YTÜ FEN EDEBİYAT FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ

RUSYA UKRAYNA SAVAŞININ DOLAR KURUNA ETKİSİNİN MODELLENMESİ

İlknur Ayça ÇİÇEK -19023045

Danışman: Doç. Dr. Serpil KILIÇ DEPREN

ÖZET

Ekonomi ve politika sürekli olarak karşılıklı bir etkileşim içerisindedir. Özellikle ekonomik değişkenlerle demokratikleşme arasında ilişki hakkında kesin bir yargıya varılamaması, politik istikrarsızlık ile ekonomik değişkenler arasında ortaya çıkan ilişkilerin incelenmesinin gerekliliğini ortaya koymaktadır. Politik istikrarsızlık birçok sebebe bağlıdır. Sebeplerden bazıları terörizm, anayasa değişiklikleri, seçim sıklığı, komşu ülkeler arasında yaşanan olumlu ve olumsuz olaylar ve savaşlardır. Bu yüzden yaşanabilecek bu durumlarda dünyanın geçmişte yaşanmış olan olaylardan yola çıkarak ekonomik olarak ne kayıplar yaşayacağının belirlemesi ve tahmin etmesi beklenmektedir. Bu duruma örnek teşkil edebilecek en büyük olaylardan bir tanesi 11 Eylül Olayı ve sonucunda oluşan Amerika - Afganistan savaşıdır. Bu araştırma önerisinde özellikle ülkemizde dolar kurundaki oynaklığı arttıran ve güncel bir konu olan Rusya Ukrayna savaşının etkisinin analiz edilmesi ve modellenmesi amaçlanmıştır. Dolar kuruna etki eden BIST 100 endeksi, VIX endeksi, CDX endeksi gibi makroekonomik göstergeler bağımsız değişkenler olarak alınmıştır. Bu değişkenlere ek olarak, Rusya – Ukrayna savaşının etkisini incelemek amacıyla farklı yerli ve yabancı ajanslara ait gazete metinlerinden faydalanılarak sayısal bir bağımsız değişken oluşturularak modele dahil edilmiştir.

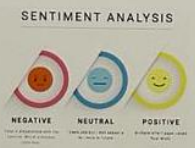
METİN MADENCİLİĞİ

Metin madenciliği (text mining), metin verilerini analiz ederek anlamlı bilgiler çıkarmayı amaçlayan bir veri madenciliği dalıdır. Metin madenciliği, doğal dil işleme, istatistiksel analiz, makine öğrenimi ve bilgisayar dilbilimi gibi disiplinlerin tekniklerini kullanarak metin verilerinin yapılandırılması, keşfedilmesi, modellemesi ve yorumlanmasını sağlar. Metin madenciliği, büyük miktarda metin verisine sahip belgelerin otomatik olarak analiz edilmesini ve anlamlı bilgilerin çıkarılmasını mümkün kılar.

METİN MADENCİLİĞİ YÖNTEMLERİ

1- DUYGU ANALİZİ (SENTIMENT ANALYSIS)

Duygu analizi (sentiment analysis), metin veya belgelere uygulanan bir doğal dil işleme tekniğidir. Amacı, bir metnin veya belgenin duygusal tonunu anlamak ve sınıflandırmaktır.Duygu analizi, metinlerdeki duygusal ifadeleri tespit ederek pozitif, negatif veya nötr gibi duygusal etiketler atar. Bu sayede,sosyal medya mesajları, müşteri yorumları, ürün incelemeleri, haber başlıkları ve diğer metin verileri gibi büyük miktardaki metinleri otomatik olarak analiz etmek ve anlamak mümkün hale gelir



2- ÖZETLEME (SUMMARIZATION)

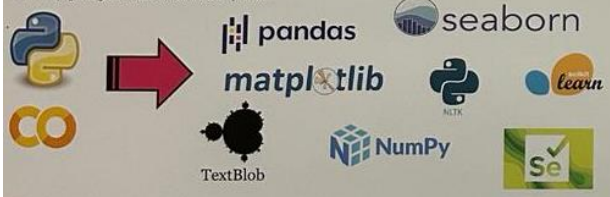
Text Summarization olarak da adlandırılan bu teknik, uzun metinlerin ana hatlarından anlamlı, kısa ve tutarlı bir özet oluşturulmasına yardımcı olur.

3- METİN SINIFLANDIRMA (TEXT CLASSIFICATION)

Metin sınıflandırma, metin verilerini farklı kategorilere veya sınıflara ayırmak için kullanılan bir makine öğrenimi tekniklerdir. Temel olarak, belirli bir metin belgesini analiz ederek, içeriğine dayanarak belgenin hangi kategoriye veya sınıfa ait olduğunu belirlemeye çalışır.Metin sınıflandırma, metin verilerinin otomatik olarak etiketlenmesini ve organize edilmesini sağlar.

VADER ANALİZİ

Doğal dil işleme alanında sıklıkla kullanılan bir duygusal analiz aracıdır. VADER, metin verilerini pozitif, negatif veya nötr duygusal değerlere göre puanlar . Ek olarak çıktı isteğine göre duygusal bileşiklik (emotional compound) adında bir değer daha bulunur. Bu değer, metnin genel duygusal yoğunluğunu temsil eder. Duygusal bileşiklik, pozitif ve negatif duygusal ifadelerin yoğunluğunu dikkate alarak hesaplanır.



- Projede python dili ve kütüphaneleri kullanılmıştır.Projede özellikle tool olarak Google colabe kullanılmıştır. Yukarıda kullanılmış olan dil ve kütüphaneler görsel olarakta verilmiştir.Selenium veri toplamak için pandas, numpy kütüphaneleri veri manipülasyonunda kullanılmıştır.Matplotlip ve seaborn görselleştirme için NLTK ve TextBlob kütüphanesi veri ön işlemede kullanılmıştır.En son olarak model kurmak için scikit learn kütüphanesi kullanılmıştır.

VERİNİN ELDE EDİLMESİ

abancı ve yerli gazeteler olmak üzere veriler sekiz ayrı kaynaktan toplanmıştır.
u haber ajanslarından 24.02.2022-24.02.2023 tarihleri arasındaki Rusya Ukrayna avaşı adına yer alan haber metinleri ve bu haber metinlerinin tarih,link,açıklama e başlık bilgileri selenium,newspapar3k kütüphaneleri sayesinde toplanıp .xlsx ormatında kayıt edilmiştir.

ERİ ÖN İŞLEME

eriyi kullanılabilecek hale getirmek için aşağıdaki adımları sırasıyla erçekleştirecek fonksiyon oluşturulmuştur.
-Genel olarak bulunan bir yazım hatasını düzeltme
- Kelimeler arası birden fazla boşluk olan yerlerin tek boşluk ile değiştirilmesi
- Noktalama işaretlerini boşluğa çevirme
-Sayıları boşluğa çevirme

nucunda ön işlemeden geçen haber metinleri vader analizi uygulanarak ört farklı değişkene çevrilmiştir.

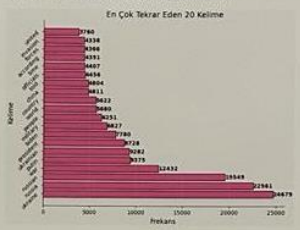
KAYNAKÇA

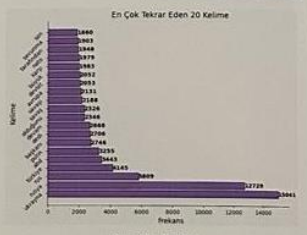
] Alpay, Ö. (2017). Ekonomi Haberlerinin BIST 100 Endeksine Etkisinin Veri Madenciliği ile İncelenmesi üzerine bir çalışma, Yüksek Lisans Tezi, Elazığ iversitesi, Sosyal Bilimler Enstitüsü, Elazığ.

] Kanago, B., & McCormick, K. (2013). The Dollar-Pound Exchange Rate During the First Nine Months of World War II. Atlantic Economic Journal, 385–

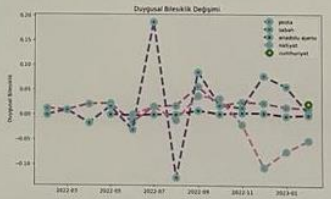
VERİLERİN GÖRSELLEŞTİRİLMESİ

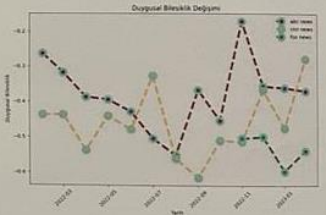
Veriler seaborn ve matplotlib kütüphaneleri yardımıyla yabancı ve Türkçe kaynaklar arasındaki farkıda iyi anlayabilmek,en çok tekrar eden ve metinleri yapısal olarak en çok hangi kelimelerden oluştuğunu tespit edebilmek için metinler görselleştirilmiştir.İlk olarak Türkçe ve yabancı haber metinleri olarak ayrıldıktan sonra word cloud kullanılarak değişim büyüklüklerine göre görselleştirilmiştir. Farklı olarak Türkçe kaynaklarda «abd» kelimesi yer alırken yabancı kaynaklarda «china» kelimesinin sıklığı göze çarpmaktadır.



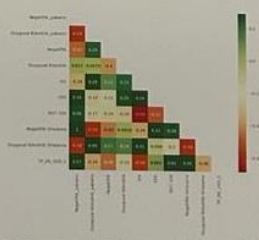


İkinci adımda ise en çok tekrar eden ilk 20 kelime frekans bazında sütun grafiği kullanılarak görselleştirilmiştir. Hem yabancı hem de Türkçe kaynaklarda kullanılan kelimelerin büyük oranda benzediğini görmekteyiz.





Üçüncü adım olarak ay bazında veri tabanımızda tuttuğumuz haber metinlerinin aylık duygusal bileşik değerlerinin çizgi grafiğini görmekteyiz.



Dördüncü adım olarak dolar kuruyla hem aylık hem de günlük olarak bağımsız değişkenler arasındaki korelasyonlar heatmap grafikleriyle görselleştirilmiştir.

MODELLEME

Regresyon tabanlı makine öğrenimi algoritmaları denenmiştir.Yanda tablo şeklinde en iyi sonuç veren algoritmadan en kötü sonuç veren algoritmaya göre sıralanmıştır.Karar verme değeri olarak RMSE kullanılmıştır.

MODEL	RMSE
RandomForestRegressor	0,225
GradientBoostingRegressor	0,235
XGBRegressor	0,251
LGBMRegressor	0,258
DecisionTreeRegressor	0,263
KNeighborsRegressor	0,279
SVR	0,372
MLPRegressor	1039

SONUÇ

Bu çalışmada çeşitli yerli ve yabancı haber ajanslarının Rusya-Ukrayna Savaşı adı altında web sitelerinde yer verdikleri haber metinleri ele alınmıştır ve bu haber metinlerine literatür dikkate alınarak çeşitli ön işleme adımları uygulanmıştır.Duygu analizi yöntemlerinden biri olan VADER yöntemi kullanılmış olup CIX endeksi, VIX endeksi ve BIST 100 endeksi ile birlikte on bir adet bağımsız değişkenle dolar kuru tahmini yapılmıştır. Haber ajanslarından alınan bu metinler kelime bulutu ve bar grafikleriyle görselleştirilmeleri yapılmış olup yabancı kaynaklı haber ajanslarının yerli haber ajanslarına göre tutumlarının daha olumsuz olduğu saptandı.En çok kullanılan kelimelere bakıldığında yerli ve yabancı haber ajansları arasında çok büyük bir fark saptanmamıştır.Haber ajanslarının metinlerinden yola çıkılarak oluşturulan sayısal değişkenlerin korelasyon analizi sonuçları göz önünde bulundurulduğunda aylık bazlı bakıldığında yabancı haber ajanslarının dolar kuruyla anlamlı ve göze alınabilecek sonuçlar ortaya çıkarken yerli haber ajanslarıyla oluşturulan sayısal değişkenlerin modelde etkisinin istediği oranda ve şekilde gerçekleştiremeyeceği ön görülmüştür .Diğer bağımsız değişkenler göz önüne alındığında modele etkisinin en kuvvetli olacağı görülen değişken BIST 100 endeksi olmuştur. Regresyon tabanlı makine öğrenmesi algoritmalarıyla oluşturulan modelde RMSE değerleri dikkate alındığında veriye en uygun modelin Random Forest

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

MAKİNE ÖĞRENMESİ İLE OECD ÜLKELERİNİN ÇEVRESEL PERFORMANSININ TAHMİN EDİLMESİ

Beste ÖZÜLKÜ

Doç.Dr. Serpil KILIÇ DEPREN

ÖZET

Bu çalışma, çevresel performans üzerinde farklı bağımsız faktörlerin etkisini derinlemesine analiz etmeyi amaçlamaktadır. Çalışma, çevresel performans endeksine odaklanarak, 181 farklı ülkeyi kapsayan 509 gözlemi içeren bir nihai veri kümesi kullanmakta veri kümesi, yedi farklı kaynaktan derlenen verilerin birleştirilmesiyle oluşturulmuş ve altı farklı makine öğrenimi algoritmasıyla modelleme için kullanılmıştır. En önemlisi, Super Learner algoritması da birden fazla makine öğrenimi algoritmasını birleştirere ensemble modeli oluşturmak için kullanılmış ve tüm bireysel makine öğrenimi algoritmalarının önüne geçmiştir.

ÇEVRESEL PERFORMANS

Çevresel performans, sürdürülebilir kalkınmanın kritik bir boyutudur ve bir ülkenin bu hedeflere ne kadar yakın olduğunu ölçer. Çevresel performans göstergeleri, çevresel politika sonuçlarına odaklanarak hükümetlerin kapsamlı kirlilik kontrolü ve doğal kaynak yönetimi hedeflerine ilerlemesini ölçmelerine yardımcı olur. Hedefler, mevcut uluslararası anlaşmalara, kirliliğin insanlar ve ekosistemler üzerindeki zararlı etkilerine dair bilimsel kanıtlara ve ekonomik olarak sürdürülebilir çevre koruma stratejilerine dayanmaktadır. Hedefler, temiz hava, içme suyu ve biyolojik çeşitliliğin korunması gibi temel çevresel konuları kapsar. Her hedef, teorik mantık, politika uyumu, ölçülebilirlik ve mevcut veri kapsamı gibi katı kriterlere göre seçilen birkaç performans göstergesiyle ilişkilidir.

Çevresel Performans Endeksi (Environmental Performance Index: EPI), öncelikle kirlilik, doğal kaynak yönetimi, çevresel sağlık, ekosistem kalitesi ve iklim değişikliği gibi çevresel konuları değerlendirmek için kullanılan bir dizi ölçüme dayanmaktadır. Bu göstergeler arasında su ve hava kalitesi, atık yönetimi, biyolojik çeşitlilik ve orman koruması gibi unsurlar yer alabilir.

Endeks, çevresel performansın yanı sıra sosyal ve ekonomik konuları da dikkate alabilir.Bu, çevresel sürdürülebilirliği diğer unsurlarla birlikte kapsamlı bir şekilde incelemeyi sağlar.Örneğin, çevresel performans sosyal eşitsizlik, yoksulluk ve ekonomik kalkınma gibi unsurlardan etkilenebilir ve değerlendirmeye dahil edilmeleri önemlidir. Bu çalışma, makine öğrenimi (ML) algoritmalarını kullanarak 2018, 2020 ve 2022 yıllarında 181 farklı ülkede çevresel performansı etkileyen faktörlerin etkilerini incelemeyi amaçlamaktadır. Özellikle, çevresel yönetim faaliyetlerini teşvik etmek için en iyi algoritmayı ve politikayı belirleme sürecinde EPI ile ilişkiler arasındaki bağlantıları incelemeye katkıda bulunmaktadır.Bu şekilde, çevresel performansı iyileştirme çabaları, ülkelerin çevresel düzenlemelere uyum sağlamasını teşvik edecektir.

Bu çalışma aşağıdaki araştırma sorularını ele almaktadır:

i. Çevresel performans üzerinde hangi faktörler önemli bir etkiye sahiptir? Etki yönü nedir?
ii. Önemli değişkenlerdeki kritik eşik değerleri nelerdir?
iii. Seçilen değişkenlerle çevresel performans modellenebilir mi? Hangi ML algoritmaları diğerlerinden daha güçlü tahminler üretir?
iv. Çevresel performans göstergesi en iyi hangi algoritma ile modellenebilir?Eğitim ve Test Verilerindeki model performansı tatmin edici midir?
v. Seçilen modelden öne çıkan öncelikli hareket alanları nelerdir? Hangi politika önerileri sunulmalıdır?

METODOLOJİ

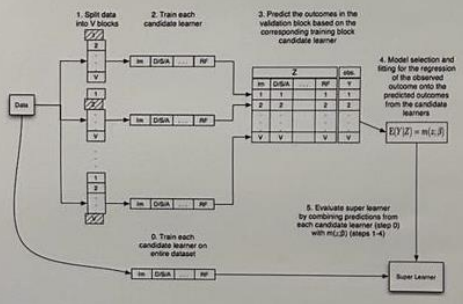
İncelenen değişkenlerin çevresel performans üzerindeki etkisini incelemek için Makine Öğrenimi (ML) algoritmalarını kullanmaktadır. Son zamanlarda birçok farklı disiplinde sınıflandırma ve tahmin yapmak tercih edilmektedir. ML, verilere dayalı olarak kendini öğrenebilen matematiksel modellerin geliştirilmesine izin veren bir yapay zeka uygulamasıdır. Bu, karmaşık görünen bir veri yığınındaki problemleri tespit etmenizi, gelecekteki senaryoları öngörmenizi ve verideki desenleri keşfetmenizi sağlar.

MAKİNE ÖĞRENİMİ (ML) ALGORİTMALARINI

Sınıflandırma ve Regresyon Ağaçları (CART)
Rastgele Orman (RF)
Aşırı Gradyan Artırma Makineleri (XGB)
Gradyan Artırma Makineleri (GBM)
Destek Vektör Makineleri (SVM)
k-En Yakın Komşular (k-NN)

SÜPER ÖĞRENİCİ ALGORİTMASI

Süper öğrenme, çeşitli öğrenme algoritmalarının optimal bir kombinasyonunu bulan bir enseble yöntemiyle bu sorunu ele alır. Bu, tahminsel modelleme problemleri için araştırabileceğiniz tüm modelleri ve yapılandırmaları birleştirerek, araştırdığınız herhangi bir tek modelden daha iyi veya ona eşit bir tahmin yapmak için kullanılan bir tekniktir.



Yukarıda görüldüğü gibi, süper öğrenici, bireysel modellerin k-katlamalı veri bölünmesinde eğitildiği bir varyasyon olan istiflemeyi veya k-katlamalı çapraz doğrulamayı içerir. Ardından, her modelin çıktısı olan kat dışı tahminlerden de anılan bir nihai meta modeli eğitilir.

KAYNAKÇA

[1] Van Der Laan, M. J., Polley, E. C., & Hubbard, A. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1).https://doi.org/10.2202/1544-6115.1309
[2] Young, S. L., Abdou, T., & Bener, A. (2018a). Deep Super Learner: A Deep Ensemble for Classification Problems. In *Lecture Notes in Computer Science* (pp. 84–95). Springer Science+Business Media. https://doi.org/10.1007/978-3-319-89656-4_7
[3] Rumao, P. (2022, March 30). Super Learner versus Deep Neural Network - Towards Data Science. *Medium*.

İSTATİSTİKLER VE BOX-JENKİNS GRAFİKLERİ

	EPI	TR	ITR	CTR	TB	GE	POP	GDPG	GDPPP	UN	INF	FDI	PUBD
Minimum	*9.8*	*0.0*	*0.0*	*0.0*	*1.9*	*16.1*	*0.1*	*-28.5*	*651.9*	*0.0*	*-2.7*	*-113500.0*	*0.0*
Q_{25}	*35.6*	*2.5*	*20.0*	*20.0*	*15.9*	*25.0*	*2.5*	*1.1*	*4186.6*	*4.1*	*0.9*	*128.2*	*87.2*
Median (Q_{50})	*46.4*	*5.4*	*50.0*	*25.0*	*20.5*	*31.4*	*9.5*	*2.6*	*12555.9*	*6.2*	*2.4*	*878.0*	*92.3*
Q_{75}	*62.0*	*9.6*	*55.0*	*30.0*	*29.1*	*39.4*	*29.6*	*4.2*	*28933.3*	*10.0*	*4.8*	*3877.0*	*76.0*
Maximum	*99.3*	*67.7*	*60.0*	*50.0*	*59.6*	*125.7*	*527.4*	*12.1*	*130475.1*	*31.4*	*88.2*	*391104.0*	*304.1*
Mean	*49.3*	*6.5*	*28.7*	*23.8*	*21.8*	*32.6*	*26.4*	*2.6*	*19855.6*	*7.9*	*4.5*	*6501.4*	*61.1*
Standard Deviation	*20.8*	*5.5*	*18.1*	*8.9*	*10.3*	*13.8*	*46.0*	*2.9*	*21174.2*	*5.7*	*8.0*	*28063.7*	*37.9*
Coefficient of Variation	*42.2*	*84.9*	*45.5*	*37.5*	*47.2*	*46.9*	*174.3*	*112.3*	*106.6*	*72.8*	*180.2*	*432.7*	*62.1*
Skewness	*0.5*	*3.3*	*-0.2*	*-0.3*	*0.4*	*2.2*	*5.4*	*-1.8*	*1.9*	*1.5*	*5.1*	*8.2*	*2.1*
Kurtosis	*-0.4*	*20.8*	*-0.3*	*0.5*	*-0.4*	*11.7*	*35.8*	*14.5*	*4.8*	*2.5*	*37.1*	*95.8*	*7.7*
Kolmogorov-Smirnov	*0.1*	*0.1*	*0.1*	*0.1*	*0.1*	*0.1*	*0.4*	*0.1*	*0.2*	*0.1*	*0.5*	*0.5*	*0.1*
p value	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*

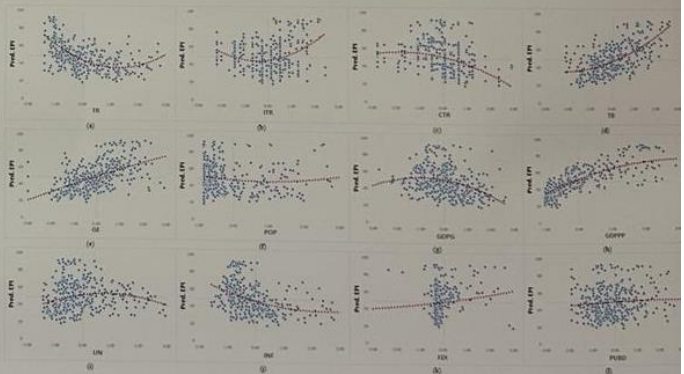
MAKİNE ÖĞRENME ALGORİTMALARININ BULGULARI

	Eğitim Seti (70% - 373 gözlem)			Test Seti (30% - 136 gözlem)		
ML Algoritmaları	R^2	RMSE	MAE	R^2	RMSE	MAE
RF	%60,19	13,1076	10,4831	%80,20	9,5091	8,0257
GBM	%58,91	13,1714	10,8988	%80,29	9,4701	8,0478
XGB	%58,55	13,3539	10,6015	%73,63	10,8146	9,0240
k-NN	%47,44	15,0998	12,3839	%72,72	10,9909	9,1736
CART	%45,30	15,4409	12,4509	%52,57	14,5279	12,1075
SVM	%43,85	18,9072	14,5938	%67,50	17,8502	14,3449

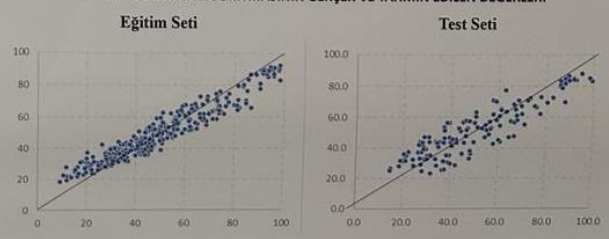
SÜPER ÖĞRENİCİ ALGORİTMASININ PERFORMANS METR

			Eğitim Seti Performansı (70% - 373 gözlem)			Test Seti Performans (30% - 136 gözlem)		
Super Learner	Risk	Katsayı	R^2	RMSE	MAE	R^2	RMSE	MA
CART	217,4	%0.0						
RF	170,6	%51,2						
XGB	196,5	%17,0	%92,23	6,3948	5,1531	%79,83	9,6707	8,2
GBM	174,4	%31,1						
SVM	221,2	%0.7						
k-NN	257,6	%0.0						

DEĞİŞKENLER ARASINDAKİ İKİLİ İLİŞKİ VE ÖNEMLİ EŞİK DEĞERLERİ



SÜPER ÖĞRENİCİ ALGORİTMASININ GERÇEK VE TAHMİN EDİLEN DEĞERLERİ



Sonuçlar

Analiz sonuçlarına göre, kişi başına düşen gayri safi yurtiçi hasıla (GSYİH), tarife oranları, vergi yükü, devlet harcamaları ve enf çevresel performansı geliştirmek için politika yapıcıların odaklanması gereken beş temel faktör olarak belirlenmiştir. Ayrıca, nüfus, gayri safi yurtiçi hasıla büyüme oranı, gelir vergisi oranı, kamu borcu, doğrudan yabancı yatırım ve kurumlar vergisi oran çevresel performans üzerinde önemli bir etkisi olduğu belirlenmiştir. Bunun yanı sıra, ilgili faktörlerin çevresel performans üze etkisinin değişimini gösteren kritik eşikler belirlenmiştir. Sonuç olarak, bu makale, politika yapıcılara çevresel perfo geliştirebilecek faktörler konusunda değerli öngörüler sunmaktadır. Yenilenebilir enerji kaynaklarına yatırım yaparak, yenilenebilir ve yeşil enerji kaynaklarına yönelik yabancı yatırımları teşvik ederek, ekonomik büyümeyi ve GSYİH'yi kontrol altına politika yapıcılar, çevresel performansı önemli ölçüde ilerletebilirler. Ayrıca, politika yapıcılar, vergi yükü, gelir vergisi ve ku vergisinin çevresel performans üzerindeki etkisini de dikkate almalı ve gerekli reformları yapmalıdır.

YTÜ FACULTY OF ARTS AND SCIENCES

DEPARTMENT OF STATISTICS

CUSTOMER CHURN ANALYSIS

EZGİ GÖKBUA 21023604

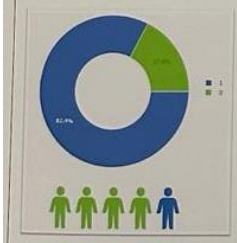
Thesis Advisor: Prof. Dr. Filiz KARAMAN

Customer churn analysis is a crucial topic for companies to succeed in today's world. The aim of this thesis is to predict customers with a high likelihood of churn by reviewing their purchasing behaviors. Anonymous data from customers who renewed and did not renew their comprehensive insurance policies in a private insurance company in Turkey was used. The study employed data balancing methods such as oversampling and undersampling. Machine learning classification methods were applied on the balanced dataset, with Random Forest, XGBoost, and LightGBM algorithms yielding the highest success rates. The data set was processed using the Python programming language.

Customer Churn Analysis in the Insurance Sector

Churn analysis involves various methods and techniques to identify the reasons behind customer attrition and make strategic decisions. Typically, data such as customer behavior, demographic information, past purchasing habits, and customer satisfaction feedback are utilized in this analysis. The insights derived can help companies offer tailored solutions to customers, increase customer loyalty, and facilitate the growth of their customer base. In this study, customer data from a specific insurance company regarding the renewal and cancellation of comprehensive insurance policies has been used. In addition to demographic information such as age and gender, the data also includes variables related to the insured vehicle, such as its age, brand, and type.

Churn Customer Percentage



Customer churn analysis is closely related to the growth strategies and revenue goals of companies. Determining why customers are leaving and identifying the factors that contribute to customer churn provides valuable insights for companies to improve customer relationships and take measures to minimize customer attrition.

During the observation, it was found that 61,014 customers have renewed their policies, while 13,001 customers have chosen not to renew their policies and ceased to be customers of the insurance company.

Before applying classification methods for customer churn prediction on the dataset, the issue of imbalanced dataset, which is an important point in classification problems, has been addressed. Two separate sampling techniques, namely Under Sampling and Over Sampling have been thoroughly examined.

Techniques to Handle Imbalanced Data

Undersampling is a sampling technique used to address the problem of imbalanced datasets in classification problems. In this method, the number of instances belonging to the majority class is reduced to achieve a more balanced dataset with the minority class.

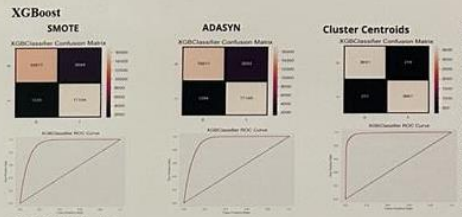
--> Random Under Sampler
--> Near Miss
--> Cluster Centroids
--> Tomek Links
--> ENN

Oversampling aims to artificially increase the number of samples in classes with a small number of instances. This way, it enhances the importance and impact of classes with a small number of instances, enabling the model to learn them better. This can potentially improve the classification performance of the model and reduce the occurrence of erroneous results due to class imbalance.

--> Random Over Sampler
--> SMOTE
--> ADASYN

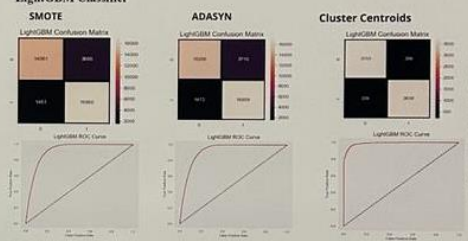
Machine Learning Classification Methods

XGBoost



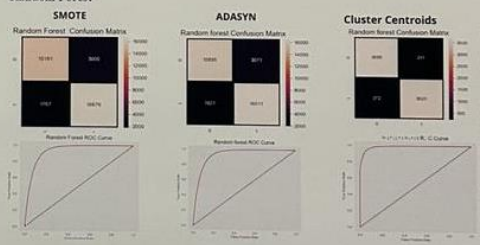
XGBoost is a tree-based learning method, which means it utilizes an ensemble of decision trees. Each tree divides the dataset using a set of decision rules to predict a target variable. The decision rule is created by comparing the features (or attributes) in the dataset with a specific threshold value. Each split further divides the dataset into more homogeneous subsets.

LightGBM Classifier



LGBM (LightGBM) is a machine learning algorithm that is lightweight and fast, and it is a gradient boosting method. It divides the dataset into rectangular regions and updates the sample weights using histograms for each region. This allows for high performance on large datasets and helps reduce overfitting.

Random Forest



Random Forest is a machine learning algorithm that consists of an ensemble of decision trees. Each decision tree processes the dataset by randomly sampling and splitting the data using random features. As a result, this method reduces high variance, improves generalization on the data, and makes stable predictions.

Conclusion

"The Lazy Classifier" method was used to evaluate all classification techniques. This evaluation was performed on the datasets obtained using balancing techniques. The evaluation identified three machine learning classification techniques, namely XGBoost, LGBM, and Random Forest, that demonstrated high performance across all balancing techniques. These methods were found to have high accuracy and F-values, indicating their effectiveness in predicting customer churn. When examining the obtained results, methods like XGBoost, LGBM, and Random Forest demonstrate themselves as powerful tools that can be used to predict churn customers in the insurance company's dataset.

Method	SMOTE	ADASYN	Cluster Centroids
XGBoost	0.87	0.88	0.94
LGBM	0.94	0.86	0.94
Random Forest	0.87	0.87	0.94

Accuracy values of Classification Methods

Recommendations

There are several strategies that can be implemented for customers who have the potential to churn:

- Customer Relationship Management (CRM)
- Personalized Marketing
- Improved Customer Experience
- Loyalty Programs
- Win-back Campaigns
- Data Analytics and Predictive Models

REFERENCES

[1] Akyiğit, H. E., & TAŞCI, T. (2021). SİGORTACILIK SEKTÖRÜNDE MAKİNE ÖĞRENMESİ İLE MÜŞTERİ KAYBI ANALİZİ. Tasarım Mimarlık ve Mühendislik Dergisi, 2(1), 66-79.

[2] AYDIN, M. A. (2021). Müşteri Kaybı Tahmininde Sınıf Dengesizliği Problemi. Politeknik Dergisi, 1-1.

[3] KAYNAR, O., TUNA, M. F., GÖRMEZ, Y., & DEVECİ, M. A. (2017). Makine öğrenmesi yöntemleriyle müşteri kaybı analizi. Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi, 18(1), 1-14.

AND SCIENCES

DEPARTMENT OF STATISTICS

Convolutional Neural Networks for Generating, Modeling and Testing Statistical Data from Traffic Images

Fikret Efe DİRİ 18023058

Advisor: Prof. Dr. Filiz KARAMAN

Abstract

As the human population continues to grow, various costs arise, one of which is traffic congestion. The blockage or complete halt of traffic flow leads to negative consequences such as time lost, increased carbon emissions, and higher fuel consumption. Efforts to minimize these adversities are made by entities such as highway administrations, municipalities, and transportation engineers, who rely on accurate traffic flow data as one of the most important sources of information. While traffic flow data can be obtained through multiple methods, cameras emerge as one of the best solutions due to their affordability and flexibility. This study aims to generate and analyze traffic flow data derived from camera recordings for the purpose of mitigating these negative impacts.

What is Traffic Flow Data?

Traffic flow data is a type of data that involves counting the number of vehicles in traffic based on certain breakdowns, as well as including vehicle speeds. These breakdowns can include vehicle types such as cars, trucks, motorcycles, etc., and information about the routes of the vehicles on the road. Below is an example of a result obtained from a study where this data was generated along a road with two-way traffic.

Cls	Direction	Kmph
2	Left	140,10
2	Right	134,44
2	Left	113,54
7	Right	88,93
2	Right	134,76

- **Cls** : It includes the types of vehicles, specifically 2 cars and 7 trucks.
- **Drection** : : It includes the route information of the vehicles. 'Left' indicates the direction of travel, while 'Right' indicates the opposite direction based on the location where the recording is taken.
- **Kmph**: It includes the speeds of the vehicles.

The Algorithm and Statistical Tests

This study utilized the Yolov5_StrongSort_OsNet Algorithm, which is based on an artificial neural network. The algorithm combines the one-stage object detection algorithm YoloV5 with the construction of an object tracking system. Similar to object tracking, the object detection and classification components of the algorithm have been pre-trained to create an enhanced algorithm. YoloV5 is a convolutional neural network (CNN) based algorithm capable of object detection and processing live images at a high speed.

Although the algorithm demonstrates success in object recognition, prediction, and detection, several modifications need to be made to adapt it for traffic flow data. The study makes use of statistical tests such as Shapiro-Wilk, Kruskal-Wallis, and Mann-Whitney to analyze the data.

Customization of the Algorithm

1. Entrance and Exit Detection:

Firstly, it is necessary to digitally detect the entrance and exit pixels of vehicles on the road from the recorded traffic image. This process is carried out by analyzing a single frame of the image, as shown in the photograph.

During the execution of the algorithm, apart from the entrance and exit detection, the video recording is played live on the screen. In order to display the calculations performed on this window, counters will be added for the vehicles entering the frame. The positions of these counters and texts have been calculated.

2. Object Count:

For this task, if-else blocks have been created as shown below. These blocks trigger the code responsible for storing various raw data when the center pixel coordinates of the detected and tracked vehicle, obtained from the algorithm, fall within the predetermined coordinates. This allows for the collection of raw data related to the vehicle.

```
def count_obj(box,w,h,id,cls):
    global ccl,ctl,ccr,ctr

    if id not in control:
        if int(box[1] +int(box[3]-box[1])/2)>(h-370):
            if int(box[0] +(box[2]-box[0])/2)>(50) and  int(box[0] +(box[2]-box[0])/2)<(615):


                control.append(id)
                dctSt['TimeSt'].append(time.time())
                dctSt['ID'].append(id)
                dctSt['Cls'].append(cls)
                dctSt['Direction'].append('Left')

                if cls==2:
                    ccl+=1
                elif cls==7:
                    ctl+=1
```

3. Processing of the Data

The collected raw data includes the route and class of the vehicle as desired. However, additional calculations are required to determine the speed of the vehicle. After the vehicles enter the designated entrance and exit areas, the computer retrieves the timestamp information, and the speed is calculated in pixel-seconds based on the time difference. Since the length of the specified region is not known, the speed data is estimated hypothetically.

Results

Vehicle	Direction	Real	Algorithm Count	Miss Rate
Car	Left	217	209	3,69%
	Right	202	188	6,93%
Truck	Left	36	31	13,89%
	Right	37	33	10,81%

The output during the execution of the algorithm appears as shown in the following figure. The vehicles on the road consist only of cars and trucks. Vehicles classified as vans and trucks are considered under the category of 'truck' at a single level. The difference between the algorithmic observations and manual road counting is evident in the images above.

```
# Perform Mann-Whitney U test
_, p_value = stats.mannwhitneyu(right_data, left_data, alternative='two-sided')

alpha = 0.05  # Significance level

if p_value < alpha:
    if right_data.median() > left_data.median():
        print("The 'Right' direction is significantly faster than the 'Left' direction.")
    else:
        print("The 'Left' direction is significantly faster than the 'Right' direction.")
else:
    print("There is no significant difference in the median 'kmph' values between the 'Right' and 'Left' directions.")
```

```
The 'Left' direction is significantly faster than the 'Right' direction.
```

Based on the conducted analysis, it can be concluded that there exists a statistically significant variation in vehicle speeds across different classes. Moreover, a noticeable discrepancy in speeds is observed between the right and left sides of the road. This study successfully facilitated the real-time generation of object detection data and speed measurements. Furthermore, an exemplified application of this data is presented, demonstrating its potential utilization.

```
# Create an empty dictionary to store the arrays for each group
group_arrays = {}

# Iterate over the groups and create arrays for each group
for cls, group in grouped_df:
    group_arrays[cls] = group['kmph'].values

# Perform Kruskal-Wallis H test
_, p_value = stats.kruskal(*group_arrays.values())

alpha = 0.05  # Significance level

if p_value < alpha:
    print("There is a significant difference in the means of the 'kmph' variable among the 'cls' groups.")
    print()

    # Calculate the median for each group
    medians = {cls: np.median(values) for cls, values in group_arrays.items()}

    # Find the fastest cls based on the highest median
    fastest_cls = max(medians, key=medians.get)

    print(f"The 'cls' with the highest median 'kmph' value is {fastest_cls}.")
    print(f"The median 'kmph' value for {fastest_cls} is {medians[fastest_cls]}.")
else:
    print("There is no significant difference in the means of the 'kmph' variable among the 'cls' groups.")
```

```
There is a significant difference in the means of the 'kmph' variable among the 'cls' groups.

The 'cls' with the highest median 'kmph' value is 2.
The median 'kmph' value for 2 is 110.51080409068724.
```

REFERENCES

[1] YoloV5_StrongSort_OsNet https://github.com/hdnh2006/Yolov5_DeepSort_Pytorch

[2] Yardım, M. S., & Akyıldız, G., (2005). Akıllı Ulaştırma Sistemleri ve Türkiyedeki Uygulamalar . 6. Ulaştırma Kongresi (pp.405-414). İstanbul, Turkey

[3] https://towardsdatascience.com/artificial-neural-networks-for-total-beginners-d8cd07abaae4

FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

Uykusuzluğa Neden Olabilecek Faktörler Ve Hayatımıza Olası Etkileri

Tuğçe Gül YILMAZ 18023030

Danışman: Doç. Dr. Atıf Ahmet EVREN

ÖZET

Araştırmada hayatımızda önemli bir faktör olan uyku ele alınmıştır. Bu ana başlık altında asıl incelenen ise uykusuzluğa neden olabilecek faktörler ve hayatımıza olası etkileri konusu olmuştur. Rastgele seçilen 256 kişilik örneklem üzerinde uygulanmıştır. Tezin amacı uykusuzluğa neden olabilecek faktörler ve uykusuzluğun hayatımıza olası etkilerini, ankette sorulan 27 alt değişken ile istatistiksel yöntemlerle araştırmak ve çıkan sonuçlar doğrultusunda ele alınan konu hakkında çıkarımlarda bulunmaktır. SPSS programı kullanılarak anket sonucundan elde edilen veriler incelenmiş ve analizler yapılmıştır. Genel olarak uykusuzluğu etkileyen faktörlerin birden fazla olduğu ve bunların özellikle; cinsiyet, medeni durum ve çalışma durumu değişkenlerine göre farklılaştıkları görülmüştür.

TEZİN AMACI

Amaç uykusuzluğa neden olabilecek faktörler ve uykusuzluğun hayatımıza olası etkilerini, ankette sorulan 27 alt değişken ile birlikte istatistiksel yöntemlerle araştırmak ve çıkan sonuçlar doğrultusunda ele alınan konu hakkında çıkarımlarda bulunmaktır.

TEZİN YÖNTEMİ

SPSS programı kullanılarak anket sonucuyla çıkan veriler incelenmiş ve analizler yapılmıştır. İlk olarak tanımlayıcı istatistikler ele alınmıştır. Burada değişkenlerin frekansları ve dağılımları incelenmiştir. İkinci olarak normallik testleri yapılarak değişkenlerin normal dağılıp dağılmadığı incelenmiştir. Üçüncü olarak hipotezler kurulup değişkenlerin bir önceki bölümde elde ettiğimiz normallik sonuçlarına göre uygun olan testler yapılarak hipotezler yorumlanmıştır. İlk olarak, iki değişken arasında ortalamanın anlamlı olup olmadığını test eden bağımsız örneklem t-testi ve mann-whitney u testi uygulanmıştır. İkinci olarak, iki değişken arasında anlamlı bir ilişki olup olmadığını test eden ki-kare testi ve anlamlı bir ilişki olup olmadığıyla beraber bu ilişkinin yönünü de test eden korelasyon testi uygulanmıştır. Sonuçları aksi belirtilmediği sürece 0,95 güven aralığında p≤0,05 değeri anlamlı olarak kabul edilerek değerlendirilmiştir.

ÖN BİLGİLER

Sonuçlar aksi belirtilmediği sürece 0,95 güven aralığında p≤0,05 değeri anlamlı olarak kabul edilerek değerlendirilmiştir.

Ki-kare testinde tablonun altında verilen bilgilerde 5'in altında frekans değeri %20'nin üzerinde bulunmadıysa Pearson Ki-Kare değerini, %20'nin üzerinde bulunduysa yine tablonun altında verilen "beklenen sayı…" kısmını ele alarak 5'in altında olması durumunda Fisher's Exact test satırı, 5 ile 25 arasında ise Continuity Correction satırı ve eğer 25 üzerinde ise yine Pearson Ki-Kare satırı ele alınarak sonuca varılmıştır. Monte Carlo simülasyonuyla elde edilen Exact testi p-değerinin (10.000 örneklem ve %99 güven aralığıyla) Exact seçeneğiyle elde edilen Exact testi p-değerine virgülden sonraki üç sıfır kadar aynı olduğu kabul edilmiştir [4].

Normallik testinde verilerin normal dağılıp dağılmadıklarına yapılan test sonucunda ortaya çıkan basıklık ve çarpıklık değerlerine bakılarak karar verilmiştir. Değerler "+1.0,-1.0" arasında ise normal dağıldıklarına karar verilmiştir [5]

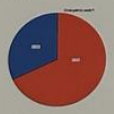
DEMOGRAFİK BULGULAR

Ankete Katılanların Cinsiyetleri

Ankete katılan 256 kişinin cinsiyet dağılımı Tablo 2.1.1 de verilmiştir ve Şekil 2.1.1 de gösterilmiştir. Buna göre anket katılımcılarının %68'inde kadınlar yer alırken, %33'ünde erkekler yer almaktadır.

Cinsiyetiniz nedir?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Erkek	82	32.0	32.0	32.0
	Kadın	174	68.0	68.0	100.0
	Total	256	100.0	100.0	

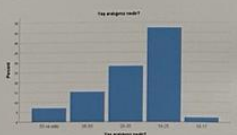


Ankete Katılanların Yaşları

Ankete katılan 256 kişinin yaş dağılımı Tablo 2.1.2 de verilmiştir ve Şekil 2.1.2 de gösterilmiştir. Buna göre anket katılımcıların %47,3'ü 18-25 yaş arası olarak en fazla katılım sağlayan yaş aralığı olurken, %2,3'le 10-17 yaşları arasındaki katılımcılar ankete en az katılım sağlamıştır.

Yaş aralığınız nedir?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	10-17	6	2.3	2.3	2.3
	18-25	121	47.3	47.3	49.6
	26-35	72	28.1	28.1	77.7
	36-50	39	15.2	15.2	93.0
	50 ve üstü	18	7.0	7.0	100.0
	Total	256	100.0	100.0	



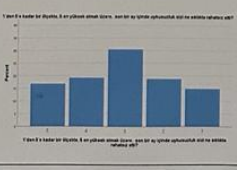
BETİMSEL BULGULAR

Son Bir Ay İçinde Uykusuzluğun Rahatsız Etme Sıklığı

Ankete katılan katılımcıların %30,5'le çoğunluk olan kısmı bir ay içinde uykusuzluğun rahatsız etme sıklığının orta düzeyde olduğunu belirtmiştir.

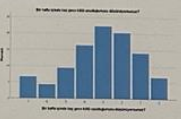
1'den 5'e kadar bir ölçekte, 5 en yüksek olmak üzere, son bir ay içinde uykusuzluk sizi ne sıklıkta rahatsız etti?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	5	43	16.8	16.8	16.8
	4	49	19.1	19.1	35.9
	3	78	30.5	30.5	66.4
	2	48	18.8	18.8	85.2
	1	38	14.8	14.8	100.0
	Total	256	100.0	100.0	



Bir Hafta İçinde Kötü Uyunan Gece Sayısı

Ankete katılan katılımcıların çoğunluk olan %22,3'ü ve %20,3'ü en çok bir hafta içinde 3 ile 2 gün kötü uyuduklarını belirtmişlerdir. Bununla birlikte bir haftanın her günü kötü uyuyan katılımcı sayısı ile hiçbir günü kötü uyumayan katılımcı sayısının %6,6'lık oranlarla eşit sayıda oldukları görülmüştür.



HİPOTEZLER VE TESTLER

MANN-WHITNEY U TESTİ

Cinsiyet * Gecenin Bir Yarısı Uyanma Nedeni

Hipotez;

H0: Cinsiyet değişkeni ile gecenin bir yarısı uyanma nedeni değişkeninin ortalamaları arasında anlamlı bir fark yoktur.

H1: Cinsiyet değişkeni ile gecenin bir yarısı uyanma nedeni değişkeninin ortalamaları arasında anlamlı bir fark vardır.

Tablo 3.2'de gösterilen test sonucundaki anlamlılık değerine bakıldığında anlamlılık değerinin 0,05'ten küçük bir değer olduğu görülmüştür (0,042<0,05). Sonuç olarak H0 hipotezi reddedilerek, cinsiyet değişkeni ile uyku problemi için yardım alma değişkeninin ortalamaları arasında anlamlı bir farkın olduğu görülmüştür.

Kİ-KARE TESTİ

Son bir ay içinde uykusuzluk sizi ne sıklıkta rahatsız etti? * Medeni durumunuz?

Hipotez;

H0: Son bir ay içinde uykusuzluğun rahatsız etme sıklığı medeni durumdan bağımsızdır.

H1: Son bir ay içinde uykusuzluğun rahatsız etme sıklığı medeni duruma bağlıdır.

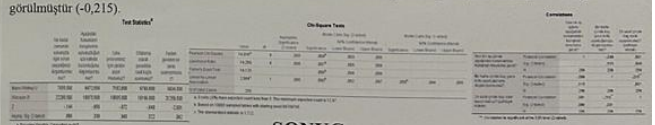
Tablo 3.3.3'te görüldüğü gibi anlamlılık değeri 0,05 anlamlılık değerinden küçük olduğu için H0 reddedilmiştir (0,005<0,05). Medeni durum ile son bir ay içinde uykusuzluğun rahatsız etme sıklığı arasında bir ilişki bulunmaktadır.

KORELASYON TESTİ

H0: Bir hafta içinde kötü geçen gece sayısı ile 24 saat içinde uyunan saat arasında ilişki yoktur.

H1: Bir hafta içinde kötü geçen gece sayısı ile 24 saat içinde uyunan saat arasında ilişki vardır.

Tablo 3.4.1'e bakıldığında 0,01 anlamlılık seviyesine göre ilişkinin anlamlı olduğu görülmüştür (0,001<0,01) ve böylelikle H0 hipotezi reddedilmiştir Bir hafta içinde kötü geçen gece sayısı ile 24 saat içinde uyunan saat arasında istatistiksel anlamda ilişki vardır. Pearson korelasyon katsayısına bakıldığında negatif yönlü bir ilişkinin olduğu görülmüştür (-0,215).



SONUÇ

Araştırmada hayatımızda önemli bir faktör olan uyku ele alınmıştır. Bu ana başlık altında asıl incelenen ise uykusuzluğa neden olabilecek faktörler ve hayatımıza olası etkileri konusu olmuştur. SPSS programı kullanılarak anket sonucuyla çıkan veriler incelenmiş ve analizler yapılmıştır. İlk olarak tanımlayıcı istatistikler ele alınmıştır. Burada değişkenlerin frekansları ve dağılımları incelenmiştir. İkinci olarak normallik testleri yapılarak değişkenlerin normal dağılıp dağılmadığı incelenmiştir. Üçüncü olarak hipotezler kurulup değişkenlerin bir önceki bölümde elde ettiğimiz normallik sonuçlarına göre uygun olan testler yapılarak hipotezler yorumlanmıştır. İlk olarak, iki değişken arasında ortalamanın anlamlı olup olmadığını test eden bağımsız örneklem t-testi ve mann-whitney u testi uygulanmıştır. İkinci olarak, iki değişken arasında anlamlı bir ilişki olup olmadığını test eden ki-kare testi ve anlamlı bir ilişki olup olmadığıyla beraber bu ilişkinin yönünü de test eden korelasyon testi uygulanmıştır. Yapılan testler sonucunda göze çarpan sonuçlar olmuştur. Genel olarak uykusuzluğu etkileyen faktörlerin birden fazla olduğu ve bunların özellikle; cinsiyet, medeni durum ve çalışma durumu değişkenlerine göre farklılaştıkları görülmüştür.

KAYNAKÇA

[1] Şenol V, Soyuer F, v.d. (2012), "Adolesanlarda Uyku Kalitesi ve Etkileyen Faktörler ", Kocatepe Tıp Dergisi, Cilt 13, No:2, s.93-102.

[4] Mehta, C. R., & Patel, N. R. (2011). IBM SPSS exact tests. Armonk, NY: IBM Corporation.

[5] Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2013). Multivariate Data Analysis: Pearson Education Limited.

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

SEÇİLİ MAKROEKONOMİK FAKTÖRLERİN S&P 500 ENDEKSİNE ETKİSİ

İREM TUĞRUL 18023010

Danışman: Doç. Dr. Serpil KILIÇ DEPREN

ÖZET

S&P 500 endeksi Amerika'da yer alan 500 büyük firmanın hisselerini taşıyan bir endeks olarak belirtilebilir. . Bu nedenle borsa üzerinden yatırım amaçlı günümüzde tasarruf sahipleri tarafından oldukça önemli bir yere sahip olduğunu söyleyebiliriz. Bu çalışmada seçili makroekonomik faktörlerden Amerika'da faiz oranları, para arzı, petrol fiyatı ve döviz kuru faktörlerinin S&P 500 endeksi üzerindeki etkisini araştırmak olacaktır. Çalışmada, zaman serileri analizi kullanılmıştır. Serilerin durağanlığı ADF, KPSS ve PP birim kök testleri ile araştırılmıştır. Bu bağlamda S&P 500 endeksine olan etkinin Johansen eşbütünleşme yaklaşımıyla uzun dönem tahminleri yapılmıştır. Değişkenler arasındaki ilişkileri Granger nedensellik testini kullanarak kısa dönemli ilişkinin varlığı incelenmiştir. Johansen eşbütünleşme testiyle değişkenler arasında uzun dönemli bir ilişkinin olduğu sonucuna ulaşılmıştır. Granger nedensellik testi sonucunda da para arzı değişkeninin S&P 500 endeksi değişkeni üzerinde anlamlı bir nedensel etkisi olduğu sonucuna varılmıştır.

S&P 500 Kavramı

S&P 500, Amerika Birleşik Devletleri borsalarında listelenen en büyük 500 şirketin hisse senedi performansını izleyen bir borsa endeksidir. En çok takip edilen hisse senedi endekslerinden birisidir. Endeks 500 ileri gelen şirketi kapsar ve mevcut piyasa değerinin yaklaşık % 80'ini oluşturur. S&P 500 endeksi, halka açık ağırlıklı bir endekstir.

Endeks Nedir?

Borsada işlem gören hisse senetlerinin fiyat ve getirilerinin performanslarını ölçmek için kullanılan göstergedir. Borsa endeksleri hisselerdeki fiyat hareketlerinden yola çıkılarak borsanın genel trendlerinin belirlenmesinde kullanılır.

S&P 500 Endeksi Nedir?

S&P 500 endeksi, Amerika Birleşik Devletleri borsasında işlem gören en büyük 500 şirketin performansını ölçen bir borsa endeksidir. Bu endeks dünyadaki en önemli hisse senedi endekslerinden biridir. ABD ekonomisinin genel sağlığına dair de önemli bilgiler verir.

Makroekonomik Göstergeler

Bir ekonominin genel sağlığı, büyümesi ve performansını ölçmek için kullanılan göstergelerdir. Hisse senetleri ve diğer finansal varlıklar üzerinde önemli etkilere neden olabilir. Çalışmada para arzı, bir ekonomideki para birimlerinin miktarını ve likiditesini belirler. Amerika Birleşik Devletleri'nde para arzı, ABD Merkez Bankası (Federal Reserve) tarafından yönetilen para politikası aracılığıyla belirlenir. ABD'de petrol fiyatları, genellikle West Texas Intermediate (WTI) ve Brent petrolü fiyatları olarak ölçülür. Petrol fiyatları zaman içinde değişebilir ve gelecekte ne olacağı belirsizdir. Döviz kuru, Amerika Birleşik Devletleri'nde döviz kuru, Amerikan dolarının başka bir para birimi karşısındaki değerini ifade eder. Faiz oranı, ABD 10 yıllık tahvil faizi denildiğinde ABD Hazine'sinin çıkardığı 10 yıllık tahvillerin piyasa faiz oranı anlaşılıyor.

ARAŞTIRMANIN VERİ SETİ

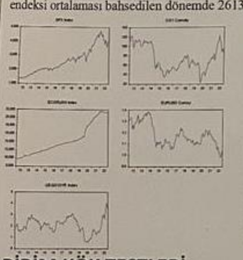
Çalışmada 31/01/2012-30/11/2022 tarihleri arasındaki aylık veriler kullanılarak seçili makroekonomik göstergelerin S&P 500 endeksine etkisi incelenmiştir. Makroekonomik değişkenler olarak faiz oranı, para arzı, petrol fiyatı ve döviz kuru seçilmiştir. Çalışmanın bağımlı değişkeni S&P 500 endeksi, bağımsız değişkenleri ise faiz oranı, para arzı, petrol fiyatı ve döviz kurudur. Bu çalışmada 131 aylık veri kullanılmıştır. Çalışmanın verisi Bloomberg'den alınmıştır. S&P 500 endeksi, faiz oranı, para arzı, petrol fiyatı ve döviz kurunun kapanış fiyatları baz alınmıştır.

TANIMLAYICI İSTATİSTİKLER

	SPX INDEX	CO1 COM...	ECORUSN...	EURUSD ...	USGG10Y...
Mean	2613.979	74.78939	14401.26	1.179315	2.081289
Median	2423.410	67.25000	13470.10	1.158000	2.124600
Maximum	4766.180	122.8800	21845.70	1.386700	4.047800
Minimum	1310.330	22.74000	9677.500	0.980200	0.528200
Std. Dev.	916.4029	28.07435	3704.126	0.097893	0.870738
Skewness	0.654475	0.260590	0.788910	0.490777	-0.059013
Kurtosis	2.459995	1.826876	2.352979	2.335495	3.160252

Yandaki tabloda 2012 – 2022 dönemine ait aylık olarak incelenen verilerin betimleyici istatistiklerine yer verilmiştir. Çarpıklık katsayısı için değerinin +1.5 ile -1.5 arasında olduğu görülmektedir

S&P 500 endeksi 2012-2015 döneminde minimum 1310.330, maksimum 4766.180 değerini almıştır. S&P 500 endeksi ortalaması bahsedilen dönemde 2613.979, standart sapması ise 916.4029'dur.



S&P 500 endeksi için 2019- 2020 yılları arasında düşüş yaşadığı 2020'den 2021'in ortalarına kadar yükselişe geçip 2021'in ortalarında zirve yapıp sonrasında hafif düşüşe geçmiştir. Petrol fiyatı (Co1 Comdty),2012'den 2014'ün ortalarına kadar belirli bir seviyede dalgalanma gösterirken 2014'ün ortalarında ani bir düşüş yaşamıştır.2020 yılında da ani bir düşüş yaşayıp sonra yükselişe geçmiştir. Para arzı (Ecorusn Index), 2012'den 2020'ye kadar sürekli olarak artmıştır. 2020'den 2022'ye kadar yükselişi hız kazanmıştır. Döviz Kuru (Eurusd Curncy), 2012'den 2014'ün ortalarına kadar belirli seviyelerde dalgalanmıştır. 2014'ten 2015'e kadar düşüş yaşamıştır.2021'den 2022'ye kadar da belirli bir seviyede düşmüştür. Faiz Oranı (Usgg10yr Index), 2018'den 2020 ortalarına kadar azalmıştır. 2020 ortalarından 2022'ye kadar genel olarak yükselişe geçtiğini söyleyebiliriz.

BİRİM KÖK TESTLERİ

Null Hypothesis: D(SPX_INDEX) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=12)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-13.30604	0.0000
Test critical values:	1% level	-3.481623	
	5% level	-2.883930	
	10% level	-2.578788	

*MacKinnon (1996) one-sided p-values.

Durağanlığı test etmek için uygulamada kullanılan yöntemlerden birisi birim kök analizidir. Serilerin durağanlığını tespit etmek için ADF birim kök testi , PP birim kök testi ve KPSS birim kök testi kullanılmıştır. Yapılan analizler sonucunda, veriler ilk halinde durağan değil fakat birinci dereceden farkları alındığında durağan olduğunu söyleyebiliriz. S&P 500 endeksi için birinci dereceden farkı alındığında durağan hale geldiği yorumu yapılabilir.

ÇOKLU REGRESYON MODELİ

$SPX = \beta_0 + \beta_1 CO1 + \beta_2 ECORUSN + \beta_3 EURUSD + \beta_4 USGG10YR + \beta_5 Ui$

Modeldeki Prob (F-Statistic) değeri. 0,05 ten küçük olduğu için model anlamlıdır. Bağımlı değişken olan S&P 500 endeksinin %21.44'lük kısmı modele dahil edilen değişkenler tarafından açıklanmıştır. Geri kalan kısım ise hata terimi aracılığıyla modele dahil edilmeyen değişkenlere aittir.

Dependent Variable: FARKSPX_INDEX
Method: Least Squares
Date: 05/31/23 Time: 00:03
Sample (adjusted): 2012M02 2022M11
Included observations: 130 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
FARKCO1_COMDTY	7.598889	1.742584	4.360701	0.0000
FARKECORUSN_INDEX	1.476996	0.035372	41.75574	0.0000
FARKEURUSD_CURNCY	888.1829	421.6952	2.106294	0.0372
FARKUSGG10YR_INDEX	0.140365	0.022453	6.251461	0.0000
C	15.76128	11.59670	1.359235	0.1765

R-squared	0.214440	Mean dependent var	21.29000
Adjusted R-squared	0.189302	S.D. dependent var	129.4420
S.E. of regression	116.5481	Akaike info criterion	12.39219
Sum squared resid	1697931	Schwarz criterion	12.50248
Log likelihood	-800.4921	Hannan-Quinn criter.	12.43700
F-statistic	8.530514	Durbin-Watson stat	2.261390
Prob(F-statistic)	0.000004		

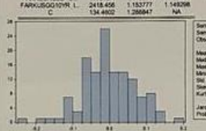
VARSAYIMLAR

Modelde sağlanması gereken varsayımlar sırayla sınanmıştır.

Variance Inflation Factors
Date: 05/30/23 Time: 14:04
Sample: 2012M01 2022M11
Included observations: 130

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
FARKCO1_COMDTY	3.036600	1.289056	1.287934
FARKECORUSN_I...	0.003514	1.310774	1.038668
FARKEURUSD_C...	177818.4	1.116390	1.109168
FARKUSGG10YR_I...	2418.456	1.153777	1.149298
C	134.4602	1.266847	NA

Çoklu doğrusal bağlantı varsayımıyla modelin incelenmesi yanda gösterilmiştir. VIF değeri 1 ile 5 arasında olması gerekmektedir. Görüldüğü üzere tüm değerler 1 ile 5 arasında değer almıştır. Bu nedenle çoklu doğrusal bağlantı sorunu yoktur. Bu şart sağlanmıştır.



Series: Residuals
Sample 2012M01 2022M11
Observations 131

Mean	-7.50e-16
Median	-0.006493
Maximum	0.175524
Minimum	-0.227909
Std. Dev.	0.070647
Skewness	-0.143226
Kurtosis	3.263761
Jarque-Bera	0.827516
Probability	0.661127

Normallik testine göre probability değeri 0.05'ten büyükse kalıntıların normal dağıldığı sonucuna ulaşılır. Bu testin sonucu probability değeri 0.6611 > 0.05 olduğu için kalıntılar normal dağılmıştır.

Breusch-Godfrey Serial Correlation LM Test:
Null hypothesis: No serial correlation at up to 2 lags

F-statistic	2.925837	Prob. F(2,123)	0.0574
Obs*R-squared	5.903824	Prob. Chi-Square(2)	0.0522

Burada ters kural mantığı geçerlidir. Probability değeri 0.0574 > 0.05'ten büyük olduğu için otokorelasyon problemi yoktur.

Heteroskedasticity Test: Breusch-Pagan-Godfrey
Null hypothesis: Homoskedasticity

F-statistic	2.321118	Prob. F(4,125)	0.0604
Obs*R-squared	8.988243	Prob. Chi-Square(4)	0.0614
Scaled explained SS	14.23540	Prob. Chi-Square(4)	0.0066

Burada da ters kural mantığı vardır. Boş hipotezi, probability değeri 0.0604 > 0.05'ten olduğu için kabul edilir. Bu demek oluyor ki, değişen varyans sorunu yoktur. Bir diğer ifadeyle sabit varyans vardır. Böylelikle bu varsayım da sağlanmış olmaktadır.

GRANGER NEDENSELLİK TESTİ

Pairwise Granger Causality Tests
Date: 05/30/23 Time: 14:07
Sample: 2012M01 2022M11
Lags: 2

Null Hypothesis:	Obs	F-Statistic	Prob.
FARKCO1_COMDTY does not Granger Cause FARKSPX_INDEX	128	0.25538	0.7750
FARKSPX_INDEX does not Granger Cause FARKCO1_COMDTY		2.30709	0.1036
FARKECORUSN_INDEX does not Granger Cause FARKSPX_INDEX	128	5.51378	0.0051
FARKSPX_INDEX does not Granger Cause FARKECORUSN_INDEX		2.28960	0.1077
FARKEURUSD_CURNCY does not Granger Cause FARKSPX_INDEX	128	0.51390	0.5994
FARKSPX_INDEX does not Granger Cause FARKEURUSD_CURNCY		0.04583	0.9552
FARKUSGG10YR_INDEX does not Granger Cause FARKSPX_INDEX	128	0.47362	0.6237
FARKSPX_INDEX does not Granger Cause FARKUSGG10YR_INDEX		0.63671	0.5306
FARKECORUSN_INDEX does not Granger Cause FARKCO1_COMDTY	128	7.49119	0.0009
FARKCO1_COMDTY does not Granger Cause FARKECORUSN_INDEX		0.93513	0.3953
FARKEURUSD_CURNCY does not Granger Cause FARKCO1_COMDTY	128	3.67423	0.0282
FARKCO1_COMDTY does not Granger Cause FARKEURUSD_CURNCY		0.14096	0.8687
FARKUSGG10YR_INDEX does not Granger Cause FARKCO1_COMDTY	128	0.56422	0.5703
FARKCO1_COMDTY does not Granger Cause FARKUSGG10YR_INDEX		0.82813	0.4393
FARKEURUSD_CURNCY does not Granger Cause FARKECORUSN_INDEX	128	0.11011	0.8958
FARKECORUSN_INDEX does not Granger Cause FARKEURUSD_CURNCY		1.57097	0.2120
FARKUSGG10YR_INDEX does not Granger Cause FARKECORUSN_INDEX	128	4.12002	0.0185
FARKECORUSN_INDEX does not Granger Cause FARKUSGG10YR_INDEX		0.15428	0.8572
FARKUSGG10YR_INDEX does not Granger Cause FARKEURUSD_CURNCY	128	2.59772	0.0785
FARKEURUSD_CURNCY does not Granger Cause FARKUSGG10YR_INDEX		0.01734	0.9828

Kısa dönemli ilişkileri test etmek için Granger-Nedensellik testi uygulanır. Nedensellik analizinin amacı değişkenler arasındaki ilişkinin varlığını test etmek ve ilişkinin yönünü belirlemektir. Granger nedensellik testi sonucunda para arzı değişkeninin S&P 500 endeksi değişkeni üzerinde anlamlı bir nedensel etkisi bulunmuştur.

JOHANSEN EŞBÜTÜNLEŞME TESTİ

Johansen Cointegration Test

Date: 05/31/23 Time: 22:00
Sample (adjusted): 2012M04 2022M11
Included observations: 128 after adjustments
Trend assumption: Linear deterministic trend
Series: FARKSPX_INDEX FARKCO1_COMDTY FARKECORUSN_INDEX FARKEURUSD_C...
Lags interval (in first differences): 1 to 1

Unrestricted Cointegration Rank Test (Trace)

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
None *	0.473870	291.8657	69.81889	0.0000
At most 1 *	0.459495	209.6631	47.85613	0.0000
At most 2 *	0.343573	130.9108	29.79707	0.0000
At most 3 *	0.288555	77.03006	15.49471	0.0000
At most 4 *	0.229981	33.45150	3.841465	0.0000

Trace test indicates 5 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Unrestricted Cointegration Rank Test (Maximum Eigenvalue)

Hypothesized No. of CE(s)	Eigenvalue	Max-Eigen Statistic	0.05 Critical Value	Prob.**
None *	0.473870	82.20254	33.87687	0.0000
At most 1 *	0.459495	78.75227	27.58434	0.0000
At most 2 *	0.343573	53.88078	21.13162	0.0000
At most 3 *	0.288555	43.57855	14.26460	0.0000
At most 4 *	0.229981	33.45150	3.841465	0.0000

Max-eigenvalue test indicates 5 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Değişkenler arasındaki uzun dönem ilişkiyi incelemek için eş bütünleşme testleri uygulanmıştır. Zaman serileri arasında bir eşbütünleşme varsa, gerçek bir ilişkinin varlığından söz edilebilir.

Johansen eşbütünleşme testi sonucunda %5 seviyesinde bağımlı (S&P 500 endeksi) ve bağımsız değişkenler (petrol fiyatı, para arzı, döviz kuru ve faiz oranı) arası uzun dönemli ilişkinin var olduğu sonucuna ulaşılmıştır.

HATA DÜZELTME MODELİ

Vector Error Correction Estimates
Date: 05/31/23 Time: 22:06
Sample (adjusted): 2012M04 2022M11
Included observations: 128 after adjustments
Standard errors in ( ) & t-statistics in [ ]

Cointegrating Eq:	CointEq1
FARKSPX_INDEX(-1)	1.000000
FARKCO1_COMDTY(-1)	-33.03054 (4.58622) [-7.19850]
FARKECORUSN_IND...	-0.831861 (0.15127) [-5.49913]
FARKEURUSD_CUR...	-710.7329 (1219.00) [-0.58304]
FARKUSGG10YR_IN...	-966.2250 (137.270) [-7.03884]
C	-115.8623

Uzun dönemdeki katsayıların olduğu denklem aşağıdaki gibidir.

SPX = 33.03 CO1 + 0.83 ECORUSN + 710.73 EURUSD + 966.22 USGG10YR +115.86

SONUÇ

Bu çalışmada, seçili makroekonomik değişkenlerden para arzı, petrol fiyatı, döviz kuru ve faiz oranının S&P 500 endeksi üzerindeki etkisi araştırılmıştır Çalışmada yapılan birim kök testleri sonucunda verilerin ilk halinde durağan olmadığı fakat birinci dereceden farkları alındığında durağan olduğu sonucuna ulaşılmıştır. Sonrasında çoklu regresyon modeli kurulmuş ve modelin anlamlılığı test edilmiştir ve model anlamlı çıkmıştır. S&P 500 endeksine etki eden bağımsız değişkenlerin petrol fiyatı, para arzı, döviz kuru ve faiz oranı olduğu sonucuna varılmıştır. Granger nedensellik testi sonucunda para arzı değişkeninin S&P 500 endeksi değişkeni üzerinde anlamlı bir nedensel etkisi bulunmuştur. Johansen eşbütünleşme testi sonucunda %5 seviyesinde bağımlı (S&P 500 endeksi) ve bağımsız değişkenler (petrol fiyatı, para arzı, döviz kuru ve faiz oranı) arası uzun dönemli ilişkinin var olduğu sonucuna ulaşılmıştır. Kısa dönem dalgalanmaların ne kadar sürede uzun dönem dengesini yakaladığını da hata düzeltme modeli ile bulunmuştur.

KAYNAKÇA

[1] Sirucek, M. (2012). Macroeconomic variables and stock market: US review
[2]Dolaşık, B. (2022). Seçili Makroekonomik Göstergelerin Borsa İstanbul (Bist-100) Endeksine Etkisi
[3]Aygün, M. (2022). Makroekonomik Değişkenler ile Borsa İstanbul Arasındaki Nedensellik İlişkisi: Toda-Yamamoto Testi

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

Kredi Kartı Başvurusu Veri Setinde Keşifsel Veri Analizi ve Makine Öğrenmesi Uygulaması

Öğrenci: Beyzagül GÖKSU - 18023031

Danışman: Prof.Dr. Ali Hakan BÜYÜKLÜ

Finansal hizmetlerin giderek dijitalleştiği günümüzde, kredi kartları bireylerin günlük harcamalarını yönetmeleri ve finansal ihtiyaçları için kolay bir ödeme aracı sağlamaları açısından önemli bir rol oynamaktadır. Bankalar ve finansal kuruluşlar, kredi kartı başvurularını değerlendirirken risk analizi yaparak başvuru sahiplerine kredi kartı verme veya reddetme kararı vermektedirler. Bu süreçte, teknolojinin de gelişmesiyle geleneksel yöntemlerin yanı sıra veri analitiği ve makine öğrenmesi gibi teknikler doğru ve hızlı kararlar almak için kullanılmaya başlamıştır.

Bu çalışma, keşifsel veri analizi ve makine öğrenmesi tekniklerini kullanarak kredi kartı başvurularını incelemeyi hedeflemektedir. Kredi kartı başvurusu veri seti üzerinde gerçekleştirilen bu analizler, başvuruların kabul edilme etkenlerinin anlaşılmasına ve onaylanma veya reddedilme olasılıklarının tahmin edilmesine yardımcı olacaktır. Bununla birlikte, bu çalışma, finansal kuruluşlara ve kredi kartı sağlayıcılarına; başvuru sürecini optimize etme, risk yönetimini iyileştirme ve müşteri memnuniyetini artırma potansiyeli sunmaktadır.

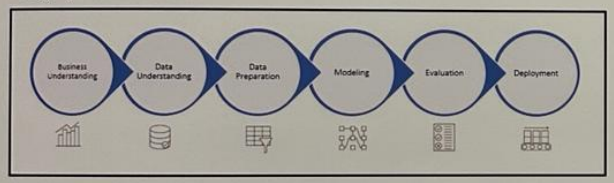
Kredi Kartı Nedir?

Kredi limiti, bir kredi kartı veya bir kredi hattı gibi kredi ürünleri için belirlenir. Kredi kartı limiti, kredi kartı sahibinin belirli bir dönemde kullanabileceği maksimum tutarı ifade eder. Kredi hattı limiti ise, müşterinin belirli bir süre boyunca kullanabileceği maksimum kredi miktarını ifade eder. Kredi limiti, müşterinin finansal durumuna ve risk profiline bağlı olarak belirlenir. Kredi limiti aynı zamanda müşterinin kredi notu, gelir düzeyi ve ödeme geçmişi gibi faktörlerle ilişkilidir. Kredi limiti, müşterinin kredi kullanma kapasitesini gösterir ve müşteri bu limit dahilinde kredi alabilir veya harcama yapabilir.

Bankalarda Risk Yönetimi

Saunders ve Cornett (2014), banka yönetiminin hissedarlar için getirileri artırmayı amaçladığını ve bunun riskin artmasıyla gerçekleştiğini belirtmektedir. Bankaların karşılaştığı birçok risk bulunmaktadır. Bunlar kredi riski, faiz oranı riski, piyasa riski, bilanço dışı risk, teknoloji ve operasyonel risk, döviz riski, ülke veya egemen riski, likidite riski, likidite riski ve iflas riski olarak sıralanabilir. Bu riskleri etkin bir şekilde yöneterek, bankalar daha iyi performans gösterebilirler.

Keşifçi Veri Analizi



Veri Keşfi: İlk adım veri setinin anlaşılması ve keşfedilmesidir. Bu adımda, verilerin kaynakları, toplama yöntemleri ve veri noktalarının neyi temsil ettiği gibi bilgilere de dikkat edilir.

Veri Temizliği: Veri setindeki hataları, eksik değerleri ve anormal verileri tespit etmek ve bunları düzeltmek için veri temizliği yapılır. Bu adımda, verilerin tutarlılığı ve doğruluğu sağlanır. Eksik veriler, aykırı değerler veya çelişkili veriler gibi problemler belirlenir ve uygun bir şekilde ele alınır.

Veri Görselleştirme: Veri setinin keşfedilmesi ve analizi için görselleştirmeler kullanılır. Grafikler, histogramlar, dağılım eğrileri ve kutu grafikleri gibi görsel araçlar kullanılarak veri setindeki desenler ve ilişkiler daha iyi anlaşılır. Görselleştirmeler, veri setindeki eğilimleri ve varyasyonları göstererek analistlere daha fazla içgörü sağlar.

İlişkilerin ve Desenlerin Analizi: Bu adımda, veri setindeki değişkenler arasındaki ilişkileri ve desenleri keşfetmek için istatistiksel analiz yöntemleri kullanılır. Korelasyon analizi, regresyon analizi, kümeleme ve faktör analizi gibi teknikler kullanılarak veriler arasındaki bağlantılar ve desenler ortaya çıkarılır. Bu adım, veri setindeki önemli faktörleri belirlemek için kullanılır.

Makine Öğrenmesi Hakkında

Makine öğrenmesi, belirli bir problemi çözmek için birçok farklı algoritma ve teknik kullanır. Bu algoritmalar, veriye dayalı örüntüleri bulmak ve modele öğrenme yeteneği kazandırmak için kullanılır. Makine öğrenmesi modelleri, genellikle eğitim veri setleri üzerinde eğitilir, ardından bu öğrenme sürecinden elde edilen bilgilerle yeni verileri tahmin etmek veya sınıflandırmak için kullanılır. Bu çalışmada makine öğrenmesi modellerinden LightGBM ve CatBoosting öne çıkmıştır.

LGBM: Ağaç tabanlı bir modeldir ve özellikle sınıflandırma ve regresyon problemlerinde kullanılır. Gelişmiş bir özellik olan "leaf-wise" büyüme stratejisini kullanarak ağaçları hızlı bir şekilde oluşturur. Bu strateji, enformasyon kazancı maksimize edilerek ağaç yapısının derinleşmesini sağlar.

CatBoosting: Hızlı ve ölçeklenebilir bir algoritmadır ve büyük veri setlerinde ve yüksek boyutlu özelliklerde etkili bir şekilde çalışabilir. Overfitting dirençlidir ve doğru hiper parametre ayarlarıyla yüksek performans sağlayabilir.

Veri Hakkında

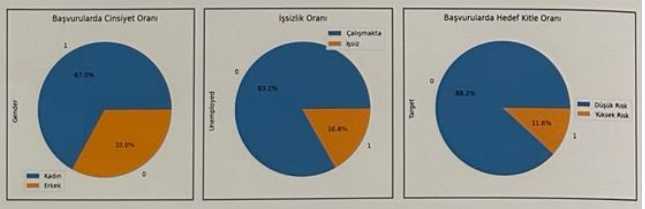
Araştırmada kullanılan veri seti Kaggle web sitesinden alınmıştır.

Veri setinin adı "Credit Card Approval Prediction"dır. İki ayrı veri setinden oluşmaktadır.

application_record.csv: 438.557 müşterinin kişisel verilerini içermekte.

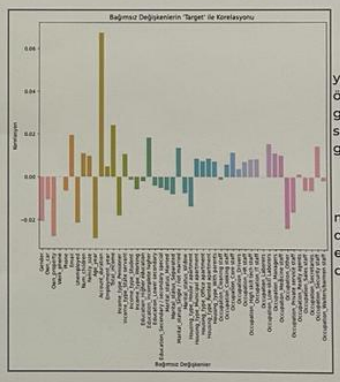
credit_record.csv: Belli müşterilerin 12 aylık kredi durumlarını içermekte.

Her müşterinin kredi durumları bilinmediğinden çalışmada kullanılmak üzere kredi durumları bilinen müşteriler üzerinden veri setleri birleştirildi. Veri işlendikten sonra çalışma 36449 müşterinin kişisel verileri ve kredi skorları ile devam etti.



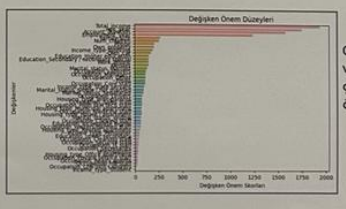
Çalışmada müşterilerin %67'si Kadın, %33'ü erkektir. Çoğunluk iş sahibi iken %16.8'i çalışmamakta dolayısıyla bu müşterilerin kredi başvurusunun reddedilmesi beklenmekte. Kredi skoru olarak belirlenen Hedef değişkenine göre yüksek risk grubunda olan %11.8 müşteri reddedilmiştir. Yani işsiz olmasına rağmen kredi başvurusu kabul edilen %5 müşteri vardır.

Makine Öğrenmesi Uygulamaları

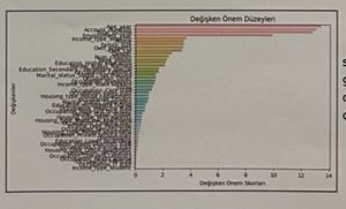


En iyi pozitif ilişkilerin sırasıyla hesap yaşı, yıllık gelir, e-posta ve yüksek öğrenim olduğu gözlemlenmekte. En güçlü negatif ilişkiler ise yaş, gayrimenkul sahipliği, işsizlik ve cinsiyet olarak gözlemleniyor.

Sonuç olarak çalışmada kullanılan makine öğrenmesi modellerinin en ilişkili değişkenler olarak hesap yaşı, yıllık gelir, e-posta ve yüksek öğrenim değişkenlerini öngörmesi bekleniyor.



LGBM: Modelin doğruluk skoru 0.8853'tür. Yıllık gelir, yaş, çalışma süresi ve hesap süresi değişkenlerini modeli en çok etkileyen değişkenler olarak öngörmüştür.



CatBoosting: Modelin doğruluk skoru 0.8855'tür. LGBM modeli gibi yıllık gelir, yaş, çalışma süresi ve hesap süresi değişkenlerini modeli en çok etkileyen değişkenler olarak öngörmüştür.

KAYNAKÇA

1. 'Credit Card Approval Prediction by Using Machine Learning Techniques' M. P. C. Pei University of Colombo School of Computing
2. 'Predicting Credit Card Approvals using Machine Learning' by Lopa Nayak, Aman Sangal
3. Credit Card Approval Prediction Data from Kaggle

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

TÜRKİYE'DE DEPREM BİLGİSİ VE BİLİNCİ ÜZERİNE BİR ARAŞTIRMA

EMİRHAN BAYRAK 18023033

Danışman: Dr. Öğr. Üyesi Doğan YILDIZ

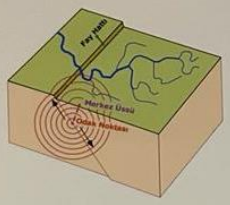
Deprem Dünya genelinde en tehlikeli ve en çok zarar veren doğal afetlerden biridir. Birçok ülke gibi Türkiye de deprem riski yüksek bir ülkedir ve sık sık depremlerle karşı karşıya kalmaktadır. Bu durum, Türkiye'de deprem bilinçliliğinin önemini artırmaktadır. Bu tezde, Türkiye'deki deprem bilinçliliği ölçülmektedir. Bu amaçla, farklı yaş gruplarından ve farklı eğitim düzeylerinden katılımcılarla toplamda 18 soruluk bir anket çalışması yapılmıştır. Ankette, deprem hakkındaki bilgi düzeyi, deprem hazırlığı yapma eğilimi, deprem anında nasıl davranılması gerektiği konusunda bilgi sahibi olma durumu gibi konulara yer verilmiştir. Ankete katılan 192 kişinin cevapları, SPSS programı kullanılarak analiz edilmiştir. Analizde, Cronbach testi ile güvenilirlik test edildi, aynı zamanda çapraz tablolar kullanılmıştır ve bu çapraz tablolar Ki-Kare testi ile desteklenmiştir. Bu yöntemlerle sorular arasındaki ikili ilişkiler arasında hipotezler test edilmiştir. Analiz sonuçları elde edilen SPSS çıktıları üzerinden yorumlanmıştır.

DEPREM

Depremler, dünya tarihinde sayısız felaketlere neden olmuştur. İnsanlar, depremlerin doğrudan ve dolaylı etkileri yoluyla binlerce yıl boyunca zarar görmüş ve yaralanmışlardır. Depremler, binaların ve diğer yapıların yıkılmasına, su kaynaklarının kirlenmesine, elektrik ve ulaşım ağlarının hasar görmesine ve hatta ekonomik çöküntülere neden olabilir.

DEPREM TÜRLERİ

Günümüzde bilinen 7 farklı deprem türü vardır bunlar tektonik depremler, volkanik depremler, çökme depremler, yer kayması depremleri, heyelan depremleri, denizaltı depremleri, ve yapay depremlerdir. Dünyadaki depremlerin yaklaşık %95'ini "Tektonik Depremler" oluşturur.



Tektonik Depremleri açıklayacak olursak;

Yer kabuğundaki levhaların sürtünmesi veya birbirinden uzaklaşması sonucu meydana gelen depremlerdir.

SİSMOGRAF

Deprem ölçümleri, sismograflar adı verilen cihazlar kullanılarak yapılır. Sismograflar, deprem sırasında yer kabuğundaki hareketleri kaydeden hassas aletlerdir. Sismografların temel işlevi, deprem anındaki titreşimleri ölçmek ve kaydetmektir. Sismograflar, yerin altındaki kayaların hareketlerini ölçmek için kullanılan özel bir dizi sensöre sahiptir. Bu sensörler, yer kabuğundaki herhangi bir titreşimi ölçmek için tasarlanmıştır.

Sismografın çalışma prensibi oldukça basittir: sensörler, deprem sırasında yer kabuğundaki titreşimleri algılar ve bu titreşimleri kâğıt üzerine çizerek kaydeder. Deprem ölçümleri, sismografların kaydettiği verilerin analiziyle yapılır. Bu veriler, depremin büyüklüğü ve merkez üssünün konumu hakkında bilgi sağlar. Deprem ölçümleri, genellikle Richter ölçeği adı verilen ölçekte ifade edilir ve bu ölçekteki bir artış, depremin şiddetinin 10 kat arttığını gösterir.

DEPREM ÖNCESİNDE YAPILMASI GEREKENLER

Deprem öncesinde kendimizi güven altında tutmak için yapılabilecek ve alınabilecek birçok tedbir vardır. Bunlardan bazıları şu şekilde sıralayabiliriz;

Deprem sigortası yaptırmak: Ev, iş yeri ve diğer mülkler için deprem sigortası yaptırmak önemlidir.
Acil durum çantası hazırlamak: Deprem sırasında veya sonrasında kullanılmak üzere bir acil durum çantası hazırlamak faydalı olabilir. Bu çanta içinde gıda, su, ilaçlar, ilk yardım malzemeleri, çakmak, el feneri ve battaniye gibi temel eşyalar bulunmalıdır.
Evdeki eşyaların güvenliğini sağlamak: Deprem sırasında eşyaların düşmesini önlemek için, dolap ve rafların sabitlenmesi gerekmektedir.
Acil durum planı yapmak: Ailenizle veya ev arkadaşlarınızla bir acil durum planı yapmak, herkesin deprem sırasında nerede olacağı ve nasıl iletişim kuracağı konusunda bilgi sahibi olmasını sağlayacaktır.
Bina güvenliğini kontrol etmek: Deprem öncesinde bina güvenliği hakkında bilgi sahibi olmak önemlidir. Bina güçlendirme çalışmaları yapılması gereken yapılar varsa, bu çalışmaların yapılması gerekmektedir.

ACİL DURUM ÇANTASININ HAZIRLANMASI

Afide bulundurulacak acil durum çantası, doğal afetler gibi beklenmedik olaylarda hızlı bir şekilde hareket etmenize yardımcı olur ve kritik durumlarda hayatta kalmayı sağlayabilir. Acil durum çantasında bulunması gerekenleri şu şekilde sıralanabilir: su, gıda, ilk yardım kitleri, hijyen malzemeleri, giyim, ışıklandırma ve iletişim malzemeleri. Bu malzemeler yaşanılan bölgeye ve kişinin ihtiyaçlarına göre düzenlenebilir. Bu ürünlerden mümkün olduğunca 3 ile 5 gün arasında tüketilebilecek miktarda bulundurulmalıdır.

DEPREM ANINDA YAPILMASI GEREKENLER

Deprem anında bireyin depremi hissetmesi, hissettikten sonra buna tepki vermesi ve hissettikten sonra buna vereceği tepkiyle birlikte yapması gereken önemli davranışlar vardır. Bunları şöyle sıralayabiliriz; sakin olmak, hemen korunmak, kapı ve pencere yanlarına yakın olmak, asansör kullanmamak ve acil çıkış kapılarına yakın durmak gibi birkaçı örnek verilebilir.

DEPREM SONRASINDA YAPILMASI GEREKENLER

Deprem sonrasında yapılması gerekenler, insanların hayatta kalmalarına, yaralanmalarının önlenmesine ve hasarın azaltılmasına yardımcı olacak önlemleri içerir. Deprem sonrasında yapılması gerekenler şu şekildedir; kendinizi ve diğerlerini güvende tutun, yaralılara yardım edin, elektrik, gaz ve su bağlantılarını kesin, haberleri takip edin, yapısal hasarın kontrolü, geçici barınak sağlayın ve acil durum çantası kullanın.

KAYNAKÇA

[1] Freed, A. M. (2007). Afterslip (and only afterslip) following the 2004 Parkfield, California, earthquake. Geophysical Research Letters, 34(6).
[2] Akkaş, H. H. (2023). Deprem Bilinci. Kahramanmaraş Merkezli Depremler Sonrası İçin, 23.

DEPREM VE TÜRKİYE

Türkiye, Kuzey Anadolu Fay Hattı, Doğu Anadolu Fay Hattı ve Batı Anadolu Fay Hattı gibi önemli fay hatları üzerinde yer alır. Bu fay hatları, sık sık depremlerin meydana geldiği bölgelerdir ve büyük depremlere neden olabilirler.

Ayrıca, Marmara Denizi bölgesi, Türkiye'nin en yüksek deprem riskine sahip bölgelerinden biridir. Bu bölgelerdeki depremler, önemli hasarlar ve kayıplara yol açabilir.

ANKET SONUÇLARI VE CRONBACH TESTİ SONUCU

DEMOGRAFİK SORULARIN SONUÇLARI

Sorular	Şıklar	N	%
Cinsiyetiniz nedir?	Erkek	103	46,4%
	Kadın	89	53,6%
Yaş aralığınız nedir?	18-25	135	70,3%
	26-35	29	15,1%
	36-50	12	6,3%
	50 ve üstü	16	8,3%
Eğitim durumunuz nedir?	İlkokul	5	2,6%
	Ortaokul	11	5,7%
	Lise	30	15,6%
	Üniversite	141	73,4%
	Lisansüstü	5	2,6%
Çalışma durumunuz nedir?	Kamu Sektörü	11	5,7%
	Özel Sektör	62	32,3%
	Öğrenci	99	51,6%
	Ev Hanımı	7	3,6%
	Kendi işimin patronuyum	3	1,6%
	Çalışmıyorum	10	5,2%
Gelir durumunuz nedir?	0-8.500	97	50,5%
	8.501-13.500	39	20,3%
	13.501-20.000	41	21,4%
	20.001-35.000	11	5,7%
	35.001 ve üstü	4	2,1%
Yaşadığınız konutun tipi nedir?	Müstakil	17	8,9%
	Apartman	133	69,3%
	Site	33	17,2%
	Yurt-Lojman	9	4,7%
Hanenizdeki kişi sayısı nedir?	1	9	4,7%
	2	33	17,2%
	3	44	22,9%
	4	68	35,4%
	5+	38	19,8%

DEPREM İLE İLGİLİ SORULARIN SONUÇLARI

Sorular	Şıklar	N	%
Depremlerin nasıl meydana geldiğini biliyor musunuz?	Evet, tam olarak biliyorum	116	60,4
	Biraz biliyorum	71	37,0
	Hiç bilgim yok	5	2,6
Kendinizi ve ailenizi bir depreme hazırlamak için neler yapmanız gerektiğini biliyor musunuz?	Evet, her şeyi biliyorum	85	44,3
	Birkaç şey biliyorum	99	51,6
	Hiçbir şey bilmiyorum	8	4,2
Deprem anında yapılması gereken doğru davranışları	Evet, tam olarak biliyorum	99	51,6
	Biraz biliyorum	86	44,8
	Hiçbir şey bilmiyorum	7	3,6
Deprem olabilecek bir bölgede yaşıyor musunuz?	Evet, deprem olabilecek bir bölgede yaşıyorum	173	90,1
	Hayır, deprem olabilecek bir bölgede yaşamıyorum	14	7,3
	Bilmiyorum	5	2,6
Binanızın depreme dayanıklı olduğunu düşünüyor musunuz?	Evet	104	54,2
	Hayır	88	45,8
Binanızın dayanıklılığı için belediyenin deprem dayanıklılık testini yaptırdınız mı?	Evet	52	27,1
	Hayır	140	72,9
Yaşadığınız bölge için belediyenizin hazırladığı olası deprem senaryosu	Evet	62	32,3
	Hayır	130	67,7
Deprem sırasında nerede olmanız gerektiği hakkında bilgi sahibi misiniz?	Evet	119	62,0
	Hayır	73	38,0
Deprem öncesinde, deprem sırasında ve sonrasında yapılması gerekenleri anlatan bir acil durum çantasına sahip misiniz?	Evet, sahibim	57	29,7
	Hayır, sahip değilim	135	70,3
Deprem öncesinde, bina veya evinizdeki eşyaların düşmesini önleyecek önlemler	Evet, aldım	61	31,8
	Hayır, almadım	131	68,2
Deprem anında cep telefonunuzu kullanmanın ne zaman uygun olduğunu biliyor musunuz?	Evet, tam olarak biliyorum	72	37,5
	Biraz biliyorum	82	42,7
	Hiçbir şey bilmiyorum	38	19,8

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,669	,664	5

Bu anket için Cronbach alfa katsayısı hesaplandığında, 0,669 olarak bulunmuştur. Bu değer, anketin iç tutarlılık düzeyinin kabul edilebilir bir seviyede olduğunu göstermektedir. Literatürde, Cronbach alfa katsayısının 0,70 veya üzeri olması genellikle kabul edilebilir bir iç tutarlılığı ifade ederken, bazı durumlarda 0,60 veya daha düşük değerler de kabul edilebilir olabilmektedir. Dolayısıyla, 0,669 Cronbach alfa değeri, anketin güvenilirliğine dair olumlu bir kanıt sunmaktadır.

ÇAPRAZ TABLO ANALİZİ VE Kİ-KARE TESTİ

Çapraz tablolar, verilerdeki değişkenlerin dağılımını anlamak ve aralarındaki ilişkiyi kontrol etmek için Ki-Kare testiyle desteklenen analiz yöntemleridir. Bu doğrultuda anket çalışmasına demografik sorular ile deprem ile ilgili sorular arasında çapraz tablo ve Ki-Kare testleri uygulandı ve bu uygulamalara örnek vermem gerekirse; 4. soru "Deprem olabilecek bir bölgede yaşıyor musunuz?" ile 9. soru olan "Deprem öncesinde, deprem sırasında ve sonrasında yapılması gerekenleri anlatan bir acil durum çantasına sahip misiniz?" arasındaki uygulandı. Aşağıda bu uygulamanın Ki-Kare test sonucu bulunmaktadır.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	7,321[a]	2	,026
Likelihood Ratio	8,214	2	,016
Linear-by-Linear Association	,118	1	,732
N of Valid Cases	192		

a. 3 cells (50,0%) have expected count less than 5. The minimum expected count is 1,48.

H_0: Soru4 ile Soru9 arasında ilişki yoktur.

H_1: Soru4 ile Soru9 arasında ilişki vardır.

$\alpha = 0.05$

Karar kuralına göre bakıldığında p-değerimiz (0.05) güvenilirlik düzeyimizden küçük olduğundan dolayı H_0'ı reddetmeliyiz. 4. soru ile 9. soru arasında ilişki vardır. Deprem bölgesinde yaşayan insanlarla sayısal bir değişim acil durum çantasına sahip olan insanların sayısını etkileyebileceğini söyleyebiliriz.

SONUÇ VE ÖNERİ

Anket çalışması, Türkiye'deki deprem bilincinin değerlendirilmesi amacıyla yapılmıştır. Katılımcılar deprem hakkında bilgi sahibi olmalarına rağmen tedbir alma konusunda yetersiz kalmaktadır. Cinsiyet, yaş ve eğitim düzeyi gibi demografik faktörler, deprem bilinci ve tedbir alma davranışları üzerinde etkilidir. Deprem bilincini artırmak için eğitim programları ve bilinçlendirme kampanyaları önemlidir. Bu çalışma, deprem bilincini artırmak ve toplumun daha iyi hazırlanmasını sağlamak için eğitim programları ve bilinçlendirme kampanyalarının gerekliliğini vurgulamaktadır.

YTÜ FEN EDEBİYAT FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ

KURAL BAZLI VE DENETİMSİZ ÖĞRENME YÖNTEMLERİ İLE MÜŞTERİ SEGMENTASYONU

Mertcan DEMİREL 20023603

Danışman: Prof. Dr. Ersoy ÖZ

ÖZET

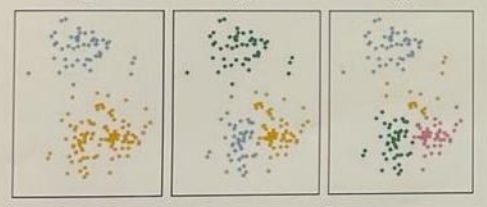
Bu tez çalışmasında kullanılan veri seti, İngiltere merkezli çevrimiçi işlem veri setidir. Veri seti, 01/12/2010 ile 09/12/2011 tarihleri arasındaki tüm işlemleri kapsamaktadır ve toptancı müşterileri içermektedir. Veri setinde 37 farklı ülke bulunmaktadır. Ancak, daha sağlıklı sonuçlar alabilmek için veri seti, "İngiltere", "Fransa", "Almanya", "İsviçre", "İspanya", "Portekiz", "Belçika" ülkeleri baz alınarak indirgenmiştir. Bu veriler, müşteri segmentasyonu için kullanılmıştır ve k-ortalama ve hiyerarşik kümeleme yöntemleri Python programlama dili üzerinde geliştirilmiştir. Her bir ülke için ortalama 5 ya da 6 ideal kümeleme sayısı belirlenmiş ve ülkeler bu sayıya göre kümelenmiştir. Segmentasyon sonucunda, bu çalışma Streamlit kütüphanesi kullanılarak bir internet sitesi olarak yayınlanmıştır. Bu çalışma, müşterileri segmente ederek hangi kümeye yatırım yapılması gerektiğini belirlemeye ve sadık müşteri grubunu bulmaya yardımcı olabilecek değerli bir kaynak olabilir.

RFM ANALİZİ

RFM analizi, "Recency" (yenilik), "Frequency" (sıklık) ve "Monetary" (para) kelimelerinin baş harflerinden oluşan bir analiz yöntemidir. Bu yöntem, müşteri satın alma davranışlarını üç kategoriye ayırarak değerlendirir. Yenilik, müşterinin son satın alma işleminden itibaren geçen süreyi ifade eder ve yeni satın almalar daha değerli kabul edilir. Sıklık, müşterinin belirli bir zaman diliminde kaç kez satın alma yaptığını gösterirken, para ise müşterinin işletmeye yaptığı toplam harcamayı ifade eder. RFM analizi, müşteri davranışlarını anlamak için kullanılan ve işletmelere pazarlama stratejilerini optimize etmede yardımcı olan bir yöntemdir.

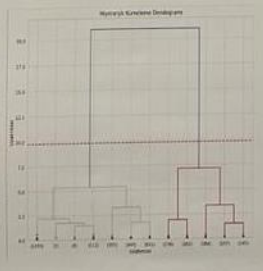
K-ORTALAMA KÜMELEME

K-ortalama kümeleme yöntemi, verileri iki aşamada kümelere ayırmak için çalışır. İlk olarak, veriler rastgele seçilen merkezi noktalara atanır ve her veri noktası en yakın merkezi noktaya atanır. Bu adım, verilere en yakın merkezi noktaya atanan kümelerin oluşmasını sağlar. Daha sonra, her kümeye ait verilerin ortalaması alınarak yeni merkezi noktalar belirlenir. Bu adım, verilerin yeniden düzenlenmesiyle yeni bir kümeleme işlemi yapılır ve merkezi noktaların konumu ve kümelerin içeriği değişerek daha doğru bir kümeleme sağlanır.



HİYERARŞİK KÜMELEME

Hiyerarşik kümeleme yöntemi, aglomeratif ve bölücü yaklaşımlar olmak üzere iki farklı şekilde uygulanabilir. Aglomeratif yaklaşımda, veriler başlangıçta ayrı kümelerde yer alır ve bu kümeler birleştirilerek daha büyük kümeler oluşturulur. Bölücü yaklaşımda ise, tüm veriler bir kümede yer alır ve bu küme alt kümeler halinde bölünür. Hiyerarşik kümeleme yöntemi, verilerin gruplandırılmasında dendrogram adı verilen bir yapıyı kullanır. Bu yöntem, verilerin doğrusal olmayan yapısı nedeniyle k-ortalama yönteminden daha doğru sonuçlar üretebilir, ancak büyük veri kümeleri için yüksek hesaplama maliyeti gibi bazı dezavantajları da bulunmaktadır.



Dendrogramlar, hiyerarşik kümeleme yöntemlerinde kullanılan bir görselleştirme aracıdır ve veri noktalarının kümeleme sürecindeki birleşmelerini gösterir. Hiyerarşik kümeleme, benzerlik veya uzaklık ölçütlerine dayanarak veri noktalarını adım adım birleştirerek veya bölerek kümeler oluşturur. Dendrogramlar, bu birleşim ve bölünme adımlarını göstererek, küme sayısı ve yapıları hakkında bilgi sağlayarak kümeleme sonuçlarını analiz etmek ve uygun küme sayısını belirlemek için kullanışlı bir araç olarak kullanılır.

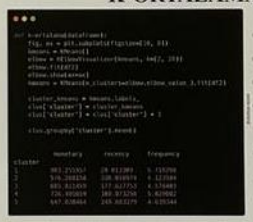
VERİ SETİ

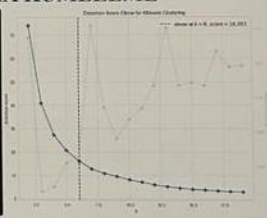
Irvine, Kaliforniya Üniversitesi'ndeki Makine Öğrenimi ve Akıllı Sistemler Merkezi'nden elde edilen "Online Retail" adlı bir veri seti kullanılmıştır. Veri seti, toplamda 541.908 veri içeren 8 değişken içermektedir. Bunlar; "InvoiceNo", "StockCode", "Description", "Quantity", "InvoiceDate", "UnitPrice", "CustomerID", "Country" adındaki değişkenlerdir.

KÜMELEME ANALİZİ

Veri setindeki en son alışveriş tarihine 2 gün eklenerek analiz tarihi belirlendi ve "today_date" değişkeninde saklandı. Ardından RFM metrikleri hesaplandı ve skorlar 1-5 aralığına indirgenerek karşılaştırılabilir hale getirildi. RFM skorlarının kullanılacağı segmentasyon için hesaplamalar yapıldı. Örneğin, R=1, F=2, M=3 ise RFM skoru "123" olacaktır. Bu işlemlerin ardından, her iki yöntemi de kullanabilmek için veri seti alt kümelerine ayrıldı ve bu alt kümeler [0,1] aralığında normalize edilerek diğer algoritmalara uygun hale getirildi.

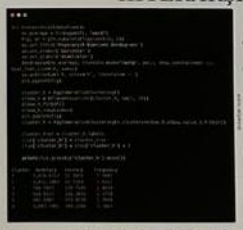
K-ORTALAMA KÜMELEME

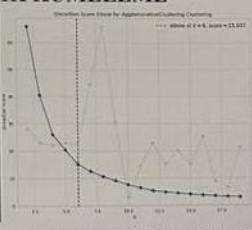




Optimal küme sayısını belirlemek için k-ortalama yöntemi kullanılarak 2'den 20'ye kadar değişen küme sayılarına göre "elbow" yöntemi grafiği oluşturuldu. Elde edilen optimal sonuçlar temel alınarak bir model oluşturuldu ve bu modele dayanarak RFM skorları elde edildi. Sonuçları yorumladığımızda, "recency" değeri en düşük olan küme grubunun sadık müşterilere ait olduğunu söyleyebiliriz. Örneğin, 1. küme en düşük "recency" değerine sahip olup ortalama olarak "29" gün önce alışveriş yapmıştır. Ayrıca, bu küme toplamda "5" ürün satın almış ve bu ürünlerin toplam değeri "903" sterlin olarak kaydedilmiştir.

HİYERARŞİK KÜMELEME





Optimal kümeleme sayısı "elbow" yöntemiyle belirlenmiş ve bu sonuca göre bir model oluşturularak RFM skorları elde edilmiştir. Sonuçların yorumlanmasına göre, en düşük "recency" değerine sahip olan küme grubu sadık müşterilere aittir. Örneğin, 1. küme en düşük "recency" değerine sahip olup ortalama olarak "12" gün önce alışveriş yapmıştır. Ayrıca, bu küme 7 adet ürün satın almış ve bu ürünlerin toplam değeri 3624 sterlin olarak kaydedilmiştir.

SONUÇ VE ÖNERİ

Bu çalışma, RFM skorlarına dayalı olarak k-ortalama ve hiyerarşik kümeleme yöntemlerinin müşteri segmentasyonunda kullanılmasını araştırmaktadır. Analizler, RFM skorlarıyla birlikte bu iki yöntemin müşterilerin segmentlere ayrılmasında etkili ve anlamlı sonuçlar sağladığını göstermektedir. K-ortalama algoritması benzer RFM skorlarına sahip müşterileri gruplandırmak için kullanılırken, hiyerarşik kümeleme yöntemi RFM skorlarına dayalı olarak müşterileri bir ağaç yapısıyla hiyerarşik olarak sınıflandırmak için kullanılmıştır. Sonuç olarak, RFM skorlarına dayalı olarak bu iki yöntemin birlikte kullanılması, müşteri segmentasyonunda önemli bir ilerleme kaydetmektedir ve şirketlere daha etkili pazarlama stratejileri oluşturma fırsatı sunmaktadır.

KAYNAKÇA

[1] Kotler, P., & Keller, K. L. (2016). Marketing Management (14th edition). Shanghai: Shanghai People's Publishing House.

[2] Mitchell, T. M. (2007). *Machine learning* (Vol. 1). New York: McGraw-hill.

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

Evrişimli Yapay Sinir Ağları ile Trafik Görüntülerinden İstatistiki Veri Üretme

Furkan Çelik 18023066

Tez Danışmanı: Prof. Dr. Gülhayat Gölbaşı Şimşek

Özet

İnsan sayısı bilindiği üzere her geçen gün artmaktadır ve bu artışın oluşturduğu çeşitli maliyetler vardır. Bunlardan biri de trafiktir. Trafik akışının tıkanması veya tamamen durması, insanların kaybettiği vakit, artan karbon emisyonu ve kullanılan yakıt miktarında artış gibi birçok olumsuz etkisi vardır. Bu olumsuzlukları minimuma indirmekle uğraşan; karayolları genel müdürlükleri, belediyeler, ulaştırma mühendisleridir ve ihtiyaç uydukları en önemli verilerden birisi trafik akışı verileridir. Trafik akışı verileri birçok yöntem ile elde edilebilir ancak az maliyeti, kullanım esnekliği gibi faktörlerden dolayı kameralar en iyi çözümler arasındadır. Bu çalışmada kamera kayıtlarından trafik akışı verisi üretilmeye çalışacaktır.

Trafik Akışı Verisi Nedir

Trafik akışı verisi trafikte bulunan araçların sayısının bazı kırılımlara göre sayılması ve ayrıca araç hızlarını da içeren bir veridir. Bu kırılımlar araçların tipleri; araba, kamyon, motosiklet vb. ve arçların yol üstündeki rotalarının bilgisidir. Aşağıda yapılan çalışmanın sonucundan bir örnektir. Çalışmada gidiş ve gelişi olan bir yol üzerinden bu veri üretilmiştir.

Cls	Direction	Kmph
2	Left	140,10
2	Right	134,44
2	Left	113,54
7	Right	88,93
2	Right	134,76

- **Cls** : araçların tiplerini içermektedir 2 araba 7 kamyon olmak üzere.
- **Drection** : Araçların rota bilgilerini içerir. Kayıtın alındığı konuma göre 'Left' gidiş 'Right' geliş tir.
- **Kmph**: Araçların hızlarını içerir.

Kullanılan Algoritma

Bu çalışmada yapay sinir ağı ile çalışan Yolov5_StrongSort_OsNet Algoritmasından faydalanılmıştır. Algoritma one-stage(tek adım) nesne tanıma algoritmasın olan YoloV5'in üzerine nesne takip sisteminin inşa edilmesi ile oluşturulmuştur. Tıpkı nesne takip de olduğu gibi nesne tanıma ve sınıflama kısımları daha önceden eğitilerek genleştirilmiş bir algoritmadır. YoloV5'e bakacak olursak CNN(Evrişimli Sinir Ağı) tabanlı, nesne tespiti yapan ve canlı görüntüleri işleyebilecek çalışma hızına sahip bir algoritmadır. Sonuç olarak eliminizdeki araç nesne tanıma, tahmin ve tespit konusunda başarılı olsa da trafik akışı verisi için çeşitli eklemler yapılmalıdır.

Algoritmanın Özelleştirilmesi

1. Giriş ve Çıkış Tespiti

Öncelikle trafik görüntüsü kaydında araçların yol üzerinde giriş ve çıkış piksellerinin sayısal olarak tespit edilmesi gerekmektedir. Bu işlem görüntünün bir karesi üzerinden fotoğrafta gözüktüğü gibi gerçekleştirilmiştir.

Giriş ve çıkışlar harici algoritma çalıştırılırken canlı olarak video kaydında ekranda oynatılmaktadır. Bu pencereye yapılan hesaplamaların gözükmesi içi görüntüye giren araçlar için sayaçlar eklenecektir. Bu sayaçların ve yazıların konumlarımda hesaplanmıştır.

2. Sayım

Bu işlem için aşağıda gözüktüğü gibi if-else blokları oluşturulmuştur. Bu bloklar algoritmanın ürettiği nesne tespiti ve takibi sonucunda elde edilen aracın merkez piksel koordinatlarının daha önceden belirlenen koordinatlar içeresinde olması halinde çeşitli ham verilerin tutulmasını sağlayan kodları tetiklenmesini ve araç ile ilgili ham verilerin tutulmasını sağlar.

```
def count_obj(box,w,h,id,cls):
    global ccl,ctl,ccr,ctr

    if id not in control:
        if int(box[1] +int(box[3]-box[1])/2)>(h-370):
            if int(box[0] +(box[2]-box[0])/2)>(50) and  int(box[0] +(box[2]-box[0])/2)<(615):


                control.append(id)
                dctSt['TimeSt'].append(time.time())
                dctSt['ID'].append(id)
                dctSt['Cls'].append(cls)
                dctSt['Direction'].append('Left')

                if cls==2:
                    ccl+=1
                elif cls==7:
                    ctl+=1
```

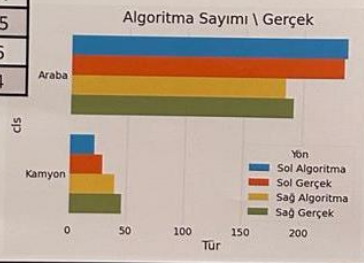
3. Verilerin işlenmesi

Tutulan ham verilerin içeresinde aracın rotası ve sınıfı istenilen halde gelmektedir ancak aracın hızı için ek hesaplamalara ihtiyaç vardır. Araçların belirlenen giriş ve çıkış alanlarına girdikten sonra bilgisayarın saat bilgisi çekilmekte ve aradaki farktan piksel saniye cinsinden hız hesaplanmaktadır. Bu hız verisi belirlenen bölgenin uzunluğu bilenmediğinden farazi olarak hesaplanmıştır.

Sonuçlar

Algoritma çalışırken oluşa çıktı şekildeki gibi gözükmektedir. Yolda bulunan araçlar sadece araba kamyonlardan oluşmaktadır. Kamyonet olarak ayrılan araçlar ile tırlar 'truck' olarak tek bir seviyede değerlendirilmiştir. Ağadaki görsellerde algoritma ile gözlem yolu sayım arasındaki fark gözükmektedir.

Araç Sınıfı	Yön	Algoritma Sayımı	Gerçek Sayım
Araba	Sağ	185	194
	Sol	238	235
Kamyon	Sağ	22	26
	Sol	40	44



KAYNAKÇA

[1] YoloV5_StrongSort_OsNet https://github.com/hdnh2006/Yolov5_DeepSort_Pytorch
[2] Yardım, M. S., & Akyıldız, G., (2005). Akıllı Ulaştırma Sistemleri ve Türkiyedeki Uygulamalar . 6. Ulaştırma Kongresi (pp.405-414). İstanbul, Turkey
[3] https://towardsdatascience.com/artificial-neural-networks-for-total-beginners-d8cd07abaae4

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

TÜRKİYE'DEKİ ONLINE ALIŞVERİŞ ALIŞKANLIKLARI

Öğrenci Adı SOYADI: Alper KINCI

Danışman: Dr. Öğr. Üyesi Elif TUNA

ÖZET

Online alışveriş, internetin herhangi bir aracı kullanılarak gerçekleştirilen alışveriş şeklidir. E-ticaret ya da çevrimiçi alışveriş gibi isimlerle de anılan bu pazarlama şekli, Türkiye'de özellikle son 5 yılda popüler bir hale gelmiştir. Online alışveriş, reklamcılığın, pazarlamanın, alım-satım ve ödeme işlemlerinin tamamının yapıldığı platformları kapsadığından, klasik pazarlamadaki kavramlara ek olarak pek çok kavramı beraberinde getirmiştir. Bu çalışmada, online alışveriş ile ilgili sorular içeren 20 soruluk bir anket hazırlanmıştır. Ankete katılan 123 kişinin cevapları, SPSS programı kullanılarak analiz edilmiştir. Sorulara verilen cevaplara, çapraz tablolar yardımıyla bir ön analiz yapılmıştır. Sonrasında ki-kare testi kullanılarak seçilen sorular arasında bir ilişkinin olup olmadığı test edilmiştir. Analiz sonuçları elde edilen SPSS çıktıları üzerinden yorumlanmıştır.

ONLINE ALIŞVERİŞİN TANIMI ve ÖZELLİKLERİ

Online alışverişi basit bir tabirle "internet üzerinden satın alma işlemi" olarak açıklayabiliriz. Günümüzde internet kullanımının giderek artması ile birlikte, online alışveriş de hızla yaygınlaşmıştır. İnternetin sunduğu kolaylık, zaman ve maliyet tasarrufu, tüketicilerin internet üzerinden alışveriş yapmasını teşvik etmektedir.

Online alışverişin özelliklerini; 7/24 açık olması, ürünlerin detaylı bir şekilde sunulabilmesi, ürünlerin fiyat ve özelliklerinin karşılaştırılabilmesi, ödeme kolaylığı ve ürünlerin ev veya iş yerine teslim edilebilmesi olarak açıklayabiliriz. Online alışverişte ödeme yöntemleri de oldukça çeşitlidir. Kredi kartı, banka havalesi, kapıda ödeme ve benzeri birçok ödeme yöntemi kullanılabilmektedir. Bunlara ek olarak, online alışveriş siteleri, binlerce ürün seçeneği sunar ve tüketiciler, mağaza gezerek bu kadar geniş bir seçeneği bulamazlar. Tüm bunları online alışverişin avantajları olarak da gösterebiliriz.

ONLINE ALIŞVERİŞİN GELİŞİMİ ve TARİHİ

Online alışverişin tarihçesi, 1980'lerin sonlarına kadar uzanmaktadır. Bu dönemde, online alışverişin temelinde telefon, televizyon veya posta siparişi ile çalışan şirketler bulunmaktaydı. İnternetin yaygınlaşmasıyla birlikte online alışveriş de önemli bir değişim geçirdi. İlk e-ticaret web siteleri, 1990'ların ortalarında açılmaya başladı.

1994 yılında, bir bilgisayar bilimcisi olan Phil Brandenberger, Pizza Hut'un web sitesi üzerinden pizza siparişi veren ilk müşteri olarak kaydedildi. Bu olay, online alışverişin doğuşu olarak kabul edilmektedir. Aynı yıl içinde, Jeff Bezos, Amazon.com'u kurmuştur. Amazon, başlangıçta sadece kitap satışı yapan bir web sitesi olarak faaliyet gösterse de, daha sonra elektronik, ev gereçleri, giyim gibi birçok farklı kategoride ürün satışına başlamıştır. 2000'li yılların başlarına kadar, online alışverişin gelişimi yavaş bir şekilde ilerlemiştir. 2000'lerin ortalarında, çevrimiçi alışveriş yapmak için gerekli olan altyapı ve teknolojik gelişmeler hızlanmaya başlamıştır. Mobil cihazların yaygınlaşması, e-ticaret web sitelerinin daha kullanıcı dostu ve erişilebilir hale gelmesi gibi faktörler, online alışverişin hızlı bir şekilde yayılmasına yol açmıştır. Günümüzde de, online alışverişin gelişimi her geçen gün hız kazanmaktadır. Özellikle, COVID-19 pandemisi nedeniyle, online alışverişe olan talep daha da artmıştır. Tüketiciler, evden çıkmadan ürünleri sipariş ederek ihtiyaçlarını karşılamaktadırlar. Ayrıca, birçok perakende şirketi, online alışverişin yaygınlaşmasıyla birlikte dijital pazarlama ve reklamcılık stratejilerine de ağırlık vermeye başlamıştır.

SOSYAL MEDYANIN ONLINE ALIŞVERİŞE ETKİSİ

Sosyal medya, online alışveriş deneyimini daha kişisel hale getirir. Birçok online alışveriş sitesi, müşterilerin alışveriş geçmişlerine ve tercihlerine dayalı öneriler sunar. Sosyal medya ise bu önerilerin daha da kişiselleştirilmesine olanak sağlar. Örneğin, bir tüketicinin sosyal medya hesabında beğendiği veya takip ettiği bir marka, kendi ürünlerini sosyal medya üzerinden önererek tüketicinin alışveriş kararını etkileyebilir.

Sosyal medya, tüketicilerin satın alma kararlarını etkilemek için markaların kullandığı bir pazarlama aracıdır. Birçok marka, sosyal medyada hedef kitlelerine özel reklamlar yayınlayarak, tüketicilerin ilgisini çekmeye ve satın alma kararlarını etkilemeye çalışır. Sosyal medya platformları ayrıca influencer marketing adı verilen bir pazarlama stratejisi için de önemli bir rol oynamaktadır. İşletmeler, sosyal medya fenomenlerini kullanarak ürünlerini tanıtabilir ve onların takipçileri aracılığıyla daha da geniş bir kitleye ulaşabilirler. Bu yöntem, tüketicilerin ürünlere daha olumlu bir şekilde yaklaşmalarına yardımcı olabilmektedir.

ANKET SONUÇLARI

Soru	Şıklar	N	%
Cinsiyetiniz nedir?	Erkek	49	39,8
	Kadın	74	64,2
Yaş aralığınız nedir?	18-25	80	65,0
	26-35	12	9,8
	36-50	11	8,9
	50 ve üstü	20	16,3
Eğitim durumunuz nedir?	İlkokul	1	0,8
	Ortaokul	3	2,4
	Lise	18	14,6
	Üniversite	93	75,6
	Lisansüstü	8	6,5
Çalışma durumunuz nedir?	Kamu Sektörü	9	7,3
	Özel Sektör	26	21,1
	Öğrenci	67	54,5
	Ev Hanımı	8	6,5
	Kendi işimin patronuyum	1	0,8
	Çalışmıyorum	12	9,8

Soru	Şıklar	N	%
İnternet üzerinden hiç alışveriş yaptınız mı?	Evet	123	100,0
	Hayır	0	0,0
En son ne zaman internet üzerinden alışveriş yaptınız?	Son 1 hafta içerisinde	63	51,2
	Son 1 ay içerisinde	37	30,1
	Son 6 ay içerisinde	18	14,6
	Son 1 yıl içerisinde	5	4,1
	İnternet üzerinden alışveriş yapmıyorum	0	0,0
Ne sıklıkla internet üzerinden alışveriş yapıyorsunuz?	Haftada bir	13	10,6
	Haftada birden fazla	14	11,4
	Ayda bir	47	38,2
	Ayda birden fazla	32	26,0
	Yılda bir	17	13,8
	İnternet üzerinden alışveriş yapmıyorum	0	0,0
İnternette gördüğünüz ürün reklamlarını tıklayarak satın alma gerçekleştirdiniz mi?	Evet	47	38,2
	Hayır	76	61,8
	İnternet üzerinden alışveriş yapmıyorum	0	0,0
İnternette gördüğünüz bir ürünün indirime girmiş olması, o ürünü almanızda etkili midir?	Evet	108	87,8
	Hayır	15	12,2
	İnternet üzerinden alışveriş yapmıyorum	0	0,0
İnternet üzerinden alışverişi hangi sektör(ler) için kullanıyorsunuz? (Birden fazla seçenek işaretlenebildiğinden yüzdelerin toplamı %100'den fazla çıkmıştır.)	Gıda Sektörü	70	56,9
	Giyim Sektörü	100	81,3
	Eğlence Sektörü	40	32,5
	Elektronik Sektörü	61	49,6
	Kozmetik Sektörü	52	42,3
	Diğer	26	21,1
	İnternet üzerinden alışveriş yapmıyorum	0	0,0
İnternet üzerinden alışveriş yaparken öncelikle, hangisine dikkat edersiniz?	Ürünün Fiyatı	20	16,3
	Ürüne Ait Kullanıcı Yorumları	39	31,7
	Güvenilir Platform	64	52,0
	Teslimat Süresi	0	0,0
	İnternet üzerinden alışveriş yapmıyorum	0	0,0
Aşağıdaki alışveriş sitelerinden hangisini veya hangilerini tercih ediyorsunuz? (Birden fazla seçenek işaretlenebildiğinden yüzdelerin toplamı %100'den daha fazla çıkmıştır.)	Trendyol	106	86,2
	Hepsiburada	66	53,7
	Getir	53	43,1
	Amazon	54	43,9
	Yemeksepeti	51	41,5
	N11	15	12,2
	Çiçeksepeti	27	22,0
	Diğer	27	22,0
	İnternet üzerinden alışveriş yapmıyorum	0	0,0
İnternet üzerinden alışverişte aşağıdaki ödeme yöntemlerinden hangisini tercih edersiniz?	Kredi/Banka Kartı	115	93,5
	Kapıda Ödeme	8	6,5
	İnternet üzerinden alışveriş yapmıyorum	0	0,0
Siz bir ürün satacak olsaydınız hangisini kullanmayı tercih ederdiniz?	Sosyal Medya	30	24,4
	Mobil Uygulamalar	88	71,5
	Diğer	5	4,1

HİPOTEZLER ve Kİ-KARE TESTİ YORUMU

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	20,141[a]	16	,214
Likelihood Ratio	17,581	16	,349
Linear-by-Linear Association	1,099	1	,294
N of Valid Cases	123		

a. 19 cells (76,0%) have expected count less than 5. The minimum expected count is ,11.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	13,899[a]	6	,031
Likelihood Ratio	13,627	6	,034
Linear-by-Linear Association	4,266	1	,039
N of Valid Cases	123		

a. 7 cells (58,3%) have expected count less than 5. The minimum expected count is ,45.

Ki-Kare İlişki Hipotezleri

H_0: Sorular arasında herhangi bir ilişki yoktur.

H_1: Sorular arasında bir ilişki vardır.

Dikdörtgen içine alınan değerler, incelememiz gereken p değerleridir. Bu p değerleri aşağıdaki gibi yorumlanır;

$p > 0.05$ ise H_0 hipotezi kabul edilir. Dolayısıyla sorular arasında herhangi bir ilişki yoktur.

$p < 0.05$ ise H_0 hipotezi reddedilir. Dolayısıyla sorular arasında bir ilişki vardır.

İLİŞKİ BULUNAN VE İLİŞKİ BULUNMAYAN SORULAR

Ki-kare analizine göre aralarında **ilişki bulunan** sorular şunlardır;

- Cinsiyet ile Sosyal Medyanın Alışverişe Etkisi Arasındaki İlişki
- Yaş ile Günlük İnternet Kullanım Sıklığı Arasındaki İlişki
- Cinsiyet ile Sosyal Medyadaki Reklamların Etkisi Arasındaki İlişki
- Cinsiyet ile İnternetten Yapılan Son Alışverişin Yapılma Zamanı Arasındaki İlişki
- Yaş ile Ürün Satışı için Kullanılan Platformlar Arasındaki İlişki
- Sosyal Medya Kullanımı ile İnternetten Yapılan Son Alışverişin Yapılma Zamanı Arasındaki İlişki
- Sosyal Medyanın Alışverişe Etkisi ile İndirimlerin Ürünü Satın Almadaki Etkisi Arasındaki İlişki

Ki-kare analizine göre aralarında **ilişki bulunmayan** sorular şunlardır;

- Yaş ile Sosyal Medyanın Alışverişe Etkisi Arasındaki İlişki
- Çalışma Durumu ile Sosyal Medya Kullanımı Arasındaki İlişki
- Çalışma Durumu ile En Çok Kullanılan Sosyal Medya Platformu Arasındaki İlişki
- Cinsiyet ile Tercih Edilen Ödeme Yöntemleri Arasındaki İlişki
- Eğitim Durumu ile İnternet Üzerinden Yapılan Alışverişin Sıklığı Arasındaki İlişki
- Sosyal Medyanın Alışverişe Etkisi ile İnternet Sitelerindeki Reklamların Alışverişe Etkisi Arasındaki İlişki
- İnternetten Yapılan Son Alışverişin Yapılma Zamanı ile İnternette Alışveriş Yaparken Dikkat Edilen Öncelikler Arasındaki İlişki

KAYNAKÇA

[1] Jarvenpaa, S. L., & Todd, P. A. (1996). Consumer reactions to electronic shopping on the World Wide Web. International Journal of electronic commerce, 1(2), 59-88.

[2] Park, C. H., & Kim, Y. G. (2003). Identifying key factors affecting consumer purchase behavior in an online shopping context. International journal of retail & distribution management.

[3] TOPAL, İ., & TEMİZKAN, V. (2016). Tüketicilerin mobil sosyal medya kullanımının marka

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

Gayri Safi Yurt İçi Hasıla ile Askeri Harcamaların İstatistiksel Analizi

Berke Hürcan GÜLDEN 18023051

Danışman: Prof. Dr. Ali Hakan BÜYÜKLÜ

Dünyada birçok ülkenin kamu harcamalarında önemli bir yere sahip olan askeri harcamalar ulusal güvenliği sağlama açısından ülkelerin önemli harcama kalemleri arasında yer almaktadır. Bu doğrultuda yapılan bu çalışmanın amacı savunma harcamaları bakımından önemli büyüklüğe sahip olan ülkelerin gayri safi yurt içi hasılaları ile askeri harcamalar değişkenlerinin hangi modellere uygun olduğunu ve değişkenler arasındaki ilişkiyi tespit etmektir. Çalışmada panel veri analizi yapılmasına imkan sağlayan Stata programı kullanılarak yapılan analizlerde öncelikle veri seti için uygun olan model tespit edilmeye çalışılmış ve Rassal Etkiler Modelinin uygun olduğu sonucuna ulaşılmıştır. Yatay kesit bağımlılığını araştırmak için Pesaran CD testine başvurulmuş ve test sonucunda yatay kesit bağımlılığının olduğu tespit edilmiş olup İkinci Nesil Panel Birim Kök Testi olan Pesaran CADF Testi uygulanmıştır. Sonuç olarak, serilerin durağan olduğu ve seçilen ülkelerde gayri safi yurt içi hasıla ile askeri harcamalar arasında pozitif bir nedensellik ilişkisinin olduğu tespit edilmiştir.

PANEL VERİ VE PANEL VERİ ANALİZİ

- Ekonometrik analizlerde herhangi bir konuda hem zamana göre hem de birimlere göre analiz yapılması gerektiğinde, genellikle bu analizler zamana ve birime göre ayrı ayrı yapılmaktadır. Zamana göre yapılan analizler zaman serileri analizi olmakta, birimlere göre yapılan analizler ise yatay kesit analizi olmaktadır. Zaman serileri ve yatay kesit analizinin birleştirilmesini ve uygun modellerin test edilmesini sağlayan yönteme panel veri analizi denilmektedir.
- Panel veriler kullanılarak oluşturulan panel veri modelleri yardımıyla ekonomik ilişkilerin tahmin edilmesi yöntemine "Panel Veri Analizi" denilmektedir. Bu analizde genelde, yatay kesit biriminin (N) sayısının dönem sayısından fazla olduğu durumlarda çalışmaktadır.

PANEL VERİNİN ÖZELLİKLERİ

- Herhangi bir yatay kesitte araştırma konusu olan birimlerin (firmalar, ülkeler vb.) davranışlarını etkileyen sayısız ölçülemeyen açıklayıcı değişken vardır. Bu değişkenlerin dışlanması sapmalı tahminlere neden olmaktadır. Benzer bir durum mikro birimlerin davranışlarını hep aynı yönde ancak her bir zaman döneminde farklı bir şekilde etkileyen zaman serisi değişkenlerinin dışlanması halinde de geçerlidir. Panel veri bu problemin giderilmesine olanak tanımaktadır.
- Panel veri bir dönemden diğerine meydana gelen değişim ile mikro birimler arasındaki değişimi birleştirmek suretiyle değişkenlik meydana getirerek çoklu doğrusallığı azaltmaktadır.

PANEL VERİ ANALİZİNİN AVANTAJLARI

- Zaman serisi ve yatay kesit analizi ile kıyaslandığında panel veri analizi, araştırmacıya daha geniş bir veri seti ile çalışma imkânı sunar.
- Yeterli bir zaman uzunluğunda, değişim dinamiklerinin çalışmasında panel veri analizi yatay kesit ve zaman serisi analizlerine göre daha avantajlı bir yöntemdir.

PANEL VERİ ANALİZİNİN DEZAVANTAJLARI

- Her birim için zaman serisi boyutunun kısa olabilmesi.
- Yatay kesit ve zaman serisi gözlemleri arasında meydana gelen parametre farklılıklarının göz önüne alınmadığı durumlarda birtakım sapmaların ortaya çıkması ve bu durumun parametrelerin tutarsız ve anlamlı olmayan tahminlerine sebep olması.

PANEL VERİ ANALİZİ MODEL TÜRLERİ

- Sabit Etkiler Modeli: Sabit etkiler modelini oluşturan katsayılar, birimlere ve/veya zamana göre farklılık göstermektedirler. Bu modelin temelinde, birimler arası farklılıkların modelde yer alan sabit terimdeki farklılıklar aracılığıyla yok edilebileceğini düşüncesi yatmaktadır. Bundan dolayı her birimi temsil edecek farklı sabit terimler kullanılmaktadır. Bu modelde yatay kesit birimleri arasındaki farklılıklar sabit terimdeki farklar ile açıklandığı ve kukla/gölge değişken yardımıyla tahmin edildiği için model kukla değişkenli model olarak da bilinmektedir.
- Rassal Etkiler Modeli: Sabit etkiler modelinde birimler arası farklılığın sabit olduğu ve bunun sabit terimdeki farklılıklarla giderileceği varsayılmaktadır. Ancak bazı çalışmalarda birimlerin tesadüfi olarak seçildiği durumlar olabilmektedir. Bu birimler arası farklılıklar da tesadüfi olduğu için rassal (tesadüfi) etkiler modeli sabit etkiler modeline alternatif olarak geliştirilmiş bir yöntemdir. Rassal etkiler modelinin avantajı, zaman değişkenlerini modele ilave edebilmektir. Rassal etkiler modelinde birim ve zaman etkileri tesadüfi değişken olarak modelde hata teriminin bileşeni olarak analize dahil edilmektedir.

VERİ TİPLERİ

Stata'da veri tipi çok önemlidir. Çalıştığımız veri sürekli bir veri ise 'float' olarak kaydedilmesi gerekir. Stata'da tam sayılar için sadece 'byte', 'int' ve 'long' kullanılır. 'float', yedi basamaklı sayı olarak depolamaktadır. Sayıların büyüklüğü önemli değildir. 9 basamaklı sayılar veya daha azı için 'long', 9 basamaktan daha fazlası ise 'double' olarak depolanabilmektedir.

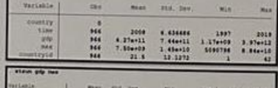
ANALİZLER

Oluşturduğum veri setinde başlangıçta toplam 1450 adet veri bulunmaktaydı. Zaman serisi bakımından 1997-2021 arası dönem ele alınmış olup, yatay kesit serisi bakımından Avrupa ve Orta Asya ülkelerindeki 49 ülke ele alınmıştır. Eksik veriler içermesinden dolayı veri setinden 2020-2021 yıllarını ve çeşitli ülkeleri çıkartarak veri setini dengeli (balanced) hale getirdiğimde veri setinde toplam 966 adet veri bulunmaktaydı. Veri setindeki bağımlı değişken Askeri Harcamalar (Military Expenditures (mex)) (current USD), bağımsız değişken Gayri Safi Yurt İçi Hasıla (gross domestic product (gdp)) (current US$) olmaktadır.

- Veri setini panel veri olarak tanımladıktan sonra sağ tarafta bulunan çıktıda yer alan strongly balanced ifadesi panel veri analizine başlayabilmemize olanak sağlar.

```
xtset countryid time
       panel variable:  countryid (strongly balanced)
        time variable:  time, 1997 to 2019
                delta:  1 unit
```

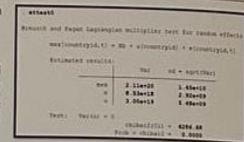
- Sağ tarafta yer alan çıktıda veri setinin özetine bakacak olursak veri setindeki değişkenleri, toplam gözlem sayısını, değişkenlerimizin ortalama, standart sapma ve minimum-maximum değerlerini görebiliriz.



- Sağ tarafta yer alan çıktıda ise veri setinde bulunan değişkenlerin detaylı özetini ve her bir değişken için toplam gözlem sayısını, yatay kesit birim sayısı ile zaman birimi sayısını görebiliriz.

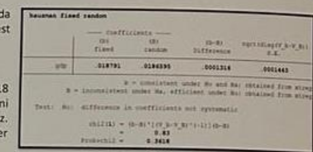
- Klasik Model ile Rassal Etkiler Modeli arasında tercih yapmak için LM testini kullanırız. Test hipotezleri;

H0: Tesadüfi etkiler sıfırdır.
H1: Tesadüfi etkiler sıfır değildir. şeklindedir.
Yan tarafta bulunan çıktıda yer alan Prob>chibar2=0.0000 ifadesi neticesinde H0 hipotezini reddetmekle herhangi bir hata yapılmayacağını söyleriz ve bu nedenle H0 hipotezi reddedilir. Dolayısıyla tesadüfi etkiler sıfır değildir, rassal etkiler modeli uygundur sonucuna ulaşılır.



- Rassal Etkiler Modeli ile Sabit Etkili Model arasında seçim yapmak için Hausman Testi uygulayacağız. Test hipotezleri;

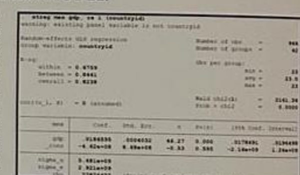
H0: Tesadüfi etkili model uygundur.
H1: Tesadüfi etkili model uygun değildir. şeklindedir.
Yan tarafta bulunan çıktıda yer alan Prob>chibar2=0.3618 ifadesi neticesinde %95 güven düzeyinde H0 hipotezini reddetmekle hata yaparız ve H0 hipotezi reddedilmez. Dolayısıyla tesadüfi etkiler sıfır değildir, rassal etkiler modeli uygundur sonucuna ulaşılır.



- Yapılan iki test neticesinde veri setimize uygun olan modelin Rassal Etkiler Modeli olduğunu sonucuna ulaşılmıştır.

RASSAL ETKİLER MODELİ

Rassal Etkiler Modelini tahmin etmek için çeşitli kodlar çalıştırarak sağ tarafta bulunan çıktıyı elde ederiz. Çıktıdan da görüleceği üzere (corr(u_i, x)=0 ifadesi) Rassal Etkiler Modeli hata terimleri ile açıklayıcı değişkenler arasındaki korelasyonun 0 olduğunu varsayar. Modelin overall R-square değeri 0.82'dir. Çıktıda gdp değişkeninin katsayısının anlamı olup sabitin katsayısının anlamsız olduğu görülmektedir. rho değeri rassal etkiler modelinin varyansının %77.8'inin birimler arasındaki farklılıktan kaynaklandığını söylemektedir. Prob>chi2=0.0000 ifadesi modelin tüm katsayılarının anlamlı olduğunu göstermektedir.



EŞ ZAMANLI KORELASYON

Eş zamanlı korelasyon, panel veri modellerinde yatay kesit birimleri arasında bağımlılık olmamasıdır. Eğer bu bağımlılık mevcut ise analizlerimiz hatalı sonuçlar verecektir. Breusch ve Pagan LM Testi ve Pesaran CD Testi ile test edilir.

PESARAN CD TESTİ

Pesaran CD Testi kısa panel verilerde (N>T ise) kullanılır. Dolayısıyla bizim panelimiz için uygundur. Test hipotezleri;
H0: Yatay kesit bağımlılığı yoktur.
H1: Yatay kesit bağımlılığı vardır. şeklindedir
Çıktıda yer alan Pr = 0.0000 ifadesi, H0 hipotezini reddetmekle hata yapmayacağımız anlamına gelmektedir. Dolayısıyla H0 hipotezi reddedilir, değişkenlerimiz arasında yatay kesit bağımlılığı vardır.

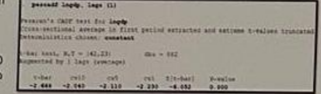
```
. xtcsd, pesaran abs

Pesaran's test of cross sectional independence =    41.438, Pr = 0.0000

Average absolute value of the off-diagonal elements =    0.574
```

- Pesaran CD Testi sonucunda modelimizde bulunan değişkenler arasında yatay kesit bağımlılığı olduğu sonucuna ulaşılmış olup yatay kesit bağımlılığını konu alan İkinci Nesil Panel Birim Kök Testi olan Pesaran CADF testine başvurulmuştur.

PESARAN CADF TESTİ

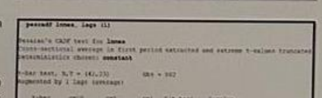
- Pesaran CADF Testi'ni gdp değişkenine uygularken test hipotezleri;

H0: lngdp değişkeni durağan değildir.
H1: lngdp değişkeni durağandır. Şeklindedir.
Yan tarafta bulunan çıktıda yer alan P-value=0.0000 ifadesi neticesinde H0 hipotezi reddedilir ve lngdp değişkeni durağandır sonucuna ulaşılır.



- Pesaran CADF Testi'ni mex değişkenine uygularken test hipotezleri;

H0: lngdp değişkeni durağan değildir.
H1: lngdp değişkeni durağandır. Şeklindedir.
Yan tarafta bulunan çıktıda yer alan P-value=0.0000 ifadesi neticesinde H0 hipotezi reddedilir ve lnmex değişkeni durağandır sonucuna ulaşılır.



SONUÇLAR

- Çalışma süresince yapılan analizler neticesinde panel veri setimiz için en uygun modelin Rassal (Tesadüfi) Etkiler Modeli olduğu sonucuna ulaşılmıştır.
- Gayri safi yurt içi hasıla ile askeri harcamalar arasında pozitif bir nedensellik ilişkisi olduğu sonucuna ulaşılmıştır.
- Pesaran CD Testi'nin sonucuna göre yatay kesit bağımlılığı vardır.
- Değişkenlerimizde yatay kesit bağımlılığı olduğundan dolayı İkinci Nesil Panel Birim Kök Testleri uygulanmıştır. İkinci Nesil Panel Birim Kök Testlerini uygulayarak yapılan analizler sonucunda değişkenlerimizin durağan olduğu sonucuna ulaşılmıştır.

KAYNAKÇA

[1] Dr. Öğr. Üyesi Elif Tuna Uygulamalı Panel Veri Analizi Ders notları
[2] Doç. Dr. Ferda Yelen Tatoğlu Panel Veri Ekonometrisi
[3] HSİAO, C., 2003, Analysis of Panel Data, Cambridge ünversity Press,ünited Kingdom.
[4] BALTAGİ, B. H.,2005, Econometric Analysis of Panel Data, Third Edition, John Wiley&Sons Inc, England.

YTÜ FEN EDEBİYAT FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ

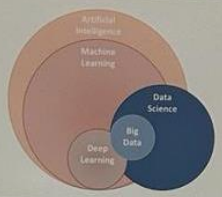
DERİN ÖĞRENME İLE FİLM YORUMLARINDAN DUYGU ANALİZİ

Alpay TUNCER 18023040

Danışman: Doç. Dr. Gülder KEMALBAY

Bilgisayar teknolojisindeki değişiklikler hayatın her alanında yenilikleri getirmiştir. Verilerin farklı tiplerinin derlenip saklanıp işlenmesi ve sosyal medyanın kullanımı geleneksel medya araçlarına bir alternatif oluşturur. Metin madenciliği, yapılandırılmamış metinleri işlemek için kullanılır. Duygu analizi, metindeki duyguları belirlemeyi amaçlar. Filmler, herkesin ilgisini çekebilen ve çeşitli amaçlar için kullanılabilen eğlence araçlarıdır. Filmlerdeki senaryo, sinematografi ve görsel öğeler izleyicilerde farklı düşünceler ve duygular yaratabilir. Bu çalışma, 83.227 film yorumunu içeren farklı filmlerde duygu analizi yapmayı amaçlamaktadır.
Sonuçlar, duygu analizinin film yorumlarını otomatik olarak analiz etmek, yorumlar ile verilen puanların arasındaki bağlantıyı anlamak ve genel izleyici duyarlılığını belirlemek için değerli bir araç olduğunu göstermektedir. Bu, sinema endüstrisinin film ticari başarısını önceden tahmin etmek ve izleyici eğilimlerini daha iyi anlamak için bu tür analizleri nasıl kullanabileceğine dair potansiyel yolları tartışır.

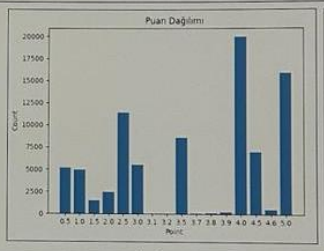
Metin madenciliği, doğal dil işleme (NLP) ve makine öğrenmesi uygulamaları için büyük veri setlerinden değerli bilgiler çıkarılması sürecidir. Bu, metinlerde belirli bilgi parçalarını, örüntüleri, duygusal durumları veya diğer önemli verileri tanımlamak ve analiz etmek için kullanılan bir tekniktir.
Doğal dil işleme (NLP), insanların doğal dillerini anlama, yorumlama, oluşturma ve manipüle etme yeteneğine sahip bilgisayar sistemlerinin geliştirilmesiyle ilgilenen bilgisayar bilimi dalıdır. Bilgisayarların metin veya sesli dil verilerini anlamalarını, yanıtlarını, etkileşimlerini otomatikleştirme ve insan dilinin karmaşıklığını ve anlamlarını modelleme amacını taşır.



Makine öğrenmesi, bilgisayarların belirli bir görevi yerine getirirken performanslarını geliştirmek için veriden bağımsız olarak öğrenme yeteneklerine sahip olmalarını sağlayan bir yapay zeka dalıdır. Bu süreç, deneyim ve örüntülerin analizine dayalı algoritma geliştirme ile otomatik model oluşturma ve sürekli iyileştirme sağlar.
Derin öğrenme, genellikle çok katmanlı yapay sinir ağları kullanarak karmaşık örüntülerin verilerden öğrenilmesi ile ilgilenen bir makine öğrenmesi alt dalıdır. Bu, görüntü ve ses tanıma, metin anlama, çeviri, oyun oynama ve diğer karmaşık görevlerde üstün performans sağlama yeteneğine sahip modellemeler oluşturur.

Duygu analizi, metinlerdeki duygusal tonları, tutumları ve yargıları belirlemek için kullanılan bir yapay zeka tekniğidir. Yani metinden insan duygusunu çıkarır. İngilizce'deki "Sentiment Analysis" teriminin Türkçe karşılığıdır. Genellikle doğal dil işleme, metin analizi, hesaplama dilbilim ve biyometrik ölçümler gibi teknikleri içerir.
Duygu analizi, metinlerin içerisinde bulunan subjektif bilgileri ve bu bilgilerin duygusal yükünü sınıflandırmak ve anlamak için kullanılan bir yapay zeka tekniğidir. Bu teknik, doğal dil işleme (NLP), metin analizi ve hesaplama dilbilimini kullanarak metinlerde bulunan pozitif, negatif veya nötr duyguları belirler.
Duygu analizi genellikle üç türde yapılır: polarite tabanlı (olumlu, olumsuz, nötr), duygu tabanlı (mutluluk, kızgınlık, üzüntü vb.) ve konu tabanlı (belli bir konu veya konular hakkında insanların duygularını analiz etme).
Duygu analizinin ham maddesi kullanılan dildir. Bu yüzden araştırmacının analizi yaptığı dile hakim olması analizin başarısı için son derece önemlidir. Literatürde özellikle İngilizce çalışmaların fazlalığı göze çarpmaktadır, bundaki en temel neden dilin yapısının basit olmasıdır. Türkçe gibi sondan eklemeli ve karmaşık yapıya sahip dillerde ise duygu analizi yapmak çok daha zordur. Türkçede bir kök; aldığı yapım ve çekim ekleri ile çok farklı anlamlara bürünebilmektedir.

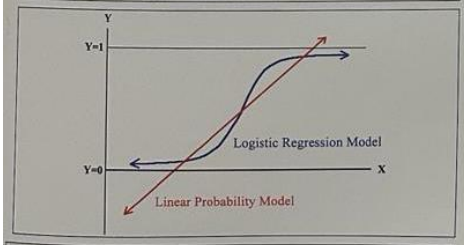
Bir film, seri halde gösterilen hareketli görüntülerden meydana gelen bir eserdir. Filmlerin oluşumu, gerçek insanların ve objelerin kamera ile çekimi veya animasyon, özel efektler gibi teknolojik yöntemler ile bu unsurların yaratılması yoluyla gerçekleşir.
Film eleştirisinin amacı filmleri çözümlemek ve değerlendirmektir. Akademik film eleştirmenliği ve gazeteci film eleştirmenliği olmak üzere iki ana kategoriye ayrılır, genellikle bu iki türden birini izler. Fakat Bu çalışmada bireysel kullanıcıların IMDb internet sitesi üzerinden yaptığı tarafsız yorumlar kullanılmıştır.
IMDb (Internet Movie Database), filmler, televizyon dizileri, TV programları, video oyunları ve internet tabanlı içerikler hakkında bilgi sunan online bir platformdur. Bu platform, oyuncular, yapım ekibi, biyografiler, özetler, trivia, derecelendirmeler ve yorumlar gibi sinema ve televizyon yapımlarıyla ilgili bilgileri içerir.
Uygulamada kullanılan veri seti, hazır verisetleri ve örnek çalışmaların olduğu Kaggle sitesinden elde edilmiştir. Elde edilen veri seti "Turkish Movie Sentiment Analysis Dataset" isimi ile Kaggle sitesinden kolayca bulunabilir. Veriler ile ilgili görselleştirmeler sağdaki gibidir.

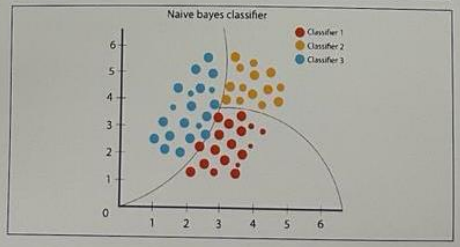


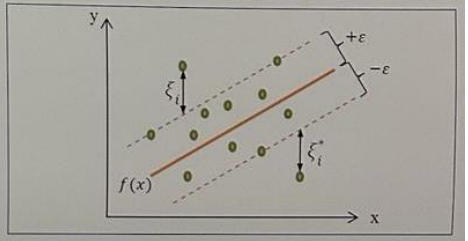
Lojistik regresyon, bir bağımlı değişkenin iki kategorik değeri (örneğin, evet/hayır, 1/0) arasındaki ilişkiyi analiz etmek için kullanılan bir istatistiksel modeldir.
Lojistik regresyon, bağımlı değişkenin olasılık dağılımını tahmin etmek için bağımsız değişkenlerin lineer kombinasyonunu kullanır.
Sonuç olarak, lojistik regresyon, sınıflandırma problemlerinde kullanılan bir modeldir ve doğrusal olmayan ilişkileri yakalamak için lojistik fonksiyonu ile bağımlı değişkenin dönüşümünü kullanır.

Naive Bayes, makine öğrenmesinde sınıflandırma problemleri için kullanılan bir olasılık tabanlı bir algoritmadır.
Naive Bayes, Bayes teoremine dayanır ve sınıflandırma yaparken bağımsız değişkenler arasındaki ilişkisizlik varsayımını kullanır.
Bu algoritma, veri setindeki özelliklerin bir sınıfa ait olma olasılıklarını hesaplayarak yeni bir veri noktasını en olası sınıfa atar ve bu nedenle genellikle metin sınıflandırması gibi doğal dil işleme problemlerinde etkilidir.

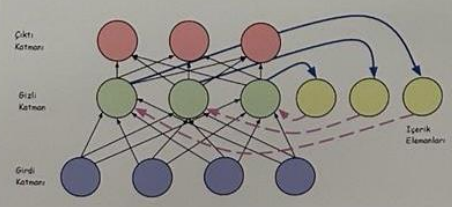
Destek Vektör Makineleri (SVM), makine öğrenmesinde sınıflandırma ve regresyon problemleri için kullanılan bir algoritmadır.
SVM, veri noktalarını birer vektör olarak temsil eder ve bu vektörlerin sınıflarını birbirinden en iyi şekilde ayıran bir hiperdüzlemi bulmaya çalışır.
SVM, veri noktalarını sınıflandırırken maksimum marjinal sınıflandırma prensibini kullanır ve aynı zamanda çekirdek fonksiyonları aracılığıyla doğrusal olmayan ilişkileri de yakalayabilir.







Yinelemeli Sinir Ağı (RNN), makine öğrenmesinde tekrar eden yapıların modellemesi için kullanılan bir yapay sinir ağı türüdür.
RNN'ler, girdi verilerinin yanı sıra önceki çıktıları da geri besleme döngüsü aracılığıyla kullanarak, geçmiş bilgiyi hafızasında tutma yeteneğine sahiptir.
Bu özellik, zaman serileri, doğal dil işleme, konuşma tanıma gibi problemlerde zaman bağımlılıklarını ve uzun vadeli bağımlılıkları modellemek için etkili bir yöntem sunar.
RNN'lerin en yaygın varyasyonları LSTM (Uzun Kısa Vadeli Hafıza) ve GRU (Geriye Dönük Birim) olarak bilinir, bu varyasyonlar RNN'lerin uzun vadeli bağımlılıkları daha iyi yakalayabilmesini sağlamak için geliştirilmiştir.



Classification Report:

	Precision	Recall	f1-Score	Support
Olumsuz	0.89	0.68	0.77	4602
Nötr	0.53	0.82	0.65	4253
Olumlu	0.71	0.75	0.72	4892
Accuracy			0.72	13747
Macro Avg	0.71	0.75	0.71	13747
Weighted Avg	0.78	0.72	0.73	13747

Bu çalışma, Imdb'deki film yorumlarından duygu analizi ile duygu tahminleme yapmayı amaçlamıştır. Uygulanan modeller arasında RNN modeli en yüksek (0.72) doğruluk oranını göstermiştir. Çalışmadaki ilk amacımız Türkçe dili için hazırlanmış Turkish Lemmatizer ile çalışmak ve literatürdeki diğer çalışmalardan farkı olup olmadığını test etmektir, ayrıca bu veriye en uygun modeli tespit etmek, bu modelin performansını incelemek ve film yorumlarından duygu analizi yapabilme kapasitesini ölçmektir. Bu şekil bir veri setine RNN modelinin uygun olabileceğini ve çalışmada kullanılan diğer modellerden daha başarılı olduğunu söyleyebiliriz. Olumlu ve olumsuz yorumların tahmini, nötr yorumlara göre daha yüksektir çünkü olumlu ve olumsuz yorumlarda belirli kelimeler öne çıkmaktadır. Veriler bireysel kullanıcılardan kontrolsüz alındığını ve verilen puanların yorumlarla doğru eşleşemeyeceğini göz önünde bulundurduğumuzda alınan doğruluk oranının normal olduğunu söyleyebiliriz ve kontrol edilmiş ve doğruluğuna emin olduğumuz bireysel setlerde bu sonuç daha olumlu olacaktır.

KAYNAKÇA

[1] Kızılkaya, Y. M. (2018). Duygu Analizi ve Sosyal Medya Alanında Uygulama (Doktora Tezi). Uludağ Üniversitesi. Sosyal Bilimler Enstitüsü, Bursa. https://acikerisim.uludag.edu.tr. (Temmuz, 2018).
[2] Elmas, Ş. Ş. (2019). Sosyal Medya Mesajlarının Veri Madenciliği Yöntemi İle Duygu Analizi (Sivas İli Örneği) (Yüksek Lisans Tezi). Sivas Cumhuriyet Üniversitesi. Sosyal Bilimler Enstitüsü, Sivas. https://acikerisim.cumhuriyet.edu.tr. (Ocak, 2019).
[3] https://aws.amazon.com/tr/what-is/sentiment-analysis/

Öznitelik Seçimi ve Boosting Algoritmalarını Kullanarak Makine Öğrenmesi Üzerinden Tahmin Modellemesi Geliştirme

Hazırlayan: 18023012 - Hürkan Üstündağ
Bitirme Çalışması Danışmanı: Doç.Dr.Öyküm Esra Yiğit
YILDIZ TEKNİK ÜNİVERSİTESİ İSTATİSTİK BÖLÜMÜ

Özet:

Günümüzde, veri biliminin ve yapay zeka teknolojilerinin gelişimi, özellikle makine öğrenmesi alanındaki yenilikler, daha önce çözülemeyen karmaşık sorunları çözme ve belirsizliği yönetme potansiyeline sahiptir. Boosting algoritmaları ve öznitelik seçimi teknikleri bu yenilikler arasındadır. Boosting algoritmaları sayesinde modelin performansı iyi derecelerde yükselebilir ve öznitelik seçimi teknikleri sayesinde de model daha doğru ve etkili bir şekilde çalışabilir Spaceship Titanic adlı veri seti bir yolcu uzay gemisidir ve amaç hangi yolcuların kaybolduğunu tahmin edebilmektir. Bu tahmin işlemi veri seti uygun bir şekilde temizlenerek bahsi geçen teknikler uygulanarak gerçekleşmiştir ve model performansının bu durumdan pozitif bir şekilde etkilendiği bu analiz sonucunda da yine görülmektedir

Öznitelik Seçimi

Öznitelik seçimi, bir makine öğrenmesi modelinin eğitim sürecinde en önemli adımlardan biridir. Modelin veriler üzerinde nasıl eğitileceğini ve hangi özelliklerin modelin tahminlerinde önemli bir rol oynayacağını belirlemek, genel model performansını ve öğrenme sürecinin etkinliğini önemli ölçüde etkiler. Modelin performansını birkaç farklı yolla artırabilir. İlk olarak, gereksiz veya ilgisiz özelliklerin kaldırılması, modelin eğitim süresini hızlandırır. Bu, özellikle büyük veri setlerinde ve yüksek boyutlu veri setlerinde önemlidir, çünkü bu tür veri setlerinde özniteliklerin sayısı genellikle verinin kendisinden daha fazla olabilir. İkincisi, öznitelik seçimi modelin anlaşılırlığını artırabilir. Az sayıda özellik, modelin çalışma mekanizmasını daha kolay anlamayı ve yorumlamayı sağlar. Son olarak, öznitelik seçimi, modelin genelleme yeteneğini artırabilir, yani modelin yeni, görülmemiş veriler üzerinde daha iyi performans göstermesini sağlar

Öznitelik Seçimi Teknikleri

Filtre teknikleri

Bu yöntemler genellikle hızlı ve etkilidir çünkü makine öğrenmesi modelleri tarafından kullanılan özelliklerin bir ön değerlendirmesini yaparlar. Bunlar, özelliklerin her birini bağımsız olarak değerlendiren istatistiksel tekniklerdir.

1) Bilgi Kazancı

Bir özelliğin hedef değişkenin belirsizliğini ne kadar azalttığını ölçer. Özellikle karar ağaçları gibi algoritmalarda kullanılır, çünkü ağacın dallanma noktalarını belirlerken hangi özelliklerin en bilgilendirici olduğunu belirlemeye yardımcı olur.

2) Ki-Kare Testi

Kategorik özellikler için kullanılır ve hedef değişken ile olan bağımsızlığını test eder. Ki-Kare testi, bir özelliğin ve hedef değişkenin bağımsız olduğunu belirleme olasılığını ölçer, bu da özelliğin hedef sınıfı tahmin etmek için yararsız olabileceğini gösterir.

3) Fisher Skoru

Özelliklerin ayrılabilirliğini ölçer. Fisher skoru yüksek olan özellikler hedef değişkeni daha iyi ayırabilir. Fisher Skoru, bir özelliğin hedef sınıflar arasında ne kadar iyi ayrım yaptığını ölçer, bu da özelliğin hedef sınıfı tahmin etmek için ne kadar yararlı olabileceğini gösterir.

4) Korelasyon Katsayısı

Bu, özelliklerin hedef değişken ile olan doğrusal korelasyonunu ölçer. Yüksek bir korelasyon katsayısı, bir özelliğin hedef değişken ile güçlü bir ilişkisi olduğunu ve dolayısıyla tahmin için yararlı olabileceğini gösterir.

5) Varyans Eşiği

Bu yöntem, düşük varyanslı özellikleri (yani verileri az değişen özellikleri) çıkarır. Çünkü düşük varyans genellikle az bilgi anlamına gelir. Bu ölçüm, bir özelliğin değerlerinin ortalama değerinden ne kadar farklı olduğunu ölçer.

Sarmal Metodlar

Bu metotlar, özellik seçimini bir arama problemi olarak ele alır. Özelliklerin farklı kombinasyonları belirlenir ve her biri ayrı ayrı değerlendirilir. Kapak metotları genellikle zaman alıcıdır çünkü birçok farklı özellik kombinasyonunu değerlendirmek gerekmektedir. Ancak, bunlar genellikle daha doğru sonuçlar verir çünkü özelliklerin seçimi belirli bir modelin performansına dayalıdır

1) İleri Doğru Seçim Tekniği

Bu yöntem, boş bir özellik seti ile başlar ve modelin performansını en çok artıran özellikleri tek tek ekler.

2) Geriye Doğru Seçim Tekniği

Bu yöntem, tüm özellikler ile başlar ve modelin performansını en az etkileyen özellikleri tek tek çıkarır.

3) Geri Arama Tekniği

Bu yöntem, modelin önemi düşük olan özellikleri teker teker çıkarır ve en iyi altkümenin hangi özelliklerden oluştuğunu belirler.

Gömülü Metodlar

Bu metotlar, modelin eğitimi sırasında özellik seçimini gerçekleştirir. Modelin karmaşıklığına bir ceza uygulayarak modelin aşırı uyumunu kontrol ederler.

1) L1 Düzenlileştirme (Lasso)

Regresyon katsayılarını sınırlayan ve bu sınırlama sonucu bazı özelliklerin katsayılarını tamamen sıfıra çeken bir yöntemdir. Bu nedenle Lasso, modelin karmaşıklığını kontrol ederken aynı zamanda otomatik bir özellik seçimi gerçekleştirir. Bu yönüyle, Lasso, özellik seçimini modelin eğitimi sırasında gömülü bir şekilde gerçekleştiren bir özellik seçim tekniğidir.

2) L2 Düzenlileştirme (Ridge)

Ridge regresyonu, regresyon katsayılarını sınırlar, ancak onları tamamen sıfıra çekmez. Bu, Ridge regresyonunun aşırı uyumu önlemesine yardımcı olur ve modelin genelleme kabiliyetini artırır. Ancak, Ridge, katsayıları tamamen sıfıra çekmediği için, tüm özellikleri modele dahil eder ve bu nedenle otomatik bir özellik seçimi sağlamaz. Ancak, özelliklerin katsayılarının büyüklüğüne bir sınırlama getirerek, önemli özelliklerin modelde daha belirgin hale gelmesine yardımcı olabilir.

3) Rassal Orman

Her özelliğin sınıflandırma performansına katkısını ölçer ve önemsiz özellikleri belirler

Boosting Algoritmaları

Boosting, makine öğrenmesinde bir topluluk öğrenme tekniği olup genellikle sınıflandırma ve regresyon problemlerini çözme amacıyla kullanılır. Boosting, birden çok zayıf öğrenici (genellikle karar ağaçları) kullanarak te bir güçlü öğrenici oluşturmayı amaçlar. Her bir öğrenici, önceki öğrenicinin hatalarını düzeltmeye odaklanır. Boosting yöntemleri genellikle genel bir modelin performansını artırmak için kullanılır ve aşırı öğrenmeyi engeller yeteneği ile bilinirler.

Boosting Algoritma Çeşitleri

AdaBoost (Adaptive Boosting)

AdaBoost, en popüler boosting algoritmalarından biridir. Bu yöntem, iterasyonları boyunca zayıf sınıflandırıcıları birleştirir ve her bir sınıflandırıcının ağırlığını hatalarına bağlı olarak ayarlar. AdaBoost, genellikle aşırı öğrenme eğilimini azaltır ve genel model performansını artırır.

Gradient Boosting

AdaBoost'un bir genelleştirmesidir ve her iterasyonda modelin hatalarını azaltmak için gradient descent algoritmasını kullanır. Genellikle ağaç tabanlı zayıf öğrenicilerle kullanılır ve güçlü bir tahmin modeli oluşturur.

XGBoost (Extreme Gradient Boosting)

XGBoost, Gradient Boosting'in hızlandırılmış ve daha etkili bir versiyonudur. Bu algoritma hem hızı hem de performansı ile popüler olmuştur ve birçok Kaggle yarışmasında kazanan modellerde kullanılmıştır.

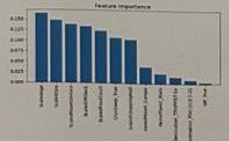
LightGBM

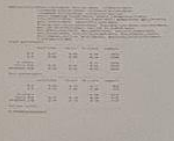
Microsoft tarafından geliştirilen LightGBM, Gradient Boosting'in bir başka uygulamasıdır. Daha hızlı eğitim süresi daha verimli hafıza kullanımı ile bilinir. Ayrıca, büyük veri setleriyle çalışma yeteneği ile de dikkat çeker.

CatBoost

Yandex tarafından geliştirilen CatBoost, özellikle kategorik verilerle çok iyi çalışabilen bir gradient boosting algoritmasıdır. Bu özellik, onu diğer boosting algoritmalarından ayırır.

Uygulama





Veri temizlendikten sonra öznitelik seçim tekniklerinden rassal orman tekniği uygulanmıştır ve bu sayede en önemli değişkenlerin hangileri olduğu ilk şekilde görülmektedir. Daha sonra bağımlı değişkeni en az etkileyen değişkenler çıkartıldı ve model tahmin edildi. 2. şekil bir lojistik regresyon modellemesidir ve gözüktüğü üzere 0.75'lik bir performansa sahiptir fakat 3.şekilde de görüldüğü üzere Xgboost uygulandığında modelde artış gerçekleşmiştir ve 0.78'lik bir skora ulaşılmıştır.

Kaynakça

[1] (Liu, H., Motoda, H. (eds) Feature Selection for Knowledge Discovery and Data Mining, 1998)

[2] https://www.kaggle.com/code/michalbrezk/xgboost-classifier-and-hyperparameter-tuning-85

[3] https://www.analyticsvidhya.com/blog/2020/02/4-boosting-algorithms-machine-learning/

İSTATİSTİK BÖLÜMÜ

MÜŞTERİ YAŞAM BOYU DEĞERİ TAHMİNİNİN PAZARLAMAYA ETKİSİ

Aleyna ATAN 18023067

Danışman: Prof.Dr.Gülhayat GÖLBAŞI ŞİMŞEK

ÖZET

Bu tez çalışmasında müşteri yaşam boyu değerinin şirketlerin yapabilecekleri pazarlama stratejileri ile müşterileri portföylerinde uzun vadede nasıl tutabilecekleri üzerine bir inceleme yapılmıştır. Bu çalışma esnasında müşteri yaşam boyu değeri tanımı, hesaplama yöntemi ve bu değerin uygulama alanları hakkında bilgiler verilmiştir. Aynı zamanda bir müşterinin bir işletme için, müşteriyi elde tutmanın ekonomik açıdan ne kadar faydalı olabileceği hakkında bir hesaplama yapılmıştır. Bu hesaplama yapılırken müşteri o işletmeden ne kadar sıklıkta satın alım yaptığına, hangi periyotlarda satın alımı yapıldığına, kaç defa satın alım olduğuna ve bu satın alımların parasal değerine ait veriler RFM metoduna tabii tutulmuştur.
Sonraki aşamada müşterinin parasal değeri, yani toplam fiyatı bulunmuştur. BG/NBD modelini kullanarak her müşteri için satın alım sayısı tahmini yapılmıştır. Bu tahminler de yapıldıktan sonra Gama-Gama modeli ile müşteri yaşam boyu değerini tahmin etmek için gereken parasal değerin tahmini yani her işlem için en olası değer tahmini yapılmıştır.

MÜŞTERİ YAŞAM BOYU DEĞERİ

Müşteri yaşam boyu değeri (Customer Lifetime Value, CLTV), bir müşterinin bir şirketle bağı süresince sağlamış olacağı toplam maddi değeri ifade eder. Müşteri yaşam boyu değeri, bir şirketin müşterisi ile kurduğu ilişki süresince müşterinin yapacağı satın almalar, vermiş olduğu hizmetlerde elde edeceği gelirleri ve diğer sağladığı tüm değerlerin toplamı olacak şekilde hesaplanır.

MÜŞTERİ YAŞAM BOYU DEĞERİNİN HESAPLANMASI

Müşteri yaşam boyu değeri (CLTV), bir işletme müşterisinin sağlayacağı tahmini değeri hesaplarken kullanılan bir değerdir. Müşteri yaşam boyu değeri, hesaplaması aşağıda verilen formül üzerinden yapılır:

Müşteri Yaşam Boyu Değeri (CLTV) = (Müşteri Değeri / Müşteri Kaybetme Oranı) * Kar Marjı

- Müşteri Değeri = Ortalama Sipariş Değeri * Satın Alma Sıklığı
- Ortalama Sipariş Değeri = Toplam Fiyat / Toplam İşlem
- Satın Alım Sıklığı = Toplam İşlem / Toplam Müşteri Sayısı
- Müşteri Kaybetme Oranı = 1- Birden Fazla Alışveriş Yapan Müşteri Oranı
- Birden Fazla Alışveriş Yapan Müşteri Oranı = Birden Fazla Alışveriş Yapan Müşteri Sayısı / Toplam Müşteri Sayısı
- Kar Marjı = Toplam Fiyat * 0.10 (.10 değişkenlik olabilir.

MÜŞTERİ YAŞAM BOYU DEĞERİ HESAPLAMA MODELLERİ

1. Berger ve Nasr'ın müşteri yaşam boyu değeri hesaplama modelleri
2. Bruhn'un müşteri yaşam boyu değeri hesaplama modelleri
3. Collings ve Baxter'in müşteri yaşam boyu değeri hesaplama modeli
4. Gelbrich ve Wünschmann'ın müşteri yaşam boyu değeri hesaplama modeli
5. RFM modeli

RFM analizi, pazarlamada özellikle direk pazarlamada en çok bilinen ve uygulanan müşteri bölümleme yöntemlerinden biridir. RFM; Recency, Frequency, Monetary kelimelerinin baş harflerinden oluşur. Recency, müşterinin son işleminin güncelliğine, frequency işlem sıklığına, monetary de müşterinin harcadığı toplam parayı ifade etmektedir. RFM analizi 1970'li yıllardan beridir kullanılan bir yöntemdir (Berry ve Linoff, 2004: 447-70). Mevcut müşterilerimiz arasında yeni bir teklife cevap vermeye en hazır olanı hangisidir sorusuna cevap veren bir tekniktir. Bu teknik genellikle doğrudan pazarlamada kullanılır.

PAZARLAMA

4P	4C
Ürün (Product)	Tüketiciye Sağlanan Değer (Customer Value)
Fiyat (Price)	Tüketiciye Maliyeti (Cost to Customer)
Fiziksel Dağıtım (Place)	Kolaylık (Convenience)
Tutundurma (Promotion)	İletişim (Communication)

Pazarlama, bir şirketin ürünlerini veya hizmetlerini hedef kitleye tanıtmak, müşteri talebini oluşturmak ve satışları artırmak için kullanılan stratejilerin ve faaliyetlerin bir bütünüdür. Pazarlama, müşteri ihtiyaçlarını anlamak, ürün veya hizmetin değerini iletmek, müşteri ilişkilerini yönetmek ve rekabet avantajı elde etmek gibi amaçları içerir.

PAZARLAMA VE MÜŞTERİ YAŞAM BOYU DEĞERİ (CLTV) İLİŞKİSİ

Pazarlama ve müşteri yaşam boyu değeri arasında yakın bir ilişki bulunmaktadır. Pazarlama, bir şirketin ürünlerini veya hizmetlerini hedef kitleye tanıtmak, marka bilinirliğini artırmak ve müşteri talebini oluşturmak için kullanılan stratejilerin ve faaliyetlerin bir bütünüdür. Müşteri yaşam boyu değeri ise, bir müşterinin şirketle olan ilişkisi boyunca sağladığı toplam gelirin, maliyetin ve karlılığın ölçüsüdür. Pazarlama faaliyetleri, müşteri yaşam boyu değerini artırmak için önemli bir rol oynar. İyi bir pazarlama stratejisi, müşteriye değer sunan ürünlerin veya hizmetlerin tanıtımını, marka bağlılığını artırıcı faaliyetleri ve müşteri deneyimini iyileştirme çabalarını içerir. Bu şekilde, müşterilerin şirketle olan ilişkileri güçlenir ve müşteri sadakati oluşturulur.

PAZARLAMADA TEMEL YAKLAŞIMLAR

Pazarlamadaki yaklaşımları temel olarak:

- İlişki pazarlaması yaklaşımı,
- Müşteri odaklı yaklaşım,
- Değer temelli yaklaşım,
- Pazar yönelimli yaklaşım diye maddelemek mümkündür.

UYGULAMA

Müşteri yaşam boyu değerini hesaplamak şirketler açısından oldukça önemli bir hesaplamadır. Müşterinin yaşam boyu değerini bulmak buna göre hareket etmek müşterileri seçmek ve müşteriye özel iletişim stratejilerine ilişkin karar vermek için iyi bir temel oluşturmaya destektir. Bu değeri hesaplarken yapacağımız analizlerden biri RFM analizidir. RFM analizi ve müşteri yaşam boyu değeri, pazarlama stratejileri ve müşteri ilişkileri yönetimi için önemli araçlardır. Bu analizler, şirketlerin müşterilerini daha iyi anlamalarına ve onlara daha kişiselleştirilmiş deneyimler sunmalarına yardımcı olur.
Bu uygulamada kullanılan veri seti, University of California Irvine bünyesindeki Machine Learning Repository'den alınmıştır. Veri kümesi, İngiltere merkezli bir e-ticaret şirketinin 01/12/2010 ve 09/12/2011 tarihleri arasında gerçekleşen, 541.910 adet online alışveriş işlemlerini içermektedir.

VERİ ANALİZİ

Veri setinde tüm düzeltmeler yapıldıktan sonrasında ülkelere göre satın alım sayıları da analiz edilmiştir. Bu analize göre en çok satın alım yapılan ülkenin İngiltere olduğu görülmektedir.
Veri setinde satın alım yapılan tüm ülkeler birbirleriyle karşılaştırıldığında, en fazla işleme sahip olan İngiltere diğer ülkeler arasından filtrelendi.
En fazla işlem yapılan İngiltere ülkesine ait toplam fiyat analizi de yapıldı.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Total Price
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6.0	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6.0	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8.0	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6.0	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6.0	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34

Veri seti ön işleme sokulduktan sonrasında Müşteri Yaşam Boyu Değerini (CLTV) tahmin etmek için bazı tahminler yapılmalıdır. Bu tahminler, geçmiş satın alma davranışlarını analiz etmeye yarar ve en iyi müşterileri bulmak için kullanılan bir pazarlama tekniğidir. Müşterilerin ne kadar sıklıkla alışveriş yaptıklarını, şu ana kadar harcadıkları toplam tutarı, işletmeden en son ne zaman alışveriş yaptıklarını gibi benzer bilgileri içerir.
Bu değerler şunlardır; sıklık, yenilik, T, parasal değerdir.
Veri setinde tüm düzenlemeler ve ön analizler yapıldıktan sonra, BG (Beta, Geo) kullanılarak, BG/NBD modeli yeni veri çerçevesine uygulandı ve her bir müşteri için satın alım sayısı tahmin edildi. BG/NBD modelinin özeti çıkarıldı.

CustomerID	frequency	recency	T	monetary_value
12747.0	10.0	367.0	369.0	375.725000
12748.0	112.0	373.0	373.0	257.314911
12749.0	3.0	210.0	213.0	999.106667
12820.0	3.0	323.0	326.0	256.573333
12823.0	4.0	222.0	296.0	252.450000

Parasal değerlerin hesaplanması için Unit Price ve Quantity değişkenleri çarpılarak yeni bir özellik oluşturuldu. Bu yeni özellik ile her bir müşteri için toplam fiyatı bulundu.

Müşteri yaşam boyu değerinin veri setindeki 6 ay için tahmin edilebilmesine ilişkin tüm modeller hazırlandı. BG/NBD ve Gama Gama modelleri kullanılarak Müşteri Yaşam Boyu Değeri tahmin edildi.

	coef	se(coef)	lower 95% bound	upper 95% bound
r	2.381731e+00	1.013873e-01	2.183012e+00	2.580451e+00
alpha	1.120749e+02	5.267568e+00	1.017504e+02	1.223993e+02
a	2.301100e-14	3.668831e-10	-7.190678e-10	7.191138e-10
b	2.398753e-05	3.819882e-01	-7.486728e-01	7.487208e-01

Müşteri yaşam boyu değeri tahmini sonrası, her müşterinin segmentine göre ürün sunmak, alt segmentte bulunan müşterilerin yaşam boyu değerlerini arttırmak amacıyla pazarlama planı oluşturmak ve müşteri kazanımı için maliyeti azaltmak amacı ile üst segmentlere odaklanmak amacı ile müşteriler yaşam boyu değerlerine göre segmentlere ve gruplara ayrıldı.

Segment	frequency	recency	T	monetary_value	expected_purc_6_months	6_Months_CLV
Hibernating	3.174713	221.337931	292.245977	147.152900	2.553847	364.667903
Need Attention	4.027650	240.140553	284.440092	265.378345	2.989161	745.573232
Loyal Customers	5.592166	242.944700	276.027650	365.362234	3.759721	1234.753807
Champions	11.108046	261.354023	280.475862	584.771128	5.994943	3095.935737

SONUÇ VE ÖNERİLER

Tahmin sonuçlarına bakıldığında, şirketlerin müşterilerinin yaşam boyu değerlerini tahmin etmelerinin gelecek dönemler için çok fazla efor sarf etmeksizin gelirlerini arttırmalarını sağlayacağını ortaya koymaktadır. Bu sonuçlardan yola çıkılarak, şirketler ilişkilerini devam ettirmesini bekledikleri müşterilerinin sadakatini sağlamak ve arttırmak için pek çok farklı yol izleyebilir. Müşterileri, yaşam boyu değerlerine göre gruplayabilir ve bu gruplar için özel kampanyalar uygulayabilir. Müşterilerin yaratacakları fayda göz önüne alınarak yapılacak kampanyalarla müşterilere özel faydalar sunulması sağlanırken bu sayede şirket gereksiz maliyetlerden kaçınmış olacaktır. Aynı zamanda yaşam boyu müşteri değeri farklı analizlere de girdi olabilir. Her müşterinin firmaya ne kadar fayda sağlayacağını bilmek firmanın gelecek yatırım kararlarından pazarlama faaliyetlerine kadar pek çok avantaj sağlayacaktır.

KAYNAKÇA

- [1] YAPRAKLI, T. Ş., & Keser, E. (2008). MÜSTERİ YAŞAM BOYU DEĞERİNİN ANALİZİ: BİR SAHA ARAŞTIRMASI. Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 12(2), 483-503.
- Tran, K. G., Nguyen, V. H., & Ho, T. Customer segmentation analysis and customer lifetime value prediction using Pareto/NBD and Gamma-Gamma model.
- Yıldırım, Y. O. PAZARLAMANIN DEĞİŞEN GÜCÜ: DİJİTAL DÖNEME GEÇİŞ VE PAZARLAMA 4.0 YAKLAŞIMLARI.

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

DEPREM ALGILARININ İŞTE ÜRETKENLİK ÜZERİNE ETKİLERİ

Elif ÇELİKTEPE 18023062

Danışman: Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK

ÖZET

Bu çalışmada, deprem sonrası travma düzeyi ve deprem riski algısının işte üretkenlik üzerindeki etkisini belirlemek için bir anket çalışması yapılmıştır. Ankette, katılımcıların deprem sonrası yaşadıkları travma düzeyi ve deprem riski algısı ölçülmüştür. Ayrıca, katılımcıların işte üretkenliklerini belirlemek için sorular sorulmuştur. Çalışma kapsamında, veri analiz yöntemleri kullanılmıştır. Veri analiz yöntemleri olarak kişisel değişkenleri incelemek için betimleyici istatistiksel yöntemler ve araştırma değişkenleri arasındaki etkiyi incelemek için çoklu regresyon analizi kullanılmıştır. Çalışma kapsamında, deprem sonrası travma düzeyinin, depresyon-anksiyete-stres seviyesinin, deprem risk algısının ve iş tatmininin; işte üretkenlik üzerinde anlamlı bir etkisi olduğu belirlenmiştir.

DEPREM SONRASI TRAVMA VE DEPREM RİSK ALGISI

Depremzedelerin yaşadıkları afet sonucunda bazı psikolojik sorunlarla karşı karşıya kaldıkları söylenebilir. Travmatik olaylar olarak adlandırılan bu sorunlar, depremzedelerin deprem sonrası uzama ile etkileşimini gerektirmektedir. Risk, olumsuz olaylar nedeniyle ortaya çıkabilecek sonuçların (kayıp veya hasar) olasılığıdır.

DASÖ-21

DASÖ, hasta tarafından doldurulabilen, anksiyete ve depresyonun temel belirtilerini içeren, yüksek psikometrik standartları karşılayan, anksiyete ve depresyonu birbirinden ayırt edebilen bir ölçek oluşturmak amacıyla geliştirilmiştir.

İŞTE ÜRETKENLİK VE İŞ TATMİNİ

İşte üretkenlik, hem işin sonuçları hem de sonuçlara ulaşma sürecindeki kritik davranışları içerir. İş tatmini, "özelliklerinin değerlendirilmesinden kaynaklanan bir işe olumlu bir duygu" olarak tanımlanmaktadır.

ANKET SONUÇLARI VE GÜVENİLİRLİK ANALİZİ

DEMOGRAFİK SORULARIN SONUÇLARI

Sorular	Şıklar	n	%
Cinsiyetiniz nedir?	Erkek	60	%52,2
	Kadın	55	%47,8
Eğitim düzeyiniz?	İlkokul	0	%0
	Ortaokul	1	%0,9
	Lise	4	%3,5
	Lisans	88	%76,5
	Lisansüstü	22	%19,1
Çalışma durumunuz?	Özel Sektör	89	%77,4
	Kamu Sektörü	16	%13,9
	Emekli	2	%1,7
	Öğrenci	2	%1,7
	Serbest Meslek	5	%4,3
	Çalışmıyorum	1	%0,9
Çalışma yılınız?	0-5 yıl	74	%64,3
	6-10 yıl	15	%13
	11-15 yıl	14	%12,2
	15-21 yıl	4	%3,5
	21+ yıl	8	%7
İşyerindeki istihdam türünüz?	Tam Zamanlı	93	%80,9
	Yarı Zamanlı	18	%15,7
	Serbest Çalışan	4	%3,5
İşyerindeki çalışma şekliniz?	Ofis	54	%47
	Hibrit	27	%23,5
	Evden(Home Office)	34	%29,5
Yaşadığınız bölge?	Akdeniz Bölgesi	19	%16,5
	Marmara Bölgesi	2	%1,7
	Karadeniz Bölgesi	89	%77,4
	İç Anadolu Bölgesi	0	%0
	Doğu Anadolu Bölgesi	2	%1,7
	Güneydoğu Anadolu Bölgesi	3	%2,6
		0	%0
Oturduğunuz bina kaç yıllık?	0-5 yıllık	17	%14,8
	6-10 yıllık	24	%20,9
	11-15 yıllık	17	%14,8
	16-21 yıllık	15	%13
	21+ yıllık	41	%35,7
	Bilmiyorum	1	%0,9

DEĞİŞKENLERE YÖNELİK BETİMLEYİCİ İSTATİSTİKLER

Değişkenler	N	Min.	Maks.	Ort.	SS
Deprem Sonrası Travma Düzeyi	115	1	5	2,4	0,78
Deprem Risk Algısı	115	1	5	3,91	0,89
DASÖ-21	115	1	5	1,68	0,57
İşte Üretkenlik	115	1	5	4,16	0,7
İş Tatmini	115	1	5	3,88	0,72

GÜVENİLİRLİK ANALİZİ BULGULARI

Değişkenler	Cronbah's Alpha	Madde Sayısı
Deprem Sonrası Travma Düzeyi	0,916	8
Deprem Risk Algısı	0,95	20
DASÖ-21	0,943	21
İşte Üretkenlik	0,96	25
İş Tatmini	0,776	5

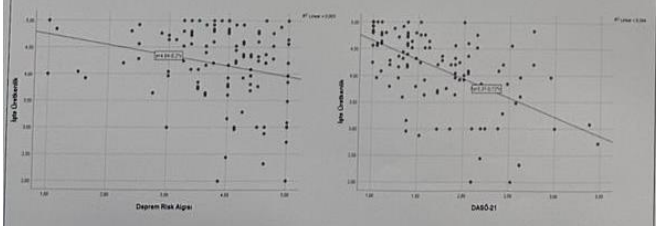
Bu değerlerden ilk dördü için; anketlerin yüksek derecede güvenilir olduğunu, iş tatmini ölçeği için ise anketlerin oldukça güvenilir olduğunu ve analizde kullanılmasında bir engelin olmadığını göstermektedir.

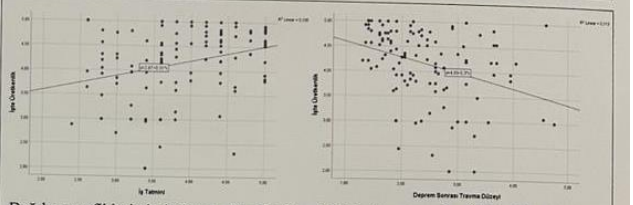
ÇOKLU REGRESYON ANALİZİ

H0:***Deprem Sonrası Travma Düzeyinin, Depresyon-Anksiyete-Stres Seviyesinin, Deprem Risk Algısının, İş Tatmininin; İşte Üretkenlik*** üzerinde etkisi yoktur.

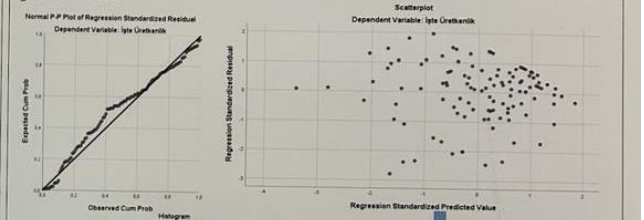
H0a: Deprem Sonrası Travma Düzeyi ve İşte Üretkenlik arasında ilişki yoktur.
H0b: Depresyon-Anksiyete-Stres Seviyesi ve İşte Üretkenlik arasında ilişki yoktur.
H0c: Deprem Risk Algısı ve İşte Üretkenlik arasında ilişki yoktur.
H0d: İş Tatmini ve İşte Üretkenlik arasında ilişki yoktur.

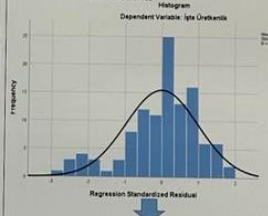
Bağımsız değişkenler ile bağımlı değişken arasındaki doğrusallık kontrolü dağılım grafiği (scatter plot) ile sağlanmıştır.





Dağılım grafiklerinde bulunan düz çizgiler değişkenler arasında doğrusallık olduğunu göstermektedir.





Kalıntı grafiği; noktaların grafik üzerinde oldukça yayıldığı ve uç değer olmadığı görülmektedir. Dolayısıyla eş varyanslılık vardır.

Normal dağılım grafikleri; tahminlere ait hatalar normal dağılmaktadır.

Çoklu Regresyon Analiz Sonuç Tablosu

Değişkenler	β	Standart Hata	Beta	t	p	VIF
Sabit*	4,744	,364	-	13,038	,000	-
DASÖ Seviyesi	-,729	,184	-,175	-6,235	,000	1,858
Travma Düzeyi	,159	,099	,638	1,640	,104	2,371
İş Tatmini	,205	,129	,036	2,954	,004	1,047
Deprem Risk Algısı	-,132	,120	,142	-1,954	,053	1,496

*İşte Üretkenlik (Bağımlı Değişken)

*R = ,649 *R^2= ,421 *Adjusted R^2 = ,399

Değişkenler	İşte Üretkenlik	Deprem Risk Algısı	DASÖ-21	İş tatmini	Deprem Sonrası Travma Düzeyi
İşte Üretkenlik	1,000	-,263	-,593	,312	-,335
Deprem Risk Algısı	-,263	1,000	,328	,045	,569
DASÖ-21	-,593	,328	1,000	-,174	,662
İş Tatmini	,312	,045	-,174	1,000	-,044
Deprem Sonrası Travma Düzeyi	-,335	,569	,662	-,044	1,000

KORELASYON TABLOSU

SONUÇ

Bu çalışmada, deprem sonrası travma düzeyi ve deprem riski algısının işte üretkenlik üzerindeki etkisini belirlemek için bir anket çalışması yapılmıştır. H0 hipotezi reddedilmiştir, yani etki anlamlıdır. Bağımsız değişkenlerin bağımlı değişken üstündeki etkilerine bakıldığında, depresyon-stres-anksiyete seviyesi ve iş tatmini değişkenlerinin etkisi olduğu görülmüş fakat deprem sonrası travma düzeyinin işte üretkenlik üzerinde etkisi için yeterli kanıt bulunamamıştır. Deprem risk algısı değişkeni için ise marjinal düzeyde etkisi görülmüş olup bu değişkenin etkisini araştırmak için çalışma geliştirilmesi önerilmektedir.

KAYNAKÇA

[1] Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. J Abnorm Psychol, 100(3), 316-336
[2] Seçer, İ. (2015). Spss ve Lisrel ile Pratik Veri Analizi, 2. Baskı. ss: 28, Anı Yayıncılık, Ankara.
[3] Endicott J, Nee J (1997) Endicott Work Productivity Scale (EWPS): A new measure to

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

ARIMA İle Türkiye Buğday Verimi Tahmini

Tuğba ÇERİK 17023062

Danışman: Doç. Dr. Gülder KEMALBAY

ÖZET

Tarım Mahsulleri Ofisi tarafından 2021 yılında yayımlanan 1938- 2021 yılları arasında kaydedilen Türkiye buğday verimi veri seti ile geleneksel doğrusal zaman serisi modellerinden biri olan ARIMA yöntemi kullanılacaktır. Araştırmanın amacına ulaşmak için buğday verim veri seti, ARIMA modeli ile elde edilen Otoregresif parametre AR mertebesi kadar gecikmeli bağımsız değişken kullanılacaktır. Uygunluk (fitness) fonksiyonu için MAE (mean absolute error; ortalama mutlak hata) kriteri seçilecektir. ARIMA tahmin modeli oluşturulacaktır.

BUĞDAY

İnsanların tüketim alışkanlıkları belirlemeye yönelik yapılan araştırmalar sonucunda, pirinçten sonra en yoğun olarak tüketilen tahıl ürününün buğday olduğu belirlenmiştir. Bu çalışmada veri seti olarak, TMO (2021) tarafından yayımlanan 1938-2021 yılları arası buğday verisi kullanılmıştır.

Tablo 1. Türkiye'de Her On Yıllık Periyot için Buğday Üretim Miktarı

YILLAR	ÜRETİM (Ton)
1941	3.483.147
1951	5.600.000
1961	7.000.000
1971	13.500.000
1981	17.000.000
1991	20.400.000
2001	19.000.000
2011	21.800.000
2021	17.650.000

ARIMA

Box-Jenkis yöntemi olarakta bilinen ARIMA, zaman serisi verilerini tahmin etmek için kullanılan popüler istatistiksel bir modeldirVerideki geçmiş verileri kullanarak noktalar arsındaki oto korelasyonu açıklamayı hedefleyen, basit ve esnek yapıda bir yöntemdir. Modelin en önemli özelliği, minimum sayıda parametre ile kurulmasıdır .

- AR (oto regresyon) gözlem değerleri ve gecikmeli veriler arasındaki bağımlı ilişkiyi kullanır.
- I(entegre), gözlemlerin farkı alınarak zaman serisi durağan hale getirilir.
- MA(hareketli ortalama) bir zaman serisinin herhangi bir dönemdeki gözlem değeri ile hata terimlerinin lineer olarak ifade edildiği yöntemdir.

Matematiksel olarak ARIMA modeli Denklem (1) gibidir:

$$\Delta d\, Zt = c + (Ø1\ \Delta d\ Zt\text{-}1 + \dots + Øp\ \Delta d\ Zt\text{-}1) - (\theta 1\ et\text{-}1 + \dots + \theta q\ et\text{-}1) + et \qquad (1)$$

DENEYSEL SONUÇLAR

Şekil 1. Yıllık buğday verimi zaman serisi grafiği

Ayırt edilebilir bazı desenler olduğu görülmektedir. Serinin, ortalama da durağan olmadığı yani ortalamasının zaman içinde sabit ilermediği ve zamana bağlı olarak farklılık gösterdiği görülmektedir.

Tablo 2. Verim serisinin birinci farkı için ADF test istatistiği

$\Delta Y_t = (1-L)Y_t$	Level								
Model	kesmesiz ve trendsiz			kesmeli ve trendsiz			kesmeli ve trendli		
Kritik	%1	%5	%10	%1	%5	%10	%1	%5	%10
Değerler	-2.6	-1.95	-1.61	-3.51	-2.89	-2.58	-4.04	-3.45	-3.15
DF Test İstatistiği		-3.7497			-7.9045			-7.9414	

DF test istatistikleri, %1, %5 ve %10 anlamlılık düzeyleri için kritik değerlerden küçüktür. DF Test İstatistiği, kritik değerlerden küçük ise H0 hipotezi reddedilir. Seri birim kök içermemektedir. Ilk farkı alınan seri durağan hale getirilmiştir. Bu nedenle, $r_t = \Delta y_t$ serisinin bir birim kök içerdiği H_0 hipotezi reddedilir, yani $\Delta Y_t = (1-L)Y_t$ serisi birinci farkta durağandır. Bu nedenle, ARIMA modelimizin entegre kısmı (I) 1'ye eşit olacak, yani d=1 olacaktır.

Tablo 3. Aday AICc ARIMA(p,d,q) modelleri için değerler

Model	AICc	Model	AICc
ARIMA(2,1,2)	614.394	ARIMA(0,1,0)	628.5677
ARIMA(0,1,0)	629.9381	**ARIMA(1,1,1)**	**610.9329**
ARIMA(1,1,0)	617.4868	ARIMA(0,1,2)	610.9345

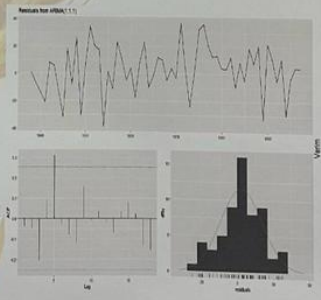
Tablo 4. ARIMA model sonuçları

	AR1	MA1
Coefficients	0.0388	-0.5689
Standard Error	0.0000	0.0000
p-value	0.0000	0.0000

Verilen çıktıya dayanarak, ARIMA(1,1,1) modelini temsil eden denklem şu şekilde ifade edilebilir:

$$r_t = 0.0388 r_t - 0.5689 r_{t-1} + e_t$$

Bu denklem, bir zaman serisi verisindeki değerlerin önceki değerleri ve hata terimi arasındaki ilişkiyi açıklar. İfade edilen model, birinci dereceden otoregresif (AR) ve birinci dereceden hareketli ortalama (MA) bileşenlerini içermektedir. ARIMA(1,1,1) modelinin, veri serisinin durağan hale getirildikten sonra (1 kez fark alındıktan sonra) uygulanan bir regresyon modeli olduğunu ifade eder.



Şekil 2. ARIMA(1,1,1) modeli için artık analizi

Şekil 3. Actual, Fit ve tahmin grafiği

Şekil 12'de verilen artık analizi, ACF grafiğinde önemli pikler olmadığını ve artıkların yaklaşık olarak normal dağıldığını göstermektedir. Seride otokorelasyon gözlenmemektedir. Bununla birlikte, artıkların varyansı zaman içinde sabit gibi görünmemektedir.

SONUÇ

Bu çalışmanın sonuçları, ARIMA(1,1,1) modelinin buğday verimi tahmini için kullanışlı bir araç olduğunu göstermektedir. Bu tür çalışmalar, on birinci kalkınma planı gibi, tarım sektöründe verimliliği artırmak amacıyla ve gelecekteki buğday üretimini daha iyi planlamak için önemli bir bilgi sağlayabilir.

KAYNAKÇA

[1] KEMALBAY, G., & KORKMAZOGLU, O. B. (2021). Sarima-arch versus genetic programming in stock price prediction. Sigma Journal of Engineering and Natural Sciences, 39(2), 110-122.
[2] T.C. Cumhurbaşkanlığı Strateji ve Bütçe Başkanlığı (2019). "On Birinci Kalkınma Planı (2019-2023)", http://www.sbb.gov.tr//wpcontent/uploads/2019/07/On-Birinci-Kalkinma-Plani.pdf (Erişim Tarihi: 27.06.2021).
[3] Bakanlığı, T. (2019). Tarımsal Ekonomi ve Politika Geliştirme Enstitüsü. Tarım Ürünleri Piyasası, Buğday, Ocak, 2021.

YTÜ FEN EDEBİYAT FAKÜLTESİ

...İSTATİSTİK.....BÖLÜMÜ

Üniversite Öğrencilerinde Spor Yapmanın Akademik Başarıya Etkisi

CENK ALTINÖZLÜ 18023054

Danışman: Doç.Dr. Atıf Ahmet EVREN

ÖZET

Tezimin özetine ünlü Amerikalı istatistikçi W. Edwards Deming'in bir sözüyle başlamak istiyorum; "İstatistikler, gerçekliğin dilidir; anlamak için dikkatli bir şekilde dinlemeli ve doğru bir şekilde yorumlamalıyız. Bu tezde belirtilen konu üzerinde SPSS kullanılarak bazı istatistiksel testler yapılarak alakalılık derecesine bakılıp bu testlerin yorumları yapılmıştır. İlk önce spor ve sporun faydaları hakkında kısa bir bilgi verilip sonrasında yapılan testler sonucu bulgular ve sonuç ile öneriler verilmiştir.

ÜNİVERSİTE ÖĞRENCİLERİNDE SPOR YAPMANIN AKADEMİK BAŞARIYA ETKİSİ

GİRİŞ

Spor, önceden belirlenmiş kurallara göre bireysel veya takım halinde yapılan, genellikle rekabete dayalı yarışma ve kişisel eğlence veya mükemmelliğe ulaşmak için yapılan fiziksel veya zihinsel aktivite. Sporları kabaca homo sapiens türünün medeniyete ulaşmadan önce doğayla veya diğer canlılarla yaptığı fiziksel mücadelelerin günümüzdeki medeni karşılığı olarak da tanımlayabiliriz.

Spor yapmanın bazı faydaları da aşağıdaki gibidir;

Kasları güçlendirir. Vücut dayanıklılığını artırır, esneklik verir. Kemikleri güçlendirir ve kemik yoğunluğunun artmasını sağlar. Vücut enerjisinin artmasını sağlar, zindelik verir. Kilonuzu dengede tutmanızı sağlar ve eklemlere binen yükü azaltır. Depresyonu ve kaygıyı engeller. Uykuyu düzenler. Rahat ve derin bir uyku sağlar. Kalp hastalıklarını önler. Tansiyonu düzenler. Yüksek kolesterol riskini azaltır. Daha kaliteli bir yaşam sunar.

Kalp rahatsızlıkları risklerini azaltma, tansiyonu normal tutma, kandaki Triglycreid oranını sabit tutma, kolesterolü düzenlenme, kalp-damar dolaşımını geliştirme gibi etkileriyle sağlık sorunları yaratacak durumları engeller.

ÇALIŞMANIN AMACI

Bu araştırmanın hedefi, üniversite öğrencilerinde sporun akademik başarıya etkisini incelemek ve sporun akademik performans üzerindeki olası faydalarını belirlemektir. Araştırma, sporla ilgili faktörlerin öğrencilerin ders başarısı ve notları üzerindeki ilişkisini anlamak için yapılmaktadır. Bu bilgiler, sporun öğrencilerin akademik performansını desteklemede nasıl kullanılabileceği konusunda stratejiler geliştirmek için kullanılabilir. Bu araştırma, üniversite öğrencilerinde sporun akademik başarı üzerindeki etkisini belirlemek için önemli bir ışık tutacak. Araştırma sonuçları, sporun akademik performansı nasıl etkileyebileceği, sporun öğrencilerin ders başarısı ve notları üzerindeki ilişkisi gibi konuları aydınlatacaktır. Ayrıca, sporun öğrencilerin konsantrasyon, bilişsel fonksiyonlar ve stres yönetimi gibi becerilerini nasıl geliştirebileceğini göstererek, üniversiteler ve eğitimciler için sporun akademik başarıyı desteklemek için kullanılabilir bir araç olduğunu vurgulayacaktır.

MATERYAL VE METODLAR

Verileri sorularını özel olarak hazırlanmış ve toplamda 150 kişiye ulaşan bir anket hazırlayarak toplanmıştır. Soruların ana konuları yapılan sporun çeşidi, akademik başarı ve demografik bilgidir. Elde edilen veriler SPSS(İstatistik paket programı) ile analiz edilmiştir.

BULGULAR

Katılımcıların spor yapma durumları ile genel akademik not ortalamaları arasındaki ilişkiye bakılmıştır. Bu ilişkiye göre GANO puanı 1-2 arasında olan 7 kişi (%50) spor yapıyorken 7 kişi (%50) spor yapmamaktadır, GANO puanı 2-3 arasında olan 58 kişi (%77.3) spor yapıyorken 17 kişi (%22.7) spor yapmamaktadır, GANO puanı 3-4 arasında olan 45 kişi (%73.8) spor yapıyorken 16 kişi (%26.2) spor yapmamaktadır.

Katılımcıların kendi okul başarı değerlendirmeleri ile spor yapmaya başladıktan sonraki akademik başarı durumları arasındaki ilişki analiz edilmiştir. Elde edilen sonuçlara göre; okul başarısını 'çok iyi' olarak değerlendiren öğrencilerin 41'i (%73.2) spor yapmaya başladıktan sonra akademik başarısının arttığını savunurken 8'i (%14.3) değişmediğini veya azaldığını düşünmektedir. Okul başarısını 'idare eder' olarak değerlendiren öğrencilerin 34'ü (%38.6) spor yapmaya başladıktan sonra akademik başarısının arttığını savunurken 27'si (%30.7) değişmediğini veya azaldığını düşünmektedir. Okul başarısını 'çok kötü' olarak değerlendiren öğrencilerden 0'ı (%0) başarısının arttığını savunurken 3'ü (%50) başarısının değişmediğini veya azaldığını düşünmektedir. Katılımcıların spor yapma durumları ile kendi okul başarı değerlendirmeleri arasındaki ilişki analiz edilmiştir. Bulgulara göre okul başarısını 'çok iyi' olarak değerlendiren öğrencilerin; 47'si (%83.9) spor yaparken 9'u (%16.1) spor yapmamaktadır. 'İdare eder' olarak değerlendiren öğrencilerin 60'ı (%68.2) spor yaparken 28'i (%31.8) spor yapmamaktadır. 'Çok kötü' olarak değerlendiren öğrencilerin 3'ü (%50) spor yaparken 3'ü (%50) spor yapmamaktadır.

Katılımcıların üniversite sınavına hazırlanırken düzenli olarak spor yapma durumları ile şu anki okul başarı değerlendirmeleri arasındaki ilişkiye bakılmıştır. Bu ilişkiye göre okul başarısını; 'çok iyi' olarak değerlendiren öğrencilerin 32'i (%57.1) üniversite sınavına hazırlanırken düzenli olarak spor yapmışken 24'ü (%42.9) yapmamıştır; 'idare eder' olarak değerlendiren öğrencilerin 29'u (%33) düzenli olarak spor yapmışken 59'u (%67) yapmamıştır; 'çok kötü' olarak değerlendiren öğrencilerin 1'i (%16.7) düzenli olarak spor yapmışken 5'i (%83.3) spor yapmamıştır.

Katılımcıların okulda göstermiş olduğu performanstan memnun olma durumları ile spor yapma durumları arasındaki ilişki analiz edilmiştir. Elde edilen sonuçlara göre; okuldaki performansından memnun olan öğrencilerin 91'i (%81.3) spor yapıyorken 21'i (%18.8) spor yapmamaktadır. Okuldaki performansından memnun olmayan öğrencilerin 10'u (%50) spor yapıyorken 19'u (%50) spor yapmamaktadır.

Katılımcıların okulda göstermiş olduğu performanstan memnun olma durumları ile spor yapmaya başladıktan sonraki akademik durumları arasındaki ilişki incelenmiştir. Bu ilişkiye göre; okuldaki performansından memnun olan öğrencilerin 71'inin (%63.4) spor yapmaya başladıktan sonra akademik başarısı artmışken 23'ünün (%20.5) akademik başarısı değişmemiştir veya azalmıştır. Okuldaki performansından memnun olmayan öğrencilerin 4'ünün (%10.5) spor yapmaya başladıktan sonra akademik başarısı artmışken 15'inin (%39.5) akademik başarısı değişmemiştir veya azalmıştır.

Katılımcıların okulda göstermiş olduğu performanstan memnun olma durumları ile kaç yıldır spor yaptıkları arasındaki ilişki analiz edilmiştir. Elde edilen bulgulara göre; okuldaki performansından memnun olan öğrencilerin 12'si (%10.7) 0-1 yıldır, 17'si (%15.2) 1-2 yıldır, 23'ü (%20.5) 3-4 yıldır ve 42'si (%37.5) 4 yıldan uzun süredir spor yapmaktadır. Okuldaki performansından memnun olmayan öğrencilerin 1'i (%2.6) 0-1 yıldır, 3'ü (%7.9) 1-2 yıldır, 3'ü (%7.9) 3-4 yıldır ve 11'i (%28.9) 4 yıldan uzun süredir spor yapmaktadır.

Katılımcıların okulda göstermiş oldukları performanstan memnun olma durumları ile haftada kaç gün antrenman yaptıkları arasındaki ilişki incelenmiştir. Elde edilen bulgulara göre; haftada 1-3 gün spor yapan öğrencilerin 51'i (%83.6) okul performansından memnunken 10'u (%16.4) memnun değildir. Haftada 3-5 gün antrenman yapan öğrencilerin 18'i (%81.8) okul performansından memnunken 4'ü (%18.2) memnun değildir.

SONUÇ VE ÖNERİLER

Bu araştırmada sporun akademik başarı üzerindeki etkisi ortaya çıkarılmıştır. Buna göre GANO puanı yükseldikçe spor yapan öğrenci oranının da yükseldiği görülmüştür. Bir başka bulguya göre okul başarısını çok iyi olarak değerlendiren öğrencilerinin büyük çoğunluğu spor yapmaya başladıktan sonra akademik başarılarının arttığını düşünmektedir. Düzenli ve aktif olarak spor yapan öğrencilerin ise kendilerini daha fazla başarılı olarak gördüğü tespit edilmiştir. Üniversite sınavına hazırlanırken düzenli olarak spor yapan öğrenciler şu an kendilerini daha fazla başarılı olarak değerlendirmektedir. Aynı zamanda şu an spor yapan öğrenciler okuldaki başarılarından daha fazla memnun oldukları saptanmıştır. Üniversitede kendisini başarılı olarak değerlendiren öğrencilerin anlamlı olarak çoğunluğu aktif olarak spor yapmaktadır. Öğrencilerin kendilerini akademik bağlamda başarılı görmeleri ile kaç yıldır spor yaptıkları arasında da anlamlı ilişki bulunmuştur. Buna göre spor yapma süresi arttıkça kendilerini başarılı olarak değerlendiren öğrencilerin oranı da artmaktadır.

Elde edilen bu sonuçlara göre, okulda memnun olan öğrencilerin spor yapma oranının daha yüksek olduğu ve spor yapmaya başladıktan sonra akademik başarılarının arttığı görülmüştür. Ayrıca, düzenli spor yapan öğrencilerin genellikle okuldaki performanslarından daha memnun oldukları gözlemlenmiştir.

Sporun akademik başarıyı olumlu yönce etkilediği yapılan testlerin sonuçları doğrultusunda ortaya çıkmıştır. Buna göre; kurumlar ve eğitim sistemleri, sporun akademik başarıyı desteklemesine olanak sağlamak için uygun ortamlar ve olanaklar sunmalıdır. Spor tesislerine erişim sağlamak, spor kulüpleri ve takımları oluşturmak, spor etkinliklerini teşvik etmek gibi adımlar atılabilir. Ayrıca, spor yapmayı teşvik eden programlar ve politikalar geliştirilmeli ve öğrencilere spor yapma fırsatları sunulmalıdır. Böylece öğrencilerin spor yapmaları teşvik edilebilir.

KAYNAKÇA

- https://tr.wikipedia.org/wiki/Spor
- https://hthayat.haberturk.com/saglik/egzersiz/haber/1011758-spor-yapmanin-

YTÜ FEN EDEBİYAT FAKÜLTESİ

İSTATİSTİK BÖLÜMÜ

DESTEK VEKTÖR MAKİNASINDA TEMEL VE KISMİ BİLEŞENLER KULLANILARAK NBA MAÇ SONUÇLARININ TAHMİNİ

Ali Can ÇAKMAK 18023024

Danışman: Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK

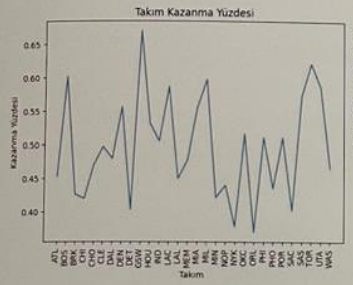
ÖZET

Bu çalışmada, NBA'nin 7 sezonuna ait maç istatistiklerini ve sonuçlarını içeren bir veri seti kullanılmıştır. Temel Bileşenler Regresyonu (PCR) ve Kısmi En Küçük Kareler Regresyonu (PLSR) yöntemleri kullanılarak bu veriler analiz edilmiş ve Destek Vektör Makinesi (SVM) algoritmasıyla sonuçlar tahmin edilmeye çalışılmıştır.

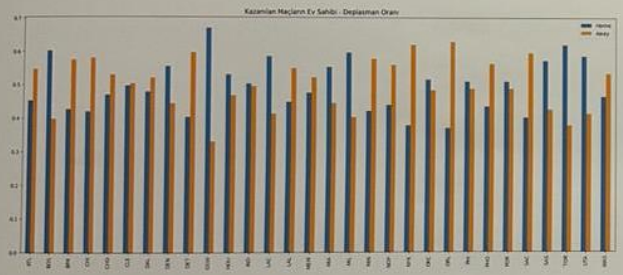
PCR ve PLSR yöntemleri kullanılarak boyut indirgeme işlemi gerçekleştirilmiş ve ardından SVM algoritmasıyla maç sonuçlarının tahmin edilmesi için bir model oluşturulmuştur. SVM, sınıflandırma ve regresyon problemlerinde etkili bir şekilde kullanılabilen bir makine öğrenimi algoritmasıdır.

PCR, çok sayıda bağımsız değişkenin olduğu durumlarda boyut indirgeme sağlamak için kullanılan bir yöntemdir. PCR, veri setindeki varyansın büyük bir kısmını açıklayan bileşenleri belirler ve bunları regresyon analizinde kullanır. Bu şekilde, model karmaşıklığını azaltarak veri setinin özünü yakalamaya çalışır. PLSR hem bağımsız değişkenlerin hem de bağımlı değişkenin boyutunu azaltmayı hedefleyen bir yöntemdir. PLSR, bağımlı değişkenle en yüksek korelasyona sahip bileşenleri bulmaya odaklanır ve bu bileşenleri kullanarak regresyon analizi yapar. Bu sayede, önemli bilgileri korurken modelin karmaşıklığını azaltır.

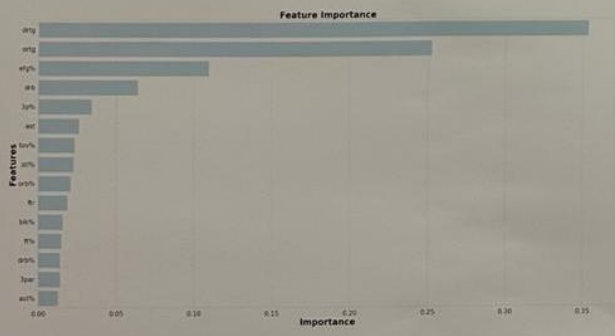
MAÇ SONUÇLARININ ANALİZİ



Yandaki grafikte tüm takımların 7 sezon boyunca oynadıkları maçların kazanma yüzdeleri gösterilmektedir. Tüm bu sezonlar boyunca en başarılı takım Golden State Warriors takımıdır.
En başarısız takım ise Orlando Magic takımıdır.



Yukarıdaki grafikte ele alınan 7 sezonluk maçlarda takımların kazandıkları maçlar ev sahibi ve deplasman olacak şekilde oranlanmıştır. Bu grafik sonucunda ev sahibi olarak en çok kazanma oranına sahip olan takım "GSW" takımıdır. "ORL" takımı ise kazandığı maçları daha çok deplasmandayken kazanmıştır.

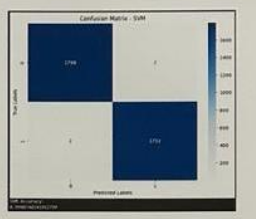


Random Forest algoritması ile tüm değişkenlerin değişken önemlilikleri hesaplanmıştır. Bir takımın maçı kazanmasına en çok etki eden 5 değişken ve oranları aşağıdaki gibidir:

- drtg(defansif reyting): 0.353
- ortg(ofansif reyting): 0.252
- efg%(efektif atış yüzdesi): 0.109
- drb(defansif ribaund sayısı): 0.064
- 3p%(3 sayı yüzdesi): 0.034

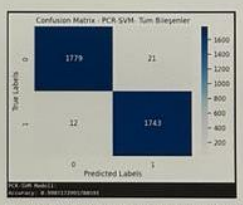
TAHMİN SONUÇLARI

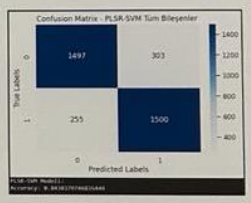
TEMEL BİLEŞENLER İLE SVM TAHMİN MODELİ



Support Vector Machine (SVM) modeli kullanılarak elde edilen sonuçlar, doğruluk oranı ile gösterilmiştir. Confusion matrix ve buna dayalı ısı haritası, modelin sınıflandırma performansını daha ayrıntılı olarak göstermektedir. Doğruluk oranı, y_test veri setindeki gerçek etiketlerle yapılan tahminlerin yüzdesini ifade etmektedir. Bu modelde veri setinde boyut indirgeme yapılmadan ham veri ile tahminleme yapılmıştır. Tahmin sonucunda 0.9988 accuracy değeri çıkmıştır

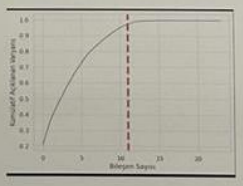
TÜM TEMEL VE KISMİ BİLEŞENLER KULLANILARAK SVM TAHMİN MODELİ

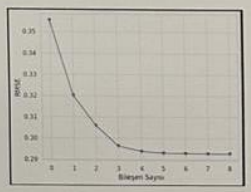




Yukarıdaki grafiklerde kurulan modellerin sonuçları confusion matris olarak gösterilmektedir. Tüm bileşenler kullanıldığında PCR-SVM modelinde accuracy değeri 0.9907 çıkmıştır. PLSR-SVM modelinde ise accuracy değeri 0.8430 olarak elde edilmiştir.

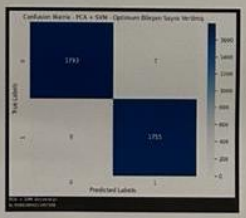
OPTİMUM BİLEŞEN SAYISI İLE SVM TAHMİN MODELİ



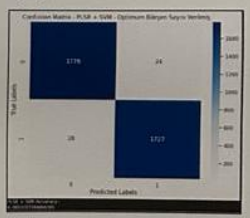


PCA ile veri modellemede kullanmak üzere bileşen sayısı seçmek için grafik oluşturulmuştur. Sonuç olarak 11 bileşenin veri setinin 0.91 ile büyük bir kısmını açıkladığı bulunmuştur ve bileşen sayısı olarak "n_component" parametresinde 11 kullanılacaktır.

Tüm bileşenlerin RMSE değerleri hesaplanarak grafikte gösterilmiştir. Optimum bileşen sayısını seçilirken RMSE değeri en düşük olan bileşen seçilmektedir. Burada 9. Bileşen en düşük RMSE değerine sahip olduğu için seçilmiştir.



Optimum bileşen sayısı kullanıldığında PCR-SVM modelinde accuracy değeri 0.9980 çıkmıştır.



Optimum bileşen sayısı kullanıldığında PLSR-SVM modelinde accuracy değeri 0.9853 çıkmıştır.

SONUÇ

Bu çalışmada, boyut indirgeme yöntemleri SVM ile birleştirilerek model performansının değerlendirilmesi yapılmıştır. Optimum bileşen sayısının seçilmesi, daha güvenilir tahminler elde etmek amacıyla model performansını artırmada yardımcı olmuştur.

KAYNAKÇA

[1] dataquestio. (2021). NBA Games Walkthrough: Get Data [Jupyter Notebook]. GitHub.
[2] Weiner, J. (2019). Predicting the outcome of NBA games with machine learning. Towards Data Science
[3] Wu, L., & Lee, Y. H. (2016). Modeling and forecasting the outcomes of NBA basketball games. Journal of Forecasting, 35(7), 646-661. doi: 10.1002/for.2411