİ 0.52'dir.Aynı şekilde sigortalının hobisinin satranç olma kategorik değişkeninin 0 sınıfının tahmin olasılığına etkisi 0.59'dur.Yerel tahminlerin açıklamasına bakarak modelin tahminlerinin her bir satır gözlem bazında nasıl yaptığını anlamak ve ilginç bulgularla karşılaşmak mümkündür.

3.Partial Dependence Plot Yöntemi

PDP, belirli bir değişkenin hedef değişken üzerindeki marjinal etkisini görselleştirir ve diğer tüm değişkenler sabit tutulduğunda bu değiskenin model tahminine olan etkisini anlamamıza katkı sağlar.



Yukarıdaki PDP belirli nümerik değişkenlere uygulanmıştır.Hedef değişken üzerindeki marjinal etki arandığı için grafiklerdeki hassas değişimler önem arz etmektedir.Bu değişimlerin olmadığı kazaya karışan araç sayısı değişkeninin tahmine ciddi bir etkisi yoktur.Bunun aksine sigortalının yaşı arttıkça hedef değişkene olan etki azalmaktadır.Aynı negatif doğrusal durum sigortalının sermaye kaybı, kazanın meydana geldği saat içinde geçerlidir.Hedef değişken üzerinde pozitif etkisi ve duyarlılığı olan değişkenler ise müşteri olunan ay sayısı,araç hasarı ile ilgili talep edilen tazminat miktarı,toplam tazminat miktarıdır.Yıllık poliçe birikimi dalgalı bir yapı gösterdiği için dış verilerle yeniden kontrol edilebilir.

4.Permutation Importance Yöntemi

Bu yöntem, bir değişkenin değerlerini rastgele karıştırarak ve bu karıştırma sayesinde bağımlılıkları yok edip model performansındaki değişikliği ölçerek, ilgili değişkenin modelin tahmin doğruluğuna etki eden ağırlığını ve sapmasını belirler.

Yandaki grafik klasik feature importance çıktılarına benzesede karar ağaçlarından gelen bilgi kazançları gibi hesaplamaları olmayıp değişkenlerin değerleri belirli oranlarda karıştırılıp hesaplandığı için daha kolay yorumlanabilir ve daha hassas çıkarımlar yapılabilmektedir. Bu grafik accuracy metriği ile oluşturulmuş olup Python'da farklı metriklerlede değerlendirme yapılabilmektedir.Grafikte en yüksek ağırlığa sahip değişkenin minimum hasar kategorisine sahip değişken olduğu görülmektedir.Sigortalının hobi kategorisininde ağırlığı yüksek olsa bile sapmasının fazla olması değişkenin öneminin veride yapılacak değişikliklere karşı hassas tepkiler verebileceğini işaret eder.

Weight	Feature
0.1176 ± 0.0070	incident_severity_Minor Damage
0.1116 ± 0.0116	incident_severity_Total Loss
0.0616 ± 0.0035	insured_hobbies_chess
0.0366 ± 0.0081	insured_hobbies_cross-fit
0.0196 ± 0.0079	insured_zip
0.0116 ± 0.0055	total_claim_amount
0.0046 ± 0.0027	property_claim
0.0016 ± 0.0016	auto_year
0.0006 ± 0.0010	policy_annual_premium
0.0006 ± 0.0010	incident_hour_of_the_day
0.0002 ± 0.0008	months_as_customer
0.0002 ± 0.0008	injury_claim
0 ± 0.0000	insured_occupation_farming-fishing
0 ± 0.0000	insured_occupation_machine-op-inspct
0 ± 0.0000	insured_occupation_prof-specialty
0 ± 0.0000	insured_occupation_protective-serv
0 ± 0.0000	insured_occupation_sales
0 ± 0.0000	insured_occupation_tech-support
0 ± 0.0000	insured_occupation_other-service
0 ± 0.0000	insured_occupation_priv-house-serv

85 more

5.Anchors Yöntemi

Anchors, belirli bir tahminin neden yapıldığını açıklamak için , o tahminin geçerli olduğu yerel kurallar oluşturan bir yöntemdir. Bu kurallar yerel olsa bile kapsamının belirli bir oranda olması beklenir ki veri genelinde yorumlarda yapılabilsin. Bu kuralları verileri diskretizasyon yüzdelikleri kullanarak aralıklara böler ve değişkenlerin bu aralıklardaki tepkilerini analiz ederek oluşturur. Kuralların güvenirliliğini ise belirli eşik değerler koyarak sağlar.

```
Anchor: incident_severity_Minor Damage > 0.00
Precision: 0.90
Coverage: 0.35
```

Yukarıdaki çıktıda ortaya çıkan basit kurala göre Doğruluğu oldukça yüksek ve kapsam değeri Verinin %35'ine tekabül etmektedir.

```
Anchor: incident_severity_Total Loss > 0.00 AND
insured_hobbies_chess <= 0.00 AND
insured_hobbies_cross-fit <= 0.00
Precision: 0.97
Coverage: 0.25
```

Yukarıdaki grafik kapsamı tüm verinin %25ine denk gelirken kuralın doğruluk değeri %97'dir.

Yandaki kural diğerlerine göre daha karmaşık, doğruluk değeri oldukça yüksek ancak kapsamı yani güvenirliliği oldukça düşüktür. Satır bazında oluşturulan bu kurallar arttırılarak veya spesifik kurallar aranarak model daha detaylı yorumlanabilir.

```
Anchor: incident_severity_Minor Damage <= 0.00 AND
vehicle_claim > 0.68 AND
incident_severity_Total Loss <= 0.00 AND
incident_hour_of_the_day <= 0.05 AND
insured_hobbies_camping <= 0.00 AND
auto_year > -0.85 AND
total_claim_amount > -0.41 AND
0.00 < policy_annual_premium <= 0.65 AND
insured_zip > -0.74 AND
property_claim > -0.13 AND
witnesses > -0.44
Precision: 0.96
Coverage: 0.00
```

Sonuç olarak, global açıklamalardaki hedef değişken üzerinde en çok etkisi olan değişkenlere (örneğin, yıllık prim miktarı, hasar derecesi, sigortalının yaşı, sigortalının hobisi, araç hasar tazminatı gibi) feature engineering işlemleri uygulanarak model başarısı arttırılabilir. Aynı işlem rare encoding mantığı ile hedef değişken üzerinde etkisi az olan değişkenlere de yapılabilir.

Yerel açıklamalar sayesinde iş biriminin istediği değişkenlerin değerleri çeşitlendirilerek modelin yaptığı tahminin nasıl yaptığı anlamlandırılabilir. Yeni eklenecek değişkenler ile yeni model kurulduktan sonra baz alınan modelin performansına göre geniş veya dar hiperparametre aralıkları ile başarı metrikleri arttırılabilir.

KAYNAKÇA

[1] Interpretable Machine Learning (christophm.github.io)
[2] 4.1. Partial Dependence and Individual Conditional Expectation plots — scikit-learn 1.5.0 documentation
[3]Hands on Explainable AI (XAI) with Python by Denis ROTHMAN

İSTATİSTİK BÖLÜMÜ

Telekom Operatörleri İçin Churn Tahmini ve Müşteri Davranış Analizi

UMUT GÖRKEM KARAOĞLU 20023604

Danışman: Prof. Dr. Ali Hakan BÜYÜKLÜ

ÖZET

Bu çalışmada, telekomünikasyon operatörlerinin müşterilerinin churn analizi gerçekleştirildi. Rassal Ormanlar modeli kullanılarak churn tahmini yapıldı ve modelin doğruluğu, karışıklık matrisi, sınıflandırma raporu ve ROC-AUC skoru değerlendirildi. Hipotez testlerinde, Senior Citizen (60 yaş üstü) olan müşterilerin churn oranlarını ki-kare ile analiz edildi ve Month-to-month sözleşmesi, ek hizmet kullanımı, partner ve dependents durumu, tenure süresi gibi faktörlerin churn üzerindeki etkilerini chi-square, Mann-Whitney U ve Kruskal-Wallis testleri ile incelendi. Verinin anlaşılmasın, yorumlanmasın ve öneri sunulabilmesinin kolaylaşması adına bu analizler çeşitli grafiklerle görselleştirildi.

CHURN ANALİZİ

CHURN NEDİR?

Churn, bir müşterinin bir hizmeti veya ürünü kullanmayı bırakması anlamına gelir. Telekomünikasyon sektöründe, churn, müşterinin hizmet sağlayıcısını değiştirmesi veya hizmetini tamamen iptal etmesi olarak tanımlanabilir. Churn analizi, işletmelerin müşteri kaybını azaltmaya yönelik stratejiler geliştirmesine yardımcı olur.

VERİ SETİ ve KULLANILABİLİR YÖNTEMLER

7043 kullanıcının Telekom bilgilerini içeren verideki değişkenler: customerID, gender, senior citizen, partner, dependents, tenure, phone service, multiple lines, internet service, online security, online backup, device protection, tech support, streaming TV, streaming movies, contract, paperless billing, payment method, monthly charges, total charges, Churn'dür.
Lojistik Regresyon (Logistic Regression), Karar Ağaçları (Decision Trees), Rastgele Ormanlar (Random Forests), Destek Vektör Makineleri (Support Vector Machines), K-En Yakın Komşu (K-Nearest Neighbors), Yapay Sinir Ağları (Artificial Neural Networks), XGBoost yöntemleri modellenerek karşılaştırılmıştır. Bunun sonucunda en uygun model Random Forests olmuştur.

```
En iyi model: Random Forest
Accuracy: 0.801277501774308
ROC-AUC Score: 0.8484245448052417

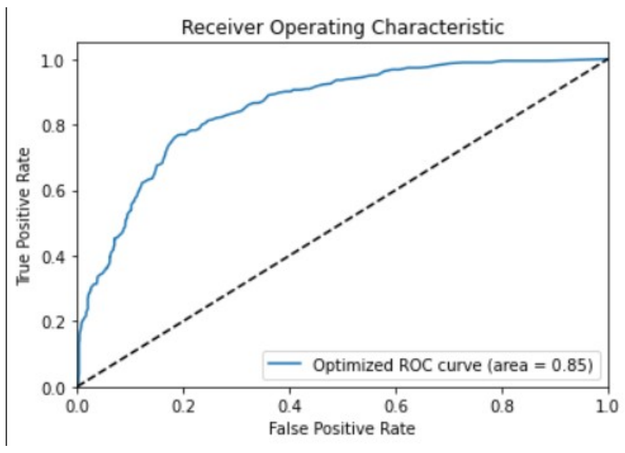
Best Parameters: {'n_estimators': 200}
Optimized Random Forest ROC-AUC Score: 0.8526879522187832
```

ROC-AUC skorları karşılaştırılmış olup en yüksek skor yandaki gibi bulunmuştur. Ardından en iyi model Random Forest Classifier optimize edilerek model tekrar test edilmiştir. Bunun sonucunda ROC-AUC skoru daha yüksek bulunmuştur.

RASSAL ORMANLAR

Random Forest, makine öğrenimi alanında yaygın olarak kullanılan bir yöntem olup, birçok karar ağacının bir araya gelmesiyle oluşan bir yöntemdir. Her bir karar ağacı, veri setinin rastgele alt kümeleri üzerinde eğitilir ve bu ağaçlar bir araya getirilerek daha güçlü ve kararlı bir model oluşturulur. Bu yöntem, sınıflandırma ve regresyon problemlerinde yaygın olarak kullanılır ve aşırı uyum (overfitting) problemini azaltarak daha iyi genelleme yapabilen modeller elde edilmesini sağlar.Bu çalışmada, Random Forest algoritması müşteri churn analizi yapmak için kullanılmıştır. Amacımız, hangi müşterilerin hizmeti bırakma olasılığının daha yüksek olduğunu tahmin etmektir. Modelin performansı, ROC-AUC eğrisi, precision , recall , ve F1-Score gibi metriklerle değerlendirilmiştir.

RASSAL ORMANLAR UYGULAMASI ve DEĞERLENDİRİLMESİ

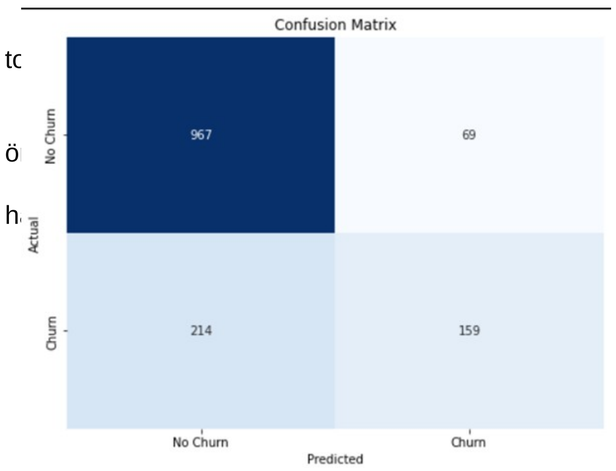


ROC eğrisi şekildeki gibidir. Eğri ne kadar sola ve yukarı yakınsa tahminleme o kadar iyidir.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

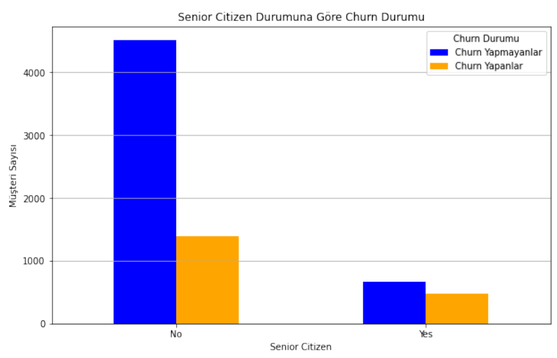


Precision (Kesinlik): Doğru pozitif tahminlerin, pozitif tahminlerin oranıdır.
Recall (Duyarlılık): Doğru pozitif tahminlerin, gerçek pozitif toplamına oranıdır.
F1-score: Precision ve recall'un ortalamasıdır.

```
Precision: 0.6973684210526315
Recall: 0.4262734584450402
F1-Score: 0.529118136439268
```

Bu sonuçlara göre modelin orta-iyi düzeyde tahmin yapabildiği gözlemlenebilir.

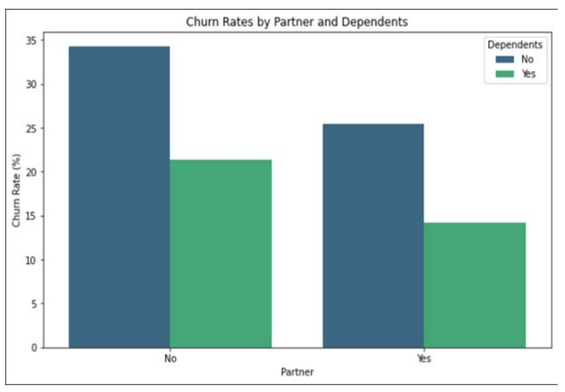
HİPOTEZLER



H0: Senior Citizen değeri 1 olan müşterilerin churn yapma ihtimali daha düşüktür

```
Chi-Square Test Statistic: 1153.9716611093477
P-Value: 6.147585925549194e-253
```

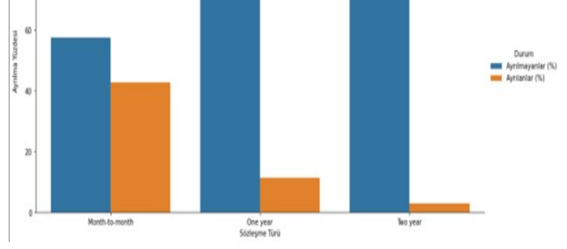
Senior Citizen değeri 1 olan müşterilerin churn yapma ihtimali istatistiksel olarak anlamlı bir fark göstermektedir.



H0: Partneri veya ona bağlı birinin churn yapma olasılığı daha düşüktür.

```
T-Statistic: -13.611788407952671
P-Value: 1.143875223735913e-41
```

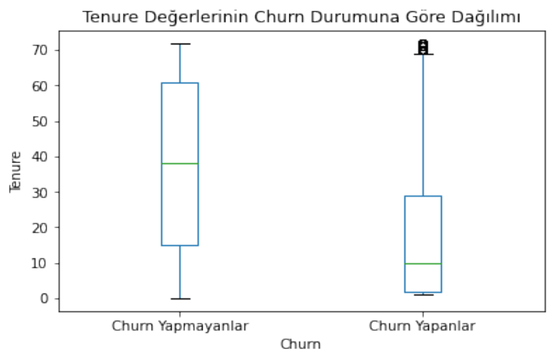
Partner ve dependents olan müşterilerin churn yapma olasılığı istatistiksel olarak anlamlı şekilde farklıdır



H0: Month-to-month kullanıcılarının churn ihtimali diğer kullanıcılarla aynıdır.

```
Chi-Square Test Statistic: 1153.9716611093477
P-Value: 6.147585925549194e-253
```

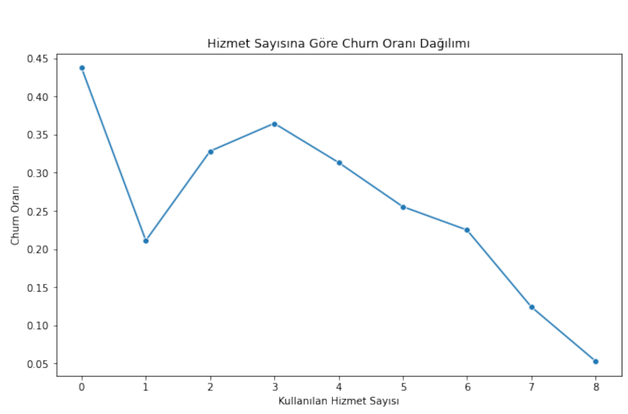
Month-to-Month sözleşmesi olan kullanıcıların churn yapma ihtimali istatistiksel olarak anlamlı bir şekilde daha yüksektir.



H0:Tenure değeri yüksek olan kullanıcıların churn yapma ihtimali daha düşüktür.

```
Kruskal-Wallis Test Statistic: 948.7996915155916
P-Value: 2.4191401818659714e-208
```

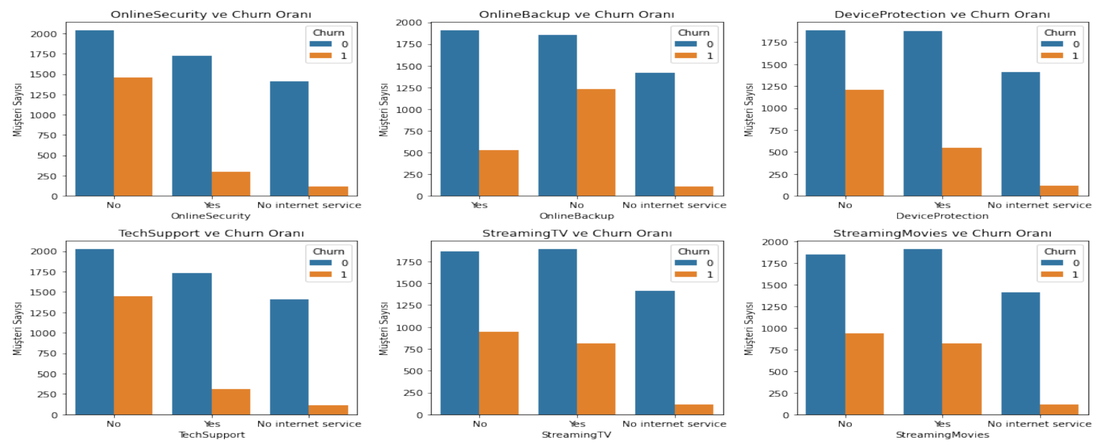
Tenure değeri yüksek olan kullanıcıların churn yapma ihtimali istatistiksel olarak anlamlı bir fark göstermektedir.



H0: Ek hizmet kullanıcılarının churn yapma ihtimali düşüktür

```
Mann-Whitney U Test Statistic: 5753171.0
P-Value: 3.742355760918728e-11
```

Ek hizmet kullanan müşterilerin churn yapma ihtimali istatistiksel olarak anlamlı bir fark göstermektedir.



KAYNAKÇA

[1] https://www.kaggle.com/datasets/datazng/telecom-company-churn-rate-call-center-data/data
[2] https://www.pythonkod.com/
[3] https://medium.com/
[4] https://www.youtube.com/watch?v=vXCjkAYy2y4
[5] https://matplotlib.org/cheatsheets/
[6] https://pandas.pydata.org/

İSTATİSTİK BÖLÜMÜ

Banka Tercihini Etkileyen Faktörlerin Belirlenmesi

Usluhan USLU

Danışman: Prof. Dr. Doğan YILDIZ

ÖZET

Bu çalışma, tüketicilerin banka tercihlerini belirleyen faktörleri araştırmayı amaçlamaktadır. Bankacılık sektöründe hızlı değişim ve artan rekabet, tüketicilerin banka seçimini etkileyen çeşitli dinamik faktörlerin ortaya çıkmasına neden olmaktadır. Bu tez, bu faktörlerin analizine odaklanarak banka seçiminin üzerindeki etkilerini incelemektedir. Çalışma kapsamında birçok kaynak incelenmiş ve elde edilen veriler detaylı bir şekilde analiz edilmiştir. Hem akademik literatür hem de pazar araştırmalarından yararlanılarak kapsamlı bir analiz yapılmıştır.

Banka Tercihini Etkileyen Faktörlerin Belirlenmesi

Bankacılık sektörü, hızlı bir değişim ve rekabet ortamında faaliyet göstermektedir. Bu sektördeki oyuncular, tüketicilerin banka tercihlerini etkileyen faktörleri anlamak ve bu dinamikleri yönetmek için sürekli olarak çaba göstermektedir. Bu tez, tüketicilerin banka tercihlerini belirleyen faktörlerin tespitini ve analizini amaçlamaktadır.

ARAŞTIRMA HEDEFLERİ

Bu çalışma, banka tercihini etkileyen faktörleri sistemli bir şekilde belirlemeyi hedeflemektedir. Bunun yanı sıra, bu faktörlerin banka tercihine olan etkisini anlamak ve değerlendirmek de araştırmanın amacı arasındadır.

BANKA

Banka, müşterilerin finansal ihtiyaçlarını karşılayan, para ve diğer finansal varlıkları kabul eden, saklayan, kredi veren, ödeme hizmetleri sağlayan ve yatırım yapmalarına yardımcı olan finansal kurumlardır. Bankalar, bireylerin, işletmelerin ve devletlerin mali işlemlerini kolaylaştırır ve ekonomik faaliyetlerin sağlıklı bir şekilde yürütülmesine katkıda bulunurlar.

Bankacılık ise, bu finansal kurumların faaliyetlerini kapsayan geniş bir sektördür. Bankacılık, paranın toplanması, saklanması, işlenmesi, yatırılması ve dağıtılması gibi bir dizi hizmeti içerir. Bankacılık, ekonomik büyümeyi destekleyen ve finansal sistemlerin sağlıklı bir şekilde işlemesini sağlayan önemli bir bileşendir. Bu sektör müşterilerin finansal ihtiyaçlarını karşılamak için çeşitli araçlar ve hizmetler sunar.

DÜNYA'DA BANKACILIK

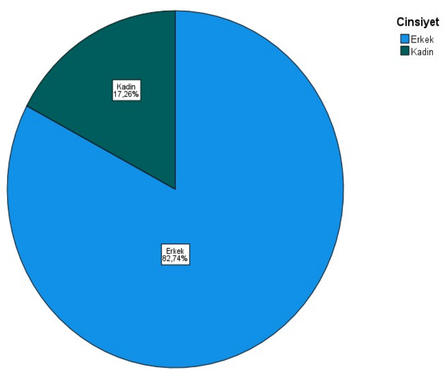
Dünya genelinde bankacılık, finansal işlemlerin yönetimi ve para ile ilgili hizmetlerin sunulmasıyla ilgilidir. Bankalar, bireylerin, şirketlerin ve hükümetlerin para birimlerini saklayabilecekleri, borç alabilecekleri, yatırım yapabilecekleri ve diğer finansal işlemleri gerçekleştirebilecekleri kurumlardır. Bankacılık, mevduat kabul etme, kredi sağlama, yatırım yapma, ödeme hizmetleri ve risk yönetimi gibi çeşitli alanları kapsar.

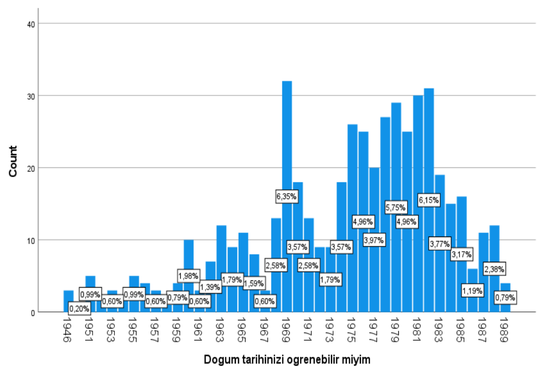
TÜRKİYE'DE BANKACILIK

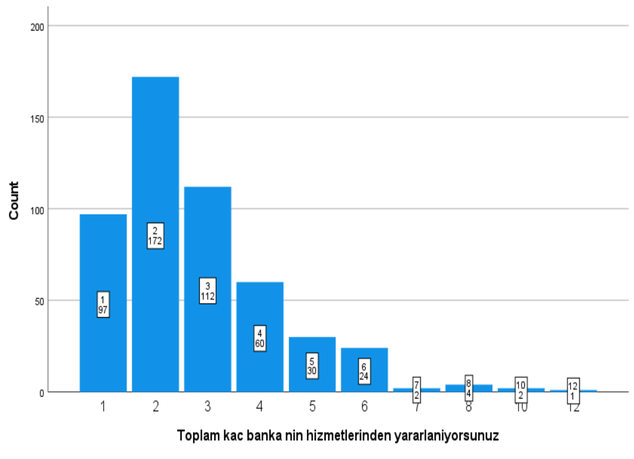
Türk bankacılık sektörü, ülkenin ekonomik yapısı, yasal düzenlemeleri ve finansal kurumlarının özelliklerine göre şekillenir. Türkiye'de bankacılık, kamu bankaları, özel bankalar, yatırım bankaları ve kalkınma ve katılım bankaları gibi çeşitli türdeki finansal kurumları içerir. Türk bankacılık sektörü, bireylere ve işletmelere kredi sağlama, mevduat kabul etme, ödeme hizmetleri sunma ve yatırım yapma gibi hizmetler sunar. Türk bankacılık sektörü, ülkenin ekonomik büyümesine ve finansal istikrarına önemli katkılar sağlar.

ARAŞTIRMA VERİLERİ

Aşağıdaki grafik, yaş, cinsiyet ve kaç banka ile çalışıldığını açıklamaktadır. Grafikte, yaş gruplarına ve cinsiyete göre banka çalışma tercihleri gösterilmiştir. Ayrıca her yaş ve cinsiyet grubunda kaç banka ile çalışıldığını gösteren dağılımı da sunmaktadır.







UYGULANAN GÜVENİRLİK TESTİLERİ

Güvenirlik testi, bir ölçüm aracının ne kadar güvenilir olduğunu belirlemek için yapılan bir değerlendirme sürecidir. Bu test, ölçüm aracının tekrarlanabilirliği ve istikrarını değerlendirir. Güvenirlik yüksekse sonuçlara daha fazla güvenebiliriz.

Cronbach Alfa Güvenilirlik Katsayısı	Yorum
≥ 0,9	Mükemmel
0,7≤ < 0,9	İyi
0,6≤ < 0,7	Kabul edilebilir
0,5≤ < 0,6	Zayıf
< 0,5	Kabul edilemez

Güvenilirlik İstatistikleri

Cronbach's Alfa	Öğe
,537	14

Ölçek İstatistikleri

Ortalama	Varyans	Standart Sapma	Öğe Sayısı
58,98	1423,908	37,735	14

UYGULANAN Kİ-KARE TESTLERİ

Parametrik olmayan istatistiksel yöntemlerden ki-kare bağımsızlık testinde ele alınan değişkenlerin farklı ölçütlere ya da belirli bir amaca göre iki ya da daha çok yönlü çapraz tablo biçiminde sınıflandırılması halinde değişkenlerin belirlenen özellikleri arasında ilişkinin olup olmadığı test edilir. Bir diğer tanımla ki- kare bağımsızlık testi, 2 x 2 ya da r x c formatındaki çapraz tablolarda gözlenen frekansların, teorik frekanslara benzerliğini test etmeyi amaçlar.

GELİR GURUPLARI

Ki-Kare Testi

	Değer	sd	Asimptotik Önem (2 taraflı)
Pearson Ki-Kare	26,869[a]	12	,008
Olasılık Oranı	21,769	12	,040
Doğrusal-Doğrusal İlişki	5,111	1	,024
Geçerli Vaka Sayısı	495		

A. 6 hücre (%30,0) 5'ten az beklenen sayıya sahiptir. Minimum beklenen sayı ,53'tür.

Ki-Kare Testi

	Değer	sd	Asimptotik Önem (2 taraflı)
Pearson Ki-Kare	16,299[a]	6	,012
Olasılık Oranı	11,392	6	,077
Doğrusal-Doğrusal İlişki	3,589	1	,058
Geçerli Vaka Sayısı	495		

A. 4 hücre (%33,3) 5'ten az beklenen sayıya sahiptir. Minimum beklenen sayı ,09'dur.

SONUÇ

Bu çalışma, bankacılık sektöründeki müşteri davranışlarını ve tercihlerini anlamak amacıyla yapılmıştır. Araştırma, bankacılık hizmetlerine olan güven düzeyi, tercihler ve müşteri memnuniyeti üzerine odaklanmıştır ve bu faktörlerin çeşitli demografik gruplar üzerindeki etkilerini incelemiştir.

Gelir seviyeleri, evde internet bağlantısı olup olmama durumu, cinsiyet ve yaş gibi demografik faktörler, bankacılık tercihlerini önemli ölçüde etkilemektedir.

Bankaların teknoloji hizmet kanallarına duyulan güven ve memnuniyet, müşteri bağlılığı üzerinde belirleyici bir etkiye sahiptir.

Ölçüm araçlarının güvenirliği incelendiğinde, bazı ölçüm araçlarının iç tutarlığının zayıf olduğu ve bu durumun düzeltilmesi gerektiği belirlenmiştir.

Bu sonuçlar, bankacılık sektöründeki hizmet sunumunun ve pazarlama stratejilerinin, müşteri profili ve tercihlerine dayalı olarak tasarlanması gerektiğini vurgulamaktadır. Bankalar, müşteri memnuniyetini artırmak ve rekabet avantajı elde etmek için, müşteri davranışlarını ve tercihlerini dikkate alarak stratejilerini belirlemelidir. Ayrıca, ölçüm araçlarının güvenirliğinin ve iç tutarlılığının artırılması, daha doğru sonuçlar elde edilmesi açısından önemlidir.

Sonuç olarak, bu tezin bulguları bankacılık sektöründe hizmet kalitesini artırmak ve müşteri memnuniyetini en üst düzeye çıkarmak adına önemli bir katkı sunmaktadır. Bankalar, elde edilen bu bulguları dikkate alarak müşteri odaklı stratejiler geliştirmeli ve hizmetlerini buna göre şekillendirmelidir..

KAYNAKÇA

[1] Bodie, Z., Kane, A., & Marcus, A. J. (2014). Investments. McGraw-Hill Education.

[2] Öztürk, E., & Çiftçi, İ. (2014). Türkiye'de Bankacılık Sektörü: Reformlar ve Gelişmeler. İstanbul: İletişim Yayınları.

[3] Berger, A. N., Molyneux, P., & Wilson, J. O. S. (2015). The Oxford Handbook of Banking. Oxford University Press.

İSTATİSTİK BÖLÜMÜ

MARKOWITZ PORTFÖY SEÇİM MODELİ VE HİSSE SENETLERİ ÜZERİNE BİR UYGULAMA

Yağmur AKASLAN 20023065

Danışman: Doç. Dr. Atıf Ahmet EVREN

ÖZET

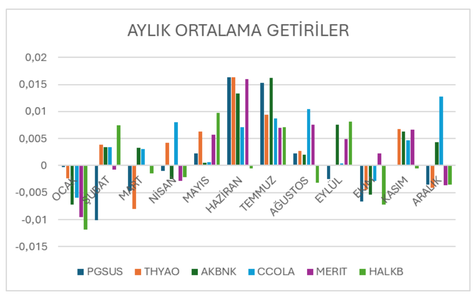
Yatırım dünyasının pek çok alanında olduğu gibi portföy seçimi ve risk yönetiminde de kararların verilmesi ve uygulamaya geçirilmesi zorlu ve uzun süreçleri gerektirmektedir. Yatırımcıların ve karar vericilerin isteklerine uygun olan portföylerin oluşturulmasında karşımıza sınırsız sayıda ve karmaşık yapıda olasılık çıkmaktadır. Bu projede Markowitz karesel programlama ile portföy seçim modelinin farklı hisse senetleri üzerinde uygulamaları yapılmıştır. Markowitz karesel programlama modeli ile portföy oluşturmak için MS Excel programı içinde yer alan solver eklentisi kullanılmıştır. Bu uygulama kapsamında Markowitz karesel programlama modeli yardımıyla 6 hisse senedi arasından en iyi risk – getiri dengesini verecek optimum portföylere ulaşılmaya çalışılmıştır.

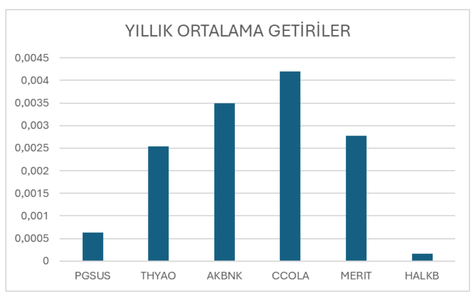
VERİLERİN GRAFİKLEŞTİRİLMESİ

Hisse senetlerinin getirilerine bakarken verilerin görselleştirilmesi bizlere daha kolay yorumlama imkanı sağlar. Grafikler, yatırımcıların belli trendleri, dalgalanmaları ve farklılıklarını tespit etmesine kolaylık sağlamaktadır.

Bu uygulamada her bir hisse senedinin günlük kapanış fiyatlarından günlük getiriler hesaplanmıştır. Hesaplanan günlük getiriler ile aylık getiriler hesaplanmış ve Excel kullanılarak veriler grafik haline getirilmiştir. Aylık hesaplanan ortalama getirilerin ortalamaları alınarak yıllık ortalama getiriler hesaplanmıştır. Hesaplanan yıllık ortalama getiriler Excel kullanılarak grafik halinde görselleştirilmiştir.

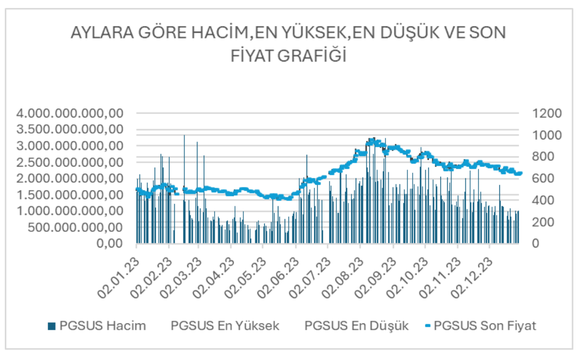
Önceki grafikte hesaplanan her bir hisse senedinin aylık ortalama getirisinin ortalamaları alınarak yıllık ortalama getiriler hesaplanmıştır. Daha sonra hesaplanan yıllık ortalama getiriler Excel kullanılarak grafik halinde görselleştirilmiştir.

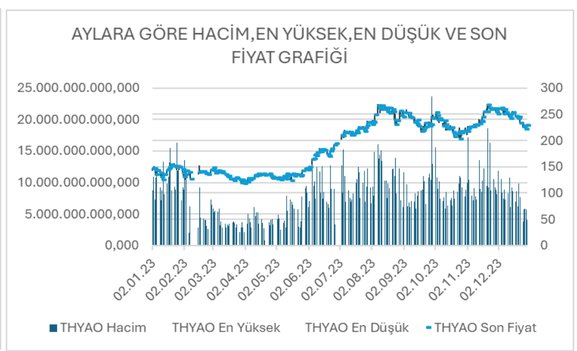


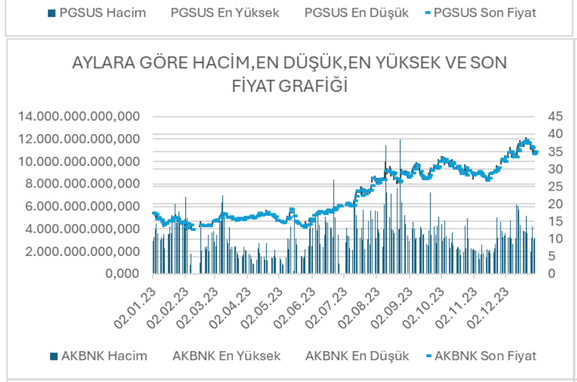


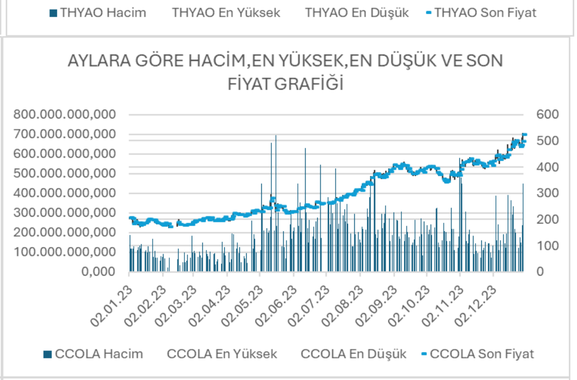
Yatırımcılar, Markowitz'e göre aynı getiri düzeyinde en düşük riski veya aynı risk düzeyinde en yüksek getiriyi elde etmeyi hedeflemektedir. Bu grafikler ile çeşitli hisse senetlerinin(PGSUS, THYAO, AKBNK, CCOLA, MERIT, HALKB) 2023 yılı için aylık ve yıllık ortalama getirilerini görebilmekteyiz.

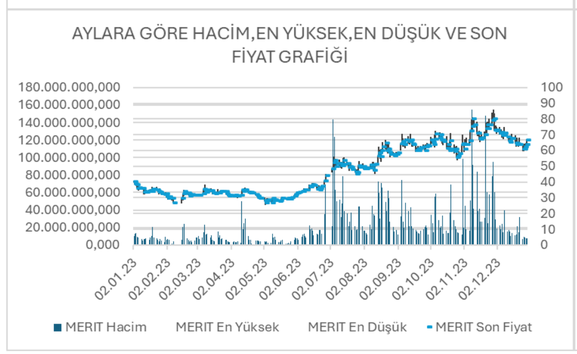
2023 yılı için aylık ve yıllık ortalama getirilere bakıldığında yüksek getiri bakımından CCOLA ve AKBNK hisselerinin öne çıktığı görülür. Tutarlılık açısından bakıldığında ise AKBNK, PGSUS VE THYAO hisselerinin öne çıkmaktadır. AKBNK hissesi yıl boyunca istikrarlıdır. PGSUS ve THYAO hisseleri ise yüksek dalgalanmalar göstermiştir fakat genel olarak pozitif getiri sağlamaktadır.

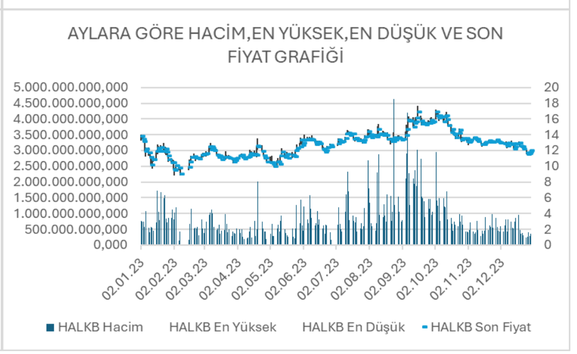












Bu uygulamada PGSUS, THYAO, AKBNK, CCOLA, MERIT, HALKB hisse senetlerinin hacim, en yüksek fiyat, en düşük fiyat ve kapanış fiyatları aylık olarak gösterilmiştir. Uygulamayı yaparken Excel'de veriler işlem hacmi- yüksek-düşüş-kapanış olarak sıralanmıştır. Excel'in "Tüm Grafikler" sekmesi altındaki "Hisse Senedi" grafiği seçeneği kullanılarak grafik oluşturulmuştur.

İşlem hacimleri bize fiyat trendlerinin kuvveti hakkında fikir verir. Yukarı trendde işlem hacmi de yüksekse yukarı trend sağlıklı bir çıkış içindedir diyebiliriz. Bunun tersi durumda yani yukarı trendde işlem hacmi zayıflıyorsa hacmin yükselişinin teyit edilmediğinden bahsedilir ve bu çıkış sağlıklı değildir. Düşüş trendinde işlem hacminin yüksek olması düşüşün kuvvetli olduğunun göstergesidir.

KORELASYON MATRİSİ

Markowitz, çalışmalarında bir portföyü çeşitlendirmede esas olanın sadece portföyde bulunan hisse senetlerinin sayısı olmadığını, bunların getirileri arasındaki korelasyon olduğunu göstermiştir. Eğer korelasyon yüksekse çeşitlendirme yapmak mümkün olmaz, aksine eğer korelasyon düşükse çeşitlendirme yapılabilir ve risk önemli ölçüde düşer.

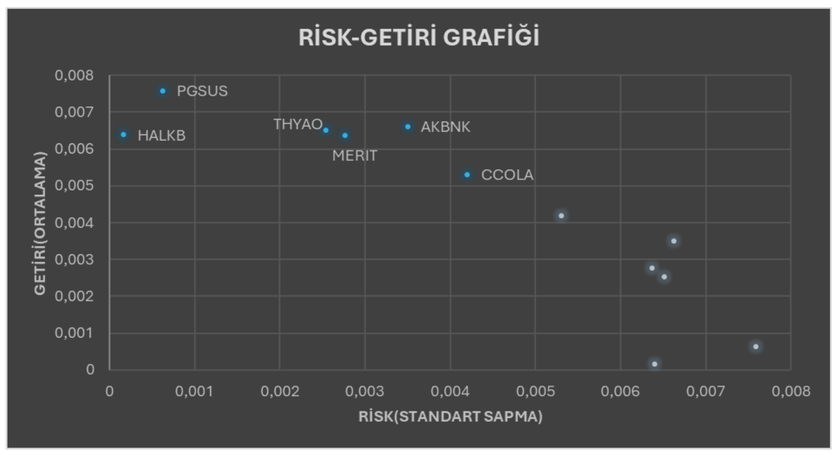
	PGSUS	THYAO	AKBNK	CCOLA	MERIT	HALKB
PGSUS	1	0,774493	0,343596	0,382156	0,349686	0,377538
THYAO	0,774493	1	0,472704	0,40821	0,37078	0,523064
AKBNK	0,343596	0,472704	1	0,388241	0,269946	0,666498
CCOLA	0,382156	0,40821	0,388241	1	0,287083	0,336769
MERIT	0,349686	0,37078	0,269946	0,287083	1	0,405585
HALKB	0,377538	0,523064	0,666498	0,336769	0,405585	1

Bu portföye göre; CCOLA, AKBNK, MERIT VE THYAO hisse senetleri yüksek getirisi ve düşük korelasyonları dolayısıyla riskin minimize edilmesini ve getirinin optimize edilmesini sağlamaktadır. Korelasyona göre portföy için bu dört hisse senedini seçmek daha doğrudur.

RİSK-GETİRİ GRAFİĞİ

Hesaplanan risk (standart sapma) ve beklenen getiri verileri Excel'de dağılım grafiği üzerinde birleştirilerek risk-getiri haritası oluşturulmuştur. Yapılan grafikte, X eksenine standart sapma (risk) ve Y eksenine ortalama getiri atanmıştır.

Yatırımları değerlendirirken, daha yüksek getiri beklenirken, daha düşük risk aranmaktadır. Yatırımcı açısından karar vermek, portföy riskinin ve getirisinin ölçülmesini gerektirmektedir. Yatırımcı için yatırım fırsatları içerisinden karar vermek yalnızca tek tek menkul kıymetler arasından seçim yapmak değildir. Çünkü, farklı oranlarda menkul kıymet bileşimleri oluşturarak çok sayıda portföy meydana getirilebilir. Fırsatların sayısı arttıkça sorun karmaşık bir hal almakta ve portföy kuramı ortaya çıkmaktadır. Ancak, yatırımcı açısından önemli olan, optimal portföyün oluşturulmasıdır. Bunun için portföyün risk ve getirisinin hesaplanması gerekmektedir.



SONUÇ

Uygulamada 6 hisse senedi ve endeks için toplam 12 şer aylık 72 getiri değeri elde edilmiştir. Ortalama getiriler elde edildikten sonra 6 hisse senedine ait kovaryans değerleri bulunmuştur, Daha sonra kovaryans değerlerin transpozeleri hesaplanmış ve varyans - kovaryans matrisi hesaplanmıştır. Bulunan değerler ile modelin çözümlenmesi sonucunda risk ve getiri değerlerine sahip hisse senetleri (PGSUS,THYAO,AKBNK,CCOLA,MERIT,HALKB) elde edilmiştir. Bu sonuçlara göre yorumlamalar yapılmıştır. CCOLA'nın en iyi risk-getiri oranına sahip olduğu görülmektedir. Markowitz karesel programlama modeline göre en iyi sonuç CCOLA hissesi olarak bulunmuştur.

KAYNAKÇA

[1] CEYLAN, Ali ;KORKMAZ, Turhan, a.g.e, s.375.

[2] Francisco Blanco Ramos, Máximo Ferrando Bolado, María Fuencisla Martínez Lobato, 2015, Teoría de la inversión [Yatırım Teorisi]. Pirámide, 2015, Madrid.

[3] Gazi Üniv. Müh. Mim. Fak. Der. Cilt 17, No 1, 43-58, 2002 BORSADA GÖSTERGELERLE TEKNİK ANALİZ İÇİN BİR KARAR DESTEK SİSTEMİ Tahsin ÇETİNYOKUŞ ve Hadi GÖKÇEN, s.7.

İSTATİSTİK BÖLÜMÜ

Makine Öğrenmesi ile Banka Kredisi Başvuru Sonucu Tahmini

Zeynep SAMUR 20023006

Danışman: Doç. Dr. Fatma Sevinç KURNAZ

ÖZET

Birçok banka mevduat olarak aldığı parayı kredi olarak kullandırarak para kazanır. Kredilerin geri ödenmesiyle aldığı mevduatları geri öder. Eğer krediler zamanında ödenmezse ya da hiç ödenmezse bankalar ciddi ekonomik sıkıntı yaşamaya başlamaktadırlar. Bu sebeple işleyişin doğru şekilde devam edebilmesi için. Kredi verilen kişilerin doğru seçilmesi gerekmektedir. Bu çalışmada bankaların müşterilerine vereceği krediler için , kredi talep eden müşterinin durum analizi yapılarak kredi başvurusunun sonucu makine öğrenmesi algoritmaları ile analiz edilmiştir. Veri seti üzerinde kullanılan algoritmalara uygun olarak düzenlemeler yapılmış, eğitilen modeller ile web arayüzü üzerinden girilen müşteri verisine göre müşterinin kredi başvuru sonucu tahmin edilmiştir.

Bankaların İşleyişi ve Gelişen Teknolojiler

Birçok banka mevduat olarak topladıkları paraları kredi olarak kullandırarak para kazanır. Bankalar için kredilerin geri ödenmesi hayati bir durumdur çünkü mevduat faizini, kredileri ödeyen kişiler sayesinde geri ödemektedirler. Bu esnada yaşayacakları sıkıntılar bankanın kapatılmasına kadar giden sonuçlara sebebiyet vermektedir. Bu sebeple bankaların kredi verecekleri kişileri özenle seçmeleri gerekmektedir insan hatasını ortadan kaldırmak için de gelişen teknolojilerden destek almaları çok önemlidir.

Makine Öğrenmesi Algoritmaları

Makine öğrenmesi algoritmaları, karmaşık veri kümelerinin keşfedilmesi, analiz edilmesi ve veriden anlam çıkarılmasında kullanılan modellerdir.

Çalışmada Kullanılan Algoritmalar:

Random Forest: Random Forest (RF), rastgele oluşturulmuş birden çok karar ağacının çıktısını birleştiren, regresyon ve sınıflandırma problemlerini çözmek için kullanılan bir makine öğrenme algoritmasıdır.
Logistic Regression: Lojistik regresyon, kategorik bir bağımlı değişkenin olasılığını tahmin etmek için kullanılan denetimli bir makine öğrenimi sınıflandırma algoritmasıdır. Birçok banka mevduat olarak topladıkları paraları kredi olarak kullandırarak para kazanır. Bankalar için kredilerin geri ödenmesi hayati bir durumdur çünkü mevduat faizini kredileri ödeyen kişiler sayesinde geri ödeyecektir.
Decision Tree: Karar ağaçları**,** özellik ve hedefe göre karar düğümleri (decision nodes) ve yaprak düğümlerinden (leaf nodes) oluşan ağaç yapısı formunda bir model oluşturan bir sınıflandırma yöntemidir.
Support Vector Machine: Support Vector Machine (Destek Vektör Makineleri diğer adıyla DVM yada SVM) sınıflandırma, regresyon ve aykırı değerleri bulmak için kullanılan denetimli (superwised) bir öğrenme tekniğidir.
Naive Bayes: Naive Bayes sınıflandırıcı bağımsız varsayımlarla Bayes teoremini temel alan olasılıklı bir sınıflayıcıdır.
KNN: K-en yakın komşu (k-nearest neighbors, KNN) algoritması, gözlemlerin birbirlerine olan benzerlikleri üzerinden tahminlerin yapıldığı gözetimli makine öğrenmesi modellerinde regresyon ve sınıflandırma problemlerinde kullanılan bir algoritmadır.

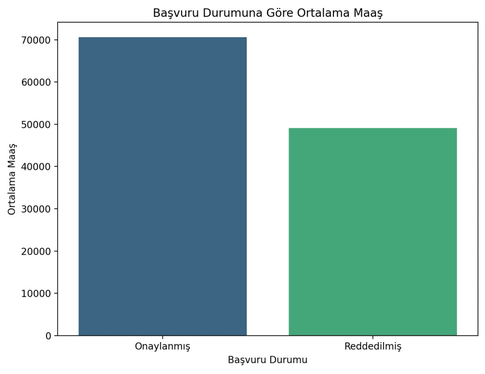
Kullanılan Veri Seti:

Kullanılan veri seti içerisinde müşterilerin yaş, maaş, ev sahipliği durumu, kredi miktarı, kredi faiz oranı ve kredi başvuru sonucu gibi veriler bulunmaktadır. Web arayüzünden erişilen veri bilgisi:

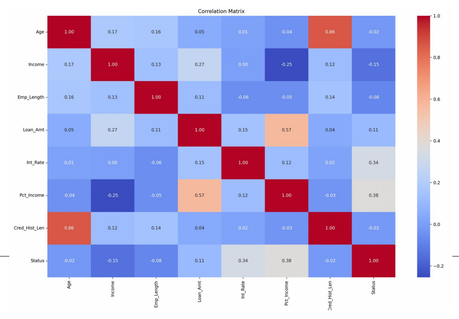
Data Info

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32581 entries, 0 to 32580
Data columns (total 11 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   person_age                  32581 non-null  int64
 1   person_income               32581 non-null  int64
 2   person_home_ownership       32581 non-null  object
 3   person_emp_length           31686 non-null  float64
 4   loan_intent                 32581 non-null  object
 5   loan_amnt                   32581 non-null  int64
 6   loan_int_rate               29465 non-null  float64
 7   loan_status                 32581 non-null  int64
 8   loan_percent_income         32581 non-null  float64
 9   cb_person_default_on_file   32581 non-null  object
 10  cb_person_cred_hist_length  32581 non-null  int64
dtypes: float64(3), int64(5), object(3)
memory usage: 2.7+ MB
```

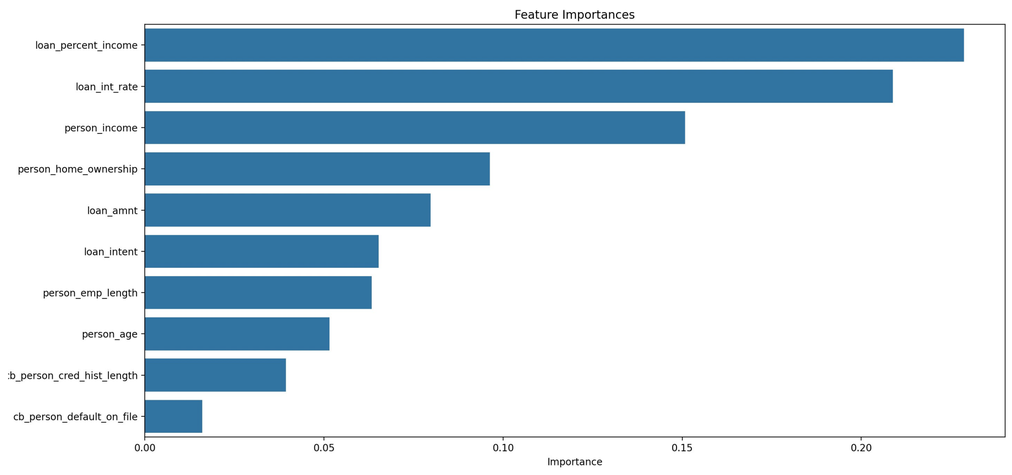
Aşağıdaki grafikte başvurusu onaylanan kullanıcıların ortalama maaşının daha yüksek olduğu görülmektedir.



Sayısal verilerin birbiriyle ilişkisini gösterebilmek için korelasyon matrisi çizilmiştir. «1» değerine yakın olan değişkenler birbiriyle en ilişkili olan değişkenlerdir.



Random Forest modeli için her bir değişkenin loan_status üzerindeki etkisi aşağıdaki grafikte gösterilmiştir.



Feature importance grafiğine göre, kişilerin kredi başvurusunun onaylanmasında veya reddedilmesindeki en büyük etken kredinin yıllık maaşın yüzde kaçı olduğudur.

6 farklı model ile test verisi üzerinde tahmin yapılmış ve yüzdelik olarak doğrulukları hesaplanmıştır. Hesaplanan doğruluk yüzdeleri web ara yüzü üzerinde «Show Model Accuracies» butonuna tıklandığında kullanıcıya gösterilmektedir.

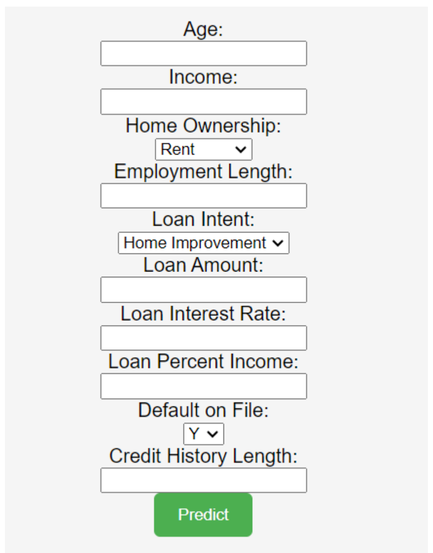
Model Accuracies

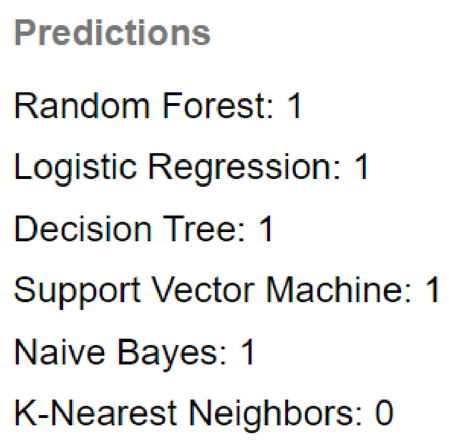
Random Forest: 92.10242171834028%
Logistic Regression: 83.01712170291532%
Decision Tree: 87.5520592318371%
Support Vector Machine: 81.38207619929045%
Naive Bayes: 81.49005090236003%
KNN: 83.29477093937992%

Kullanılan veri seti üzerinde en başarılı tahmin yapan model %92 başarı oranı ile Random Forest algoritması olduğu tespit edilmiştir.

Aynı zamanda web ara yüzü üzerinden kullanıcıdan alınan veri üzerinde de tahmin yapılmaktadır.

Kullanıcıdan veri alınması:
«Predict» butonuna basıldığında her algoritma ile tahmin yapılması ve web ara yüzünde gösterilmesi:





Kullanılan modeller girilen müşteri verisini öğrendiği veri seti üzerinden analiz ederek kredi başvuru sonucunu olumsuz(1) veya olumlu(0) olarak tahmin etmektedir.

KAYNAKÇA

[1] https://www.kaggle.com/datasets/urvishvekariya/credit-risk-assessment/data
[2] https://www.kaggle.com/code/colecola/kredi-tahmini-1
[3] https://www.kaggle.com/code/inoshishidolo/riskassessment-international-data-v3

İSTATİSTİK BÖLÜMÜ

SOSYAL MEDYANIN ÜNİVERSİTE ÖĞRENCİLERİ ÜZERİNDEKİ OLUMSUZ ETKİLERİ

Zeynep TAN - 18023074

Danışman: Doç. Dr. Reşit ÇELİK

ÖZET

Bu çalışma, sosyal medyanın üniversite öğrencileri üzerindeki olumsuz etkilerini incelemektedir. Sosyal medya, öğrencilerin akademik, sosyal ve psikolojik alanlarda sorunlar yaşamasına neden olmaktadır. Aşırı kullanımı, gerçek hayattan kopma, özgüvensizlik, asosyallik, karamsarlık ve depresyon gibi etkiler yaratmaktadır. Bu tez, sosyal medyanın öğrencilerde kaygı ve depresyon gibi psikolojik sorunlara yol açtığını vurgulamaktadır. Sonuç olarak, sosyal medyanın olumsuz etkilerine dikkat çekerek farkındalık yaratmayı amaçlamaktadır.

GİRİŞ

Sosyal medya, kullanıcıların içerik aradığı, paylaştığı ve ürettiği interaktif bir platformdur. İnsanlar burada bilgiye kolayca ulaşabilir, iletişim kurar ve yardım alabilirler. Haber aktarımında sosyal medya, hızlı bilgi sağlayarak geleneksel gazeteciliği geride bırakmış. Eğitimde de internet üzerinden kolay erişim imkanı sunarak yenilikler getirmiştir. Ayrıca, demokratik fikir paylaşımı için uygun bir ortam sağlayarak sanal toplulukların oluşmasına katkıda bulunmuştur. Sosyal medya, her yaştan insanın hayatında merkezi bir rol oynamaya başlamıştır. Ve bu durumun birçok olumlu ve olumsuz yan etkileri olmaya başlamıştır.

ÇALIŞMANIN AMACI

Bu araştırmanın amacı, sosyal medyanın üniversite öğrencileri üzerindeki olumsuz etkilerini incelemektir. Sosyal medyanın aşırı kullanımı, öğrencilerin akademik performansı, sosyal ilişkileri ve psikolojik sağlığı üzerinde ciddi olumsuz etkiler yaratmaktadır. Araştırma, bu etkileri belirleyerek farkındalık yaratmayı ve eğitim kurumları ile ailelere rehberlik edecek stratejiler geliştirmeyi hedeflemektedir.

ÖN BİLGİLER

Bu araştırmanın örneklemi, Türkiye'nin farklı üniversitelerinde öğrenim gören ön lisans, lisans ve yüksek lisans öğrencilerinden oluşmaktadır. Basit tesadüfi örneklem yöntemi kullanılarak seçilen ve açıklayıcı faktör analizi uygulanan 219 öğrenci, bu çalışmanın örneklem grubunu oluşturmuştur. Araştırmadan elde edilen verilerin analizinde IBM SPSS 25.0 programı kullanılmıştır.

BULGULAR

Bağımlı Değişken	Cinsiyet	N	Ort.	SS	t	p
Sosyal Medyanın Olumsuz Etkilerine Yönelik Tutum	Kadın	110	2,8148	,93507	,607	,545
	Erkek	99	2,7399	,83884		

- Bu tablonun sonucu kişilerin sosyal medyanın olumsuz etkilerine yönelik tutumlarının cinsiyete göre farklılık göstermediği, gruplar arasında anlamlı bir farklılık göstermemektedir.

Bağımlı Değişken	Yaş	N	Ort.	SS.	F	p	Post-Hoc
Sosyal Medyanın Olumsuz Etkilerine Yönelik Tutum	a) 18-25 Yaş	118	2,8453	,85621	1,701	,185	
	b) 26-35 Yaş	83	2,7334	,90829			
	c) 36 Yaş ve Üzeri	8	2,2813	1,09941			
	Toplam	209	2,7793	,88945			

- Kişilerin sosyal medyanın olumsuz etkilerine yönelik tutumları yaşa göre farklılık göstermemektedir.

Bağımlı Değişken	Okuduğunuz Şehir	N	Ort.	SS.	F	p	Post-Hoc
Sosyal Medyanın Olumsuz Etkilerine Yönelik Tutum	a) Adana	2	3,0625	,61872	1,160	,325	
	b) Ankara	13	2,6154	1,07464			
	c) Antalya	2	2,1875	,26517			
	d) Edirne	3	1,9583	1,13422			
	e) Eskişehir	15	3,0917	,90197			
	f) İstanbul	149	2,7911	,88166			
	g) İzmir	16	2,6953	,84775			
	h) Kocaeli	7	3,0714	,64895			
	I) Trabzon	2	1,8125	,79550			
	Toplam	209	2,7793	,88945			

-Kişilerin sosyal medyanın olumsuz etkilerine yönelik tutumları okudukları şehire göre farklılık göstermemektedir.

Bağımlı Değişken	Eğitim Düzeyi	N	Ort.	SS.	F	p	Post-Hoc
Sosyal Medyanın Olumsuz Etkilerine Yönelik Tutum	a) Ön Lisans	20	2,6125	,86117	3,255	**,023**	a>d b>d
	b) Lisans	167	2,8496	,85921			
	c) Yüksek Lisans	20	2,5125	1,03388			
	d) Doktora	2	1,2500	,35355			
	Toplam	209	2,7793	,88945			

- Kişilerin sosyal medyanın olumsuz etkilerine yönelik tutumları okudukları eğitim düzeyine göre farklılaşmaktadır.

Bağımlı Değişken	Sosyal medya platformlarını günlük olarak kaç saat kullanıyorsunuz?	N	Ort.	SS.	F	p	Post-Hoc
Sosyal Medyanın Olumsuz Etkilerine Yönelik Tutum	a) 1 saatten az	11	2,7045	,89839	1,641	,181	
	b) 1-2 saat	47	2,6170	,81486			
	c) 2-4 saat	88	2,7358	,85757			
	d) 4 saatten fazla	63	2,9742	,96803			
	Toplam	209	2,7793	,88945			

-Kişilerin sosyal medyanın olumsuz etkilerine yönelik tutumları günlük sosyal medya kullanım süresine göre farklılık göstermemektedir.

Bağımlı Değişken	Sosyal medyayı en çok hangi amaçla kullanıyorsunuz?	N	Ort.	SS.	F	p	Post-Hoc
Sosyal Medyanın Olumsuz Etkilerine Yönelik Tutum	a) Arkadaş edinmek, yeni insanlar tanımak için.	4	1,9688	1,14735	2,030	,076	
	b) Gündemdeki gelişmeleri takip etmek.	57	2,7303	,79437			
	c) Fotoğraf, video ve anılarımı paylaşmak için.	23	2,9130	,89382			
	d) Eğlence odaklı içerikler izleyip kafa dağıtmak için.	103	2,8653	,85170			
	e) Ürünler, markalar hakkında bilgi edinmek için.	6	3,0417	1,02368			
	f) Kişisel gelişim içerikleri okumak ve motive edici içerikler takip etmek.	16	2,3125	1,17260			
	Toplam	209	2,7793	,88945			

Kişilerin sosyal medyanın olumsuz etkilerine yönelik tutumları sosyal medya kullanım amacına göre farklılık göstermemektedir.

Bağımlı Değişken	Sosyal medya etkisiyle ihtiyacınız olmayan gereksiz alışveriş yaptığınızı düşünüyor musunuz?	N	Ort.	SS.	F	p	Post-Hoc
Sosyal Medyanın Olumsuz Etkilerine Yönelik Tutum	a) Evet	53	3,1132	,89512	8,081	**,000**	a>c b>c
	b) Bazen	68	2,8493	,80824			
	c) Hayır	88	2,5241	,87856			
	Toplam	209	2,7793	,88945			

-Kişilerin sosyal medyanın olumsuz etkilerine yönelik tutumları sosyal medyanın etkisiyle gereksiz alışveriş yapma durumuna göre farklılaşmaktadır.

	Kesinlikle Katılmıyorum %	Katılmıyorum %	Kararsızım %	Katılıyorum %	Kesinlikle Katılıyorum %
1. Sosyal medya kullanımı sonrasında uykusuzluk yaşıyorum.	12,9%	**28,2%**	22,5%	25,8%	10,5%
2. Sosyal medya kullanımı beni bazen karamsarlığa itiyor.	11,5%	20,1%	14,8%	**39,7%**	13,9%
3. Sosyal medya kullanırken bazen duygusal rahatsızlık hissediyorum.	12,9%	17,2%	20,1%	**34,9%**	14,8%
4. Sosyal medyanın bazı şeyleri düşünmemin, sorumluluklarımı yerine getirmemin önüne geçtiğini düşünüyorum.	14,8%	**28,2%**	15,8%	25,4%	15,8%
5. Sosyal medyada geçirdiğim zamanın gerçek hayatta sosyal ilişkilerimi olumsuz etkilediğini düşünüyorum.	26,8%	**44,0%**	14,8%	8,1%	6,2%
6. Sosyal medya kullanımının dikkat dağınıklığına yol açtığını düşünüyorum.	7,2%	12,9%	13,9%	**38,3%**	27,8%
7. Sosyal medya kullanımının bir zaman kaybı olduğunu düşünüyorum.	9,6%	21,1%	27,8%	**28,2%**	13,4%
8. Sosyal medya kullanımının benlik saygımı olumsuz etkilediğini düşünüyorum.	28,7%	**40,2%**	17,7%	9,1%	4,3%
9. Sosyal medyanın bize sunduğu kusursuz hayatlar beni kendi hayatıma karşı agresif bir tutuma itiyor.	21,1%	**27,3%**	16,7%	24,9%	10,0%
10. Sosyal medyanın beni doyumsuz ve hep daha fazlasını isteyen mutsuz bir insan yaptığını düşünüyorum.	23,0%	**29,2%**	16,7%	20,1%	11,0%

-Sosyal Medyanın Olumsuz Etkilerine Yönelik Tutumlara İlişkin Dağılımlar

SONUÇ

Kişilerin büyük çoğunluğu sosyal medya kullanımı sonrası uykusuzluk problemi, sosyal medyanın bazı düşüncelerinin önüne geçtiğini, sosyal medyanın gerçek hayatta sosyal ilişkileri olumsuz etkilediğini, sosyal medyanın benlik saygısını olumsuz etkilediğini, sosyal medyanın bize sunduğu kusursuz hayatların kendi hayatını agresif bir tutuma ittiğini ve sosyal medyanın kendilerini daha doyumsuz ve daha fazlasını isteyen mutsuz bir insan yaptığını düşünmezken, sosyal medya kullanımı sonrası karamsarlık ve duygusal rahatsızlık hissettiğini, sosyal medyanın dikkat dağınıklığına yol açtığını ve sosyal medyanın bir zaman kaybı olduğunu düşünmektedir. Sonuç olarak sosyal medya kullanımının dozunda olması gerektiği ve bilinçli kullanımla çok faydalı bir mecra olabilecek olmasına rağmen bilinçsiz kullanımda birçok olumsuz duruma sürükleyeceği bu sebeple herkesin bu konuda kontrollü olması gerektiğini söylemek gerekmektedir.

KAYNAKÇA

[1] Tabachnick and Fidell, (2013) B.G. Tabachnick, L.S. Fidell Using Multivariate Statistics (sixth ed.) Pearson, Boston.
[2] https://www.bibliomed.org/mnsfulltext/8/8-1457291453.pdf?1716135801
[3] https://enessaritac.blogspot.com/2014/11/faktor-analizinde-kmo-ve-bartlett-testi.html