Makine Öğrenmesi ile Uçuş Rötar Tahmini

Burak Gün

Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Bilgisayar ve Bilişim Mühendisliği Ana Bilim Dalı Sakarya Üniversitesi, Türkiye

e-posta: burak.gun1@ogr.sakarya.edu.tr

Abstract—This article delves into the domain of flight delay prediction, employing machine learning methodologies to develop an accurate and reliable model. By leveraging a diverse dataset encompassing variables such as scheduled departure times, distances, months, days, and airline information, the study explores data preprocessing techniques, model training, and performance evaluation. The predictive model, constructed using tools like pandas, scikit-learn, and random forest regressor, aims to enhance understanding and forecasting capabilities in the context of flight delays.

Index Terms—Flight Delay Prediction; Machine Learning;

Abstract—Bu makale, uçuş gecikme tahmininin alanına odaklanarak, doğru ve güvenilir bir model geliştirmek için makine öğrenimi yöntemlerini kullanmaktadır. Planlanan kalkış zamanları, mesafeler, aylar, günler ve havayolu bilgileri gibi değişkenleri içeren çeşitli bir veri kümesi kullanılarak, çalışma veri ön işleme tekniklerini, model eğitimini ve performans değerlendirmeyi araştırmaktadır. Pandas, scikit-learn ve random forest regresör gibi araçlar kullanılarak oluşturulan tahmin modeli, uçuş gecikmeleri bağlamında anlama ve tahmin yeteneklerini artırmayı amaçlamaktadır.

Anahtar Kelimeler—Uçuş Gecikme Tahmini; Makine Öğrenmesi;

I. GİRİŞ

H AVA taşımacılığı, günümüzde hız ve erişim açısından önemli bir taşıma biçimi olarak öne çıkmaktadır. Bu öylesine büyük bir sektördür ki,bugün Amerika Havacılık İdaresi günlük 46.000 uçuşla 26 milyon yolcuyu taşıyarak yılda 15 milyon uçuş idare etmektedir [1]. Ancak, hava yolculuklarının zamanında başlamaması, pek çok seyahat planını olumsuz etkileyen önemli bir sorundur. Bu bağlamda, uçuş kalkış gecikmelerini anlamak ve tahmin etmek, hem havayolu şirketleri hem de yolcular için kritik bir konudur. 2007 yılında yapılan araştırmaya göre, uçuşların %23'ü 15 dakikadan fazla gecikerek, ABD'yi 32.9 milyar dolarlık zarara uğratmıştır ve bunların yaklaşık %17'si hava durumundan kaynaklanmıştır [2]. 2017'de bu zarar 26.6 milyar dolar [3], 2019'da ise bu değer 33 milyar dolar olarak belirtmiştir [4]. Buna ek olarak, hava trafik durumu ve hatta hem uçuş hem de havayolu görevlilerinin grev yapması da uçuş gecikmeleri üzerinde etkili olmuştur [5].

Bu makalede, 2015 yılına ait ABD'de gerçekleşmiş 5.8 milyon uçuş verisi kullanılarak, uçuş kalkış gecikmelerini öngörmede supervised makine öğrenimi modelleri incelenmiştir. Veri setindeki çeşitli özellikleri (saat, mesafe, ay, gün vb.) kullanarak, rastgele orman regresyonu (random forest

regression) ve lineer regression (linear regression) algoritmaları ile modeller eğitilmiş ve sonuçlar karşılaştırılmıştır.

A. Denetimli(Supervised) Makine Öğrenmesi

Denetimli öğrenme, belirli bir eşleştirilmiş girdi-çıktı eğitim örnekleri kümesine dayalı olarak bir sistemin girdi-çıktı ilişkisi bilgisini elde etmeye yönelik bir makine öğrenimi paradigmasıdır. Çıktı, girdi verisinin veya denetimin etiketi olarak kabul edildiğinden, bir girdi-çıktı eğitim örneği etiketli eğitim verisi veya denetimli veri olarak da adlandırılır. Bazen öğretmenle öğrenme (learning with a teacher), etiketli verilerden öğrenme veya tümevarımlı makine öğrenimi olarak da adlandırılır. Denetimli öğrenmenin amacı, girdi ve çıktı arasındaki eşlemeyi öğrenebilen ve yeni girdiler verildiğinde sistemin çıktısını tahmin edebilen yapay bir sistem oluşturmaktır [6].

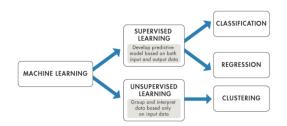


Fig. 1. Makine Öğrenmesi Modelleri

1) Lineer Regresyon Algoritması: Lineer regresyon, bir veya daha fazla bağımsız değişken ile bağımlı değişken arasındaki ilişkiyi regresyon formülü kullanarak analiz eden bir yöntemdir. Bir veya daha fazla geri dönen modelin doğrusal kombinasyonudur [7]. Figür 2'de örnek gösterim sağlanmıştır.

$$Y = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_n X_n + e \tag{1}$$

Burada;

 $Y = C_1kt_1$

X = Bağımlı/Bağımsız Girdi Değişkenler,

B = Bağımlı/Bağımsız Girdi Değişken Katsayıları,

e = Hata Değeri olarak ifade edilmiştir.

1

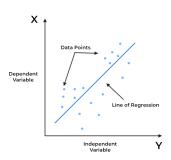


Fig. 2. Lineer Regresyon Algoritması

2) Rastgele Ağaç Regresyon Algoritması: Rastgele ağaç regresyonu, bütünleşik bir yöntemdir ve karar ağacı adı verilen temel birime sahiptir [7]. Bu yöntem, topluluk öğrenmesini kullanarak birden fazla karar ağacını bütünleştirir. Yani burada temel olarak amaç, basit bir girdi yapmak ve her bir karar ağacından bir sonuç almaktır. Daha sonra ise her bir karar ağacından gelen sonuçlar entegre edilmektedir. Figür 3'de örnek gösterim sağlanmıştır.

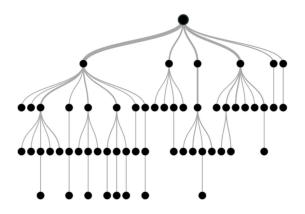


Fig. 3. Rastgele Ağaç Algoritması

II. LİTERATÜR İNCELEMESİ

Havayolu ulaşımına olan talebin artması sebebiyle, uçuş gecikme tahmini, araştırmacıların yoğun olarak ilgilendiği önemli bir konudur. Bu çalışmaların çoğu veri madenciliği ve makine öğrenmesi tekniklerine dayanmaktadır çünkü havacılık operasyonlarına ilişkin büyük miktarlarda operasyonel verinin toplanması, depolanması ve işlenmesi artık mümkündür [8]. Farklı araştırma yöntemlerine göre, uçuş gecikmesi tahmin yöntemleri üç ana kategoriye ayrılabilir:

- 1) Geleneksel İstatistiksel Analiz
- 2) Modelleme Simülasyonu ve Kuyruk Teorisi
- 3) Makine Öğrenimi Yöntemleri

Bunların arasında makine öğrenme yöntemleri şu şekilde ayrılabilir:

- 1) Karar Ağacı
- 2) Rastgele Orman
- 3) Bayesian Ağı

- 4) K-en Yakın Komşu (KNN)
- 5) Genetik Algoritması
- 6) K-Ortalamalar Kümeleme Algoritması
- Destek Vektörü Makinesi (SVM) ve derin öğrenme yöntemleri de dahil olmak üzere Derin Sinir Ağı (DNN), Tekrarlayan Sinir Ağı (RNN) ve Konvolüsyonel Sinir Ağı (CNN) [9].

Nigam R. ve Govinda K. bulut tabanlı lojistik regresyon kullanarak %80 başarı elde etmişlerdir [10]. Jingyi Qu ve ark. sundukları Hafif Ağ ECA-MobileNetV3 sayesinde %98.90 gibi maksimum başarı oranına ulaşabilen, daha hızlı işlem kapasitesine sahip, daha az kaynak kullanan ve daha hızlı gerçek-zamanlı performans sergileyebilen bir model geliştirmişlerdir [11]. Peng Hu ve arkadaşları rastgele ağaç algoritmasında 50 ağaç kullarak yaklaşık %90 gibi başarı oranına sahip bir model oluşturmuşlardır [12]. Yongjie Yan ve arkadaşları standart destek vektör makineleri (SVM) üzerine inşa ettikleri geliştirilmiş destek vektör makineleri çözümüyle normalden daha verimli sonuç ürettikleri bir model tasarlamışlardır [13]. Bin Yu ve arkadaşları ile destek vektör makineleri ve derin öğrenme algoritmalarını birlikte kullanmışlardır [14].

Literatür taraması, bu konuda bir dizi araştırmanın gerçekleştirildiğini ve çeşitli algoritmaların konuya çözüm bulma amacıyla kullanıldığını göstermektedir.

III. METODOLOJI

Bu araştırma, python progralmama dilini kullanarak uçuş rötar süresini öngörmek için iki farklı regresyon yönteminin kullanılmasını içermektedir. İlk adım, çeşitli faktörleri içeren geniş bir veri setinin toplanmasıdır. Bu faktörler arasında uçuş tarihleri, uçuş saatleri, planlanan kalkış zamanı, uçuş mesafesi gibi çeşitli değişkenler yer almaktadır. Bu veri seti, regresyon modellerinin karmaşık ilişkileri anlamasına ve uçuş rötar süresini etkili bir şekilde tahmin etmesine olanak tanımaktadır.

Veri toplama süreci ardından, veri ön işleme adımları eksik veya anormal değerleri ele alarak ve sayısallaştırma işlemleri uygulayarak gerçekleştirilmiştir. Daha sonra, veri seti eğitim ve test setleri olarak bölünmüş, böylece regresyon modellerinin eğitimi ve performanslarının değerlendirilmesi mümkün hale getirilmiştir. Lineer Regresyon ve Rastgele Orman Regresyonu gibi modeller bu aşamada seçilmiş ve bu yöntemlerin uçuş rötar süresini tahmin etme yetenekleri karşılaştırmalı bir analize tabi tutulmuştur.

A. Veri Özellikleri

1) Uçuş Verisi: Bu makalede kullanılan veri seti 2015 yılında ABD'de gerçekleşen 5.8 milyon uçuş bilgisini kapsamaktadır. Veri seti içerisinde bulunan uçuş bilgileri aşağıda belirtilen değişkenleri içermektedir. Year, *month, *day, *day_of_week, airline, flight_number, tail_number, origin_airport, destination_airport, *scheduled_departure, departure_time, departure_delay, taxi_out, wheels_off, scheduled_time, elapsed_time, air_time, *distance, wheels_on, taxi_in, scheduled_arrival, arrival_time, arrival_delay, diverted, cancelled, cancellation_reason, air_system_delay, security_delay, airline_delay, late_aircraft_delay, weather_delay.

İleride görüleceği üzere yıldız(*) ile belirtilen değişkenler makine öğrenmesi sürecinde kullanılan değişkenlerdir.

2) Havayolu Verisi: Uçuşlar 322 havaalanında 14 farklı havayolu şirketi tarafından gerçekleştirilmişr. Veri setinde bulunan havayolu şirket bilgileri aşağıda verilmiştir.

United Air Lines Inc. UA American Airlines Inc. AA US US Airways Inc. F9 Frontier Airlines Inc. **B6** JetBlue Airways 00 Skywest Airlines Inc. AS Alaska Airlines Inc. Spirit Air Lines NK Southwest Airlines Co. WN DΙ Delta Air Lines Inc. E۷ Atlantic Southeast Airlines Hawaiian Airlines Inc. HA MQ American Eagle Airlines Inc. VX Virgin America

Fig. 4. Havayolu Şirketleri

B. Veri Düzenleme

Bu süreç, uçuş verilerindeki bozuk veya eksik bilgilerin tespit edilmesini, hatalı verilerin düzeltilmesini ve eksik bilgilerin doldurulmasını içerir.Bu konu makine öğrenimi modellerinin sağlıklı ve güvenilir tahminler yapabilmesi için kritik bir öneme sahiptir.

C. Veri Görselleştirme ve Yorumlanması

Veri görselleştirmenin ana amacı, insanların büyük miktardaki verileri anlamasına yardımcı olmasıdır. Burada kullanılan veri hacimleri yüksektir ve insanların hatırlamak ve anlamak için bilişsel algılama kapasiteleri sınırlıdır [17]. Veri görselleştirme insanların daha etkili kararlar vermeleri için mümkün olduğunca basit gösterilmelidir ve bu veri analizi ve tahmini için önemli bir yol oynamakdatır.

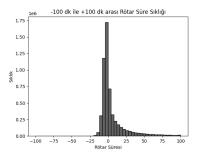


Fig. 5. Rötar Süreleri Histogramı

Figür 5'de rötar süreleri gösterilmiştir. Burada -20dk ile +20dk arasında bir yığılma olduğu görülmektedir.

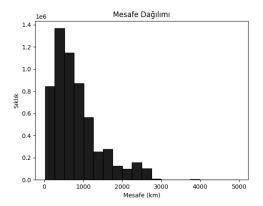


Fig. 6. Uçuş Mesafeleri Histogram

Figür 6'da uçuşların mesafeleri gösterilmiştir ve uçuş yoğunluğunun 500km ile 1000km olduğu gözlemlenmiştir.

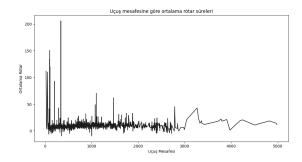


Fig. 7. Mesafeye Bağlı Olarak Gecikme Karşılaştırması

Figür 7'de mesafeye bağlı olarak rötar süreleri gösterilmiştir. Figür 6 ve Figür 7 ortak olarak değerlendirildiğinde ise, uçuş mesafesinin uçuş rötar süresine olan etkisi net bir şekilde görülmektedir.

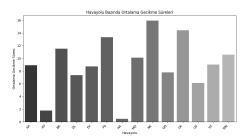


Fig. 8. Havayolu Şirketlerinin Uçuş Rötarına Etkisi

Gecikmelerin sadece hava durumuna ya da dış etkilere bağlı olmadığını Figür 8'te görebilmekteyiz. Burada, havayolu şirketlerinin yıl içerisinde ortalama gecikme süreleri paylaşılmıştır.

Figür 9'da veri setinde bulunan değişkenlerin uçuş rötar gecikmesi üzerinde ne kadar ağırlıklı oldukları gösterilmiştir. Görülmektedir ki uçuş mesafesi ve planlanan uçuş kalkış zamanı, rötar süresini maksimum düzeyde etkilemektedir.

Mean Squared Error: 1623.408069		
	Feature	Importance
4	DISTANCE	0.463999
3	SCHEDULED_TIME	0.403706
1	DAY	0.068872
2	DAY_OF_WEEK	0.040281
0	MONTH	0.023142

Fig. 9. Uçuş Rötarına Etkisi Olan Değişkenler

D. Makine Öğrenmesinde Model Seçimi ve Eğitimi

Model seçimi, bir veri seti ve belirli bir problemin özellikleri gözetilerek optimal bir tahminleme veya sınıflandırma modelinin belirlenmesini içerir. Farklı algoritmaların ve modellerin performansı, genellikle veri setinin karmaşıklığı, boyutu ve içeriği gibi faktörlere bağlı olarak değişiklik gösterir. Doğru model seçimi, veri setinin içsel yapısını ve hedef problemin niteliklerini en iyi şekilde yansılayan bir modelin seçilmesini amaçlar. Bu tercih, modelin genelleme yeteneğini artırabilir ve aşırı uyum veya yetersiz uyum gibi istenmeyen durumların önüne geçebilir.

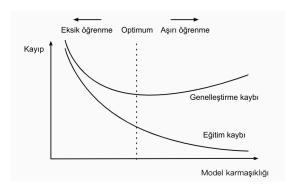


Fig. 10. Model Seçim Kriterleri

Figür 10'da görüleceği gibi seçtiğimiz model eksik öğrenme ya da aşırı öğrenme gibi sorunlarlarla karşılaşabilir ve sonuç tahmin değerleri gerçeği yansıtmayabilir. Bahsedilen konulara ek olarak, model seçimi sistemin performansını ve maliyetini de ciddi oranda etkilemektedir. Bunların önüne geçebilmek için probleme uygun modelin seçilmesi ve eğitim kaybı olmadan optimum bir öğrenme yapılması gerekmektedir.

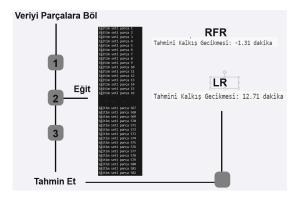


Fig. 11. Model Egitim Aşamakarı

Figür 11'de eğitim aşamaları gösterilmiştir. Veri seti çok büyük olduğu için aynı anda işlemek ve eğitmek çok zordur. Bu yüzden veri seti parçalara bölünüp kısım kısım eğitilmektedir. Farklı 2 model ile eğitildikten sonra ise rötar süresi tahmin edilmiştir. Rastgele orman regresyonu gecikme süresini -1.3 dakika tahmin etmişken bu süre lineer regresyon için 12.7 dakika olarak bulunmuştur.

IV. SONUÇLAR VE PERFORMANS DEĞERLENDİRMESİ

Bu bölümde, doğrusal regresyon ve rastgele orman modellerinin tahmin performansları karşılaştırılmıştır. MSE (Mean Squared Error) ve RSME(Root Mean Squared Error) ölçütleri kullanılarak performans değerlendirmesi yapılmıştır.

Ortalama Kare Hatası (MSE)

MSE, bir regresyon modelinin tahminlerinin gerçek değerlere olan uzaklığını ölçen bir metriktir. Her bir gözlem için hata karelerinin ortalaması alınır. MSE formülü şu şekildedir:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Burada n veri noktası sayısını, y_i gerçek değeri, \hat{y}_i tahmin edilen değeri ifade eder. MSE'nin değeri sıfıra ne kadar yakınsa, modelin tahminleri gerçek değerlere o kadar yakındır.

Kök Ortalama Kare Hatası (RMSE)

RMSE, MSE'nin karekökü alınarak elde edilen bir metriktir ve ölçüm birimiyle uyumlu bir hale getirilir. Bu, tahmin hatalarının orijinal veri setinin biriminde daha anlamlı bir ifade sağlar. RMSE formülü şu şekildedir:

$$RMSE = \sqrt{MSE}$$

RMSE'nin değeri de sıfıra ne kadar yakınsa, modelin tahminleri gerçek değerlere o kadar yakındır. RMSE, hata terimlerinin varyansını ölçerek modelin tahmin gücünü değerlendirmekte etkili bir metriktir.

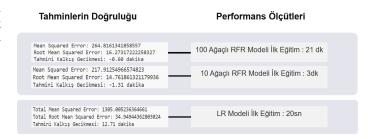


Fig. 12. Tahmin Doğruluğu ve Performans Ölçütleri

Bu çalışmada görülmüştür ki rastgele orman regresyonu(RFR) tahminleri daha az hata ile daha doğru tahminler yapmıştır. Rastgele ormanı regresyonunda ağaç sayısı sonuç tahminlerine etki etmiştir. Figür 12'de görüldüğü üzere MSE ve RMSE değerleri değişmektedir. Modelde kullanılan ağaç sayının performansa olan etkisi ise çok büyüktür. 100 ağaçlı modelde eğitim süresi yaklaşık olarak 21 dakika iken, bu

süre 10 ağaçlı modelde 3 dakikaya kadar düşmektedir. Buna ek olarak, lineer model 20 saniye gibi çok kısa bir sürede eğitilirken sonuç tahminlerinde ise çok büyük sapmalar tespit edilmiştir. Model seçimi ve eğitimi bölümünde bahsedildiği gibi, seçilen modelin performansa ve maliyete olan etkisi burada net bir şekilde gözlemlenmiştir.

Sonuç olarak; LR algoritması kısa sürelerde eğitilip gerçekten uzak tahminler üretirken, RFR algoritması ise daha çok ağaçlı yapıyla uzun süre eğitilip daha doğru tahminler üretmektedir.

V. GELECEK İÇİN ÖNERİLER

Gelecek çalışma kapsamında, sistemin daha isabetli tahminler yapabilmesi için gerçek zamanlı uçuş bilgilerinin API'ler aracılığıyla alınması planlanmaktadır. Bu sayede hem küresel hem de daha doğru çıktılar elde edilmesi hedeflenmektedir.

VI. VERİ ULAŞILABİLİRLİĞİ

Bu makalede kullanılan veri setine https://www.kaggle.com/code/fabiendaniel/predicting-flight-delays-tutorial/notebook adresinden ulaşılabilir. Erişim tarihi: Ocak 2, 2024.

REFERANSLAR

- [1] Olutimehin, Damilare (2023). Predicting Delayed Flights in the United States of America Using Machine Learning. DOI: https://doi.org/10.13140/RG.2.2.27466.13764.
- [2] Kumar, B. H. Manjunatha and Prasad, N. Achyutha and Kalshetty, Jagadevi and V. S., Rekha and Nirmala, G. (2022). Business Analysis and Modelling of Flight Delays using Artificial Intelligence. International Journal of Health Sciences, 6(S1), 10-15. DOI: 10.53730/ijhs.v6nS1. 6735
- [3] Shao, W., Prabowo, A., Zhao, S., Tan, S., Koniusz, P., Chan, J., Hei, X., Feest, B., Salim, F. (2019). Flight Delay Prediction using Airport Situational Awareness Map.
- [4] Mamdouh, H., Diab, A. A., Hefny, H., Fahmy, A. (2023). Title of the Article. Journal of Big Data, 10(179). DOI: https://doi.org/10.1186/ s40537-023-00854-w.
- [5] Fernandes, Nuno & Moro, Sérgio & Costa, Carlos & Aparicio, Manuela. (2019). Factors influencing charter flight departure delay. Research in Transportation Business & Management, 34, 100413. DOI: 10.1016/j. rtbm.2019.100413.
- [6] Liu, Q. ve Wu, Y. (2012). Supervised Learning. DOI: https://doi.org/10. 1007/978-1-4419-1428-6 451.
- [7] Guan, R. (2023). Predicting Forest Fire with Linear Regression and Random Forest. Highlights in Science, Engineering and Technology, 44, 1-7. DOI: https://doi.org/10.54097/hset.v44i.7159.
- [8] Mehmet Atlıoğlu, Mustafa Bolat, Murat Şahin, Volkan Tunalı, Deniz Kılıç. Supervised Learning Approaches to Flight Delay Prediction. Sakarya University Journal of Science, 2020, vol. 24, ss. 1223-1231. http://dx.doi.org/10.16984/saufenbilder.710107.
- [9] Teng Wang, Yufan Zheng, and Haiwen Xu. A Review of Flight Delay Prediction Methods. In 2022 2nd International Conference on Big Data Engineering and Education (BDEE), IEEE, 2022. http://dx.doi.org/10. 1109/bdee55929.2022.00029.
- [10] Rahul Nigam and K. Govinda. Cloud based flight delay prediction using logistic regression. In 2017 International Conference on Intelligent Sustainable Systems (ICISS), IEEE, 2017. http://dx.doi.org/10.1109/iss1. 2017.8389254.
- [11] Jingyi Qu, Bo Chen, Chang Liu, Jinfeng Wang. Flight Delay Prediction Model Based on Lightweight Network ECA-MobileNetV3. Electronics, 2023, vol. 12, s. 1434. http://dx.doi.org/10.3390/electronics12061434.
- [12] Peng Hu, Jianping Zhang, Ning Li. Research on Flight Delay Prediction Based on Random Forest. 2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), 2021. http://dx.doi.org/10.1109/iccasit53235.2021.9633476.
- [13] Weinan Wu, Kaiquan Cai, Yongjie Yan, Yue Li. An Improved SVM Model for Flight Delay Prediction. 2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC), 2019. http://dx.doi.org/10.1109/dasc43569. 2019.9081611.

- [14] Bin Yu, Zhen Guo, Sobhan Asian, Huaizhu Wang, Gang Chen. Flight delay prediction for commercial air transport: A deep learning approach. Transportation Research Part E: Logistics and Transportation Review, 2019, vol. 125, ss. 203-221. https://doi.org/10.1016/j.tre.2019.03.013. https://www.sciencedirect.com/science/article/pii/S1366554518311979.
- [15] Gui, G. et al. (2020). Flight Delay Prediction Based on Aviation Big Data and Machine Learning. *IEEE Transactions on Vehicular Technology*, 69, 140-150. http://doi.org/10.1109/TVT.2019.2954094
- [16] Kim, Y. J. et al. (2016). A deep learning approach to flight delay prediction. 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), 1-6. http://doi.org/10.1109/DASC.2016.7778092
- [17] Llaha, Olta ve Aliu, Azir (2023). Application of Data Visualization and Machine Learning Algorithms for Better Decision Making.