Teknoloji Fakültesi

MARMARA ÜNİVERSİTESİ

TEKNOLOJİ FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

BİTİRME PROJESİ 2. ARA RAPORU

Makine Öğrenimi ile Uçuş İptal Gecikme Tahmini ve Akıllı Uçak Bilet Sistemi

PROJE YAZARI

Eda Nur Mutlu - 170421843

Nuri Can Birdemir - 171421013

DANIŞMAN

Dr. Öğr. Üyesi EYÜP EMRE ÜLKÜ

İL, TEZ YILI

İstanbul, 2025

Teknoloji Fakültesi

MARMARA ÜNİVERSİTESİ

TEKNOLOJİ FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

BİTİRME PROJESİ 2. ARA RAPORU

Makine Öğrenimi ile Uçuş İptal Gecikme Tahmini ve Akıllı Uçak Bilet Sistemi

PROJE YAZARI

Eda Nur Mutlu - 170421843

Nuri Can Birdemir - 171421013

DANIŞMAN

Dr. Öğr. Üyesi EYÜP EMRE ÜLKÜ

İL, TEZ YILI

İstanbul, 2025

İÇİNDEKİLER

Sayfa

ŞEKİL LİSTESİ

TABLO LİSTESİ

ÖZET

Bu çalışma, uçuş iptali ve gecikmelerini tahmin edebilen bir makine öğrenimi modeli geliştirmeyi amaçlamaktadır. Hava yolu taşımacılığında, beklenmedik iptaller ve gecikmeler hem yolcular hem de havayolu şirketleri için önemli sorunlara yol açmaktadır. Bu bağlamda, geçmiş uçuş verileri, hava durumu bilgileri ve uçuş detayları kullanılarak tahminleme modelleri oluşturulmuştur.

Çalışmada öncelikle, Amerika Birleşik Devletleri Ulaştırma Bakanlığı (DOT) tarafından sağlanan açık uçuş verileri incelenmiş ve uygun veri setleri bir araya getirilmiştir. Daha sonra veri temizleme, eksik verileri giderme, zaman formatlarını düzenleme ve özellik mühendisliği gibi veri ön işleme teknikleri uygulanmıştır.

Makine öğrenimi aşamasında XGBoost, Random Forest (Rastgele Orman), Decision Tree (Karar Ağacı), KNN (k-en yakın komşu) ve Gradient Boosting (Gradyan Güçlendirme) algoritmaları kullanılarak tahmin modelleri geliştirilmiştir. Modelin performansını artırmak amacıyla SMOTE (Synthetic Minority Over-sampling Technique) yöntemi ile veri dengelenmiş ve modellerin accuracy (doğruluk), precision (kesinlik), recall (duyarlılık), f1-score ve ROC-AUC gibi metrikleri hesaplanmıştır. Son olarak, geliştirilen modelin uçuş iptali ve gecikme tahmininde ne kadar başarılı olduğu değerlendirilmiştir.

Çalışmanın sonucunda, uçuş iptali ve gecikmeleri tahmin edebilen başarılı bir model oluşturulmuş ve bu modelin uçak bileti satın alacak yolculara yardımcı olabilecek bir web platformuna entegre edilmesi sağlanmıştır.

ABSTRACT

This study aims to develop a machine learning model capable of predicting flight cancellations and delays. In air transportation, unexpected cancellations and delays cause significant problems for both passengers and airline companies. In this context, prediction models have been created using historical flight data, weather information, and flight details.

Initially, open flight data provided by the U.S. Department of Transportation (DOT) was examined, and relevant datasets were compiled. Various data preprocessing techniques were applied, including data cleaning, handling missing values, adjusting time formats, and feature engineering.

During the machine learning phase, XGBoost, Random Forest, Decision Tree, KNN, and Gradient Boosting algorithms were used to develop prediction models. To improve model performance, the SMOTE method was applied to balance the dataset, and evaluation metrics such as accuracy, precision, recall, f1-score, and ROC-AUC were calculated. Finally, the effectiveness of the developed model in predicting flight cancellations and delays was assessed.

As a result of this study, a successful model capable of predicting flight cancellations and delays was created. Future work includes integrating this model into a web platform to assist passengers in making informed flight booking decisions.

BÖLÜM 1. GİRİŞ

Havayolu taşımacılığı, küresel ölçekte hızla büyüyen bir sektör olmasına rağmen, uçuş gecikmeleri endüstrinin karşılaştığı en önemli sorunlardan biri olmaya devam etmektedir. Uçuş gecikmeleri, yolcu memnuniyetini doğrudan etkilerken, operasyonel aksaklıklar nedeniyle havayolu şirketlerini ve havalimanlarını ekonomik kayıplarla karşı karşıya bırakmaktadır. Bununla birlikte, gecikmeler havalimanı trafiğini yoğunlaştırarak zincirleme gecikmelere yol açmakta ve hava sahası yönetimini karmaşık hale getirmektedir [1].

ABD Ulaştırma Bakanlığı tarafından yayımlanan 2023 yılı verilerine göre, ABD'de büyük hava yolu şirketlerine ait iç hat uçuşlarının yalnızca %78'i zamanında kalkış gerçekleştirebilmiştir. Diğer bir ifadeyle, uçuşların yaklaşık %22'si planlanan saatten 15 dakika veya daha fazla gecikmeli olarak kalkmıştır [3].

Benzer şekilde, Birleşik Krallık'ta da uçuş gecikmeleri ciddi bir sorun teşkil etmektedir. 2023 yılında, Birleşik Krallık'tan seyahat eden yolcuların %34'ü uçuşlarında gecikme veya iptalle karşılaşmıştır. Toplamda 45 milyon yolcu bu tür aksaklıklardan etkilenmiş, 3,8 milyon yolcunun uçuşu ise tamamen iptal edilmiştir. Özellikle Londra Gatwick Havalimanı, %42'lik gecikme oranıyla en düşük zamanında kalkış performansına sahip havalimanı olarak öne çıkmıştır [4]. Bu durum, uçuş gecikmelerinin sektörde hâlâ yaygın ve çözülmesi gereken bir problem olduğunu ortaya koymaktadır.

Pandemi döneminde havayolu taşımacılığına olan talepte yaşanan büyük dalgalanmalar, uçuş gecikmeleri ve iptallerine yönelik analizlerin ve tahminleme çalışmalarının önemini bir kez daha gözler önüne sermiştir. Bu karmaşık sorunlara çözüm üretmek için geleneksel yaklaşımların ötesine geçilerek, yapay zekâ ve makine öğrenimi gibi ileri düzey teknolojilerin kullanımı kaçınılmaz hale gelmiştir. Bu bağlamda geliştirilecek akıllı sistemler hem yolcuların bireysel memnuniyetini artıracak hem de havayolu şirketlerinin operasyonel verimliliğini iyileştirecek stratejik faydalar sağlayacaktır.

Bu çalışma kapsamında, uçuş gecikmeleri ve iptallerini tahmin edebilen makine öğrenimi tabanlı bir model geliştirilmiştir. Modelin en önemli yenilikçi yönlerinden biri, geleneksel çalışmalarda sıklıkla göz ardı edilen meteorolojik verilerin uçuş bilgileriyle entegre edilmesidir. Özellikle Meteostat API aracılığıyla elde edilen hava durumu parametrelerinin (sıcaklık, yağış, rüzgâr yönü/hızı vb.) modele dahil edilmesiyle, tahminlerin doğruluk ve güvenilirliği önemli ölçüde artırılmıştır. Literatürde genellikle yalnızca geçmiş uçuş verileri veya operasyonel değişkenler kullanılırken, bu çalışmada çevresel koşulların etkisi detaylı

biçimde modellenerek daha bütüncül ve gerçekçi bir yaklaşım sunulmuştur. Bu özgün entegrasyon sayesinde hem yolcu memnuniyeti odaklı hem de operasyonel verimliliği destekleyen öngörüler sağlanabilmiştir.

Geliştirilen model, uçuş gecikme tahminlerini doğrudan bilet satış platformuna entegre ederek kullanıcı deneyimini iyileştirmeyi amaçlamaktadır. Yolcular, uçuş bilgilerini girdiklerinde, uçuşlarının gecikme riski hakkında tahminler alabilecek ve bu doğrultuda seyahat planlamalarını optimize edebilecektir. Alternatif uçuş önerileri, bilet fiyat analizi ve gecikme riskine göre daha iyi planlama seçenekleri sunularak yolculara değer katılacaktır. Aynı zamanda, havayolu şirketleri, operasyonel süreçlerini daha verimli hale getirebilecek, gecikme kaynaklarını daha iyi yönetebilecek ve gelir kayıplarını minimize edebilecektir.

Makine öğrenimi yöntemleri, büyük ve karmaşık veri setlerinden anlamlı bilgiler çıkarma konusunda güçlü bir araçtır [5]. Bu çalışma kapsamında, uçuş gecikme ve iptallerini tahmin etmek için XGBClassifier algoritması tercih edilmiştir. XGBClassifier, yüksek doğruluk oranı ve genellenebilirlik özellikleriyle öne çıkmaktadır [6]. Bunun yanı sıra, bu algoritmanın esnek yapısı, modelin sürekli olarak yeni verilerle güncellenebilmesini sağlayarak gelecekteki performansını artırmakta ve dinamik bir çözüm sunmaktadır. Geleneksel istatistiksel yöntemlere kıyasla, XGBClassifier daha karmaşık ilişkileri modelleyebilme kabiliyeti sayesinde, uçuş gecikmeleri ve iptalleri gibi çok faktörlü problemlerde daha üstün bir performans sergilemektedir.

Çalışma, üç ana hedef kitleye değer yaratmayı amaçlamaktadır:

1. Yolcular: Gecikme riskine dair erken uyarılar alarak seyahatlerini daha bilinçli planlayabilecek ve alternatif uçuş seçeneklerine erişim sağlayabileceklerdir.
2. Havayolu Şirketleri: Gecikme kaynaklarını daha iyi analiz ederek operasyonel süreçlerini optimize edebilecek ve müşteri memnuniyetini artırabileceklerdir.
3. Havalimanı Yönetimi ve Sivil Havacılık Otoriteleri: Gecikme tahminlerine dayalı olarak uçuş planlamasını daha verimli hale getirebileceklerdir.

Sonuç olarak, bu çalışma, uçuş gecikmelerine yönelik çözüm arayışında makine öğrenimi teknolojilerinin sağladığı yenilikçi imkanlardan faydalanmaktadır. Çalışmanın küresel havayolu taşımacılığına katkı sağlayacak bir model sunması ve özellikle pandemi döneminden sonraki süreçte karşılaşılan operasyonel zorluklara [8] cevap verebilmesi hedeflenmiştir.

1.1. Proje Çalışmasının Amacı ve Önemi

Bu çalışmanın temel amacı, havayolu taşımacılığında karşılaşılan uçuş gecikmeleri, iptaller ve iptal nedenlerini tahmin edebilen bir makine öğrenimi modeli geliştirmektir. Çalışma kapsamında, uçuş operasyonlarına etki eden çok sayıda faktör sistematik olarak analiz edilmiş ve bu faktörler arasındaki karmaşık ilişkiler makine öğrenimi algoritmaları aracılığıyla modellenmiştir. Özellikle hava durumu verilerinin etkisini derinlemesine incelemek üzere, Meteostat API ile elde edilen meteorolojik parametreler uçuş verileriyle entegre edilmiştir. Bu sayede, tahmin modelinin doğruluğu ve genellenebilirliği önemli ölçüde artırılmış hem yolcu memnuniyetini artırabilecek hem de havayolu şirketlerinin operasyonel karar alma süreçlerine katkı sağlayabilecek bir karar sistem oluşturulmuştur.

Çalışmanın teknik boyutunda, farklı makine öğrenimi algoritmalarının performansları karşılaştırmalı olarak değerlendirilmiştir. Bu kapsamda XGBClassifier, Random Forest Classifier, Decision Tree Classifier, KNN Classifier ve Gradient Boosting Classifier algoritmaları kullanılmıştır. Algoritmaların performans karşılaştırması sonucunda en yüksek başarıyı gösteren model seçilerek, bu model üzerinde optimizasyon çalışmaları gerçekleştirilmiştir. Ön değerlendirmelerde XGBClassifier algoritmasının karmaşık veri setlerindeki örüntüleri etkili bir şekilde öğrenebilme ve yeni verilere hızla adapte olabilme özelliğiyle öne çıkması beklenmektedir. Geliştirilen model, çeşitli performans metrikleriyle değerlendirilerek, tahmin doğruluğunun sürekli iyileştirilmesi sağlanmıştır.

Çalışmanın pratik uygulama boyutunda, geliştirilen tahmin modelinin bilet satış platformlarına entegrasyonu gerçekleştirilmiştir. Bu entegrasyon sayesinde, yolculara uçuş seçimi aşamasında gecikme risk değerlendirmesi sunulmuş, alternatif uçuş rotaları önerilmiş ve gecikme riskine dayalı dinamik fiyatlandırma stratejileri geliştirilmiştir. Bu yaklaşım, yolcuların seyahat planlamalarını daha bilinçli yapabilmelerini sağlarken, havayolu şirketlerinin müşteri memnuniyetini artırmasına da katkıda bulunulması planlanmaktadır.

Havayolu şirketleri açısından çalışma, operasyonel verimliliği artırmaya yönelik bir karar destek sistemi oluşturmayı amaçlamaktadır. Bu sistem, gecikmelerin kök nedenlerini analiz ederek, operasyonel maliyetlerin optimize edilmesine ve gecikmelerin önlenmesine yönelik erken müdahalelerin yapılmasına olanak sağlayacaktır. Böylece havayolu şirketleri, veri odaklı kararlar alarak operasyonel süreçlerini iyileştirebilecek ve rekabet avantajı elde edebilecektir.

Bu çalışma, akademik literatüre özellikle meteorolojik verilerin uçuş tahmin modellerine entegrasyonu ve çok sınıflı iptal nedenlerinin tahmini gibi alanlarda somut katkılar sunmaktadır. Literatürde sıkça göz ardı edilen hava durumu değişkenleri, bu projede kapsamlı biçimde işlenerek model performansına doğrudan etki edecek biçimde değerlendirilmiştir. Ayrıca, iptal nedenlerini sınıflandırabilen yapısıyla çalışma, yalnızca gecikme olup olmadığını değil, nedenini de öngörebilen yapay zekâ tabanlı sistemlerin önünü açmaktadır. Sektörel açıdan ise, geliştirilen model; havayolu şirketlerinin gecikme risk analizlerini iyileştirmelerine, yolculara daha şeffaf bilgi sunmalarına ve uçuş planlamasında karar destek sistemi olarak kullanılmasına olanak tanımaktadır. Bu yönleriyle çalışma hem akademik araştırmalar hem de havacılık sektöründeki dijital dönüşüm uygulamaları için güçlü bir referans niteliği taşımaktadır.

1.2. Literatür Taraması

Günümüzde havayolu taşımacılığına olan talebin hızla artması, uçuş gecikmelerinin analizini önemli bir araştırma alanı haline getirmiştir [2]. Araştırmacılar, uçuş gecikmelerini tahmin etmek için makine öğrenimi ve veri madenciliği tekniklerinden yaygın olarak yararlanmaktadır. Literatürdeki çalışmalar genellikle havalimanı tesislerinin konumu, hava durumu ve havalimanı kapasitesi gibi faktörlere odaklanmıştır. Makine öğrenimi teknikleri, büyük ölçekli veri setlerinin depolanmasını ve işlenmesini mümkün kılarak bu alanda önemli katkılar sunmaktadır [9]. Bununla birlikte, mevcut araştırmaların çoğu belirli bir coğrafi bölgeye veya sınırlı sayıda faktöre odaklanmış olup, gecikme tahmininde bütüncül yaklaşımların geliştirilmesi konusunda eksik kalmıştır.

Bu çalışma, uçuş gecikmeleri için tahmin modellerine yeni bir perspektif kazandırmayı hedeflemektedir. Literatürdeki boşluğu doldurmak amacıyla, uçuş verilerinin yanı sıra meteorolojik veriler gibi çeşitli faktörlerin entegre edildiği kapsamlı bir veri seti kullanılmaktadır. Özellikle Meteostat gibi geniş kapsamlı meteorolojik veri kaynaklarının kullanımı, hava durumu etkilerinin daha hassas bir şekilde modellenmesine olanak tanımaktadır. Delahaye ve Puechmorel'in çalışmasında vurgulandığı gibi, literatürde genellikle ihmal edilen veya yüzeysel olarak ele alınan hava durumu faktörleri, bu çalışma kapsamında hava durumu faktörlerinin uçuş gecikmeleri üzerindeki etkisi derinlemesine analiz edilerek modelin doğruluğu ve güvenilirliği artırılmaktadır. [10].

Metodolojik açıdan, çalışmada Random Forest ve XGBoost gibi gelişmiş algoritmaların yanı sıra veri dengesizliği sorununu çözmek için SMOTE yöntemi uygulanmaktadır.

Chawla ve diğerlerinin ortaya koyduğu gibi, veri dengesizliği problemi, gecikmelerin ve iptallerin nadir gözlemler içermesi nedeniyle literatürde önemli bir sorun olarak öne çıkmaktadır [11]. Bu çalışma, veri dengesizliği problemini sistematik bir şekilde ele alarak daha dengeli, genellenebilir ve yüksek doğrulukta tahmin modelleri geliştirmeyi amaçlamaktadır.

Khaksar ve Sheikholeslami yaptıkları çalışmada, hava yolu gecikmelerini tahmin etmek için çeşitli makine öğrenme algoritmaları uygulamaktadır [12]. ABD ve İran uçuş ağlarından elde edilen veriler üzerinde Decision Trees, Random Forest, Clustering (kümeleme) ve Bayesian sınıflandırma gibi yöntemler denenmiş ve gecikmelerin nedenlerini daha doğru tahmin etme üzerine yoğunlaşılmıştır. Özellikle ABD uçuş ağında görüş mesafesi ve rüzgar hızının gecikmeleri önemli ölçüde etkilediği, İran uçuş ağında ise filo yaşı ve uçak tipi gibi faktörlerin öne çıktığı belirtilmiştir. Elde edilen tahminlerde Decision Trees ve Clustering yöntemleri ile hibrit bir sınıflandırma kullanılarak yaklaşık olarak %70 doğruluk oranına ulaşılmıştır.

Bojia Ye ve arkadaşları yaptıkları çalışmada, havaalanlarındaki uçuş gecikmelerini tahmin etmek için denetimli öğrenme yöntemlerini kullanan bir metodoloji önermektedir [13]. Çalışmada, Nanjing Lukou Uluslararası Havaalanı'ndan elde edilen operasyonel uçuş verileri ve hava durumu bilgileri işlenmiş ve tahmin modelleri için dört tür havaalanı ile ilgili özellik oluşturulmuştur. Çalışma, özellikle 1 saatlik gecikme tahminlerinde yüksek doğruluk sağlamış ve LightGBM modeli %86,55 doğruluk oranıyla en iyi sonuçları vermiştir. Bu modelin sonuçları, operasyonel ve hava koşullarına göre tahmin yapılmasının önemini vurgulamaktadır. Benzer bir çalışmada Atlıoğlu, Türkiye'nin önde gelen bir havayolu şirketinden elde edilen operasyonel veri setini kullanarak, 11 farklı makine öğrenimi modeliyle değerlendirme yapmıştır [14]. Her model için çeşitli performans ölçütlerini karşılaştırarak, en yüksek doğruluğa ulaşmak için veri setindeki en uygun özellikleri belirlemeye çalışmıştır.

Shahinaz M. Al-Tabbakh ve arkadaşları, Mısır Hava Yolları'nın uçuş gecikmesi verilerini analiz etmek için çeşitli makine öğrenmesi tekniklerini uygulamışlardır [15]. Çalışmanın temel amacı, uçuş gecikmelerinin tahmin edilebilmesi için en uygun sınıflandırma algoritmasını belirlemektir. Araştırmacılar, veri hazırlama, sınıflandırma algoritmaları kullanma ve model performansını değerlendirme adımlarını içeren bir metodoloji izlemişlerdir. Sekiz farklı sınıflandırma algoritması (Decision Tree, Random Forest, REPTree, PART, Decision Table (Karar Tablosu) , OneR, JRip) WEKA veri madenciliği

aracında uygulanarak karşılaştırılmıştır. Sınıflandırma modellerinin performansı accuracy, precision, recall, F1-skoru ve ROC alanı metrikleri kullanılarak değerlendirilmiştir. Analiz sonuçlarına göre, PART algoritması %83,1 accuracy oranıyla en yüksek performansa sahip olmuştur. Diğer bir yandan, REPTree algoritması da %80,3 accuracy oranı ve en hızlı çalışma süresiyle öne çıkan ağaç tabanlı sınıflandırıcı olarak belirlenmiştir. Çalışma, uçuş gecikmesi tahmininde makine öğrenmesi tekniklerinin etkinliğini göstermiş ve Mısır Hava Yolları'na uçuş operasyonlarını iyileştirmek için önemli bilgiler sağlamıştır. Araştırmacılar, gelecekte daha büyük veri kümeleri kullanarak büyük veri madenciliği teknolojilerini uygulamayı planlamaktadırlar.

Kurt'un çalışmasında, ABD iç hat uçuş verilerini kullanarak uçuş gecikmelerinin öngörülmesi amaçlanmıştır [16]. Bu doğrultuda, Decision Trees, Random Forest, Bagging (Torbalama), Extra Trees (Ekstra Ağaçlar), Gradient Boosting ve XGBoost sınıflandırıcı gibi çeşitli makine öğrenmesi yöntemleri denenmiştir. Bu modellerin doğruluk, F1 skoru ve recall gibi başarı ölçütleriyle değerlendirildiği çalışmada, en yüksek accuracy oranı %71,72 ile Gradient Boosting algoritması tarafından sağlanmıştır. Çalışma, farklı veri özelliklerini modele dahil ederek tahmin performansının iyileştirilebileceğini önermektedir.

Tang'ın araştırmasında ise New York JFK Havalimanı'ndan kalkan uçuşların bir yıllık verileri kullanılarak uçuş gecikmelerini tahmin etmek amacıyla yedi farklı sınıflandırma algoritması değerlendirilmiştir [17]. Bu algoritmalar arasında Decision Tree, %97,78 accuracy oranı ile en yüksek başarıyı sağlamıştır. Özellikle ağaç tabanlı sınıflandırıcılar olan Random Forest ve Gradient Boosting yöntemlerinin, diğer temel sınıflandırıcılara göre daha yüksek performans sergilediği tespit edilmiştir. Çalışmada verilerin dengesiz dağılımının ağırlıklı doğruluk gibi ölçütlerle giderildiği belirtilmiştir.

Meteorolojik verilerin entegrasyonu, gelişmiş makine öğrenimi algoritmalarının kullanımı ve veri dengesizliği probleminin sistematik şekilde ele alınmasını içeren bu çok yönlü yaklaşımla, havayolu sektöründeki operasyonel süreçleri iyileştirecek daha hassas ve genellenebilir tahmin modellerinin geliştirilmesi hedeflenmektedir. Literatürdeki mevcut çalışmaların sınırlamaları göz önüne alındığında, “Makine Öğrenimi ile Uçuş İptal Gecikme Tahmini ve Akıllı Uçak Bilet Sistemi” isimli bitirme tezi kapsamında hem teorik hem de pratik açıdan önemli katkılar sağlayacağı öngörülmektedir. Özellikle, meteorolojik verilerin kapsamlı entegrasyonu ve modern makine öğrenimi tekniklerinin kullanımı, gelecekteki araştırmalar için yeni bir çerçeve sunma potansiyeli taşımaktadır.

Bu araştırma kapsamında, 2018-2024 yılları arasında yayınlanan ve uçuş gecikmeleri üzerine makine öğrenimi yaklaşımlarını inceleyen çalışmalar sistematik olarak analiz edilmiştir. İncelenen çalışmaların çoğunluğunun tek bir havayolu veya havalimanına odaklandığı, meteorolojik verileri sınırlı düzeyde kullandığı ve veri dengesizliği problemini yeterince ele almadığı tespit edilmiştir. Çoğu mevcut çalışma, yalnızca belirli coğrafi bölgelerle sınırlı verileri veya tek bir faktör grubunu ele almakta ve geniş veri entegrasyonunu yeterince dikkate almamaktadır. Ayrıca, veri dengesizliği sorunu literatürde sıkça karşılaşılan bir engel olmakla birlikte, modern tekniklerin bu bağlamda yeterince uygulanmadığı gözlemlenmiştir.

Bu çalışmada, diğer çalışmalardan farklı olarak uçuş gecikmesi tahmininde kullanılan parametreler çeşitlendirilmiş ve genişletilmiştir. Tahmin modeli geliştirilirken uçuş kodu, hava şartları (örneğin, rüzgar hızı, sıcaklık, yağış durumu), önceki uçuşların iptal/gecikme oranları, uçak tipi ve uçuş yoğunluğu gibi parametreler dikkate alınmıştır. Bununla birlikte, uçuş gecikmelerini sadece "var" ya da "yok" şeklinde sınıflandırmak yerine, belirli dakika aralıklarında tahmin yapılması sağlanmıştır. Örneğin, bir uçuşun 15-30 dakika veya 30 dakika üzeri gecikme yaşama ihtimali, modelin güven oranına dayalı olarak tahmin edilebilmektedir. Bu yaklaşım, yalnızca uçuş iptali üzerine yoğunlaşan önceki çalışmalardan farklılaşarak, gecikme tahmini konusunda daha detaylı ve eyleme geçirilebilir çıktılar sunmaktadır.

Ayrıca, bu çalışmada geniş veri entegrasyonu sağlanarak uçuş verileri ve Meteostat gibi kapsamlı meteorolojik veri kaynakları bir araya getirilmiştir. Bu yaklaşım, hava durumu faktörlerinin uçuş gecikmeleri üzerindeki etkisini daha hassas bir şekilde modellemeye olanak tanımaktadır. Veri dengesizliği sorunu ise SMOTE gibi modern veri işleme teknikleri kullanılarak ele alınmış ve modelin performansı artırılmıştır.

Sonuç olarak, bu çalışma literatürdeki boşluğu doldurarak uçuş gecikmelerini tahmin etmek için dengeli, yüksek doğruluklu ve güvenilir bir model sunmayı hedeflemektedir. Bunun yanı sıra, önerilen modelin detaylı tahmin kabiliyetleri ve geniş veri entegrasyonu sayesinde gelecekteki araştırmalar için önemli bir temel oluşturacağı öngörülmektedir.

BÖLÜM 2 – MATERYAL VE YÖNTEM

Bu bölümde, çalışmanın gerçekleştirilmesinde kullanılan veri seti, veri ön işleme aşamaları ve denetimli öğrenme modelleri hakkında detaylı bilgi verilecektir. İlk olarak, uçuş ve hava durumu verilerinden oluşan veri seti tanıtılacak ve bu verilerin nasıl entegre edildiği

açıklanacaktır. Ardından, veri ön işleme süreci ele alınacak; eksik veri, aykırı değerler ve dengesiz veri gibi sorunlara yönelik uygulanan adımlar anlatılacaktır. Son olarak, tahmin modeli oluşturulmasında kullanılan denetimli öğrenme algoritmalarına (örneğin, XGBClassifier , Random Forest ve XGBoost) odaklanılacak ve bu modellerin nasıl eğitildiği ve değerlendirildiği açıklanacaktır. Bu sayede, çalışmanın metodolojik yaklaşımı ve kullanılan teknikler kapsamlı bir şekilde sunulacaktır.

2.1. Veri Seti

Bu çalışmada kullanılan veri seti, Amerika Birleşik Devletleri Ulaştırma Bakanlığı ve Ulaştırma İstatistikleri Bürosu (Bureau of Transportation Statistics) tarafından sağlanan ve 2016 ile 2024 yılları arasındaki uçuş gecikme ve iptal verilerini içermektedir. Veri seti, DOT'nun "On-Time: Reporting Carrier On-Time Performance" (1987-günümüz) veritabanından alınmış olup, uçuş güzergahları (kalkış ve varış noktaları), olay zaman aralıkları (dakika, yerel saat), gecikme ve iptal nedenleri gibi değişkenleri içermektedir. Bu veriler 3 açık kaynak verileri birleştirilerek oluşturulmuştur. İlk olarak, Patrick Zelaya tarafından sağlanan açık veri kümesinden iptal edilen ve edilmeyen uçuş verileri elde edilmiştir. Veri seti, 1.360.878 satır ve 32 sütun uçuş bilgisi içermektedir [18]. Diğer bir veri seti, yalnızca iptal edilen uçuşları içeren ve 64.097 satırdan oluşan Threnjen tarafından paylaşılan veri setinden alınmıştır [19]. Son veri seti ise, Shubham Singh tarafından sağlanan açık verilerden yalnızca iptal edilen uçuşlara ait 21.824 satırlık uçuş bilgilerini içermektedir [20]. Toplamda 3 veri setinden elde edilen son birleşik veri seti 1.446.799 satırdan oluşmaktadır. Bu 3 veri setine ilişkin ayrıntılı öznitelikler Tablo 1’de sunulmuştur.

Hava durumu, özellikle rüzgâr hızı, sıcaklık, yağış, kar ve görüş mesafesi gibi parametreler aracılığıyla uçuşların zamanında gerçekleşmesini doğrudan etkileyen başlıca dışsal faktörlerden biridir [26]. Literatürde bu tür çevresel verilerin çoğu zaman sınırlı kullanılması, tahmin modellerinin gerçek dünya koşullarına duyarlılığını azaltmaktadır. Bu nedenle, çalışmada uçuş verileri ile Metostat kütüphanesinden elde edilen hava durumu verileri de birleştirilmiştir. Metostat kütüphanesi, her bir havalimanı için coğrafi koordinatlar kullanılarak günlük bazda toplanan meteorolojik veriler ile uçuş gecikme ve iptalleri üzerinde hava koşullarının etkisini incelenmesine olanak tanımaktadır. Bu iki veri setinin entegrasyonu, uçuş performansını daha kapsamlı bir şekilde analiz edebilmek için önemli bir adım olmuştur.

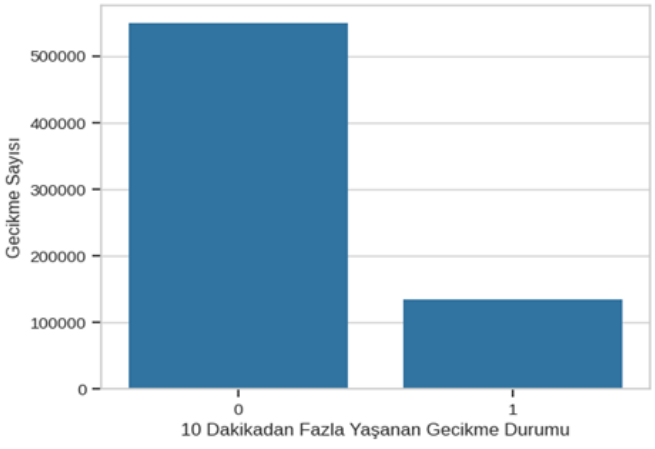
Çalışmada kullanılan veri kaynağı olan Meteostat kütüphanesi, çeşitli meteorolojik istasyonlardan elde edilen hava durumu verilerini sağlamaktadır. Bu veriler, günlük minimum ve maksimum sıcaklık (°C), toplam yağış miktarı (mm), kar yağışı miktarı (mm), rüzgar yönü (0-360°), ortalama ve en yüksek rüzgar hızı (km/saat), ortalama deniz seviyesi basıncı (hPa) ve toplam günlük güneşlenme süresi (dakika) gibi havacılık operasyonlarını doğrudan etkileyebilecek temel meteorolojik parametreleri içermektedir. Her bir havalimanı için coğrafi koordinatlar kullanılarak günlük bazda toplanan bu veriler, uçuş gecikmeleri ve iptalleri üzerinde hava koşullarının etkisini analiz etmek için ana veri setiyle entegre edilmiştir. Bu entegrasyon, uçuş operasyonlarının performansını etkileyen faktörlerin daha kapsamlı bir şekilde incelenmesine olanak tanımaktadır.

Tablo 1 Orijinal Veri Setinin Yapısı

	Güncellenmiş Başlık	Kaynak Başlık	Veri Tipi	Açıklama
0	FL_DATE	FlightDate	object	Uçuş tarihi
1	AIRLINE	Airline	object	Havayolu şirketinin adı
2	AIRLINE_DOT	AirlineDot	object	Havayolu için DOT tanımlayıcısı
3	AIRLINE_CODE	Reporting_Airline	object	Havayolu şirketinin adı
4	DOT_CODE	DOT_ID_Reporting_Airline	int64	Havayolu için DOT tanımlayıcısı
5	FL_NUMBER	Flight_Number_Reporting_Airline	int64	Uçuş numarası
6	ORIGIN	Origin	object	Çıkış havalimanı kodu
7	ORIGIN_CITY	OriginCityName	object	Çıkış havalimanı şehri
8	DEST	Dest	object	Hedef havaalanı kodu
9	DEST_CITY	DestCityName	object	Varış havalimanının bulunduğu şehir
10	CRS_DEP_TIME	CRSDepTime	int64	Planlanan kalkış saati
11	DEP_TIME	DepTime	float64	Gerçek kalkış saati
12	DEP_DELAY	DepDelay	float64	Kalkış gecikmesi
13	TAXI_OUT	TaxiOut	float64	Taksi yaparken harcanan zaman
14	WHEELS_OFF	WheelsOff	float64	Uçağın tekerleklerinin yerden ayrıldığı zaman
15	WHEELS_ON	WheelsOn	float64	Uçağın tekerleklerinin yere değdiği zaman
16	TAXI_IN	TaxiIn	float64	Taksi yaparken harcanan zaman
17	CRS_ARR_TIME	CRSArrTime	int64	Planlanan varış saati
18	ARR_TIME	ArrTime	float64	Gerçek varış zamanı
19	ARR_DELAY	ArrDelay	float64	Varış gecikmesi
20	CANCELLED	Cancelled	float64	Uçuşun iptal edilip edilmediğinin göstergesi (iptal için 1, iptal değil için 0)
21	CANCELLATION_CODE	CancellationCode	object	İptal nedeni (varsa)
22	DIVERTED	Diverted	float64	Uçuşun yönlendirilip yönlendirilmediğinin

				göstergesi (yönlendirildi için 1, yönlendirilmedi için 0)
23	CRS_ELAPSED_TIME	CRSElapsedTime	float64	Planlanan geçen süre
24	ELAPSED_TIME	ActualElapsedTime	float64	Gerçek geçen süre
25	AIR_TIME	AirTime	float64	Havada geçirilen zaman
26	DISTANCE	Distance	float64	Katedilen mesafe
27	DELAY_DUE_CARRIER	CarrierDelay	float64	Taşıyıcı nedeniyle gecikme
28	DELAY_DUE_WEATHER	WeatherDelay	float64	Hava koşulları nedeniyle gecikme
29	DELAY_DUE_NAS	NASDelay	float64	Ulusal Hava Sahası Sistemi (NAS) nedeniyle gecikme
30	DELAY_DUE_SECURITY	SecurityDelay	float64	Güvenlik nedeniyle gecikme
31	DELAY_DUE_LATE_AIRCRAFT	LateAircraftDelay	float64	Uçağın geç varması nedeniyle gecikme

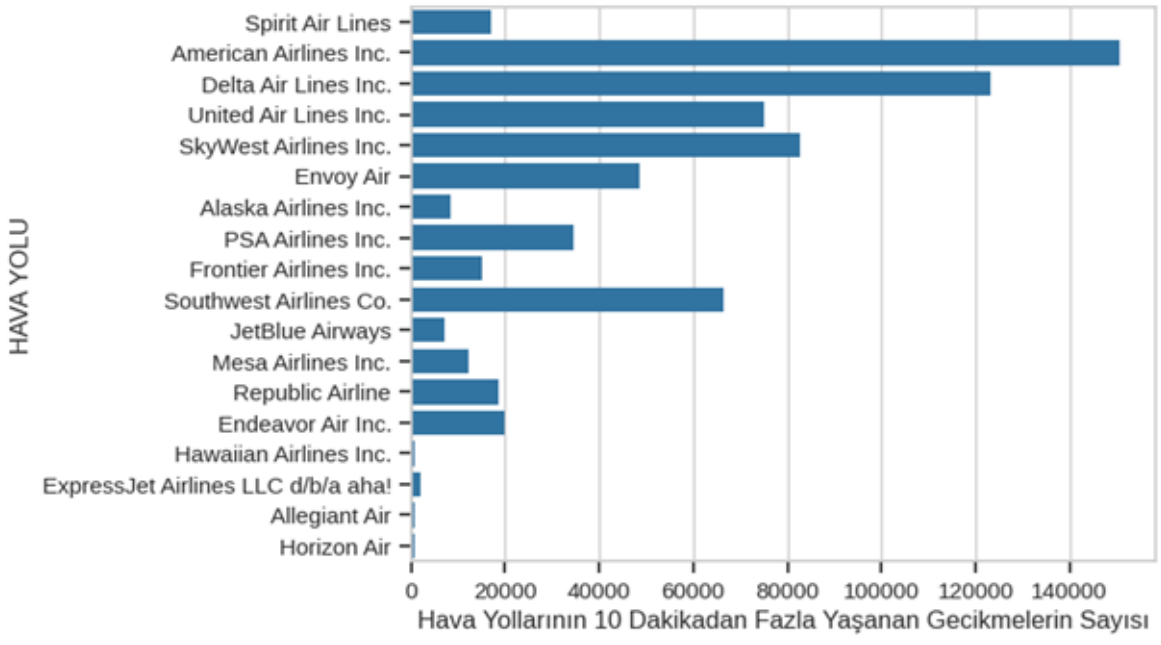
Orijinal veri setinde yapılan ön analizler sonucunda, uçuşların gecikme durumlarını sınıflandırmak amacıyla 10 dakikalık eşik değeri kullanılmıştır. Bu yaklaşıma göre, 10 dakikadan fazla gecikme yaşayan uçuşlar "1" (gecikmeli), 10 dakika veya daha az gecikme yaşayanlar ise "0" (zamanında) olarak etiketlenmiştir. Şekil 1'de bu ikili sınıflandırmaya dayalı olarak gecikme durumlarının dağılımı görsel olarak sunulmuştur. Bu görselleştirme, modelin hedef değişkenini oluştururken hangi kriterlerin esas alındığını göstermesi açısından önemlidir.



Şekil 1 Gecikme Durumu Grafiği

Şekil 2'de, farklı hava yolu şirketlerine ait 10 dakikadan fazla gecikme yaşayan uçuş sayılarının dağılımı yatay bar grafiği biçiminde görselleştirilmiştir. Grafik, hangi hava yollarının gecikmeli uçuş sayısında öne çıktığını karşılaştırmalı olarak göstermektedir. Özellikle American Airlines Inc., Southwest Airlines Co. ve Delta Air Lines Inc. gibi büyük

ölçekli hava yolu şirketlerinin 10 dakikadan fazla gecikme yaşayan uçuş sayılarında diğer şirketlere kıyasla belirgin bir şekilde yüksek değerlere sahip olduğu görülmektedir. Bu durum, söz konusu firmaların uçuş hacminin yüksek olmasıyla birlikte, operasyonel süreçlerinde yaşanan gecikmelere de işaret etmektedir. Buna karşılık, Horizon Air, Allegiant Air ve ExpressJet Airlines gibi daha küçük ölçekli firmaların gecikme sayıları oldukça sınırlı kalmıştır. Grafik, gecikme analizlerinin hava yolu bazında değerlendirilmesinde önemli bir görsel veri sunmaktadır.

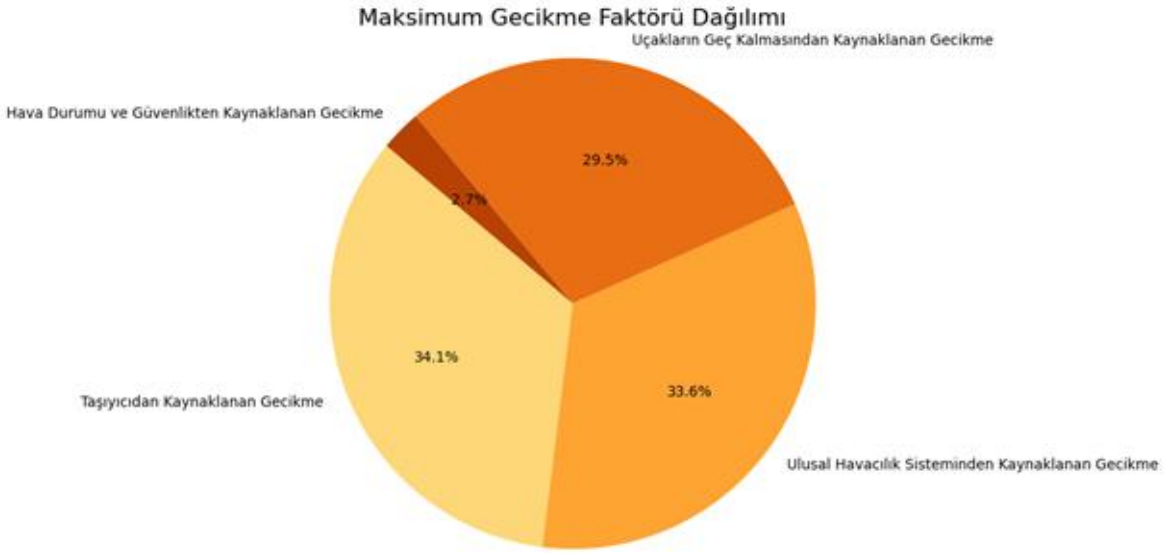


Şekil 2 Hava Yolu Şirketlerine Göre Gecikme Durumu Grafiği

Elimizdeki veri setine göre Şekil 3'teki gibi maksimum gecikme faktörü dağılımı incelenmiştir. Bunun sonucunda, %34.1 taşıyıcıdan kaynaklanan gecikme, %33.6 ulusal havacılık sisteminden kaynaklanan gecikme, %29.5 uçakların geç kalkmasından kaynaklanan gecikme, %2.7 hava durumu ve güvenlikten kaynaklanan gecikmelerin olduğu tespit edilmiştir.

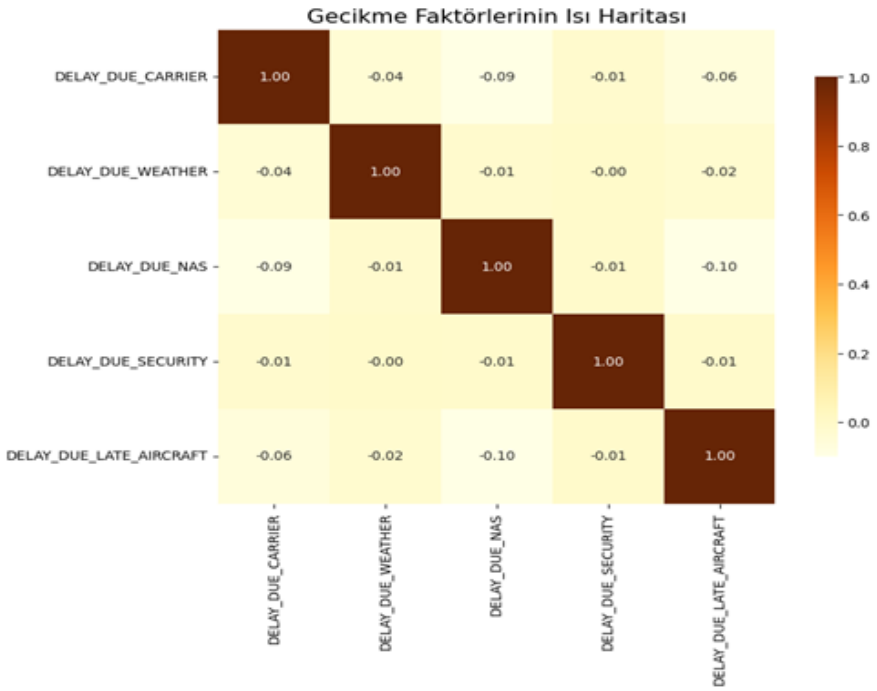
Şekil 3'teki gecikme faktörü dağılımı, makine öğrenimi modellerinin performansını etkileyecek önemli içgörüler sağlamaktadır. Taşıyıcıdan kaynaklanan %34.1'lik gecikme oranı, model eğitiminde hassas bir özellik çıkarımı gerektirmektedir. Ulusal havacılık sisteminden kaynaklanan %33.6'lık gecikme, modellerin karmaşık sistemsel ilişkileri yakalama kapasitesini test etmektedir. Uçak gecikmelerinin %29.5'i, zaman serisi ve ardışık bağımlılık modellemelerinin önemini vurgulamaktadır. Düşük oranlı hava durumu ve

güvenlik gecikmeleri (%2.7) ise sınıf dengesizliği sorunlarına işaret eder, bu da SMOTE gibi veri dengeleme tekniklerinin gerekliliğini ortaya koymaktadır.



Şekil 3 Gecikme Faktörü Dağılımı

Şekil 4'teki gecikme faktörlerinin ısı haritası, uçuş gecikmelerinin birbirleriyle olan korelasyonlarını görsel olarak temsil etmektedir. Renkli hücrelerin yoğunluğu, faktörler arasındaki ilişkilerin şiddetini göstermekte olup, makine öğrenimi modellerinin özellik seçiminde ve çoklu değişken etkileşimlerinin anlaşılmasında kritik bir rol oynamaktadır.



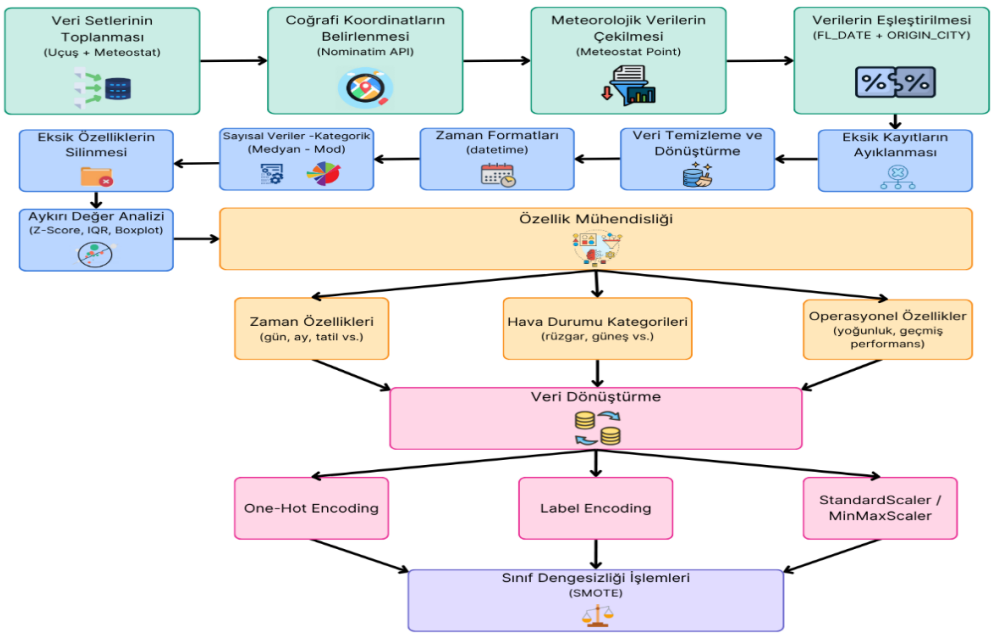
Şekil 4 Gecikme Faktörlerinin Isı Haritası

2.2. Veri Ön İşleme

Makine öğrenimi modellerinin başarısı, büyük ölçüde veri ön işleme aşamasının etkinliğine bağlıdır [7]. Bu çalışmada, uçuş gecikme tahminlerinin doğruluğunu artırmak amacıyla kapsamlı bir veri ön işleme stratejisi uygulanmıştır. Özellikle, meteorolojik verilerin entegrasyonu ve veri kalitesinin iyileştirilmesi konularına odaklanılmıştır.

Veri ön işlemenin ilk aşamasında, uçuş verilerine meteorolojik verilerin entegrasyonu gerçekleştirilmiştir. Bu süreçte Meteostat kütüphanesi kullanılarak her bir uçuş noktası için detaylı hava durumu bilgileri elde edilmiştir. Meteorolojik veri entegrasyonu için öncelikle Nominatim API aracılığıyla havalimanlarının coğrafi koordinatları belirlenmiştir. Elde edilen koordinat bilgileri, Meteostat'ın Point sınıfı kullanılarak meteorolojik veri noktalarına dönüştürülmüştür. Her uçuş tarihi için minimum sıcaklık, maksimum sıcaklık, yağış miktarı, kar yağışı, rüzgar yönü, rüzgar hızı, rüzgar hamlesi, atmosfer basıncı ve güneşlenme süresi gibi kritik meteorolojik parametreler toplanmıştır.

Meteorolojik verilerin uçuş verileriyle birleştirilmesi sürecinde, veri setinin büyüklüğü göz önünde bulundurularak parçalı (chunk) işleme stratejisi benimsenmiştir. Bu yaklaşım, bellek kullanımını optimize ederken veri işleme sürecinin kesintisiz devam etmesini sağlamıştır. Veriler, 'FL_DATE' ve 'ORIGIN_CITY' sütunları üzerinden eşleştirilmiş, eşleşmeyen kayıtlar ise veri kalitesi takibi için ayrı bir dosyada saklanmıştır. Tüm bu süreç Şekil 5'teki akış şemasında gösterilmiştir.



Şekil 5 Veri Ön İşleme Akış Diyagramı

Meteostat kütüphanesi kullanılarak her bir uçuş noktası için detaylı hava durumu bilgilerinin veri setine eklenmesi ile birlikte veri setinin son hali Tablo 2'de gösterilmiştir.

Tablo 2 Model Eğitim Veri Setinin Yapısı

	Güncellenmiş Başlık	Kaynak Başlık	Veri Tipi	Açıklama
1	FL_DATE	FlightDate	object	Uçuş tarihi
2	AIRLINE_CODE	Reporting_Airline	object	Havayolu şirketinin adı
3	ORIGIN_CITY	OriginCityName	object	Çıkış havalimanı şehri
4	DEST_CITY	DestCityName	object	Varış havalimanının bulunduğu şehir
5	CRS_DEP_TIME	CRSDepTime	int64	Planlanan kalkış saati
6	DEP_TIME	DepTime	float64	Gerçek kalkış saati
7	CRS_ARR_TIME	CRSArrTime	int64	Planlanan varış saati
8	CANCELLED	Cancelled	float64	Uçuşun iptal edilip edilmediğinin göstergesi (iptal için 1, iptal değil için 0)
9	CANCELLATION_CODE	CancellationCode	object	İptal nedeni (varsa)
10	DISTANCE	Distance	float64	Katedilen mesafe
11	TMIN	tmin	float64	Minimum sıcaklık (Celsius)
12	TMAX	tmax	float64	Maksimum sıcaklık (Celsius)
13	PRCP	prcp	float64	Yağış miktarı (mm)
14	SNOW	snow	float64	Kar yağışı miktarı (mm)
15	WDIR	wdir	float64	Rüzgar yönü (derece)
16	WSPD	wspd	float64	Rüzgar hızı (m/s)
17	WPGT	wpgt	float64	Rüzgar hamlesi (m/s)
18	PRES	pres	float64	Atmosfer basıncı (hPa)
19	TSUN	tsun	float64	Toplam güneşlenme süresi (saat)

Veri setindeki eksik değerler için kapsamlı bir analiz ve temizleme süreci uygulanmıştır. Eksikliklerin rastgele veya sistematik olup olmadığını belirlemek amacıyla eksik veri dağılımları ve paternleri incelenmiştir. Sayısal değişkenlerdeki eksik değerler medyan ile doldurulurken, kategorik değişkenlerde mod kullanılmıştır. %30'dan fazla eksik veri içeren özellikler veri setinden çıkarılmış, meteorolojik verilerdeki eksiklikler ise zaman serisi karakterine uygun olarak en yakın zaman damgasındaki değerlerle doldurulmuştur.

Aykırı değerlerin tespiti ve işlenmesi için çok yönlü bir yaklaşım benimsenmiştir. Z-skoru ve IQR metodu gibi istatistiksel yöntemler, box plot ve scatter plot gibi görsel analiz teknikleri ve domain bilgisi bazlı kontroller bir arada kullanılmıştır. Gecikme süreleri için 24 saatten fazla olan değerler ve meteorolojik verilerdeki aykırı değerler detaylı bir şekilde incelenmiştir. Analiz sonucunda, veri setine anlamlı katkı sağlayan ve istatistiksel olarak mantıklı kabul edilen aykırı değerler korunmuş, hatalı veri girişlerinden kaynaklanan anormallikler ise uygun yöntemlerle düzeltilmiştir.

Model performansını artırmak amacıyla kapsamlı bir veri analizi ve tahmin için açıklayıcı değişkenlerin türetilmesi ve işlenmesi süreci uygulanmıştır. Zaman bazlı özellikler arasında uçuş saatinin günün hangi dilimine denk geldiği, haftanın günü, tatil günü bilgisi ve sezon bilgisi yer almaktadır. Meteorolojik özellikler, hava durumu verilerinden türetilen kategorik değişkenler, rüzgar şiddeti kategorileri ve görüş mesafesi sınıflandırması gibi yeni özellikleri içermektedir. Operasyonel özellikler kapsamında ise havalimanı yoğunluk göstergeleri, önceki uçuşların gecikme durumları ve rotaya özgü geçmiş performans metrikleri oluşturulmuştur.

Veri standardizasyonu ve normalizasyon aşamasında, sayısal özelliklerin ölçek farklılıklarından kaynaklanan sorunlar ele alınmıştır. Sürekli değişkenler için StandardScaler kullanılarak standardizasyon uygulanırken, sınırlı aralıktaki değişkenler MinMaxScaler ile [0,1] aralığına normalize edilmiştir. Kategorik değişkenler ise One-Hot Encoding yöntemi ile sayısallaştırılmıştır.

Gecikme sınıfları arasındaki dengesizlik problemi ele alınmıştır. SMOTE algoritması kullanılarak azınlık sınıfı örnekleri sentetik olarak artırılmış ve sınıf ağırlıkları parametresi ayarlanarak dengeli bir veri seti elde edilmiştir. Bu kapsamlı ön işleme adımları, modelin eğitim verilerini daha etkin kullanmasını ve gerçek dünya uygulamalarında daha güvenilir tahminler yapmasını sağlamıştır.

İptal ve gecikme nedeni tahmin modeli için veri ön işleme sürecinde, öncelikle zamansal verilerin ayrıştırılması gerçekleştirilmiştir. 'FL_DATE' sütunu datetime formatına dönüştürülerek yıl, ay ve gün bilgileri ayrı özellikler olarak çıkarılmıştır. Kategorik değişkenler için LabelEncoder kullanılarak havayolu kodu (AIRLINE_CODE), kalkış şehri (ORIGIN_CITY) ve varış şehri (DEST_CITY) sayısal değerlere dönüştürülmüştür. İptal kodu (CANCELLATION_CODE) sütunundaki eksik değerler 'N' (iptal yok) değeri ile doldurularak encode edilmiştir.

Veri setindeki dengesizlik problemi, SMOTE algoritması kullanılarak ele alınmıştır. İptal durumu için sampling_strategy=0.5 parametresi ile azınlık sınıfı örnekleri sentetik olarak artırılırken, iptal nedeni sınıflandırması için 'auto' stratejisi kullanılmıştır. Eksik verilerin doldurulması için SimpleImputer ile medyan stratejisi uygulanmış, ardından StandardScaler ile özellikler ölçeklendirilmiştir.

Gecikme tahmin modeli için veri ön işleme sürecinde, öncelikle iptal edilmiş uçuşlar veri setinden çıkarılmıştır. Zamansal veriler benzer şekilde ayrıştırılmış ve kategorik değişkenler LabelEncoder ile dönüştürülmüştür. Gecikme süreleri dört farklı sınıfa ayrılmıştır:

- Sınıf 0: Zamanında veya erken (≤ 0 dakika)
- Sınıf 1: Hafif gecikme (1-15 dakika)
- Sınıf 2: Orta gecikme (16-30 dakika)
- Sınıf 3: Ciddi gecikme (>30 dakika)

Eksik değerler SimpleImputer kullanılarak medyan değerleri ile doldurulmuş ve StandardScaler ile özellikler normalize edilmiştir. Sınıf dengesizliği problemi, k_neighbors=5 parametresi ile SMOTE algoritması kullanılarak çözülmüştür. Bu süreçte, her bir gecikme sınıfı için eşit sayıda örnek oluşturularak dengeli bir veri seti elde edilmiştir.

Her iki model için de özellik önem dereceleri analiz edilmiş ve görselleştirilmiştir. Bu analiz, modellerin tahmin performansını etkileyen en önemli faktörlerin belirlenmesine ve özellik seçimi stratejilerinin değerlendirilmesine olanak sağlamıştır. Veri ön işleme adımları, scikit-learn kütüphanesinin Pipeline yapısı kullanılarak sistematik ve tekrarlanabilir bir şekilde uygulanmıştır.

Bu detaylı veri ön işleme süreci, modellerin eğitim verilerini daha etkin kullanmasını sağlamış ve tahmin performanslarını önemli ölçüde artırmıştır. Özellikle sınıf dengesizliği probleminin çözülmesi ve özellik mühendisliği adımları, modellerin gerçek dünya uygulamalarında daha güvenilir tahminler yapmasına katkıda bulunmuştur.

2.3. Denetimli Öğrenme Modelleri

Makine öğrenimi, bilgisayarların veriyi analiz etmesini, olası desenleri keşfetmesini ve bu desenleri kullanarak tahminlerde bulunmasını sağlayan algoritmaların genel adıdır. Farklı ortamlarda öğrenmenin göreceli zorluğu hakkında bilgi verebilen bu algoritmalar çeşitli kategorilere ayrılmaktadır [21]. Makine öğrenimi algoritmalarının en yaygın iki türü ise denetimli ve denetimsiz öğrenmedir. Denetimli öğrenme algoritmaları, girdileri istenen çıktılara dönüştüren bir işlev oluşturur. Denetimli öğrenmenin temel türleri arasında regresyon ve sınıflandırma yer alır. Denetimsiz öğrenme ise etiketlenmiş örnekler olmadan bir girdi kümesini modellemeye odaklanır.

Bu çalışmada makine öğrenmesi, uçuş iptalleri ve gecikmelerinin tahmin edilmesi amacıyla kullanılmıştır. Havayolu operasyonlarındaki karmaşık ilişkileri modellemek ve çok sayıda değişkenin etkisini analiz etmek için makine öğrenmesi algoritmaları ideal araçlardır [22].

Decision Tree algoritması, veriyi hiyerarşik bir yapıda bölerek sınıflandırma ve regresyon problemlerini çözen bir makine öğrenimi yaklaşımıdır. Her bir düğüm, bir özellik üzerinde karar verme sürecini temsil eder ve ağacın yapısı, en ayırt edici özelliklere dayalı olarak oluşturulur. Uçuş gecikmesi tahmininde, farklı faktörlerin etkisini net bir şekilde görselleştirme imkânı sağlar [23].

LightGBM, Gradient Boosting çerçevesinde çalışan, yüksek performanslı ve verimli bir makine öğrenimi algoritmasıdır. Büyük veri setlerinde hızlı eğitim yapabilme ve bellek kullanımını optimize etme özellikleriyle öne çıkar. Uçuş gecikmesi gibi karmaşık ve yüksek boyutlu veri setlerinde, diğer algoritmalara kıyasla daha yüksek doğruluk oranları elde edebilir [24].

Random Forest, birden fazla Decision Tree birleştiren bir ensemble öğrenme metodudur [25]. Her bir ağaç, bootstrap örneklemesi ve rastgele özellik seçimi ile eğitilir, bu sayede aşırı uyum (overfitting) riskini azaltır. Uçuş gecikmesi tahmininde, farklı alt veri setleri üzerinden tahminler yaparak daha güvenilir sonuçlar elde edilmesini sağlar [22].

Clustering algoritmaları, benzer özelliklere sahip veri noktalarını gruplandırarak, gizli desenler keşfetmeye olanak sağlar. Bayesian sınıflandırma ise olasılıksal bir yaklaşımla, öncül bilgileri ve gözlemsel verileri birleştirerek sınıflandırma yapar. Bu yöntemler, uçuş gecikmelerindeki karmaşık ilişkilerin anlaşılmasında tamamlayıcı rol oynar [23].

XGBoost, makine öğrenimi alanında oldukça popüler ve güçlü bir algoritmadır, özellikle büyük veri setleri üzerinde yüksek doğrulukla tahmin yapma yeteneği ile tanınır [27]. Gradient Boosting tekniklerinin geliştirilmiş bir versiyonu olan XGBoost, her bir iterasyonda modelin hatalarını düzeltmek için yeni ağaçlar ekleyerek modelin genel doğruluğunu artırır. Bu özelliği, özellikle karmaşık ilişkiler ve çoklu faktörlerin etkisi altında olan uçuş tahminleri gibi problemler için oldukça faydalıdır. Uçuş iptali ve uçuş gecikmesi gibi tahminler söz konusu olduğunda, XGBoost algoritması, verinin çok sayıda faktörden etkilenen karmaşık yapısını modellemek için etkili bir araçtır [21].

Uçuş iptali ve uçuş gecikmesi tahminlerinde doğru sonuçlar alabilmek için, algoritma seçiminde dikkatli olunması gereklidir. Burada kullanılan XGBoost algoritması, özellikle iki önemli özelliği ile tercih edildi:

Sınıf dengesizliğiyle baş etme, uçuş iptalleri ve gecikmeleri gibi olaylar genellikle dengesiz veri setleriyle ilişkilidir. Örneğin, iptal edilen uçuşların sayısı, gerçekleştirilen uçuşlara göre çok daha az olabilir. XGBoost, bu tür dengesizlikleri etkili bir şekilde yönetebilir ve azınlık sınıfını daha doğru şekilde öğrenebilir.

Yüksek performans ve esneklik; XGBoost, büyük veri setlerinde hızlı ve doğru tahminler yapabilme yeteneğine sahip bir algoritmadır. Özellikle uçuş gibi çok sayıda faktöre bağlı olan veri setlerinde, XGBoost'un güçlü özellik mühendisliği ve model optimizasyonu yetenekleri oldukça faydalıdır.

Bu nedenle, XGBoost, hem dengesiz veri setlerinde yüksek doğruluk sağlamak hem de karmaşık ilişkileri modellemek için ideal bir seçenek olarak tercih edilmiştir.

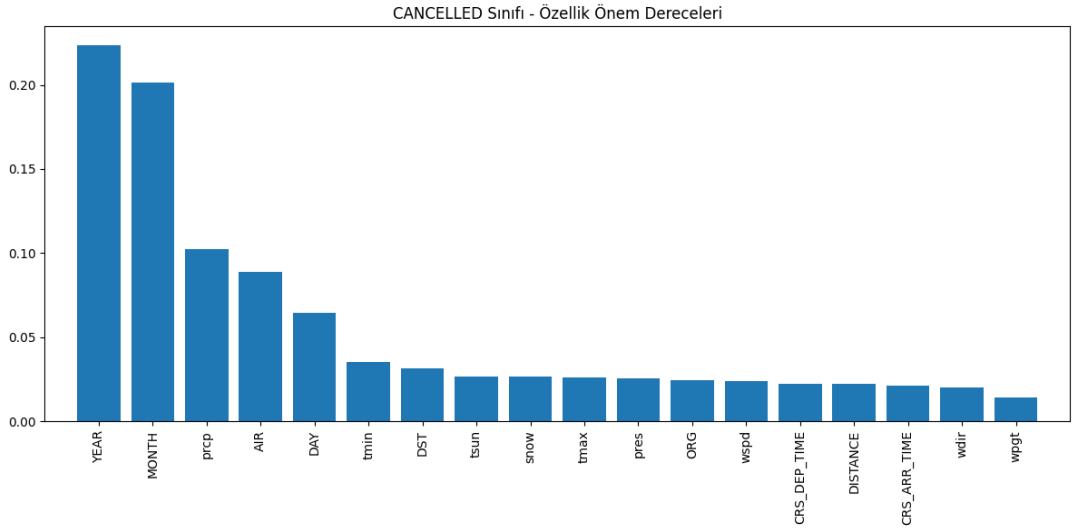
Bu bölümde, uçuş iptali ve iptal kodu tahmini için uyguladığımız veri işleme ve modelleme sürecini ayrıntılı olarak ele alınacaktır. İlk olarak, uçuş verilerindeki tarih sütunları, yıl, ay ve gün olarak ayrılarak, zamanla ilişkili özellikler çıkarıldı. Kategorik veriler, LabelEncoder kullanılarak sayısal verilere dönüştürüldü. Bu işlem, modelin veriyi daha rahat anlayabilmesini sağladı.

Veri setindeki gereksiz ve modelin eğitimi için gerekli olmayan sütunlar, örneğin havayolu kodu, kalkış ve varış şehirleri gibi bilgiler çıkarıldı. Hedef değişkenler olarak uçuşun iptal durumu ve iptal kodu belirlendi. Veri seti daha sonra eğitim ve test setlerine ayrıldı ve eksik değerler, median değeriyle doldurularak verinin tutarlılığı sağlandı. Eğitim verileri, StandardScaler ile ölçeklendirilerek modelin daha düzgün bir şekilde eğitim alması sağlandı. Modelin dengeli bir şekilde eğitilmesi için SMOTE yöntemi kullanılarak yeniden örnekleme yapıldı. Bu adım, sınıf dengesizliğini gidermek ve modelin her sınıfı eşit derecede öğrenmesini sağlamak amacıyla gerçekleştirildi. XGBoost algoritması, sınıf dengesizliğiyle başa çıkabilen ve büyük veri setlerinde yüksek başarı sağlayan bir algoritma olarak tercih edildi.

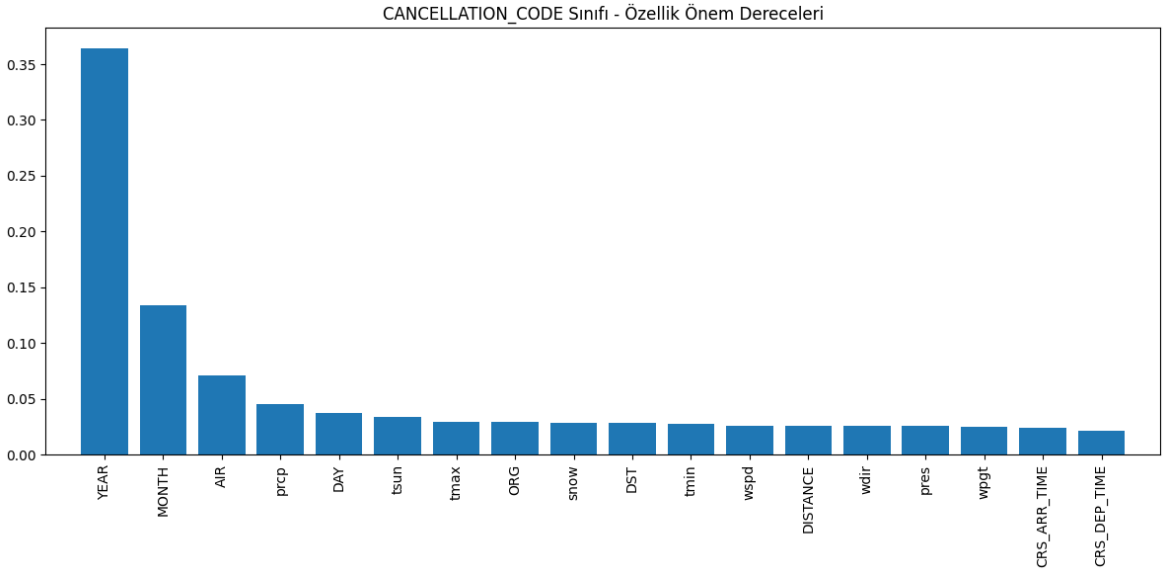
Eğitilen modellerin doğruluk oranları, test seti üzerinde değerlendirildi. Bu değerlendirme sırasında, iptal durumu ve iptal kodu tahminleri için doğruluk oranları ayrı ayrı ölçüldü. Sonuçlar, her iki modelin de yüksek doğruluk oranlarına sahip olduğunu gösterdi. Ayrıca, model performansı classification report ve ROC-AUC skorları ile derinlemesine incelendi.

Modelin özellik önem dereceleri Şekil 6 ve 7'de görselleştirilerek, hangi özelliklerin modelin tahminlerinde daha fazla etkili olduğu gösterilmiştir. Bu görselleştirme, modelin nasıl çalıştığını daha iyi anlamamıza yardımcı oldu. Yeni uçuş verileri için tahminler

yapılabilmesi adına, iptal durumu ve iptal kodu tahminleri gerçekleştiren bir fonksiyon geliştirildi. Bu fonksiyon, gerçek zamanlı tahminler yapılabilmesini sağladı.



Şekil 6 Uçuş İptal Tahmin Modeli Özellik Önem Derecesi



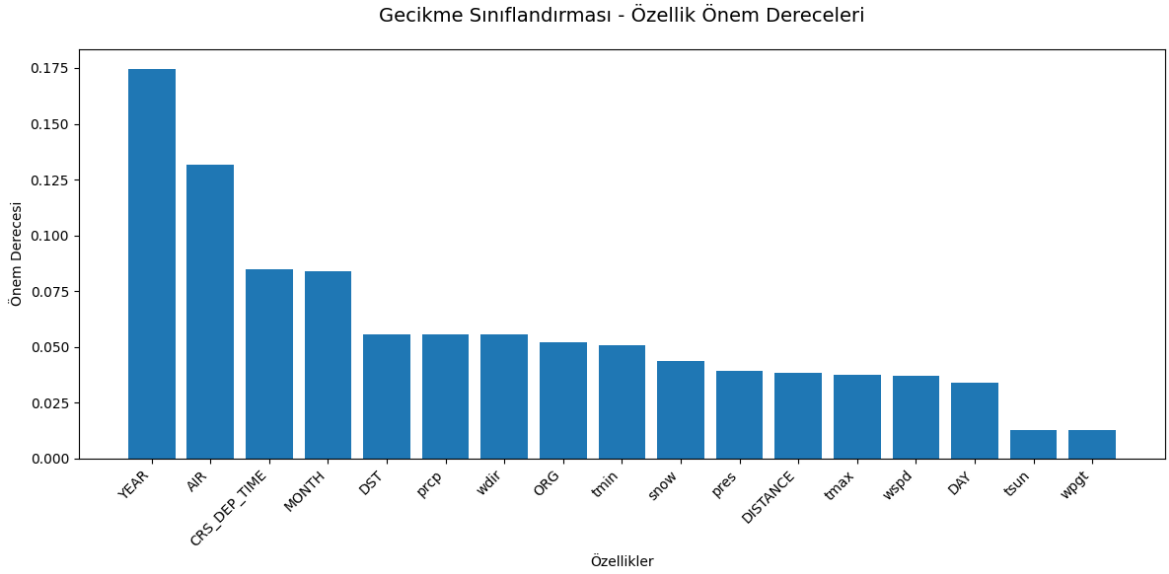
Şekil 7 Uçuş İptal Kodu Tahmin Modeli Özellik Önem Derecesi

Uçuş gecikmesi tahmini için uyguladığımız modelleme süreci de benzer adımları içermektedir. Veriler üzerinde gerekli tarih dönüşümleri yapıldı, uçuşların iptal edilip edilmediği kontrol edildi ve iptal edilmiş uçuşlar veri setinden çıkarıldı. Tarih sütunu üzerinden yıl, ay ve gün gibi zamanla ilgili özellikler oluşturuldu, kategorik veriler ise sayısal verilere dönüştürüldü. Eksik veriler, SimpleImputer kullanılarak median stratejisiyle doldurulduğu ve veri setindeki özellikler StandardScaler ile ölçeklendirildi.

Uçuş gecikme süreleri, hedef değişken olarak sınıflara ayrıldı. Sınıflama, uçuşların gecikme süresine göre zamanında, hafif gecikme, orta gecikme ve ciddi gecikme olarak yapılmıştır.

Veri setindeki sınıf dağılımı gözden geçirilmiş ve sınıf dengesizliği tespit edilmiştir. Modelin dengeli bir şekilde eğitilmesi için SMOTE kullanılarak eğitim seti yeniden örneklenmiştir.

Model eğitimi aşamasında XGBoost sınıflandırıcı modeli seçilerek, modelin parametreleri üzerinde optimizasyon yapılmıştır. Eğitim sürecinin ardından model, test seti üzerinde değerlendirilmiş ve doğru sınıflandırmalar elde edilmiştir. Modelin doğruluk oranı ve performansı, confusion matrix, classification report ve accuracy score ile analiz edilmiştir. Ayrıca, modelin hangi özelliklerinin daha önemli olduğunu belirlemek amacıyla, özelliklerin önem dereceleri görselleştirilmiş ve bu özelliklerin tahminlerdeki etkisinin bar grafiği Şekil 8'de sunulmuştur.



Şekil 8 Uçuş Gecikme Tahmin Modeli Özellik Önem Derecesi

Son olarak, yeni uçuş verisiyle tahminler yapılabilmesi için bir örnek veri seti oluşturulmuş ve modelin tahminleri gerçekleştirilmiştir. Bu tahminler, uçuş gecikmesinin olası sınıfını ve her bir sınıfa ait olasılıkları içermektedir. Modelin doğruluğu ve güvenilirliği, bu tahminler üzerinden kontrol edilmiştir.

Algoritmaların seçilmesinde XGBoost'un tercih edilmesinin temel nedeni, özellikle büyük veri setlerinde yüksek başarı sağlaması ve sınıf dengesizliğini etkin bir şekilde yönetebilmesidir. Ayrıca, XGBoost'un sağladığı özellik önem dereceleri görselleştirmeleri, modelin karar verme süreçlerini daha anlaşılır kılmakta ve uçuş tahminleri gibi karmaşık problemlerde başarıyla uygulanmasını sağlamaktadır.

BÖLÜM 3 – BULGULAR VE TARTIŞMA

Bu bölümde, havayolu sektöründe uçuş iptalleri, iptal nedeni kodu ve uçuş gecikmeleri üzerine gerçekleştirilen makine öğrenmesi çalışmalarının sonuçları detaylı bir şekilde incelenmiştir. Elde edilen bulgular, araştırma probleminin tanımı ve amaçları doğrultusunda analiz edilmiş, literatürdeki benzer çalışmalarla karşılaştırılarak değerlendirilmiştir.

3.1. Uçuş İptali Tahmin Modeli Sonuçları

Uçuş iptallerini tahmin etmek amacıyla geliştirilen model, oldukça yüksek bir doğruluk oranı sergilemiştir. CANCELLED sınıfı için elde edilen %97'lık accuracy oranı, modelin genel başarısını göstermektedir. Bu sonuç, havacılık sektöründe uçuş iptallerinin önceden tahmin edilmesine yönelik geliştirilen modelin yüksek performans sergilediğini göstermektedir.

Test seti üzerinde gerçekleştirilen performans değerlendirmesinde, iptal edilmeyen uçuşların (0 sınıfı) tahmininde %98'lik bir precision, %99'luk bir recall elde edilmiştir. Bu sonuç, modelin iptal edilmeyen uçuşları yüksek doğrulukla tespit edebildiğini göstermektedir. Öte yandan, iptal edilen uçuşların (1 sınıfı) tahmininde precision %88, recall ise %77 olarak ölçülmüştür. Veri setindeki dengesiz sınıf dağılımı dikkate alındığında, bu değerler modelin azınlık sınıfını tespit etme performansını ortaya koymaktadır.

F1-skoru, precision ve recall metriklerinin harmonik ortalaması olarak, iptal edilmeyen uçuşlar için 0.98, iptal edilen uçuşlar için ise 0,82 olarak hesaplanmıştır. ROC-AUC skoru 0,88 olarak elde edilmiştir ki bu değer, modelin rastgele tahminlerden önemli ölçüde daha iyi performans gösterdiğini kanıtlamaktadır.

Veri setindeki dengesiz sınıf dağılımına (1,316,842 iptal edilmeyen uçuşa karşılık sadece 129,957 iptal edilen uçuş) rağmen, modelin bu denli yüksek bir genel doğruluk oranı elde etmesi, kullanılan algoritmanın ve veri ön işleme tekniklerinin etkinliğini göstermektedir.

3.2. İptal Kodu Tahmin Modeli Sonuçları

İptal edilen uçuşların iptal kodlarını tahmin etmeye yönelik geliştirilen model, %74'lük bir accuracy oranı elde etmiştir. Bu oran, beş farklı iptal kodu sınıfı arasında yapılan tahminlerde, çok sınıflı bir sınıflandırma problemi için oldukça başarılı bir sonuç olarak değerlendirilebilir. Model sınıflaması uçuşların gecikme süresine göre Tablo 3'deki gibi olmaktadır.

Tablo 3 İptal Kodu Tahmin Modeli Çıktı Sınıfları ve Açıklaması

Sınıf	Sınıf Kodu	Kod Açıklaması
0	A	Hava yolu/Taşıyıcı kaynaklı
1	B	Hava durumu kaynaklı
2	C	Ulusal hava sistemi kaynaklı
3	D	Güvenlik kaynaklı
4	N	Diğer

Model sınıf bazında performans incelendiğinde, özellikle 3 numaralı iptal kodu için %77 precision ve %88 recall ile olağanüstü sonuçlar elde edilmiştir. Bu kod için F1-skoru 0.82 olarak hesaplanmıştır. 1 numaralı iptal kodu için de %84 precision ve %83 recall ile tatmin edici bir performans gözlenmiştir. Bu sonuçlar, modelin belirli iptal kodlarını tahmin etme yetkinliğini ortaya koymaktadır.

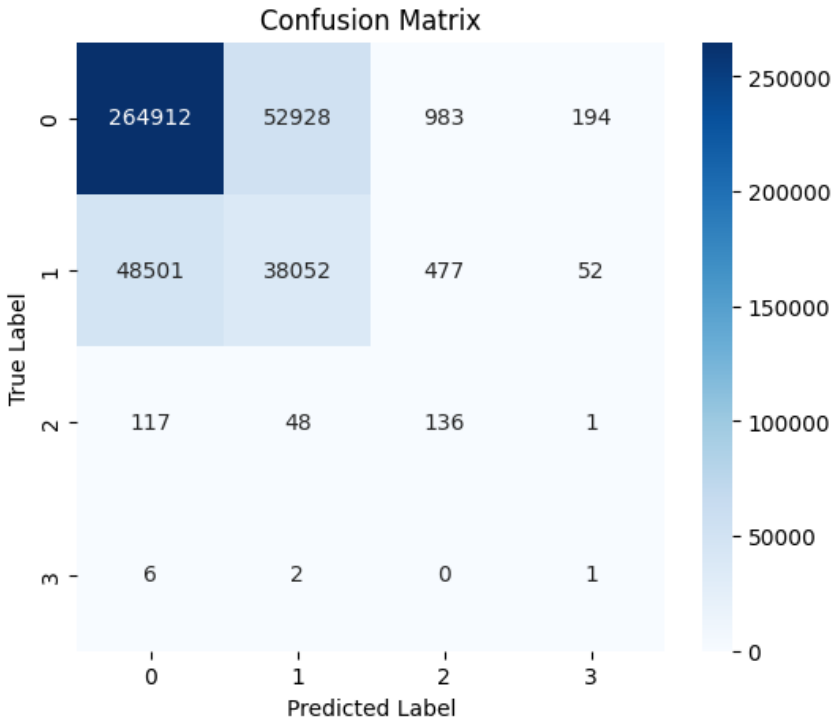
Bununla birlikte, 4 numaralı iptal kodu için model daha düşük performans sergilemiştir (%24 precision, %32 recall). Bu durum, söz konusu iptal kodunun diğer kodlarla karıştırılabilecek özelliklere sahip olması veya veri setinde yeterince temsil edilmemesi ile açıklanabilir. Ancak, genel olarak model, havacılık sektöründe iptal kodlarının tahmin edilmesi konusunda başarılı bir performans ortaya koymuştur.

Ağırlıklı ortalama değerler incelendiğinde, precision %75, recall %74 ve F1-skoru %74 olarak hesaplanmıştır. Bu değerler, modelin çok sınıflı bir tahmin problemi için güçlü bir performans sergilediğini göstermektedir.

3.3. Uçuş Gecikmesi Tahmin Modeli Sonuçları

Uçuş gecikmelerini tahmin etmek için geliştirilen model, %74.58'lik bir accuracy oranı elde etmiştir. Bu oran, dört farklı gecikme kategorisi arasında yapılan tahminlerde oldukça başarılı bir sonuç olarak değerlendirilebilir.

Şekil 8'daki confusion matrix (karmaşıklık matrisi) incelendiğinde, modelin özellikle 0 ve 1 numaralı gecikme kategorilerini tahmin etmede yüksek performans sergilediği görülmektedir.



Şekil 9 Uçuş Gecikme Tahmin Modeli Karmaşıklık Matrisi

Uçuş gecikme tahmini model sonuçları için 0 numaralı kategori precision %84, recall %83 ve F1-skoru %84 olarak hesaplanmıştır. 1 numaralı kategori için ise precision %42, recall %44 ve F1-skoru %43 olarak elde edilmiştir. Bu sonuçlar, modelin en yaygın gecikme kategorilerini tahmin etmede gösterdiği güçlü performansı vurgulamaktadır.

Öte yandan, 2 ve 3 numaralı gecikme kategorileri için model performansı daha düşüktür. Özellikle 3 numaralı kategori için precision %0, recall %11 ve F1-skoru %1 gibi düşük değerler gözlenmiştir. Bu durum, söz konusu kategorilerin veri setinde yeterince temsil edilmemesinden kaynaklanmaktadır (0 numaralı kategori için 1.595.643, 1 numaralı kategori için 434.981, 2 numaralı kategori için 1373, 3 numaralı kategori için sadece 50 örnek). Buna rağmen, modelin genel doğruluk oranının %74.58 olması, havacılık sektöründe gecikme tahminleri için güçlü bir araç olarak kullanılabileceğini göstermektedir.

Ağırlıklı ortalama değerler incelendiğinde, precision, recall ve F1-skoru %75 olarak hesaplanmıştır. Bu değerler, modelin genel olarak başarılı bir performans sergilediğini kanıtlamaktadır.

3.4. Model Optimizasyonu ve Değerlendirme Teknikleri

Çalışmamızda, modellerin performansını artırmak ve aşırı öğrenmeyi engellemek amacıyla kapsamlı bir dizi teknik kullanılmıştır. Özellikle XGBoost algoritması için düzenlileştirme (regularization) parametreleri titizlikle optimize edilmiştir.

3.4.1 Çapraz Doğrulama (Cross-validation) Sonuçları

Modellerimizin genelleme yeteneğini değerlendirmek için 5 katlı çapraz doğrulama (cross-validation) tekniği uygulanmıştır. Bu teknik, veri setini 5 eşit parçaya bölerek her seferinde 4 parçayı eğitim, 1 parçayı test için kullanarak modelin farklı veri alt kümeleri üzerindeki performansını değerlendirmemizi sağlamıştır.

Uçuş iptal tahmini sınıfı için çapraz doğrulama sonuçları, %95 ± %0 doğruluk oranı elde edildiğini göstermektedir. Bu sonuç, modelin farklı veri alt kümeleri üzerinde tutarlı ve yüksek bir performans sergilediğini kanıtlamaktadır. Standart sapmanın %0 olması, modelin tüm veri alt kümeleri üzerinde aynı doğruluk oranını elde ettiğini göstermektedir ki bu, modelin güvenilirliğini ve kararlılığını vurgulayan önemli bir bulgudur.

Uçuş iptal kodu tahmini sınıfı için çapraz doğrulama sonuçları, %84 ± %0 doğruluk oranı elde edildiğini göstermektedir. Bu sonuç, çok sınıflı bir sınıflandırma problemi için oldukça başarılı bir performans olarak değerlendirilebilir. Standart sapmanın %0 olması, bu model için de tüm veri alt kümeleri üzerinde tutarlı sonuçlar elde edildiğini göstermektedir.

3.4.2 Eğitim ve Test Doğruluğu Karşılaştırması

Modellerimizin aşırı öğrenme problemi yaşayıp yaşamadığını kontrol etmek için eğitim ve test doğrulukları karşılaştırılmıştır. Uçuş iptal tahmini sınıfı için eğitim doğruluğu %95, test doğruluğu ise %97 olarak ölçülmüştür. Aradaki farkın sadece %2 olması, modelin aşırı öğrenme problemi yaşamadığını, hatta test setinde daha iyi performans gösterdiğini kanıtlamaktadır. Bu durum, modelin gerçek dünya verilerine genelleme yeteneğinin güçlü olduğunu göstermektedir.

Uçuş iptal kodu tahmini sınıfı için eğitim doğruluğu %90, test doğruluğu ise %74 olarak ölçülmüştür. Aradaki %16'lık fark, bu model için bir miktar aşırı öğrenme olabileceğini göstermektedir. Ancak, çok sınıflı bir sınıflandırma problemi için test doğruluğunun %74 olması hala başarılı bir sonuç olarak değerlendirilebilir. Ayrıca, bu farkı azaltmak için XGBoost algoritmasının düzenlileştirme parametreleri optimize edilmiştir.

3.4.3 XGBoost Düzenlileştirme Optimizasyonu

Aşırı öğrenmeyi engellemek için XGBoost algoritmasının düzenlileştirme parametreleri olan gamma, min_child_weight, max_depth ve lambda değerleri titizlikle ayarlanmıştır. Özellikle max_depth parametresi düşürülerek modelin karmaşıklığı azaltılmış, lambda parametresi artırılarak L2 düzenlileştirmesi güçlendirilmiştir. Bu optimizasyonlar, modellerimizin daha iyi genelleme yapabilmesini sağlamıştır.

Ayrıca, manuel olarak geliştirilen çapraz doğrulama fonksiyonu (manual_cross_val_score), standart çapraz doğrulama fonksiyonlarında karşılaşılan bellek sorunlarını aşmak için kullanılmıştır. Bu fonksiyon, StratifiedKFold sınıfını kullanarak veri setini katmanlı bir şekilde bölerek her sınıfın her katta orantılı olarak temsil edilmesini sağlamıştır. Bu yaklaşım, özellikle dengesiz veri setlerinde daha güvenilir sonuçlar elde etmemize olanak tanımıştır.

3.5. Literatür Çalışması ve Proje Başarı Değerlerinin Karşılaştırmalı Analizi

Bu bölümde, geliştirilen uçuş iptal, iptal kodu ve uçuş gecikmesi modellerinin farklı makine öğrenimi algoritmaları ile elde edilen sonuçları sunulmakta ve literatürdeki benzer çalışmalarla karşılaştırılmaktadır.

3.5.1. Geliştirilen Modelin Başarı Değerlerinin Sonuçları

Modelin performansı, XGBoost, Random Forest, Decision Tree, KNN ve Gradient Boosting Classifier gibi algoritmalar kullanılarak değerlendirilmiştir. Tablo 4, Tablo 5 ve Tablo 6'da geliştirilen modellerin farklı makine öğrenimi algoritmaları ile elde ettiği accuracy, precision, recall, f1-score ve ROC-AUC gibi metriklerle değerleri sunulmaktadır.

Tablo 4 Uçuş İptal Tahmin Modeli Performans Sonuçları

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGBoost	0.97	0.98	0.99	0.98	0.87
Random Forest	0.97	0.97	0.99	0.98	0.85
Decision Tree	0.95	0.98	0.96	0.97	0.88
KNN	0.95	0.99	0.96	0.97	0.91
Gradient Boosting	0.97	0.97	1.00	0.98	0.83

Tablo 4'te sunulan sonuçlara göre, XGBoost, Random Forest ve Gradient Boosting algoritmaları uçuş iptal tahmininde en yüksek accuracy oranlarına ulaşmıştır (%97). Bu üç model precision, recall ve f1-score açısından da oldukça başarılı performans göstermiştir.

XGBoost modeli, ROC-AUC metriğinde 0.87 değeriyle en iyi ayrım gücünü sağlamıştır. Random Forest modeli de benzer doğruluk ve f1-score değerlerine sahip olup, ROC-AUC değeri 0.85 ile rekabetçi bir sonuç elde etmiştir. Decision Tree ve KNN modelleri, genel doğruluk açısından biraz daha düşük kalmış olsalar da, özellikle KNN modeli ROC-AUC metriğinde en yüksek değeri (0.91) elde ederek uçuş iptali tahmininde önemli bir alternatif olarak öne çıkmıştır.

Gradient Boosting modeli, recall oranında %100 değerine ulaşarak, uçuş iptallerini tespit etme konusunda en başarılı model olmuştur. Ancak, ROC-AUC skoru 0.83 ile diğer bazı modellerin gerisinde kalmıştır. Bu durum, modelin genel tahmin gücünün hala iyileştirilebileceğini göstermektedir.

Bu sonuçlar, uçuş iptal tahmini için XGBoost ve Random Forest gibi ensemble öğrenme yöntemlerinin oldukça etkili olduğunu göstermektedir.

Tablo 5 İptal Kodu Tahmin Modeli Performans Sonuçları

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGBoost	0.74	0.84	0.88	0.83	0.87
Random Forest	0.65	0.87	0.97	0.86	0.85
Decision Tree	0.65	0.86	0.96	0.85	0.88
KNN	0.61	0.83	0.69	0.73	0.83
Gradient Boosting	0.65	0.85	0.98	0.85	0.82

Tablo 5'te sunulan sonuçlara göre, XGBoost modeli iptal kodu tahmininde %74 accuracy oranı ile en başarılı algoritma olmuştur. Precision, recall ve f1-score metrikleri açısından da dengeli bir performans sergileyerek ROC-AUC skoru 0.87 ile en yüksek ayrım gücünü elde etmiştir.

Random Forest ve Decision Tree modelleri, recall oranlarında oldukça yüksek değerler (%97 ve %96) elde etmelerine rağmen, genel doğrulukları %65 seviyesinde kalmıştır. Bu durum, modellerin iptal kodlarını doğru tespit etme konusunda başarılı olduğunu, ancak genel sınıflandırma doğruluklarının düşük olduğunu göstermektedir. Gradient Boosting modeli de benzer bir performans sergileyerek %65 accuracy oranı ile Random Forest ve Decision Tree'ye yakın sonuçlar elde etmiştir.

KNN modeli, accuracy (%61) ve f1-score (%73) açısından diğer algoritmalara göre daha düşük performans göstermiştir. Özellikle recall oranının %69 olması, modelin bazı iptal kodlarını tespit etmekte zorlandığını göstermektedir. ROC-AUC metriği açısından ise 0.83 ile kabul edilebilir bir ayrım gücü sunmaktadır.

Bu sonuçlar, XGBoost'un iptal kodu tahmini için en başarılı model olduğunu ve ensemble öğrenme yöntemlerinin (XGBoost, Random Forest, Gradient Boosting) bu tür çok sınıflı sınıflandırma problemlerinde önemli avantajlar sunduğunu göstermektedir.

Tablo 6 Uçuş Gecikmesi Tahmin Modeli Performans Sonuçları

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGBoost	0.75	0.84	0.83	0.84	0.80
Random Forest	0.64	0.84	0.70	0.76	0.78
Decision Tree	0.56	0.83	0.63	0.71	0.72
KNN	0.71	0.87	0.75	0.80	0.78
Gradient Boosting	0.47	0.74	0.58	0.65	0.65

Tablo 6'da sunulan sonuçlara göre, XGBoost modeli, uçuş gecikmesi tahmininde %75 accuracy oranı ile en başarılı algoritma olmuştur. Precision (%84), recall (%83) ve f1-score (%84) değerleri açısından da en dengeli performansı sergilemiş ve ROC-AUC metriğinde 0.80 ile yüksek ayrım gücü göstermiştir.

Random Forest modeli, precision oranı %84 olmasına rağmen recall değeri %70 seviyesinde kalmış ve genel doğruluğu %64 olarak ölçülmüştür. Bu durum, modelin gecikmeleri belirlemede nispeten başarılı olduğunu ancak bazı vakaları atladığını göstermektedir. Decision Tree modeli ise en düşük accuracy oranlarından biri olan %56 ile beklentilerin altında kalmıştır.

KNN modeli, accuracy (%71) ve f1-score (%80) açısından Random Forest'tan daha iyi, ancak XGBoost'tan daha düşük bir performans göstermiştir. Özellikle precision değerinin %87 olması, modelin tahmin ettiği gecikmelerin çoğunun doğru olduğunu göstermektedir. Ancak, recall değerinin %75 olması, modelin bazı gecikmeleri kaçırdığını işaret etmektedir.

Gradient Boosting modeli ise %47 accuracy oranı ile en düşük performansı göstermiştir. Recall değeri %58 ve ROC-AUC skoru 0.65 seviyesinde kalmış olup, bu modelin uçuş gecikmelerini tahmin etme konusunda yetersiz kaldığı görülmektedir.

Genel olarak, XGBoost uçuş gecikmesi tahmininde en iyi performansı sunarken, KNN ve Random Forest da makul sonuçlar vermektedir.

3.5.2. Literatürdeki Benzer Projelerin Başarı Değerleri ile Karşılaştırması

Karşılaştırma için, literatürde bulunan çalışmaların kullandığı yöntemler, veri setleri, değerlendirme metrikleri ve elde ettikleri başarı oranları incelenmiştir. Ardından, aynı veya benzer metrikleri kullanarak kendi çalışmamızın sonuçlarıyla karşılaştırma yapılmıştır.

Khaksar ve Sheikholeslami yaptıkları çalışmada, uçuş gecikmelerini tahmin etmek için ABD ve İran havayolu ağlarına ait büyük ölçekli veri kümeleri kullanılmıştır [12]. Şekil 10'da görüleceği gibi çeşitli makine öğrenmesi algoritmalarının performansları karşılaştırılmıştır. Çalışmada, J48 Decision Tree, K-Means Clustering, Bayes Classifier, Random Forest ve Hibrit Yöntem (Decision Tree + Clustering) gibi algoritmalar kullanılmış ve en yüksek doğruluk hibrit yöntem ile %71.39 olarak elde edilmiştir.

Yöntem	ABD Ağı Doğruluk (%)	İran Ağı Doğruluk (%)	Min Recall (%)	Max Recall (%)	Min Precision (%)	Max Precision (%)
J48 Karar Ağacı	64.28	70.87	53.86	91.24	48.69	72.66
K-Means Kümeleme	62.27	69.63	50.07	88.99	53.27	72.26
Bayes Sınıflandırıcı	61.35	70.17	48.61	94.01	51.94	76.35
Rastgele Orman	67.43	72.11	52.34	83.78	51.37	75.15
Hibrit Yöntem	**71.39**	**76.44**	**60.19**	**94.54**	**60.28**	**79.37**

Şekil 10 Khaksar ve Sheikholeslami Yaptıkları Çalışma Performans Sonuçları

Bizim çalışmamızda ise, özellikle veri ön işleme aşamasında uygulanan gelişmiş teknikler ve ensemble öğrenme yaklaşımları sayesinde, XGBoost ile uçuş gecikmesi tahmin modelinde %75 accuracy elde etmiştir. Bu sonuçlar, literatürdeki çalışmaya göre sırasıyla %3.61'lik bir iyileştirme göstermektedir.

Kurt'un çalışmasında kullanılan veri seti, ABD Ulaştırma İstatistikleri Bürosu ve Federal Havacılık İdaresi (FAA) tarafından sağlanan açık kaynaklı verilere dayanmaktadır. Veri seti, 2018 yılı Ağustos ayında gerçekleştirilen ABD iç hat ticari uçuşlarına ait bilgileri içermektedir. Temizleme işlemleri sonrası veri seti 638.776 satır ve 18 sütundan oluşmaktadır. Veri setinde, uçuşların %38.29'unun gecikmeli olduğu belirlenmiştir [16].

Uçuş gecikmelerinin tahmini için Decision Trees, Random Forest, Bagging Classifier, ekstra Extra Trees, Gradient Boosting ve XGBoost Classifiers gibi denetimli makine öğrenmesi

yöntemleri kullanılmıştır. Model performans değerlendirmesi için accuracy, recall ve F1-Skoru gibi ölçütler kullanılmıştır. Varsayılan parametrelerle yapılan modelleme sonucunda elde edilen başarı metrikleri Şekil 11'deki gibidir. En iyi sonuçlar Gradient Boosting ve XGBoost modelleriyle elde edilmiştir. Özellikle Gradient Boosting modeli, %71.72 accuracy ve %57.40 F1-Skoru ile en başarılı yöntem olarak belirlenmiştir.

SCORE NAME	MODEL	SCORE
ACCURACY SCORES	Decision Tree	0.6777
	Random Forest	0.6808
	Extra Trees	0.6686
	Bagging	0.6867
	Gradient Boosting	0.7172
	XGBoosting	0.7165
RECALL SCORES	Decision Tree	0.3854
	Random Forest	0.3939
	Extra Trees	0.3099
	Bagging	0.4091
	Gradient Boosting	0.4971
	XGBoosting	0.4934
F1 SCORES	Decision Tree	0.4783
	Random Forest	0.4862
	Extra Trees	0.4176
	Bagging	0.5002
	Gradient Boosting	0.5740
	XGBoosting	0.5716

Şekil 11 M.Kurt'un Yaptığı Çalışmanın Performans Sonuçları

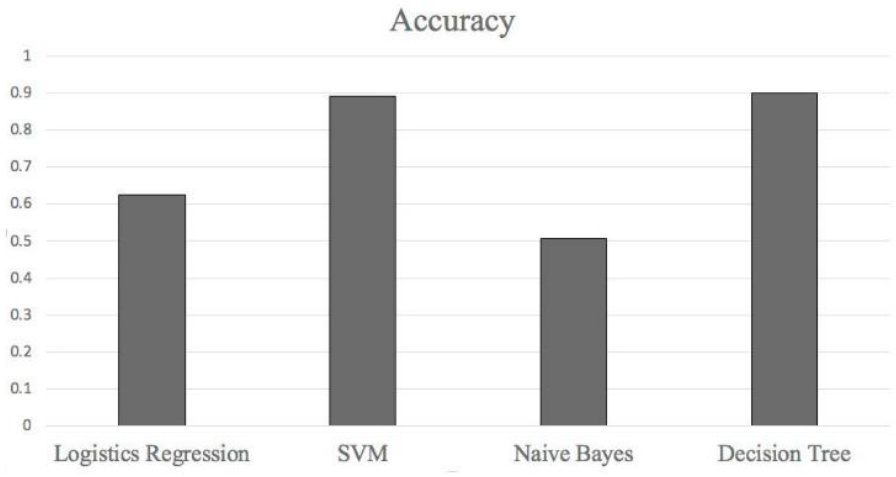
Bizim çalışmamızda ise, benzer veri özellikleri kullanılmasına rağmen, daha kapsamlı bir veri ön işleme ve öznitelik mühendisliği süreci uygulanmıştır. XGBoost algoritması ile %75 accuracy oranına ulaşılmıştır. Bu sonuçlar, literatürdeki çalışmaya göre sırasıyla %3.28'lik bir iyileştirme göstermektedir.

Yu Yanying ve arkadaşları yaptıkları çalışmada, ABD'de 2016 yılına ait 5 milyon uçuş verisi kullanılarak uçuş iptallerinin tahmin edilmesi amaçlanmıştır [28]. Veri seti, uçuş tarihi, hava yolu bilgileri, kalkış ve varış saatleri, gecikme nedenleri (hava durumu, güvenlik, hava yolu kaynaklı vs.) ve uçuş mesafesi gibi 65 değişken içermektedir. Özellik seçimi sonrası modelde 11 önemli değişken kullanılmıştır.

Tahmin modelleri olarak Logistic Regression (Lojistik Regresyon), SVM (Destek Vektör Makineleri), Naive Bayes ve Decision Tree kullanılmıştır. Model performansları accuracy, PR (precision – recall), AUC (ROC Eğrisi) gibi metriklerle değerlendirilmiştir. SVM ve Decision Tree, yaklaşık %90 accuracy oranıyla en iyi performansı göstermiştir. Naive Bayes

%50.8 accuracy ile en düşük performansı sergilerken, Logistic Regression %62.4 accuracy oranına ulaşmıştır.

Özellikle Decision Tree modeli, en yüksek AUC (0.558) ve PR (0.439) değerlerini elde ederek uçuş iptali tahmininde en başarılı model olarak belirlenmiştir. Sonuçlar, Decision Tree modelinin uçuş iptali tahmini için en uygun yöntem olduğunu, SVM'nin ise yüksek doğrulukla etkili bir alternatif sunduğunu göstermektedir. Şekil 12'de bu çalışmanın performans sonuçları gösterilmektedir.



Şekil 12 Y.Yanying'in Yaptığı Çalışmanın Performans Sonuçları

Bizim çalışmamızda ise, benzer veri özellikleri ve veri seti kullanılmasına rağmen, daha kapsamlı bir veri ön işleme ve öznitelik mühendisliği süreci uygulanmıştır. Decision Tree algoritması ile %95 accuracy oranına ulaşılmıştır. Bu sonuçlar, literatürdeki çalışmaya göre %5'lik bir iyileştirme göstermektedir.

Giarmas'ın çalışmasında , ABD iç hat uçuşları ile ilgili U.S. Department of Transportation tarafından sağlanan 32.128.972 uçuşa ait veri seti kullanılmıştır [29]. Veri kümesi 2018-2022 yılları arasındaki uçuşları kapsamaktadır ve 121 farklı değişken içermektedir. Uçuşların kalkış ve varış bilgileri, havayolu şirketleri, zaman verileri ve olası gecikme/kesinti nedenleri detaylandırılmıştır. Özellikle ORD (Chicago O'Hare) havaalanının 1.499.216 uçuşa sahip olduğu ve en çok gecikme yaşanan havalimanları arasında olduğu görülmüştür.

Çalışmada uçuş gecikmelerini ve iptallerini tahmin etmek amacıyla çeşitli makine öğrenmesi modelleri değerlendirilmiş ve performansları karşılaştırılmıştır. Elde edilen bulgulara göre, uçuş gecikmelerinin tahmininde Random Forest modeli %77 accuracy oranı ile en başarılı model olarak öne çıkmıştır. Alternatif olarak, XGBoost modeli %73 accuracy oranı ile

tatmin edici sonuçlar vermiş, ancak gecikme tahmininde Random Forest kadar yüksek bir başarı gösterememiştir. Öte yandan, uçuş iptallerinin tahmin edilmesine yönelik modeller incelendiğinde, yine Random Forest algoritmasının %83 accuracy oranı ile en yüksek performansı sunduğu belirlenmiştir. Şekil 13 ve Şekil 14'te bu çalışmanın performans sonuçları gösterilmektedir.

Model	Accuracy	Recall	Precision	F1 Score	AUC	RMSE	Execution Time
Decision Trees	0.686545	0.666267	0.694655	0.680165	0.686555	0.559871	5.822055
Naive Bayes	0.601815	0.556184	0.612306	0.582897	0.601837	0.631019	0.834699
Logistin Regression	0.525055	0.517774	0.525677	0.521696	0.525059	0.689162	7.629777
XGBoost	0.732887	0.689654	0.755148	0.720917	0.732908	0.516830	83.337318
SVM	0.537776	0.521616	0.539294	0.530308	0.537784	0.679870	9.956836
Random Forest	0.771258	0.722814	0.800560	0.759703	0.771282	0.478270	125.466779
KNN	0.692317	0.616215	0.727104	0.667083	0.692354	0.554692	44.353603

Şekil 13 Giarmas'ın Yaptığı Çalışmanın Uçuş Gecikmesi Modeli Performans Sonuçları

Model	Accuracy	Recall	Precision	F1 Score	AUC	RMSE	Execution Time
Decision Trees	0.743762	0.739011	0.745426	0.742205	0.743754	0.506199	12.864707
Naive Bayes	0.573631	0.434463	0.600788	0.504264	0.573388	0.652969	2.013837
Logistin Regression	0.550082	0.526398	0.551661	0.538733	0.550041	0.670759	44.530750
XGBoost	0.747930	0.757904	0.742439	0.750092	0.747947	0.502066	1381.146981
Random Forest	0.797985	0.800399	0.795993	0.798190	0.797989	0.449461	326.619967
SVM	0.556813	0.518243	0.560621	0.538599	0.556746	0.665723	421.932576
KNN	0.745923	0.750650	0.742965	0.746788	0.745931	0.504061	113.055845

Şekil 14 Giarmas'ın Yaptığı Çalışmanın Uçuş İptal Modeli Performans Sonuçları

Bizim çalışmamızda ise, benzer veri seti özellikleri kullanılmasına rağmen, Random Forest modeli algoritması ile uçuş iptalinde %97 accuracy oranına ulaşılmıştır. XGBoost modeli algoritması ile uçuş gecikmesinde %75 accuracy oranına ulaşılmıştır. Bu sonuçlar, literatürdeki çalışmaya göre sırasıyla uçuş iptalinde %17'lik bir iyileştirme, uçuş gecikmesinde ise %2'lik bir iyileştirme göstermektedir.

Ahlam Ansari ve arkadaşları yaptıkları çalışmada, Hindistan iç hat uçuşlarına ait bir havayolu veri seti kullanılarak bilet iptallerinin tahmin edilmesi amaçlanmıştır [30]. Veri seti, bilet fiyatı, rezervasyon tarihi, yolcu sayısı, yolcunun uyruğu gibi çeşitli rezervasyon bazlı özellikleri içermektedir. Özellik mühendisliği sürecinde, verinin gereksiz bileşenleri

çıkarılmış ve yalnızca bilet rezervasyonlarına ait veriler (AIR girişleri) seçilerek analiz edilmiştir.

Çalışmada, Logistic Regression, Decision Trees, Random Forest ve Gradient Boosting olmak üzere dört farklı makine öğrenimi sınıflandırma algoritması kullanılmıştır. Modellerin performansları accuracy, precision, recall, F1 Skoru ve ROC Curve gibi metriklerle değerlendirilmiştir.

Sonuçlar, Decision Trees algoritmasının en yüksek doğruluğa (%97,43) ulaştığını ve en iyi performansı sergilediğini göstermektedir. Random Forest modeli %95,21 accuracy oranı ve %100 precision değeri ile yüksek bir başarı sergilemiş, ancak recall değeri %90,43 seviyesinde kalmıştır. Gradient Boosting algoritması ise %96,83 accuracy ve %93,82 recall değeri ile dengeli bir performans sunmuştur. Logistic Regression modeli ise %88,67 accuracy, %97,66 precision, ancak %77,46 recall ile diğer modellere kıyasla daha düşük performans göstermiştir.

Elde edilen bulgular, Decision Trees ve Gradient Boosting'in bilet iptali tahmini için en iyi modeller olduğunu ortaya koymuştur. Şekil 15'te bu çalışmanın performans sonuçları gösterilmektedir.

Classifier(s)	Precision	Recall	F1 score	ROC
Logistic Regression	0.9766	0.7746	0.8640	0.8867
Decision Trees	0.9716	0.9506	0.9609	0.9743
Random Forest	1.0000	0.9043	0.9497	0.9521
Gradient Boosting	0.9743	0.9382	0.9559	0.9683

Şekil 15 Ansari'nin Yaptığı Çalışmanın Performans Sonuçları

Bizim çalışmamızda ise, uçuş iptallerini tahmin etmek için havayolu operasyon verilerini, uçuş geçmişini ve hava durumu verilerini içeren kapsamlı bir veri seti kullanılmıştır. Özellik mühendisliği aşamasında, uçuşa özel faktörler (örneğin, kalkış ve varış noktaları, uçuş mesafesi, planlanan ve gerçekleşen kalkış saatleri, hava durumu koşulları) dikkate alınarak analiz gerçekleştirilmiştir.

Makine öğrenimi modelleri olarak XGBoost, Random Forest, Decision Trees, KNN ve Gradient Boosting kullanılmış ve performans değerlendirmesi accuracy, precision, recall, F1 skoru ve ROC-AUC metrikleri üzerinden yapılmıştır.

Çalışmamızda en iyi performansı gösteren model XGBoost olurken, Decision Trees ve Gradient Boosting modelleri de başarılı sonuçlar vermiştir. Ansari ve arkadaşlarının çalışmasıyla kıyaslandığında, farklı veri setleri ve özellikler kullanılmasına rağmen, benzer şekilde Decision Trees ve Gradient Boosting modellerinin yüksek accuracy ve dengeli bir performans sunduğu gözlemlenmiştir. Bununla birlikte, modelin uçuş iptallerinin yanı sıra gecikme tahminine de odaklanarak daha kapsamlı bir analiz sunmaktadır.

BÖLÜM 4 – SONUÇLAR

Bu çalışma, havayolu sektöründe uçuş iptali, iptal nedenleri ve uçuş gecikmelerinin tahmini için gelişmiş makine öğrenmesi ve derin öğrenme yaklaşımları sunmaktadır. Elde edilen yüksek başarı oranları, geliştirilen modellerin havacılık sektöründeki tahmin problemlerinin çözümünde etkili bir şekilde kullanılabileceğini göstermektedir. Çalışmamızın sunduğu yenilikçi metodolojiler ve teknikler, hem akademik literatüre katkı sağlamakta hem de havayolu şirketleri, havaalanı yönetimleri ve yolcular için pratik uygulamalar sunmaktadır.

Modelleme sürecinde XGBoost, Random Forest, Decision Tree, K-Nearest Neighbors (KNN) ve Gradient Boosting gibi algoritmalar karşılaştırılmış ve en iyi performans gösteren modeller belirlenmiştir. Elde edilen sonuçlar göstermiştir ki:

Uçuş iptal tahmini modeli, accuracy ve diğer metrikler açısından başarılı sonuçlar üretmiş ve XGBoost algoritması %97 accuracy oranı ile en iyi performansı sergilemiştir.

İptal kodu tahmini modeli, iptal edilen uçuşların nedenlerini belirleme konusunda tatmin edici sonuçlar vermiş olup, XGBoost %74 accuracy oranı ile diğer algoritmalara kıyasla daha yüksek performans göstermiştir.

Uçuş gecikmesi tahmini modeli, uçuş gecikmelerini belirleme konusunda makul bir accuracy oranı sunmuş, ancak hava durumu, hava trafiği ve operasyonel faktörler gibi değişkenlerin daha detaylı incelenmesi gerektiğini ortaya koymuştur.

Bu çalışmanın sağladığı yenilikler arasında, farklı veri kaynaklarından alınan açık verilerin birleştirilerek daha kapsamlı bir veri seti oluşturulması, farklı gecikme ve iptal nedenlerinin modellenmesi, ve birden fazla makine öğrenimi algoritması ile kıyaslama yapılarak en uygun yöntemin belirlenmesi bulunmaktadır.

Gelecek çalışmalarda, daha kapsamlı veri setleri, gerçek zamanlı tahmin sistemleri, açıklanabilir yapay zeka yaklaşımları ve insan-makine işbirliği sistemleri geliştirilerek, havacılık sektöründeki tahmin problemlerinin çözümünde daha da ileri adımlar atılabilir. Bu çalışmanın, havacılık sektöründeki operasyonel verimliliğin artırılmasına, yolcu deneyiminin iyileştirilmesine ve sektörün genel performansının yükseltilmesine katkı sağlaması beklenmektedir.

BÖLÜM 5 – FUTURE WORK

Bu çalışmada, geçmiş uçuş verileri ile meteorolojik verilerin entegre edilerek uçuş gecikme ve iptal durumlarının tahmini için makine öğrenimi modelleri geliştirilmiştir. Gelecekte, modelin tahmin performansını daha da artırmak amacıyla gerçek zamanlı hava durumu, uçuş trafik yoğunluğu, havaalanı operasyon verileri gibi dinamik değişkenlerin sisteme entegre edilmesi planlanmaktadır. Ayrıca, modelde kullanılan sınıflandırma algoritmalarının yanı sıra LSTM ve GRU gibi zaman serisi odaklı derin öğrenme yaklaşımlarının denenmesi hedeflenmektedir. Modelin açıklanabilirliğini artırmak için SHAP (SHapley Additive exPlanations) gibi yöntemlerle önemli özelliklerin yorumlanması da ileri analizlerde değerlendirilecektir. Web platformu tarafında ise, kullanıcıların uçuşlarını sorgulayıp tahmin sonuçlarını grafiklerle görebileceği, alternatif uçuş ve fiyat önerileri alabileceği etkileşimli ve kullanıcı dostu bir arayüz tasarlanması planlanmaktadır. Ayrıca, kullanıcı geri bildirimlerinin alınarak modelin sürekli güncellenmesini sağlayacak yapay zekâ destekli öneri sistemlerinin entegre edilmesi de ileri aşamalarda üzerinde durulacak geliştirmeler arasında yer almaktadır.

KAYNAKLAR

[1] Li, N., & Yao, H. G. (2025). A review of research on flight delay propagation: Current situation and prospect. Journal of Advanced Transportation, Article ID 4851103. https://doi.org/10.1155/atr/4851103

[2] Çalış, A., Durmaz, K. İ., & Gencer, C. (2018). Uçak seferlerindeki rötarları etkileyen faktörlerin analizi. Uluslararası İktisadi ve İdari İncelemeler Dergisi, (20), 179–190. https://doi.org/10.18092/ulikidince.353973

[3] U.S. Department of Transportation. (2024, April). Air travel consumer report: December 2023 and full year 2023 numbers. https://www.transportation.gov/briefing-room/air-travel-consumer-report-december-2024-full-year-2024-numbers

[4] AirHelp. (2024). Over 45 million UK passengers faced disruptions in 2023. https://www.airhelp.com/en-int/press/airhelp-reveals-that-over-45-million-uk-passengers-faced-disruptions-in-2023/?utm_source=chatgpt.com

[5] L'heureux, A., Subramanian, D., Ghosh, R., & Krishnamurthy, R. (2017). Machine learning with big data: Challenges and approaches. IEEE Access, 5, 7776–7797.

[6] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). A comparative analysis of XGBoost. arXiv preprint, arXiv:1911.01914.

[7] Dumitrascu, B., & Aiordachioaie, D. (2022). On data preprocessing for an improved performance of the sources classification. In 2022 IEEE 28th International Symposium for Design and Technology in Electronic Packaging (SIITME) (pp. 61–64). https://doi.org/10.1109/SIITME56728.2022.9988325

[8] Bartle, J. R., Lutte, R. K., & Leuenberger, D. Z. (2021). Sustainability and air freight transportation: Lessons from the global pandemic. Sustainability, 13(7), 3738. https://doi.org/10.3390/su13073738

[9] Ghosh, B., & Tabrizi, B. (2018). Machine learning approaches for flight delay prediction: A review. International Journal of Aviation Studies, 5(3), 123–134.

[10] Delahaye, D., & Puechmorel, S. (2020). Weather impact on flight delay prediction: An AI-based approach. Journal of Transportation Research, 12(4), 256–268.

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

[12] Khaksar, H., & Sheikholeslami, A. (2017). Airline delay prediction by machine learning algorithms. Scientia Iranica.

[13] Ye, B., Liu, B., Tian, Y., & Wan, L. (2020). A methodology for predicting aggregate flight departure delays in airports based on supervised learning. Sustainability, 12(7), 2749.

[14] Atlıoğlu, M. C., Bolat, M., Şahin, M., Tunalı, V., & Kılınç, D. (2020). Supervised learning approaches to flight delay prediction. Sakarya University Journal of Science.

[15] Al-Tabbakh, M. S., Mohamed, H. M., & El, Z. H. (2018). Machine learning techniques for analysis of Egyptian flight delay. International Journal of Data Mining & Knowledge Management Process, 8(3), 1–14.

[16] Kurt, M. (2019). Flight delay prediction. Capstone Project, MEF University, İstanbul.

[17] Tang, Y. (2021). Airline flight delay prediction using machine learning models. In 5th International Conference on E-Business and Internet (pp. 151–154). Singapore.

[18] Zelaya, P. (2023). Flight delay and cancellation dataset (2019–2023). Kaggle. https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023

[19] Threnjen. (2019). 2019 airline delays and cancellations. Kaggle. https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations

[20] Singh, S. (2024). Flight delay dataset (2018–2024). Kaggle. https://www.kaggle.com/datasets/shubhamsingh42/flight-delay-dataset-2018-2024

[21] Oladipupo, O. T. (2010). Types of machine learning algorithms. In New Advances in Machine Learning.

[22] Kumar, R., & Singh, N. (2020). A survey on data mining and machine learning techniques for flight delay prediction. International Journal of System Assurance Engineering and Management.

[23] Breiman, M. (1984). Classification and regression trees. Wadsworth and Brooks/Cole.

[24] Ke, G., Meng, Q., Zhang, T., Chen, W., & Liu, T. (2017). LightGBM: A highly efficient gradient boosting decision tree. Microsoft Research.

[25] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[26] Kim, Y., Mavris, D. N., & Zachariah, J. (2021). Impacts of weather on airline performance metrics: A data-driven analysis. Transportation Research Part D: Transport and Environment, 92, 102740.

[27] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

[28] Yu, Y., Mo, H., & Li, H. (2019). A classification prediction analysis of flight cancellation based on Spark. In 7th International Conference on Information Technology and Quantitative Management (ITQM 2019), Procedia Computer Science, 162, 480–486. https://doi.org/10.1016/j.procs.2019.12.014

[29] Giarmas, N. (2025). Flight delay and cancellation prediction using machine learning models (Master's thesis, Department of Business Administration, Business Analytics and Data Science).

[30] Ansari, A., Shaikh, A., Mapkar, S., & Khan, M. (2019). Cancellation prediction for flight data using machine learning. In 2nd International Conference on Advances in Science & Technology (ICAST-2019), K. J. Somaiya Institute of Engineering & Information Technology, University of Mumbai, Maharashtra, India. SSRN.