İSTANBUL TOPKAPI ÜNİVERSİTESİ

MÜHENDİSLİK FAKÜLTESİ

VERİ MADENCİLİĞİ

ABD Trafik Kazalarında Şiddet Düzeyi Öngörüsü

Ahmet KOÇ - 22040101033 - ahmetkoc1@stu.topkapi.edu.tr – BM

İremnur ERBAŞ – 23010101081 - iremnurerbas@stu.topkapi.edu.tr - BM

Rabia Nur AKDAŞ – 22040101054 - rabianurakdas@stu.topkapi.edu.tr – BM

Sıla KARAHAN - 22040101037 - silakarahan@stu.topkapi.edu.tr - BM

İlkay ÖZKAN- 22040101022 - ilkayozkan@stu.topkapi.edu.tr – BM

GITHUB LİNKİ: github.com/ahmetk60/usa-acc

2) Problem Tanımı

Bu proje, trafik kazalarının şiddet derecesini önceden tahmin etmeyi amaçlayan bir makine öğrenimi çalışmasıdır. Kullanılan veri seti, Amerika Birleşik Devletleri'nde meydana gelen milyonlarca trafik kazasına ait kayıtları içermektedir. Her kayıt; hava durumu, yol tipi, görüş mesafesi, sıcaklık, günün saati, şehir bilgisi ve trafik yoğunluğu gibi çevresel ve durumsal özellikleri kapsamaktadır.

Proje kapsamında yanıt aranılan temel soru şudur: "Bir trafik kazası meydana geldiğinde, o kazanın çevresel ve yol koşullarına bakarak şiddet derecesi önceden tahmin edilebilir mi?" Bu soru doğrultusunda proje, kazaların ciddiyetini belirlemeye yönelik sınıflandırma (classification) türünde bir makine öğrenimi problemidir. Model, her kazayı belirli bir şiddet sınıfına atamayı öğrenmektedir. Bu sınıflar genellikle dört ana kategoriden oluşmaktadır: hafif, orta, ciddi ve çok ciddi kazalar. Projenin hedef değişkeni, "kaza şiddeti" (Severity) olarak adlandırılmıştır. Modelin tahmin etmeye çalıştığı bu değişken, kazanın ne kadar ağır sonuçlara yol açtığını ifade etmektedir. Bağımsız değişkenler ise hava koşulları, yol tipi, sıcaklık, görünürlük, trafik yoğunluğu, günün saati gibi çeşitli çevresel ve durumsal faktörlerden oluşmaktadır.

Projenin başarısını değerlendirmek için Macro F1-Skoru kullanılmıştır. Bu metrik, dengesiz sınıf dağılımlarında tüm sınıfları eşit önemde değerlendirir ve modelin genel performansını daha adil bir şekilde ölçer. Ek olarak, modelin performansı sınıf bazlı precision, recall ve F1-score değerleriyle desteklenmiştir. Sonuç olarak bu proje, trafik kazalarının şiddetini önceden tahmin edebilmek amacıyla veri madenciliği ve makine öğrenimi tekniklerini bir araya getirmektedir. Bu tür bir model, ilerleyen süreçlerde riskli bölgelerin belirlenmesi, acil durum yönetiminin iyileştirilmesi ve trafik güvenliği stratejilerinin geliştirilmesine katkı sağlayabilecek potansiyele sahiptir.

3) Proje Yönetimi

Kilometre Taşları ve Zaman Çizelgesi

1. Hafta: Veri Seti ve Proje Konusu Seçimi

Görev: Uygun trafik kazası veri setlerinin incelenmesi, proje amacının belirlenmesi ve hedeflerin netleştirilmesi.

Sorumlular: Ahmet Koç, İlkay Özkan, Sıla Karahan, Rabia Nur Akdaş, İremnur Erbaş

2. Hafta: Veri Ön İşleme, Veri Hazırlama ve Keşifsel Veri Analizi (EDA)

Görev: Eksik ve aykırı değerlerin giderilmesi, kategorik verilerin dönüştürülmesi, veri görselleştirme ve değişken analizi.

Sorumlular: Ahmet Koç, İlkay Özkan, Sıla Karahan, Rabia Nur Akdaş, İremnur Erbaş

3. Hafta: XGBoost Modeliyle Deneme ve Özellik Mühendisliği

Görev: Projede XGBoost modeli ile deneme yapılmasına karar verildi; veri setindeki 46 özellik ekip arasında paylaştırılarak özellik mühendisliği, data cleaning ve EDA çalışmalarına başlandı.

Sorumlular: Ahmet Koç, İlkay Özkan, Sıla Karahan, Rabia Nur Akdaş, İremnur Erbaş

4.–5. Hafta: Model Araştırması ve Netleştirme

Görev: İleride kullanılacak modellerin performans tahminlerinin araştırılması, model optimizasyonu ve en uygun modellerin final projesi için seçimi yapıldı.

Sorumlular: Ahmet Koç, İlkay Özkan, Sıla Karahan, Rabia Nur Akdaş, İremnur Erbaş

6. Hafta: Performans Analizi ve Değerlendirme

Görev: ROC AUC, Accuracy, Macro F1-Score gibi metriklerin uygulanması ve ekip içi değerlendirme.

Sorumlular: Ahmet Koç, İlkay Özkan, Sıla Karahan, Rabia Nur Akdaş, İremnur Erbaş

7. Hafta: Raporlama

Görev: Vize proje raporunun hazırlanması, bulguların derlenmesi ve sunuma hazır hale getirilmesi.

Sorumlular: Ahmet Koç, İlkay Özkan, Sıla Karahan, Rabia Nur Akdaş, İremnur Erbaş

Roller ve Sorumluluklar :

Rabia Nur Akdaş (Model: XGBoost + Logistic Regression)

Trafik ve tarih-saat ile ilgili özellikleri kullanarak veri setinin incelenmesi ve değişkenlerin seçimi ile temel sınıflandırma modelinin oluşturulmasından sorumludur.

Model sonuçlarını diğer ekip üyeleriyle karşılaştırarak genel değerlendirmeye katkı sağlamıştır.

Sıla Karahan (Model: XGBoost + Logistic Regression)

Hava durumu ile ilgili özellikleri kullanarak verinin hazırlanması, modelin eğitilmesi ve hiperparametre ayarlarının yapılmasında görev almıştır.

Model performansını ROC AUC ve F1-Score metrikleriyle değerlendirmiştir.

İremnur Erbaş (Model: XGBoost + Logistic Regression)

Trafik ve tarih-saat ile ilgili özellikleri kullanarak modelin eğitimi, test edilmesi ve sonuçların görselleştirilmesi üzerinde çalışmıştır.

Elde edilen performans sonuçlarını raporun değerlendirme kısmına eklemiştir.

Ahmet Koç (Model: XGBoost + Baseline)

Tüm özellikleri kullanarak modelin eğitim, test ve doğrulama süreçlerini yürütmüştür.

Sonuçları diğer modellerle karşılaştırarak genel performans analizine katkı sunmuştur.

İlkay Özkan (Model: XGBoost + Baseline)

Description ve açıklayıcı metin özelliklerini kullanarak modelin parametre ayarlamaları ve hata analizi üzerinde çalışmıştır.

Tüm modellerin karşılaştırıldığı genel sonuç tablosunun hazırlanmasında rol almıştır.

Genel Sorumluluklar (Tüm Ekip Üyeleri)

- Veri ön işleme (eksik/aykırı değerlerin temizlenmesi)
- Özellik mühendisliği ve değişken seçimi
- Model performanslarının karşılaştırılması (Accuracy, F1-Score, ROC AUC)
- Sonuçların yorumlanması ve proje raporunun hazırlanması
- Proje çıktılarının ortak değerlendirilmesi

Çalışmanın tüm çıktıları için: https://github.com/ahmetk60/usa-acc

4) İlgili Çalışmalar (Mini Literatür İncelemesi)

Temel Referanslar ve Karşılaştırmaları:

1. Rathnayaka, R. M. A. L., et al. (2017). "Predicting accident severity: A comparative study of machine learning methods."

o Kapsam: Kaza şiddetini (ikili ve çok sınıflı) tahmin etmek için Lojistik Regresyon (LR), kNN, Naive Bayes (NB) gibi tekli modeller ile Random Forest (RF) ve XGBoost gibi topluluk (ensemble) modellerini karşılaştırmıştır.

o Ölçütler: Accuracy, Precision, Recall, F1-score ve AUROC.

o Sonuçlar: Topluluk yöntemlerinin (özellikle RF ve XGBoost) kaza şiddetini tahmin etmede tekli modellere göre daha yüksek doğruluk sağladığı raporlanmıştır.

2. Amini, M., et al. (2022). "A Novel Explainable Artificial Intelligence Approach for Road Accident Severity Prediction." (MDPI)

o Kapsam: Kaza şiddeti tahmini için Açıklanabilir Yapay Zeka (XAI) yaklaşımı sunarak modelin "neden" o tahmini yaptığını anlamaya odaklanmıştır.

o Yöntemler: Model tahmini için Artık Sinir Ağı (ResNet) ve model kararlarını yorumlamak için SHAP değerleri kullanılmıştır.

o Sonuçlar: SHAP'ın, kazanın şiddetini etkileyen en önemli faktörleri belirlemek için başarılı bir şekilde kullanılabileceği gösterilmiştir.

3. Ramya, A., & Reshma, S. (2019). "Accident Severity Prediction Using Data Mining Methods." (IEEE)

o Kapsam: Kaza şiddetini tahmin etmek için veri madenciliği tekniklerini, özellikle Random Forest (RF) algoritmasını kullanmıştır.

o Sonuçlar: RF algoritmasının, kaza şiddetini etkileyen çevresel faktörler gibi özellikleri kullanarak ilgili örüntüleri tanımlamada ve şiddeti sınıflandırmada etkili olduğu bulunmuştur.

Projemizin Doldurduğu Boşluklar ve Farklı Yönleri:

• Veri Seti ve Kapsam: Projemiz, Amerika Birleşik Devletleri'nin tüm eyaletlerini kapsayan, yaklaşık 7 milyon kaza verisi içeren geniş bir Kaggle veri setini kullanacaktır. (İşlem gücüne bağlı olarak bu setin bir alt kümesi kullanılacaktır.)

• Çok Modelli ve Karşılaştırmalı Yaklaşım: Literatürdeki çalışmalar genellikle 1-2 modele odaklanırken, projemizde her ekip üyesi farklı bir model geliştirecek (Lojistik Regresyon, Random Forest, XGBoost, Decision Tree, Gradient Boosting) ve bu modeller doğrudan aynı veri seti üzerinde karşılaştırılacaktır.

• Özellik Mühendisliği Odaklı Yaklaşım: Diğer çalışmalar PCA ile boyut indirgeme veya kısıtlı özellik setleri kullanırken, projemiz zaman, hava durumu, yol tipi ve coğrafi konum gibi daha zengin bir değişken setini işleyerek gerçek dünya senaryolarına daha uygun bir analiz sunmayı hedeflemektedir.

• Açıklanabilirlik ve Değerlendirme: Projemiz sadece tahmin performansına değil, aynı zamanda hangi faktörlerin kaza şiddetini en çok etkilediğini karşılaştırmalı olarak analiz etmeye de odaklanacaktır.

• Standartlaştırılmış Başarı Ölçütleri: Tüm aday modeller, önceden belirlenmiş (ROC AUC >0.80, F1-score >0.75) ve standartlaştırılmış başarı metrikleri kullanılarak şeffaf bir şekilde değerlendirilecektir.

5) Veri Tanımı ve Yönetimi

- **Veri Seti:**
    - **Ad:** US Accidents (2016 - 2023)
    - **Kaynak:** Kaggle: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents
    - **Lisans/Kullanım Hakları:** Creative Commons (CC BY-NC-SA 4.0). Bu lisans, veri setinin yalnızca ticari olmayan, araştırma veya akademik amaçlı uygulamalar için kullanılmasına izin vermektedir.
- **Veri Şeması (Ön Analiz):**
    - Severity (Hedef Değişken): Sayısal/Kategorik (1-4 arası). 1 en az etki, 4 en ciddi etki.
    - Start_Time / End_Time: Kaza başlangıç/bitiş zamanı (Zaman Damgası).
    - Start_Lat / Start_Lng: Kaza konumu (Sayısal, GPS Koordinatları).
    - Distance(mi): Kazadan etkilenen yolun uzunluğu (Sayısal, mil).
    - Weather_Condition: Hava durumu (Kategorik).
    - Sunrise_Sunset: Gündoğumu/Günbatımı bilgisi (Kategorik, Gündüz/Gece).
    - Traffic_Signal, Stop, Junction, Crossing: Çevresel faktörler (Boolean).
- **Boyut:**
    - Satır / Sütun: Yaklaşık 7.7 milyon satır ve 46 sütun.
    - Sınıf Dengesi (Hedef Değişken: Severity): Veri setinde ciddi bir sınıf dengesizliği mevcuttur:
        - Derece 1 (En Hafif): 67.366 (~%0.9)
        - **Derece 2:** 6.156.981 (~%79.7)
        - Derece 3: 1.299.337 (~%16.8)
        - Derece 4 (En Ciddi): 204.710 (~%2.6)
- Veri Erişim Planı: Veriler Kaggle platformundan .csv formatında indirilecek ve yerel makinelerde (veya gerekirse Google Colab gibi bulut ortamlarında) işlenecektir.
- Etik, Gizlilik, Önyargı: Veri seti kişisel kimlik bilgileri içermemektedir. Kaza noktalarını belirten coğrafi koordinatlar (Enlem/Boylam) anonimleştirilmemiştir ancak doğrudan kişisel adres bilgisi sağlamamaktadır.

Değişken Adı	Veri Tipi	Beklenen Değerler / Aralık	Açıklama
Severity	int	1–4	Kaza şiddeti seviyesi (1 = Hafif, 4 = Ağır)
Temperature(F)	float	0–120	Fahrenheit cinsinden sıcaklık (şehir medyanı ile doldurulduu)
Wind_Speed(mph)	float	0–100	Rüzgar hızı, mph cinsinden (aykırı değerler kırpıldı)
Distance(mi)	float	0–100	Kaza mesafesi (mile)
Duration(min)	float	0–1000	Kaza süresi (dakika), negatif değerler 0 olarak ayarlandı
Pressure(in)	float	25–32	Basınç (inch Hg), aykırı değerler kırpıldı
Humidity(%)	float	0–100	Nem oranı (%)
Visibility(mi)	float	0–20	Görüş mesafesi (mile)
Month	int	1–12	Kaza ayı
Hour	int	0–23	Kaza saati
DayOfWeek	object	'Monday'–'Sunday'	Kaza günü
Is_Weekend	bool	True/False	Hafta sonu mu?
Is_Rush_Hour	bool	True/False	Yoğun saatlerde mi?
Was_Precipitation	bool	True/False	Yağış oldu mu?
Road_Type	object	'Highway/Main_Road', 'Local_Street'	Yol tipi
Is_Low_Speed_Zone	bool	True/False	Düşük hız bölgesi mi?
FEAT_Is_Road_Closed	bool	True/False	Yol kapalı mı?
FEAT_Is_Lane_Blocked_On_Highway	bool	True/False	Şerit kapalı mı?
FEAT_Is_Local_Slowdown	bool	True/False	Yerel yavaşlama var mı?
FEAT_Is_Minor_Crash	bool	True/False	Küçük kaza mı?
Is_Stop_and_Go_Traffic	bool	True/False	Dur kalk trafiği mi?
Is_High_Energy_Zone	bool	True/False	Yüksek enerji bölgesi (ana yol + kavşak)
Is_Night_Weekend	bool	True/False	Gece + hafta sonu mu?
Bad_Weather	bool	True/False	Kötü hava durumu (yağmur, kar, sis, fırtına vb.)
Is_High_Energy_Accident	bool	True/False	Uzun mesafe ve uzun süreli kaza
Is_Major_Disruption	bool	True/False	(Sadece Deney B) Yol kapalı veya şerit kapalı

6) Keşifsel Veri Analizi (Exploratory Data Analysis – EDA)

6.1. Veri Kalitesi Kontrolleri

Veri seti öncelikle kalite açısından incelenmiştir. Eksik değerler, veri tipine uygun imputation stratejileriyle ele alınacaktır. Yinelemeler tekilleştirilmiş ve veri setindeki tekrarlar temizlenmiştir. Aykırı değerler hem istatistiksel yöntemler hem de görsel incelemeler (boxplot, z-score) ile tespit edilmiştir. Ayrıca, veri sızıntısı riskleri değerlendirilmiş ve ileride modelleme sırasında veri sızıntısını önleyecek önlemler planlanmıştır.

6.2. Dağılımlar ve Denge

Veri setindeki değişkenlerin dağılımları histogram ve density plotlar ile incelenmiştir. Hedef değişkenin dağılımı analiz edilerek sınıf dengesizlikleri tespit edilmiştir. Grup bazlı dağılımlar ve dökümler de incelenmiş, belirli grupların aşırı temsil edilip edilmediği değerlendirilmiştir. Bu analizler, model performansını etkileyebilecek dengesizlikleri ortaya koymaktadır.

6.3. Özellik – Hedef İlişkileri

Özellikler ile hedef değişken arasındaki ilişkiler korelasyon analizleri, karşılıklı bilgi ölçümleri ve basit tek değişkenli testler kullanılarak değerlendirilmiştir. Bu analizler, hangi özelliklerin hedef değişken üzerinde güçlü etkiler yaratabileceğini ve hangi değişkenler arasında yüksek ilişki olduğunu göstermektedir. Özellikler arası yüksek korelasyon, çoklu doğrusallık (multicollinearity) risklerini de ortaya koymaktadır.

6.4. Görselleştirme Planı

Verilerin farklı yönlerini anlamak için kapsamlı görselleştirmeler planlanmıştır. Boxplot ve violin plotlar, aykırı değerleri ve dağılım asimetrilerini gözlemlemek için; pairplot ve scatterplotlar, değişkenler arası ilişkileri ve olası grupları analiz etmek için kullanılmıştır. Heatmap görselleştirmeleri ise korelasyon matrislerini görselleştirerek özellikler arası ilişkileri hızlıca değerlendirmeye olanak sağlamaktadır.

EDA görselleştirmeleri, verilerin farklı yönlerini anlamak için planlandı:

Analiz Türü	**Görselleştirme Önerisi**
Zamansal Yoğunluk	Saat, gün, ay bazında histogram ve lineplot
Coğrafi Dağılım	Heatmap veya Choropleth map
Altyapı Etkisi	Junction, Stop, Traffic_Signal bazlı barplot
Hava Durumu Etkisi	Weather_Condition, Was_Precipitation countplot
Sınıf Dengesizliği	Severity countplot ve oran grafikleri
Metin Özellikleri	Description üzerinden çıkarılan yeni değişkenler için barplot

7) Veri Hazırlama Planı

7.1. Temizleme

Veri setindeki tekrar eden kayıtlar tekilleştirilmiş ve olası anomaliler tespit edilerek işlenmiştir. Ayrıca farklı birimlerde kaydedilmiş ölçümler standardize edilmiştir. Bu adım, verinin tutarlılığını ve modelin doğruluğunu artırmak için kritik öneme sahiptir.

7.2. İmputasyon Stratejisi

Eksik veriler, veri tipine göre ele alınacaktır:

Sayısal değişkenler: Ortalama, medyan veya ileri-doldurma (forward fill) yöntemleriyle imputasyon.

Kategorik değişkenler: En sık görülen kategori ile doldurma veya uygun şekilde “missing” etiketi ekleme.

Zaman değişkenleri: Gerekli durumlarda lineer interpolasyon veya zaman bazlı trend analizleri ile doldurma.

7.3. Dönüşümler

Veri setindeki değişkenler, model performansını artırmak için uygun dönüşümlere tabi tutulacaktır:

Ölçeklendirme ve normalleştirme (StandardScaler, MinMaxScaler).

Log veya Box-Cox transformasyonu ile asimetrik dağılımların düzeltilmesi.

Kategorik değişkenlerin kodlanması: one-hot encoding veya label encoding yöntemleri kullanılarak sayısal forma dönüştürülmesi.

7.4. Özellik Mühendisliği

Yeni özellikler, veri setinin etki alanı bilgisi ve analiz gereksinimlerine göre üretilecektir:

Mevcut özellikler arasında etkileşimler ve toplama/çıkartma gibi kombinasyonlar.

Zaman serisi özellikleri için zaman aralığı ve trend göstergeleri.

Diğer alanlardan elde edilebilecek türetilmiş özellikler.

7.5. Özellik Seçimi ve Boyut İndirgeme

Özellik setinin gereksiz veya yüksek korelasyonlu değişkenlerden arındırılması ve boyut indirgeme yöntemleri:

Filtre yöntemleri: Mutual Information (MI), ANOVA F-testleri.

Wrapper yöntemleri: Recursive Feature Elimination (RFE).

Embedded yöntemler: Lasso veya Tree-based feature importance.

Boyut indirgeme: Principal Component Analysis (PCA) gibi yöntemlerle modelin daha verimli ve hızlı çalışması sağlanacaktır.

8) Model Planlama

8.1 Kullanılabilecek Modeller

Bu projede hedef değişkenin çok sınıflı bir yapıda olması nedeniyle, farklı karmaşıklık seviyelerine sahip çeşitli sınıflandırma algoritmaları değerlendirilmiştir. Başlangıçta **Logistic Regression (Multinomial)** modeli temel bir benchmark olarak kullanılacaktır. Bu model basit, yorumlanabilir ve hızlı bir başlangıç noktası sunar.
Ardından **Random Forest Classifier** gibi ağaç tabanlı yöntemler uygulanarak doğruluk oranı ve özellik önemleri analiz edilecektir. Daha ileri seviyede, **Gradient Boosting algoritmaları (XGBoost, LightGBM, CatBoost)** test edilerek daha yüksek tahmin performansı elde edilmeye çalışılacaktır.
Karmaşık karar sınırları veya doğrusal olmayan ilişkilere sahip veri yapılarında **Support Vector Machine (SVM)** modeli değerlendirilecektir. Büyük ve karmaşık veri kümelerinde ise **Yapay Sinir Ağları (MLPClassifier)** non-lineer ilişkileri modelleyebilme yeteneğiyle kullanılacaktır.

8.2 Hiperparametre Planlaması

Model performansını en üst düzeye çıkarabilmek için hiperparametre optimizasyonu gerçekleştirilecektir. Bu süreçte **Grid Search** ve **Random Search** yaklaşımlarından biri veya her ikisi kullanılacaktır.

- **Logistic Regression:** C (regularization), solver, max_iter
- **Random Forest:** n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features
- **XGBoost:** n_estimators, max_depth, learning_rate, subsample, colsample_bytree
- **LightGBM:** num_leaves, max_depth, learning_rate, n_estimators, min_data_in_leaf
- **MLPClassifier:** hidden_layer_sizes, activation, solver, alpha, learning_rate

Bu parametrelerin optimize edilmesiyle modelin hem doğruluk hem de genellenebilirlik performansı artırılacaktır.

8.3 Sınıf Dengesizliği Stratejisi

Veri setinde sınıf dengesizliği mevcut olduğu için, modelin azınlık sınıfları doğru tanıyabilmesi amacıyla **SMOTE (Synthetic Minority Over-sampling Technique)** yöntemi uygulanacaktır. Bu yöntem, azınlık sınıfları için sentetik örnekler üreterek sınıflar arasındaki dengesizliği azaltır.
Bu sayede modelin yalnızca baskın sınıfları değil, nadir görülen sınıfları da doğru şekilde tahmin etmesi sağlanır. SMOTE yöntemi, model eğitiminden önce yalnızca eğitim verisine uygulanarak test verisinde sızıntı riskinin önüne geçilecektir.

8.4 Model Seçim Stratejisi

Modelleme süreci aşamalı olarak yürütülecektir. İlk adımda baseline performans ölçümü için Logistic Regression modeli çalıştırılacak, ardından Random Forest ve XGBoost gibi güçlü modellerin sonuçlarıyla karşılaştırılacaktır.
Model performansları doğruluk (accuracy), F1-score ve ROC AUC gibi metriklerle kıyaslanacak, sınıf dengesizliği dikkate alınarak SMOTE ile yeniden örnekleme yapılacaktır. En iyi performans veren model, hiperparametre optimizasyonu ve olası **ensemble (model birleşimi)** yaklaşımlarıyla nihai hale getirilecektir.

9) Değerlendirme Tasarımı (Evaluation Design)

9.1 Kullanılan Metrikler

Model performansını ölçmek için birden fazla değerlendirme metriği kullanılacaktır. Çünkü tek bir metrik, özellikle dengesiz sınıflı veri setlerinde, modeli doğru değerlendirmede yetersiz kalabilir.
Bu nedenle proje kapsamında **F1-score**, **Precision (Kesinlik)**, **Recall (Duyarlılık)** ve **ROC-AUC** metrikleri temel alınacaktır.
Dengesiz veri yapısı göz önünde bulundurulduğunda, **PR-AUC (Precision-Recall Area Under Curve)** metriği de özellikle küçük sınıfların performansını ölçmede önemli bir kriter olarak kullanılacaktır.
Doğruluk (Accuracy) metriği referans amaçlı izlenecek, ancak tek başına karar ölçütü olmayacaktır.

Her modelin çıktısı bu metriklerle karşılaştırılarak, hem genel doğruluk hem de azınlık sınıfları üzerindeki başarı dengesi değerlendirilecektir.

9.2 Doğrulama (Validation) Protokolü

Model değerlendirme sürecinde **train-test bölünmesi** ve **cross-validation (çapraz doğrulama)** stratejileri uygulanacaktır.
Veri seti rastgele, ancak sınıf oranlarını koruyacak şekilde **stratified train-test split** yöntemiyle %80 eğitim, %20 test olarak ayrılacaktır.
Eğitim sürecinde modelin genellenebilirliğini artırmak amacıyla **5-fold Cross Validation** yöntemi kullanılacaktır.
Böylece modelin farklı veri bölümleri üzerinde tutarlı performans gösterip göstermediği gözlemlenecektir.

Zaman bileşeninin önemli olduğu durumlarda (örneğin, kazaların zaman serisi etkileri inceleniyorsa), **Time Series Split** yöntemi de alternatif bir doğrulama planı olarak değerlendirilecektir.
Veri sızıntısı (data leakage) riskine karşı, SMOTE gibi veri dengesizliği teknikleri yalnızca eğitim verisine uygulanacaktır. Test setine herhangi bir yapay örnek eklenmeyecektir.

9.3 Hata Analizi

Modelin yanlış sınıflandırma yaptığı örneklerin yapısını anlamak için ayrıntılı bir **hata analizi (error analysis)** yapılacaktır.
Her modelin tahmin sonuçları için **Confusion Matrix (Karışıklık Matrisi)** oluşturulacak ve bu matriste her sınıfın doğru ve yanlış tahmin oranları incelenecektir.
Özellikle azınlık sınıflarına ait örneklerin ne sıklıkta diğer sınıflarla karıştırıldığı değerlendirilecek ve bu hataların nedenleri araştırılacaktır.

Ayrıca, **false positives (yanlış pozitifler)** ve **false negatives (yanlış negatifler)** üzerinde odaklanılarak, modelin hangi koşullarda hataya eğilimli olduğu tespit edilecektir.
Bu analizden elde edilen bulgular, modelin yeniden eğitilmesi, parametre ayarlarının yapılması ve özellik mühendisliği aşamalarında geri besleme olarak kullanılacaktır.

10) Riskler ve Azaltma Yöntemleri

10.1 Veri Riskleri

Projede kullanılan veri setiyle ilgili karşılaşılabilecek temel riskler arasında veri erişimi, veri kalitesi, veri boyutu ve sınıf dengesizliği yer almaktadır. Veri erişim riski, kaynaklara sürekli ve güvenli erişim sağlanamaması veya güncel verilerin alınamaması durumunda ortaya çıkar. Veri kalitesi riski ise eksik, hatalı veya tutarsız veri kayıtlarının modelin yanlış öğrenmesine yol açabilmesidir. Veri boyutu riski, veri setinin çok büyük olması durumunda işlem süresi ve bellek kullanımının artmasına neden olabilir. Ayrıca, sınıf dengesizliği riski, bazı sınıfların yeterince temsil edilmemesi sonucu modelin belirli sınıfları göz ardı etmesiyle performans kaybına yol açabilir.

10.2 Yöntem Riskleri

Modelleme ve algoritma kullanımında karşılaşılabilecek riskler arasında overfitting, uzun çalışma süresi, yorumlanabilirlik sorunları ve covariate shift bulunur. Overfitting, modelin eğitim verisine aşırı uyum sağlayarak test verisinde düşük performans göstermesi durumudur. Uzun çalışma süresi, modelin eğitimi veya tahmin süresinin çok uzun olmasıyla ortaya çıkar ve proje verimliliğini düşürebilir. Yorumlanabilirlik sorunu, özellikle karmaşık modellerde modelin karar mekanizmasının anlaşılmasının zor olmasıdır. Covariate shift ise eğitim ve test veri dağılımlarının farklı olması sonucu model performansını olumsuz etkileyebilir.

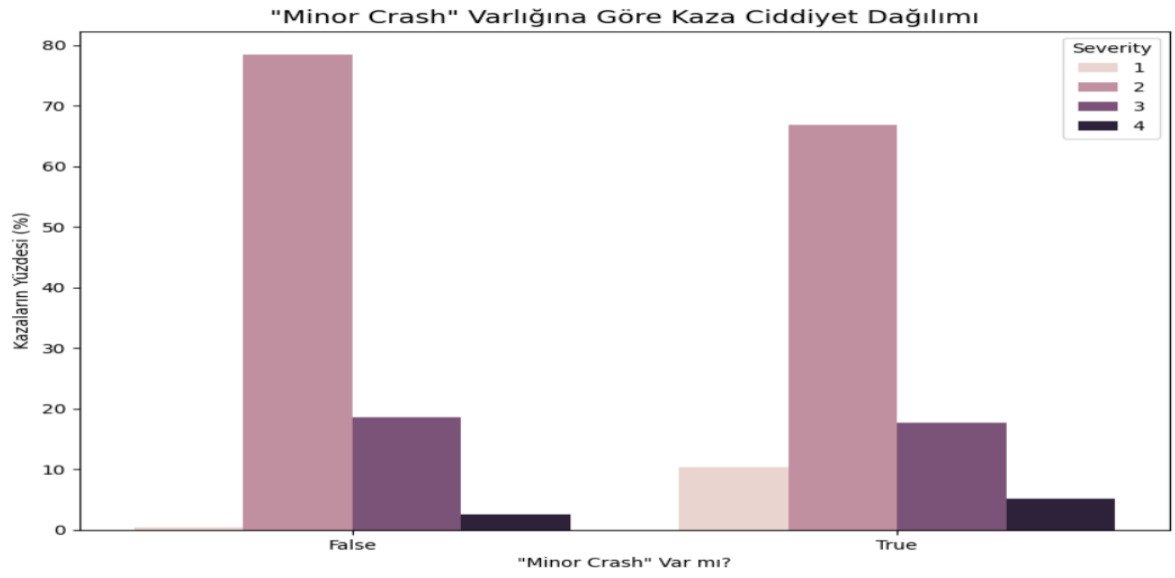
10.3 Azaltıcı Yöntemler

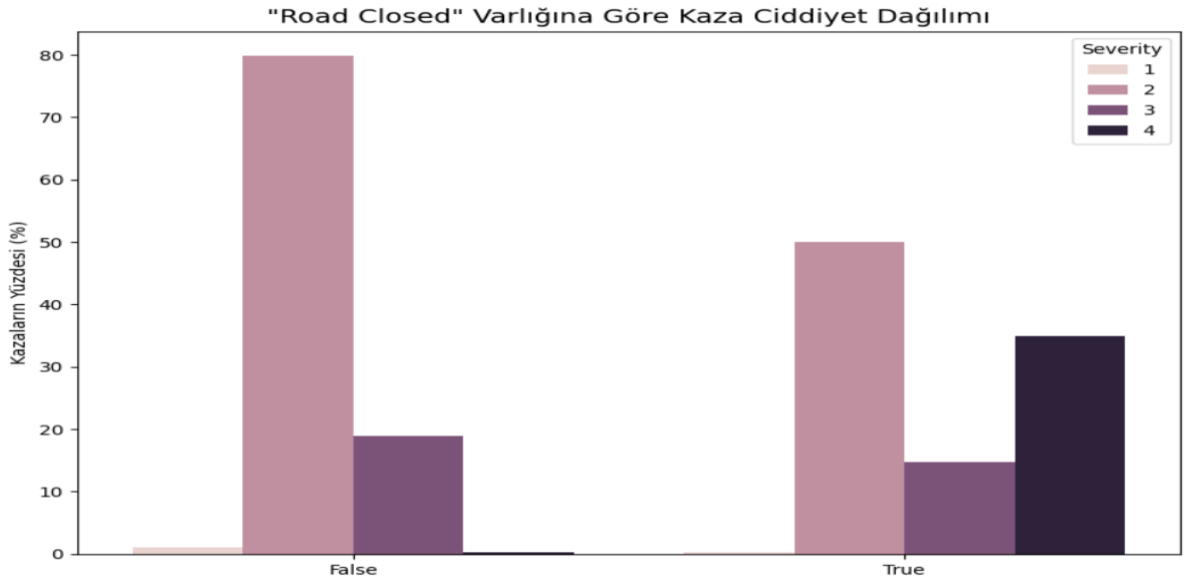
Bu riskleri minimize etmek için çeşitli yöntemler uygulanacaktır. Büyük veri setleriyle çalışırken küçük prototip veri setleri üzerinde hızlı deneyler yapmak, hem işlem süresini azaltacak hem de overfitting riskini düşürecektir. Özellik sayısını azaltmak için PCA veya feature selection yöntemleri kullanılacak, böylece hem overfitting hem de runtime riski kontrol altına alınacaktır. L1 ve L2 düzenlileştirme yöntemleriyle modelin aşırı öğrenmesi önlenecektir. Sınıf dengesizliği için SMOTE veya benzeri oversampling teknikleri uygulanacak, böylece azınlık sınıfların yeterince temsil edilmesi sağlanacaktır. Eğitim ve test veri dağılımları düzenli olarak izlenerek covariate shift riskleri takip edilecektir. Ayrıca, eksik değerler, aykırı değerler ve veri tutarsızlıkları temizlenerek veri kalitesi riskleri azaltılacaktır.

11) Kullanılan Araçlar

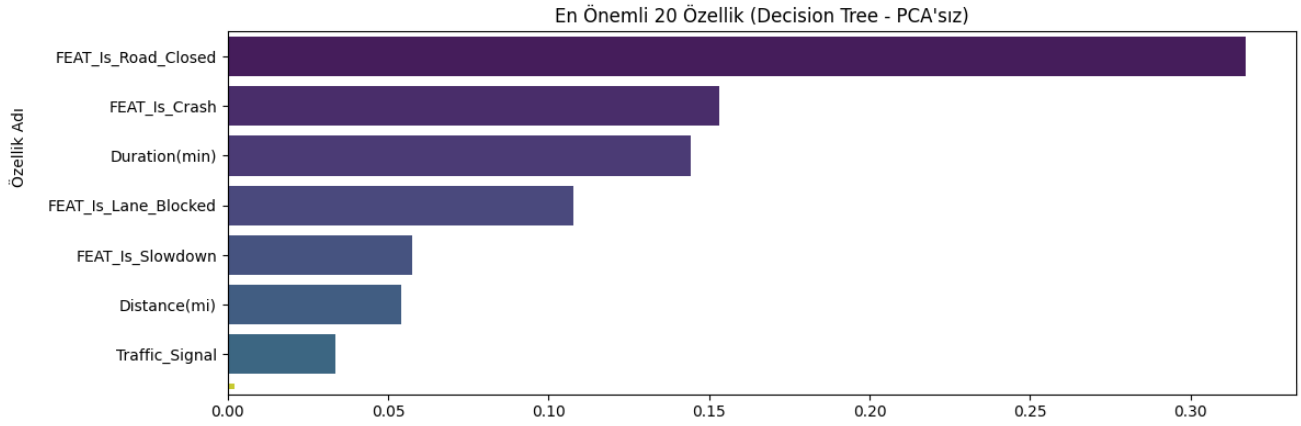
Projede geliştirme ve analiz sürecinde kullanılan ortam Python tabanlıdır. Python'un [versiyon bilgisi buraya eklenebilir] sürümü tercih edilmiş olup, veri işleme, modelleme ve görselleştirme için Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn ve benzeri kütüphaneler kullanılmıştır. Analiz ve model sonuçlarının tekrarlanabilirliği için random seed değerleri sabitlenmiştir. Donanım olarak, model eğitim ve veri işleme süreçleri için [CPU/GPU bilgisi buraya eklenebilir] kullanılması planlanmaktadır.Geliştirilen tüm kodlar ve proje dokümantasyonu GitHub üzerinde depolanmıştır. Bu sayede ekip üyeleri ve paydaşlar kodları inceleyebilir, güncelleyebilir ve projeyi bağımsız olarak çalıştırabilir.Projede kullanılan modellerin eğitimi ve veri işleme adımları için gerekli hesaplama süresi, veri setinin boyutu ve model karmaşıklığı göz önünde bulundurularak planlanmıştır. Büyük veri setleri ve derin öğrenme tabanlı modellerin eğitimi belirli bir süre alabilir; bu nedenle, prototip deneyleri için küçük veri alt kümesi üzerinde hızlı testler yapılması önerilmektedir. Hesaplama kaynaklarının sınırlı olduğu durumlarda, veri boyutu azaltma, özellik seçimi ve erken durdurma (early stopping) yöntemleri kullanılacaktır.

12) Beklenen Sonuçlar ve Görselleştirme Planı

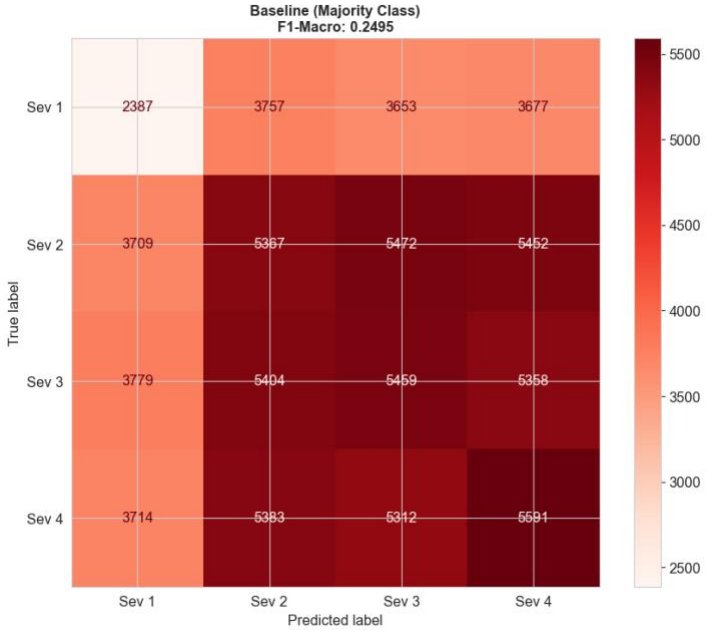


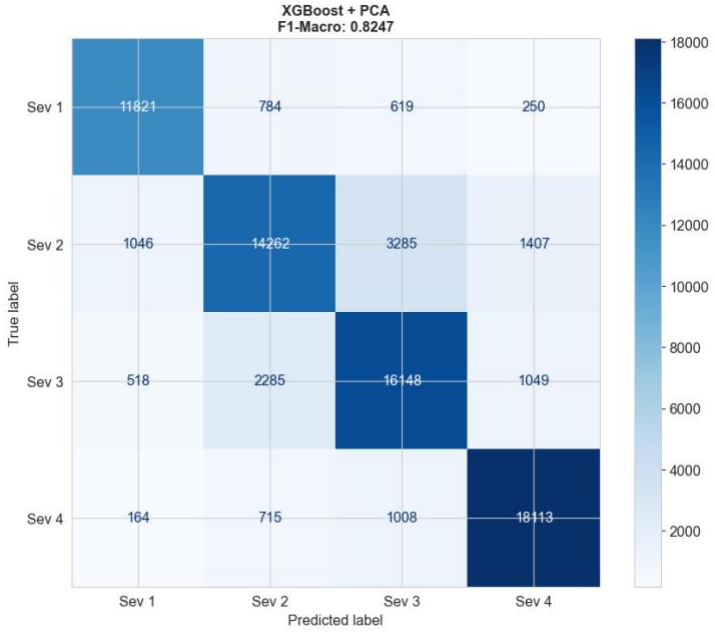


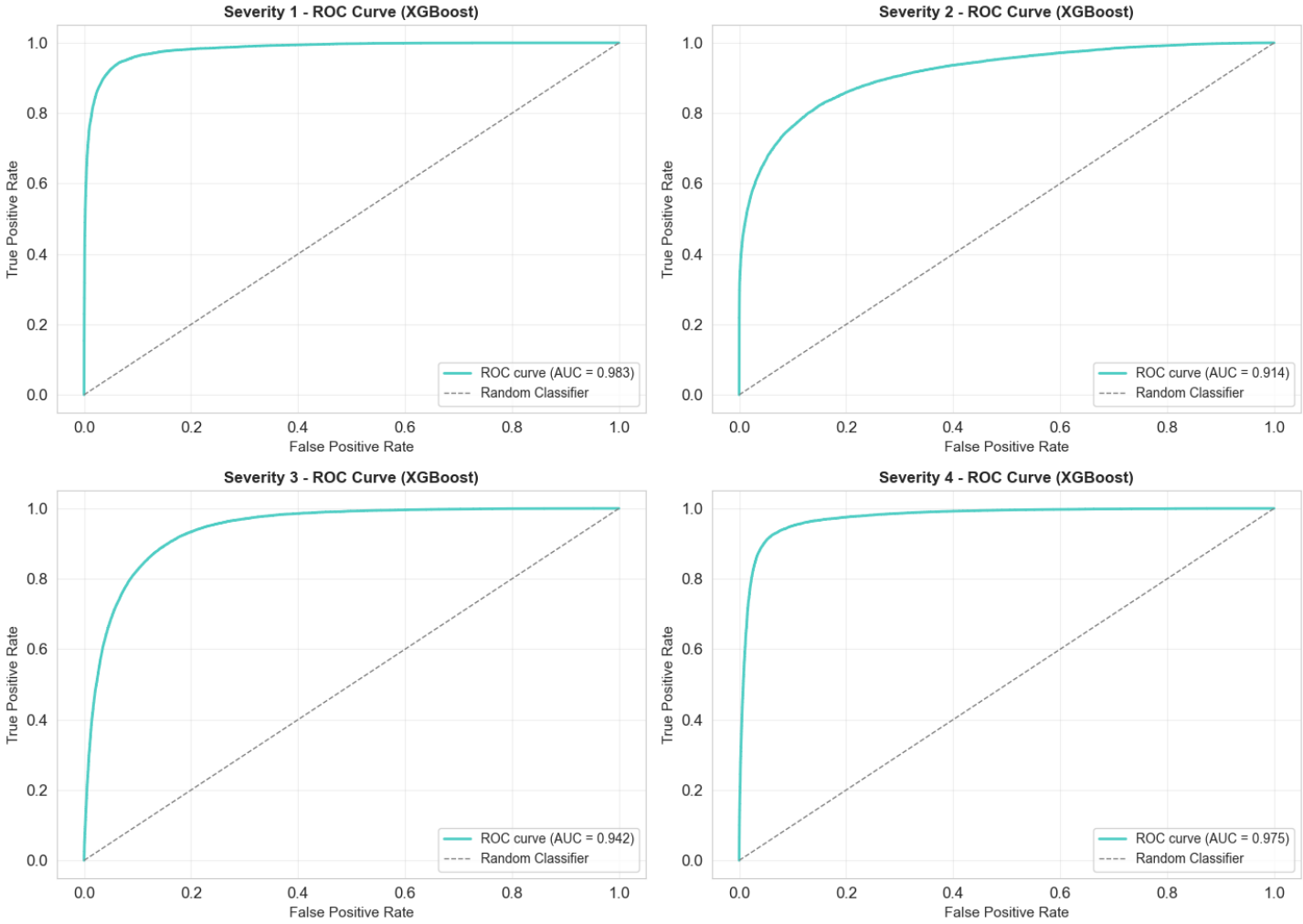
Severity 4 için önemli metrik bulundu.



Üretilen özelliklerin ağırlığının önemi ve 1 milyon sample için karmaşıklık matrisleri.







Dengesiz dağılan verisetinde baskın olan grubun false positive değerlerinin karşılaştırılma plot

13) Referanslar

[1] S. Moosavi, M. H. Samavatian, S. Parthasarathy ve R. A. S. R. B., “A Countrywide Traffic Accident Dataset,” arXiv preprint arXiv:1906.05409, 2019.

[2] S. Moosavi, “US Accidents (2016 - 2023),” Kaggle, 2019. [Çevrimiçi]. Available: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

[3] R. M. A. L. Rathnayaka, et al., “Predicting accident severity: A comparative study of machine learning methods,” 2017.

[4] J. K., et al., “A Machine Learning Approach for Motor Vehicle Crash Severity Prediction,” MDPI, 2020.

[5] M. Amini, et al., “A Novel Explainable Artificial Intelligence Approach for Road Accident Severity Prediction,” MDPI, 2022.

[6] A. Ramya ve S. Reshma, “Accident Severity Prediction Using Data Mining Methods,” IEEE, 2019