İSTANBUL TOPKAPI ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ

-VERİ MADENCİLİĞİ (1.DÖNEM)-
UÇUŞ GECİKME TAHMİNİ PROJE RAPORU
-GRUP ADI: TÜRBÜLANS (TURBULENCE)-

Öğrenci Adı : EMİNE GÜNEŞ
Öğrenci No : 22040101036
E-mail : eminegunes@stu.topkapi.edu.tr

Öğrenci Adı : OĞUZHAN ÖZDEMİR
Öğrenci No : 22040101017
E-mail : oguzhanozdemir2@stu.topkapi.edu.tr

Öğrenci Adı : MUHAMMET ENES İNAL
Öğrenci No : 22040101023
E-mail : muhammetenesinal@stu.topkapi.edu.tr

Öğrenci Adı : MUHAMMED MERT ORUÇ
Öğrenci No : 22040101035
E-mail : muhammedmertoruc@stu.topkapi.edu.tr

Öğrenci Adı : ZEKERİYA DENİZ UĞURLU
Öğrenci No : 22040101034
E-mail: zekeriyadenizugurlu@stu.topkapi.edu.tr

GİTHUB REPO LİNK: https://github.com/Oguzhan1511/Turbulance

1) PROBLEM TANIMI

İş/Bilimsel Soru: Projemiz, Amerika'daki 2015 yılına ait mevcut uçuş verileri kullanılarak, bir uçuşun varış noktasına planlanan saatten 15 dakikadan fazla gecikmeli olarak varıp varamayacağını önceden yüksek doğrulukla tahmin edebilir mi? sorusuna cevap aramaktadır. Uçuş sektöründe rötarlar hem uçuş şirketlerini finansal olarak zarara sokup hem de yolcu memnuniyetsizliğine yol açmaktadır. Bu proje, uçuşların gerçekleşmeden önce gecikme olup olmayacağını tahmin ederek havayolu şirketleri için planlama ve yolcular için bilgilendirme sağlamayı amaçlamaktadır.

Görev Türü : İkili Sınıflandırma (Binary Classification)

Hedef Değişken : 'IS_DELAYED' Uçuşun 15 dakikadan fazla geç kalıp kalmadığını gösteren ikili (0 veya 1) değişkendir. '0 (negatif)' zamanında veya erken varış ('ARRIVAL_DELAY' <= 15 dakika). '1 (pozitif)' gecikmeli varış ('ARRIVAL_DELAY' > 15 dakika).

Başarı Kriterleri : 'ROC AUC >= 0.70' = Modelin sınıfları ayırt etme yeteneği

'Doğruluk (Accuracy) >= %80 '

'F1-Score >= 0.60' = Gecikmeli uçuşların doğru tespiti ile yanlış tespiti arasındaki dengeyi ölçme

2) PROJE YÖNETİMİ

1.BÖLÜM (21-30 Ekim) : Literatür taraması, Kaggle 'Flight Delays' veri setinin seçimi ve proje konusunun netleştirilmesi.

2.BÖLÜM (1-10 Kasım) : Verilerin tanınması (EDA) ve görselleştirmesi. Eksik verilerin analizi. Yüksek miktardaki 'null' verilen olduğu sütunların tespit edilmesi ve veri temizleme yöntemlerinin belirlenmesi.

3.BÖLÜM (10-17 Kasım) : Veri ön işleme (Data Preprocessing) pipeline'ının kurulması. Kategorik değişkenler, havalimanları için Target Encoding ve havayolları için One-Hot Encoding işlemlerinin uygulanması.

4.BÖLÜM (17-30 Kasım) : Basit 'Baseline' modellerin (Dummy, Decision Tree, Gaussian Naive Bayes vs.) kurulması. Basit modeller üzerinde denenmek üzere bazı feature selection ve dimension reduction yöntemlerinin uygulanması.

5.BÖLÜM (1-30 Aralık) : Gelişmiş modellerin planlanması ve uygulanması. Gelişmiş modellerden en iyi sonuçların alınabilmesi için feature selection ve dimension reduction yöntemlerinin uygulanması. Hiper-parametre optimizasyonlarının planlanması ve uygulanması.

6.BÖLÜM (1-7 Ocak) : Gelişmiş modellerden çıkarılan sonuçların performans analizlerinin yapılması. Modellerin tanımlanan performans metriklerine göre karşılaştırılması. Confusion Matrix analizi ile hata türlerinin incelenmesi.

7.BÖLÜM (7-16 Ocak) : Final proje raporunun derlenmesi ve sunum hazırlaması.

Roller ve Sorumluluklar:

Zekeriya Deniz Uğurlu (Veri Entegrasyonu ve Ön İşleme) : ‘flights.csv’, ‘airlines.csv’ ve ‘airports.csv’ veri setlerinin birleştirilmesi. Eksik veri analizi ve temizleme stratejilerinin yapılması. Encoding işlemlerinin uygulanması. ‘preprocess_pipeline’ ve ‘encode_features’ fonksiyonlarının geliştirilmesi.

Emine Güneş (Veri Analizi) : Verilerin incelenmesi. Veri setindeki kategorik ve sayısal değişkenlerin görselleştirilmesi. Histogramlar ve dağılım grafiklerinin oluşturulması. Sayısal değişkenler ile hedef değişken arasındaki korelasyon matrisinin çıkarılması.

Muhammet Enes İnal (Temel Base Model İşlemleri) : Projenin başarısını ölçmek için bir referans noktası oluşturulması. Karmaşık olmayan, hızlı çalışan ve yorumlanabilir modellerin kurulması ve eğitilmesi. Bu modellerin ham veri üzerindeki ilk Accuracy ve ROC-AUC gibi skorlarının raporlanması.

Muhammed Mert Oruç (Özellik Seçimi İşlemleri) : Yüksek boyutlu veri setinde, oluşturulan base modellere en çok katkı sağlayan niteliklerin istatistiksel yöntemlerle belirlenmesi ve gürültülü verilerin ayıklanması. İşlemlerden sonra baseline model ile test edilip performans etkilerinin test edilmesi.

Oğuzhan Özdemir (Boyut İndirgeme İşlemleri) : Özellikle encoding işlemleri sonrası artan sütun sayılarının yönetilmesi ve hesaplama maliyetlerinin düşürülmesi. Belirlenen boyut indirgeme yöntemleri ile veri setinin varyansını koruyan daha az sayıda bileşene indirgemesi. İndirgenmiş veri seti ile Baseline modellerinin tekrar test edilerek işlem süresi ve başarı farkının analiz edilmesi.

Çıktılar:

GitHub Repo Link: https://github.com/Oguzhan1511/Turbulance

Final Proje Raporu: FET445_[OgrenciNumarası]_[Turbulence]_FinalReport.pdf

Kod Dosyaları: FET445_[22040101034]_[Turbulence]_1.ipynb

FET445_[22040101035]_[Turbulence]_2.ipynb

FET445_[22040101036]_[Turbulence]_3.ipynb

FET445_[22040101017]_[Turbulence]_4.ipynb

FET445_[22040101023]_[Turbulence]_5.ipynb

Veri Seti: ‘flights.csv’ , ‘airlines.csv’ ve ‘airports.csv’ tabloları.

Veri seti linki (Kaggle) : https://www.kaggle.com/datasets/usdot/flight-delays

Sunum: Proje bulgularını özetleyen sunum dosyası (.pptx)

3) İlgili Çalışmalar

Abhishek Banerjee tarafından yapılan çalışma uçuş gecikmesi tahmini odaklıdır. Bizim projemizde olduğu gibi uçuşun geç kalıp kalmayacağını tahmin etmeye çalışmıştır ama bizim projemizden farkı bu çalışmadan uçağın kalkış noktasından geç kalkmayı tahmin etmektedir. Linear regression, ridge, random forest, decision tree, boosted linear modellerini kullanmıştır ve hepsinin doğru tahmin oranlarını karşılaştırmıştır. Veri setindeki tüm verileri kullanmak yerine bazı havalimanlarının uçuşlarını kullanmıştır.

Link: https://www.kaggle.com/code/abhishek211119/2015-flight-delays-and-cancellation-prediction

Manasi Chhibber tarafından yapılan çalışmada yine uçağın geç kalıp kalmayacağı hesaplanmıştır. Tek model olarak Decision Tree kullanılmıştır ve kodda doğruluk oranını %99.8 olarak gösterilmiştir. Bizden farklı olarak Training ve Testing verilerini 70 30 olarak ayırmıştır. Data leakage engellemek için ekstra bir şey yapmamıştır.

Link: https://www.kaggle.com/code/manasichhibber/flight-delay-predictions.

Levaniz tarafından yapılan çalışma diğer iki çalışmaya göre daha detaylıdır. Model olarak random forest kullanılmıştır. Date leakage önlenmesi için gereken kolonları silmiştir. Doğruluk (accuracy) metriğini kullanarak verileri karşılaştırmıştır. Veri kullanımını azaltmak için bazı yerlerde memory kısıtlamıştır.

Link: https://www.kaggle.com/code/levaniz/machine-learning-analysis-of-flights-data.

Bizim çalışmamızda ise hepsinin eksiklikleri bulunarak en iyi hale getirilmeye çalışılmıştır. Öncelikle data leakage engellenerek modelin gerçek sonuçlar verecek şekilde öğrenmesi sağlanmıştır. Farklı base modeller denenmiş ve en iyi doğruluk oranına sahip modeli seçilmiştir. Veri setinin bazı bölümleri değil hepsi kullanılarak hiçbir hafıza kısıtlanmamıştır. Aynı veri seti üzerinden yapılmış diğer çalışmalara göre farkımız feature selection, feature engineering, dimension reduction kullanmamızdır. Aynı modellerde farklı parametreler kullanılarak yine en iyi sonucu verecek parametreler bulunmuştur.

4)Veri Tanıtımı ve Yönetimi

1-Veri Seti

Adı: 2015 Uçuş Gecikmeleri ve İptalleri (2015 Flight Delays and Cancellations)

Kaynak: Kaggle

Bağlantı: https://www.kaggle.com/datasets/usdot/flight-delays

Lisans/Kullanım Hakları: Bu veriler ABD Hükümeti tarafından üretilmiştir. Bu yüzden telif hakları kısıtlaması yoktur akademik ve ticari kullanılabilir.

2-Veri Şeması

Veri seti 3 ana dosyadan oluşmaktadır ve birbirleriyle ilişkilidir.

1. **flights.csv (Ana Veri Tablo)**

   Değişkenler:

   -Zaman Bilgileri: YEAR, MONTH,DAY,DAY_OF_WEEK (Tam sayı- Integer)

   -Uçuş Bilgileri: AIRLINE(String),FLIGHT_NUMBER(Integer),TAIL_NUMBER(String)

   -Lokasyon: ORIGIN_AIRPORT, DESTINATION_AIRPORT(String)

   -Zamanlama: SCHEDULED_DEPARTURE,DEPARTURE_TIME,DEPARTURE_DELAY(Float-Int)

   -Hedef Değişken:ARRIVAL_DELAY >15 dakika ise “1” (Gecikti) değilse “0” (gecikmedi) anlamına gelir.

   -İptal Durumu:CANCELLED (0 veya 1 Binary)

   -Birimler: Zamanlar HHMM formatında, gecikmeler dakika cinsinden

   -Beklenen Aralıklar: Gecikmeler negatif(erken varış) olabilir veya pozitif (geç kalma) olabilir.

2. **airlines.csv (Havayolu Tablosu)**

   IATA _CODE: Havayolunun 2 haneli kod kısaltmasıdır(Örneğin; AA;AS…)

   AIRLINE: Havayolunun tam adıdır(Örneğin; American Airlines)

3 **.airports.csv (Havaalanı Tablosu)**

   IATA_CODE: 3 haneli havaalanı kodu(Örneğin; AJX,LAX)

   AIRPORT,CITY,COUNTRY,STATE: Lokasyon bilgileri (String)

   LATITUDE,LONGITUDE: Coğrafi konumlar (Float-Harita görselleştirmesi için)

3- Boyut(Size ve Scale)

-Satırlar(Rows): flights.csv dosyasında yaklaşık 5.8 milyon satır vardır.

-Sütunlar(Columns): flights.csv dosyasında 31 adet sütun vardır.

-Beklenen Sınıf Dengesi(Class Balance): Veri seti dengesizdir

Uçuşlar çoğunlukla zamanında kalkar(%80)

Gecikme yaşayan uçuşlar %20 dağılımındadır

4-Veri Erişim Planı(Data Access Plan)

Elde Etme: Veriler kaggle platformundan elde edilecektir.

Depolama: Proje geliştirme aşamasında veriler diskte depolanacaktır.

Güncelleme: Veri grubu 2015 yılına aittir herhangi bir aktif veri akışı olmadığı için güncellemeye ihtiyaç yoktur.

5-Etik, Gizlilik ve Önyargı

-Gizlilik(Privacy): Veri seti anonimdir. Etik açıdan kullanımı güvenlidir.

-Rıza: ABD Ulaştırma Bakanlığı'nın yayınladığı OpenData olduğu için bireysel rızaya gerek yoktur.

5) Keşifsel Veri Analizi (Exploratory Data Analysis)

Veri Kalitesi Kontroller : Veri setinin güvenilirliğini sağlamak amacıyla temizlik işlemleri uygulanmıştır. İlk olarak 'CANCELLATİON_REASON', 'AIR_SYSTEM_DELAY' ve 'WEATHER_DELAY' gibi sütunların veri tanıma (EDA) işlemlerinden sonra %80'in üzerinde eksik veri içerdiği tespit edilmiş ve bu sütunlar veri setinden temizlenmiştir.

'ARRIVAL_DELAY', 'DEPARTURE_DELAY' , 'ACTUAL_ELAPSED_TIME' , 'WHEELS_OFF' ve 'TAXI_IN/OUT' gibi değişkenler sızıntı riski (data_leakage) taşıdığı gerekçesiyle veri setinden temizlenmiştir.

Ayrıca, iptal edilen (CANCELLED = 1) ve yönlendirilen (DİVERTED = 1) gibi uçuşlar filtrelenerek veri tutarlılığı sağlanmıştır.

Dağılımlar ve Denge : Hedef değişken analizi sonucunda, veri setinde sınıf dengesizliği olduğu görülmüştür. Gecikmeyen uçuşların sayısı geciken uçuşlara kıyasla oldukça fazladır. Sayısal değişkenlerin 'DISTANCE' ve 'SCHEDULED_TIME' özelliklerinin sağa çarpık bir dağılım sergilediği histogramlar aracılığıyla gözlemlenmiştir.

Özellik-Hedef İlişkileri : Sayısal değişkenler arasındaki ilişkileri anlamak için kategorik değişkenler(Havayolu ve Havalimanı) ile hedef değişken arasındaki ilişkiyi modele yansıtabilmek için, bu değişkenlerin ortalama gecikme oranlarına göre kodlanması yani Target Encoding stratejisi benimsenmiştir. 'SCHEDULED_DEPARTURE' gibi zaman bilgileri, günün bölümlerine (Sabah, öğle, akşam vs.) ayrılarak kategorik hale getirilmiş ve hedef üzerindeki etkisi daha belirgin hale getirilmiştir.

Görselleştirme Planı : Verinin özelliklerini ortaya koymak için aşağıdaki grafikler kullanılmıştır:

Countplot : Hedef değişkenin(IS_DELAYED) sınıf dağılımını ve dengesizliğini göstermek için.

Histogram & KDE: Uçuş mesafesi ve sürelerinin dağılımını ve çarpıklığını analiz etmek için.

Heatmap (Isı Haritası) : Değişkenler arası korelasyon matrisini görselleştirmek ve çoklu bağlantı riskini değerlendirmek için.

6) Veri Hazırlama Planı

Temizleme : Analiz sonucunda tespit edilen veri sızıntısı kaynakları (‘DEPARTURE_TIME’ , ‘ARRIVAL_DELAY’ , ‘TAXI_IN’ , ‘TAXI_OUT’ vs.) eğitim setinden temizlenmiştir. Gürültü yaratan ve yüksek kardinaliteli ‘FLIGHT_NUMBER’ ve ‘TAIL_NUMBER’ gibi sütunlar silinmiştir. Analiz odağını sadece gerçekleşen uçuşlar olarak daraltmak adına ‘CANCELLED’ ve ‘DIVERTED’ gibi gerçekleşmeyen veya yönlendirilen uçuş kayıtları veri setinden temizlenmiştir.

Imputasyon Stratejisi : Eğitim sırasında mevcut olan fakat test setinde görülmeyen havalimanı kodları için Target Encoding işlemi yaparken ‘Global Ortalama (mean)’ değeri atanarak, modelin yeni verilere karşı hata üretmesi veya yanlış tahmin yapmasının önüne geçilmiştir.

Dönüşümler : 300’ den fazla benzersiz değeri olan ‘ORIGIN_AIRPORT’ ve ‘DESTINATION_AIRPORT’ değişkenleri için, her havalimanının ‘ortalama gecikme oranı’ hesaplanarak sayısal bir değere dönüştürülmüştür. Bu sayede yüzlerce yeni sütun oluşturulmadan kategorik bilgi korunmuştur.

‘AIRLINE’ (14 sınıf) ve ‘TIME_OF_DAY’ (5 sınıf) gibi düşük kardinalite değişkenler için One-Hot Encoding uygulanarak modelin bu kategoriler arasındaki farkı daha iyi öğrenmesi sağlanmıştır.

Özellik Mühendisliği: Ham olan ‘SCHEDULED_DEPARTURE’ verisi, sayısal bir değerden anlamlı zaman dilimlerine dönüştürülmüştür. Saat bilgisi belirli aralıklar kullanılarak ‘Sabah’, ‘Öğle’ , ‘Akşam’, ‘Gece’ ve ‘Gece Yarısı’ olmak üzere 5 farklı kategoriye (‘TIME_OF_DAY’) ayrılmış ve modelin zaman etkenini daha iyi öğrenmesi sağlanmıştır.

‘flights.csv’ tablosu, ‘airlines.csv’ ve ‘airports.csv’ tabloları ile birleştirilerek, sadece kodlar yerine havayolu ve havalimanı karakteristiklerinin de analize dahil edilmesi sağlanmıştır.

Özellik Seçimi ve Boyut İndirgeme : Değişkenler arasındaki çoklu bağlantıyı çözmek için Korelasyon Matrisi kullanılmış ve hedef değişkenle (IS_DELAYED) ilişkisi en düşük olan öznitelikler belirlenmiştir.

Özellikle One-Hot Encoding sonrası artan sütun sayısını yönetmek ve hesaplama maliyetini düşürmek için, veri setindeki varyansın büyük kısmını koruyan PCA yöntemi uygulanarak veri daha düşük boyutlu bir uzaya taşınmıştır.

7) Modelleme Planı

Muhammed Mert Oruç: Dummy Classifier : Modelin hiçbir şey öğrenmeden sadece çoğunluk sınıfını tahmin ederek minimum başarı alt sınırını belirlemek için kullanılmıştır.

Ridge Classifier : One-Hot Encoding sonrası artan sütun sayısı ve çoklu bağlantı riskiyle başa çıkmak için doğrusal model olarak seçilmiştir.

Zekeriya Deniz Uğurlu: Decision Tree : Veri setindeki doğrusal olmayan ilişkileri yakalayabilen basit bir yapı sunduğu için tercih edilmiştir.

Linear SVC : Büyük ve seyrek veri setlerinde, karmaşık hesaplamalara girmeden hızlı ve etkili bir ayrım düzlemi çizebildiği için seçilmiştir.

Oğuzhan Özdemir: Logistic Regression: İkili sınıflandırma problemlerinde endüstri standardı olması ve her bir sütunun üzerindeki etkisini olasılıksal olarak yorumlar.

Gaussian Naive Bayes: Özelliklerin birbirinden bağımsız olduğu varsayımıyla çalışan, eğitim süresi çok kısa olan ve büyük veri setlerinde hızlı olan modeldir.

Emine Güneş: Dummy Classifier: Sınıf dengesizliğinin model başarısını yanıltıp yanıltmadığını kontrol etmek ve rastgele tahmin başarısını raporlamak için eklenmiştir.

Gaussian Naive Bayes: Hesaplama maliyeti düşük olduğu için ve karmaşık modellerden önce verinin genel dağılımına dayalı basit tahminleme performansını görüntülemek amacıyla seçilmiştir.

Muhammet Enes İnal: Logistic Regression: Modelin çıktılarını (0 veya 1) olasılık değeri olarak verebilmesi ve basit yapısı nedeniyle base model olarak seçilmiştir.

Decision Tree: Veri ön işleme gerektirmemesi ve özniteliklerin hiyerarşik önemini basitçe ortaya koyduğu için seçilmiştir.

Aday Modeller:

Oğuzhan Özdemir: LightGBM : 5 milyon satırlık büyük veri setimizde, yaprak odaklı büyüme stratejisi sayesinde diğer boosting algoritmalarına nazaran hızlı eğitildiği ve yüksek başarı sağlayacağı, Base model olarak seçilen Logistic Regression'un aksine doğrusal olmayan örüntüleri de yakalayabileceği düşünüldüğü için aday model olarak seçilmiştir.

Extra Trees Classifier : RandomForest'a kıyasla bölünme noktalarını rastgele seçerek modelin aşırı öğrenme riskini daha fazla düşürdüğü ve parazitli verilerde daha sağlam çalışacağı düşünüldüğü için seçilmiştir.

Zekeriya Deniz Uğurlu: Random Forest : Çok sayıda karar ağacını birleştirerek,hataları minimize etmesi ve hem kategorik hem de sayısal verilerle güçlü performans vermesi düşünüldüğü için seçilmiştir. DecisionTree Classiffier modelinin en büyük sorunu olan yüksek varyans ve ezberleme problemini çözeceği için aday olarak görülmüştür.

SVC(RBF Kernel ile) : Veriyi daha yüksek boyutlu bir uzaya taşıyarak, doğrusal olarak ayrılamayan karmaşık sınıf sınırları tespit etme yeteneği nedeniyle ve Linear SVC base modelin sadece düz bir çizgi ile ayırabildiği sınıfları, RBF çekirdeği ile eğrisel ve çok boyutlu sınırlar çizerek ayırır. Bu sayede Linear SVC'den çok daha üstün ayrım gücüne sahiptir. Bu yüzden aday olarak seçilmiştir.

Emine Güneş: Naive Bayes: Hiper-parametre optimizasyonu ile basit bir modelin maksimum kapasitesini test edebileceği için seçilmiştir. Base model olarak seçilen DummyClassifier ve Gaussian NB'nin varsayılan ayarlarından daha hassas bir olasılık hesabı yapmaktadır.

CatBoost: Özellikle "Airline" ve "Airport" gibi kategorik değişkenleri One-Hot Encoding işlemine tabi tutmadan kendi içinde işleyebilmesi ve veri setimizin büyük problemi olan sınıf dengesizliğine karşı dirençli olması sebebiyle en güçlü adaylardan biridir.

Muhammed Mert Oruç: QDA : Sınıfların kovaryans matrislerini ayrı ayrı öğrenerek doğrusal olmayan karar sınırları çizebilmesi, onu basit linear modellerden ayıran en önemli özelliğidir. Ayrıca RidgeClassfier base modeli doğrusal bir model olduğundan dolayı veriyi düz bir çizgiyle ayırmaya çalışır. QDA ise verinin dağılımına göre eğrisel sınırlar çizer ve eksik öğrenme riskini ortadan kaldırır. Bu sebeple aday olarak kullanılabilir.

ADA Boost: Önceki modellerin yanlış sınıflandırdığı zor uçuş örneklerine daha fazla ağırlık vererek, modelin zayıf yönlerini iteratif olarak güçlendirdiği için aday olarak seçilmiştir.

Muhammed Enes İnal: XGBoost: İçerdiği regülarizasyon terimleri ile aşırı öğrenmeyi engellemesi nedeniyle aday olarak düşünülmüştür. Decision Tree modelinin aksine, ağaçları bağımsız değil sıralı kurar ve hatayı optimize ederek ilerler. Logistic Regression'a göre çok daha esnek bir yapı sunduğu için ana model olarak seçilebilir.

Stacking Classifier: Tek bir modelin başarısına güvenmek yerine, XGBoost, RandomForest ve KNN gibi farklı modellerin tahminlerini birleştirip genelleştirme yeteneği en yüksek modeli oluşturmak için seçilmiştir.

8) Değerlendirme Tasarımı

Kullanılan Metrikler: Ekibimiz geliştirdiği tüm modellerin performansını karşılaştırmak için birden fazla metrik kullanmıştır. F1, Precision, Recall, ROC-AUC, PR-AUC, MAE gibi karşılaştırma metrikleri kullanılmıştır. Bu metrikler 6 base modelde denenerek karşılaştırılmıştır.

Zekeriya Deniz Uğurlu (PROJE-1):

Model	Accuracy	F1 Score	Precision	Recall	ROC AUC	PR AUC	MAE (Hata)
Decision Tree (Normal)	60.45	36.88	0.26	0.65	66.48	30.12	0.40
Decision Tree (RFE)	61.26	36.86	0.26	0.63	66.50	30.05	0.39
Decision Tree (PCA)	60.34	35.82	0.25	0.62	65.06	27.61	0.40
Linear SVM (Normal)	58.82	35.12	0.24	0.62	63.56	26.07	0.41
Linear SVM (PCA)	58.10	34.70	0.24	0.62	62.93	25.61	0.42
Linear SVM (RFE)	58.59	34.82	0.24	0.62	63.11	25.45	0.41

Emine Güneş (PROJE-3):

Model	**accuracy**	**F1 Score**	**precision**	**recall**	**ROC AUC**	**PR AUC**	**MAE**
Dummy Classifier	82.09	0.00	0.00	0.00	50.00	17.91	17.91
Dummy Classifier +CA	82.09	0.00	0.00	0.00	50.00	17.91	17.91
Dummy Classifier +LDA	82.09	0.00	0.00	0.00	50.00	17.91	17.91
Gaussian NB	72.87	0.25	0.25	0.26	61.42	24.07	33.84
Gaussian NB +CA	72.36	0.26	0.25	0.28	61.48	24.12	33.90
Gaussian NB+LDA	82.09	0.00	0.00	0.00	63.55	26.12	28.44

Muhammed Mert Oruç (PROJE-2):

Model	**Accuracy**	**F1 Score**	**Precision**	**Recall**	**ROC AUC**	**PR AUC**	**MAE**
Dummy Classifier	70.59%	0.18	0.18	0.18	49.98%	17.90%	29.41%
Dummy Classifier+ ANOVA F	70.59%	0.18	0.18	0.18	49.98%	17.90%	29.41%
Dummy Classifier +SVD	70.59%	0.18	0.18	0.18	49.98%	17.90%	29.41%
Ringe Classifier	58.82%	0.35	0.24	0.62	63.56%	26.07%	41.18%
Ringe Classifier+A NOVAF	58.60%	0.35	0.24	0.62	63.30%	25.79%	41.40%
Ringe Classifier + SVD	58.09%	0.35	0.24	0.62	62.94%	25.63%	41.91%

Oğuzhan Özdemir (PROJE-4):

Model	Accuracy	F1 Score	Precision	Recall	ROC AUC	PR AUC	MAE
Logistic Regrasyon	%82.09	0.00	0.00	0.00	0.60	0.23	0.18
Logistic Regrasyon + Mutual Information	%82.09	0.00	0.00	0.00	0.64	0.25	0.18
Logistic Regrasyon + MU + LDA	%82.09	0.00	0.00	0.00	0.65	0.26	0.18
Naive Bayes	%75.04	0.25	0.25	0.25	0.51	0.18	0.21
Naive Bayes + Mu	%79.05	0.25	0.26	0.28	.053	0.21	0.21
Naive Bayes + MU + LDA	%82.09	0.25	0.26	0.28	0.54	0.23	0.21

Muhammet Enes İnal (PROJE-5):

Model	Accuracy	F1 Score	Precision	Recall	ROC AUC	PR AUC	MAE (Hata)
Logistic Reg.	58.95	35.12	0.2449	62.04	63.57	26.07	0.4105
Logistic Reg. (SelectFromModel)	59.00	35.08	0.2448	61.85	63.48	25.96	0.4100
Logistic Reg. (TruncatedSVD)	58.28	34.70	0.2411	61.91	62.93	25.62	0.4172
Logistic Reg. (Sel.FromModel + SVD)	58.90	35.04	0.2444	61.91	63.43	25.90	0.4110
Decision Tree	66.89	39.97	0.2959	61.57	68.81	34.34	0.3311
Decision Tree (SelectKBest - ANOVA F)	60.64	36.85	0.2585	64.12	66.51	30.05	0.3936
Decision Tree (FastICA)	66.13	39.07	0.2882	60.65	67.16	31.08	0.3387
Decision Tree (SelectKBest ANOVA F+ FastICA)	60.39	36.71	0.2571	64.15	66.16	29.51	0.3961

Desicion Tree Base Modeli: Bu modele iki farklı strateji ile yaklaşılmıştır. Seçilen selection ve boyut indirgeme yöntemlerine göre modelin doğruluk skoru da değiştiği görülmüştür. Bunun sonucunda seçilen selection modellerinden RFE yönteminin daha iyi olduğu görülmüştür. Seçilen boyut indirgeme modellerine göre ise FastICA yönteminin daha başarılı olduğu görülmüştür.

Dummy Classifier: Dummy modellerinde stratejiler farklı olarak kullanılmıştır. Proje 3 de most_frequent Proje 2 de ise stratified kullanmıştır. most_frequent stratejisi her örneği çoğunluk sınıfı olarak tahmin ettiği için model 1 tahmini yapmaz. Bu yüzden F1, precision ve recall sıfır çıkar. Stratified stratejisi ise sınıf oranlarına göre rastgele tahmin ürettiğinden zaman zaman 1 sınıfını da tahmin eder. Böylece düşük de olsa F1, precision ve recall değerleri oluşur. Bu da stratified stratejisinin daha başarılı olduğunu göstermiştir.

Gaussian NB: Bu modelde farklı olarak proje 4'te var_smoothing parametresi kullanılmıştır. var_smoothing kullanıldığında, modelin olasılık tahminleri yumuşatıldığı için dağılımlar daha stabil hale gelirmiştir, böylece bu yöntem daha başarılı olmuştur.

Logistic Regresyon: Bu modelde farklı parametreler kullanılmıştır. Proje 5 veri dengesizliğinin önüne geçmek için dengeleme parametresi kullanmıştır ve bunun sonucunda doğruluk skorunun düştüğü ama daha gerçekçi bir skor olduğu görülmüştür. Seçilen selection yöntemleri karşılaştırıldığında Mutual Information yöntemi daha başarılı olmuştur. Boyut indirgeme yöntemlerinde ise LDA yönteminin daha başarılı olduğu görülmüştür.

9) Riskler ve Azaltma Yöntemleri

Veri Riskleri: Proje kapsamında kullanılan uçuş veri seti yaklaşık 5.8 milyon satır içerdiğinden, çalıştırma sırasında yüksek RAM kullanımı oluşturmuş ve işlem süresini artırmıştır. Ayrıca veri setindeki ARRIVAL_DELAY ve DEPARTURE_DELAY sütunları, modelin tahmin etmeye çalıştığı "gecikme" bilgisini doğrudan içerdiği için data leakage riski yaratmıştır. Bu durum modelin gerçek performansını bozabileceğinden, preprocessing aşamasında bu sütunlar tamamen kaldırılmıştır.

Aynı şekilde ORIGIN_AIRPORT ve DESTINATION_AIRPORT sütunlarında çok sayıda benzersiz havalimanı kodu bulunduğundan, yüksek boyut problemi oluşmuştur. Bu nedenle bu kategorik değişkenler silinmiştir. Bu işlemlerle hem sızıntı engellenmiş hem de bellek kullanımı azaltılmıştır.

Azaltıcı Yöntemler (Mitigations): Azaltıcı yöntemler olarak ekip üyeleri çoğunlukla boyut indirgemeleri kullanmıştır:

Emine Güneş – LDA

LDA, gecikme durumunu (IS_DELAYED) tahmin eden sınıflandırma modellerinde (Dummy ve Gaussian Naive Bayes) performansı değerlendirmek için kullanılmıştır. Bu yöntem sınıflar arasındaki ayrımı güçlendirdiği için, modelin gecikme tahminindeki ayrım gücünü gözlemlemeyi sağlamıştır.

Oğuzhan Özdemir – LDA

LDA, Logistic Regression ve Naive Bayes modelleri üzerinde uygulanarak, farklı denetimli algoritmaların sınıf ayrımı yapan eksenlerde nasıl performans gösterdiği karşılaştırılmıştır.

Zekeriya Deniz Uğurlu – PCA

PCA, veri setindeki varyansı en iyi temsil eden bileşenleri çıkararak boyutu azaltmak

için kullanılmıştır. Gürültüyü azaltıp veri setini daha yönetilebilir hale getirerek hem hız hem de model stabilitesi açısından iyileşme sağlamayı amaçlamıştır.

Muhammet Enes İnal – FastICA

FastICA, veri setindeki bileşenleri istatistiksel olarak bağımsız hale getirerek yeni bir özellik temsili oluşturmak için kullanılmıştır. Böylece karmaşık yapılar daha ayrıştırılabilir forma getirilmiş ve gizli bağımsız bileşenlerin modeller üzerindeki etkisi değerlendirilmiştir.

Muhammed Mert Oruç – Truncated SVD

Truncated SVD, encoded özelliklerin çok yüksek boyutlu yapısını daha yönetilebilir hale getirmek için uygulanmış ve özellikle ilk 10 bileşen kullanılarak modelin hem hız hem de kaynak tüketimi optimize edilmiştir.

10) Kullanılan Araçlar

Python sürümü: Python 3.11

Ana Kütüphaneler: Pandas, Numpy, Matplotlib, Seaborn, SkLearn, Warnings, GridSearchCV

random seeds: random_state=42, random_state=45

Geliştirilen kodlar: https://github.com/Oguzhan1511/Turbulance

11) Beklenen Sonuçlar ve Görselleştirme Planı

Göstermeyi beklediğiniz tablolar/grafikler:

Performans Karşılaştırma Tabloları: Geliştirilen tüm modellerin Train ve Test verileri üzerindeki Accuracy , F1-Score , Recall, ROC-AUC ve PR-AUC yan yana gösteren özet bir veri çerçevesi (DataFrame) sunulacaktır. Bu tablo hangi modelin en dengeli sonucu verdiğini gösterecektir.

Karışıklık Matrisi(Confusion Matrix): Sınıf dengesizliği nedeniyle modelin “Gecikme var” sınıfını ne kadar doğru tahmin net bir şekilde gösterir.

ROC Eğrileri: Modellerin pozitif sınıfı ayırt etme yeteneğini göstermek için ROC eğrileri çizdirilecek ve eğri altında kalan (AUC) üzerinden modeller arası sıralama yapmaktadır..

PCA Varyans Grafiği: Boyut indirgeme işlemi uygulanan modeller için, seçilen bileşen sayısının toplam varyansın ne kadar açıkladığını gösteren “Kümülatif Varyans Grafiği” sunularak veri kaybı oranı gösterilmektedir.

Yorumlanabilirlik Yaklaşımı(Interpretability Approach):

Özellik Önem Düzeyleri: Ağaç Tabanlı modeller kullanıldığında, modelin karar verirken hangi değişkenlere daha fazla ağırlık verdiğini gösteren “Feature Importance” çubuk grafikleri gösterilecektir.

Katsayı Analizi: Lojistik Regresyon ve Ridge Classifier gibi doğrusal modeller için,özniteliklerin katsayıları incelenerek, hangi değişkenini gecikme olasılığını arttırdığı veya azalttığı yorumlanacaktır.

12)Referanslar

1-) U.S. Department of Transportation (DOT), "2015 Flight Delays and Cancellations," Kaggle, 2016. [Online]. Erişim adresi: https://www.kaggle.com/datasets/usdot/flight-delays.

2-)M. Chhibber, "Flight Delay Predictions," Kaggle, [Online]. Erişim adresi: https://www.kaggle.com/code/manasichhibber/flight-delay-predictions.

3-)Levaniz, "Machine Learning Analysis of Flights Data," Kaggle, [Online]. Erişim adresi: https://www.kaggle.com/code/levaniz/machine-learning-analysis-of-flights-data.

4-)Abhishek, "2015 Flight Delays and Cancellation Prediction," Kaggle, [Online]. Erişim adresi: https://www.kaggle.com/code/abhishek211119/2015-flight-delays-and-cancellation-prediction