FET445 Veri Madenciliği

ABD Trafik kazalarında Şiddet Düzeyi Öngörüsü

GRUP: THE MINE STORM CREW

Youtube Linki:
https://www.youtube.com/watch?v=rnm9h_mRxrI
Github Linki:
https://github.com/ahmetk60/usa-acc

TARİH: 25.12.2025

Problem Tanımı

Trafik kazaları, can ve mal kaybına neden olan, trafik akışını ve şehir lojistiğini doğrudan etkileyen küresel bir sorundur. Bu proje, **ABD Trafik Kazaları (US Accidents)** veri setini kullanarak, bir kazanın gerçekleştiği andaki çevresel, zamansal ve mekânsal koşullara dayanarak **kaza ciddiyetini (Severity)** tahmin etmeyi amaçlamaktadır.

Problem, çok sınıflı bir sınıflandırma (Multi-Class Classification) problemi olarak ele alınmıştır. Hedef değişken olan **"Severity"**, kazanın trafik üzerindeki etkisini ve şiddetini **1 (En Düşük)** ile **4 (En Yüksek)** arasında derecelendirmektedir.

Proje Amacı

Temel amaç, kazaların şiddetini etkileyen faktörleri (hava durumu, yol yapısı, zaman dilimi vb.) analiz etmek ve makine öğrenmesi modelleri ile yüksek doğruluklu tahminler yapmaktır.

(7728394, 46)

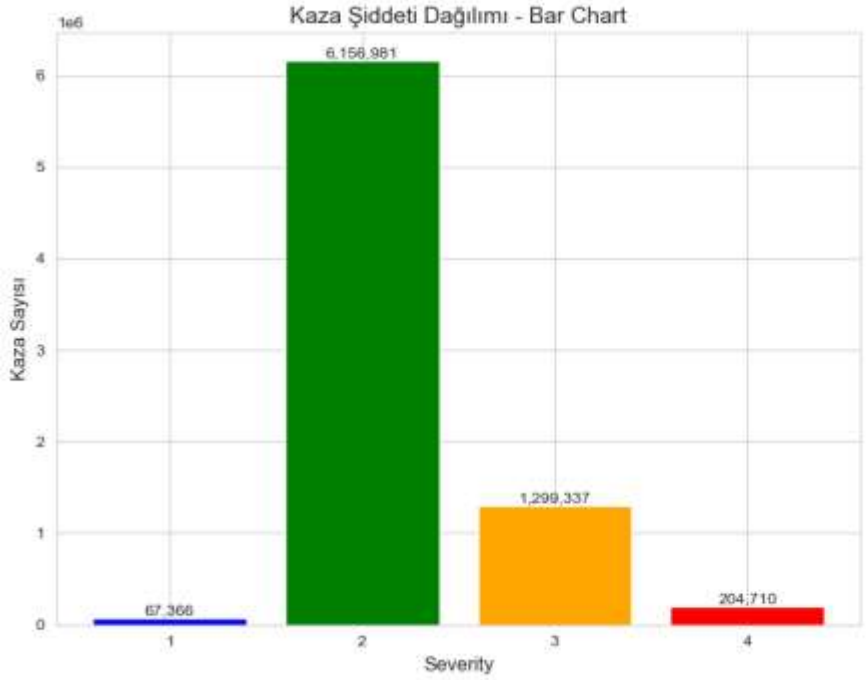
Veriseti: **US Accidents (2016 - 2023)**

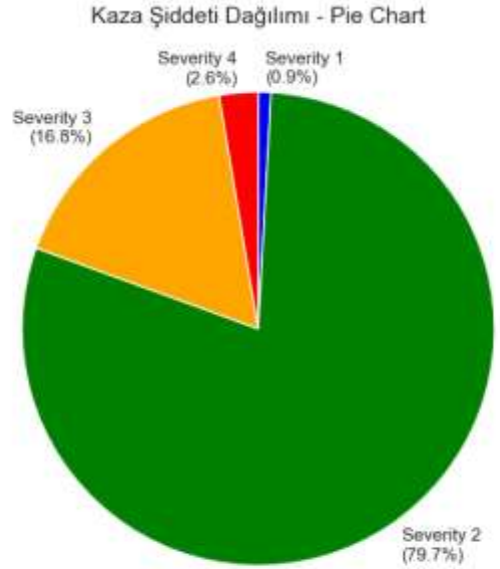
ABD'nin 49 eyaletini kapsayan ülke çapında bir trafik kazası veri setidir. Kaza verileri, anlık trafik olayı verileri sağlayan birden fazla API kullanılarak Şubat 2016 ile Mart 2023 tarihleri arasında toplanmıştır. Söz konusu API'ler; ABD ve eyalet ulaştırma departmanları, kolluk kuvvetleri, trafik kameraları ve yol ağlarındaki trafik sensörleri dahil olmak üzere çeşitli kaynaklar tarafından kaydedilen trafik verilerini iletmektedir. Veri seti, güncel haliyle yaklaşık 7.7 milyon kaza kaydı içermektedir.

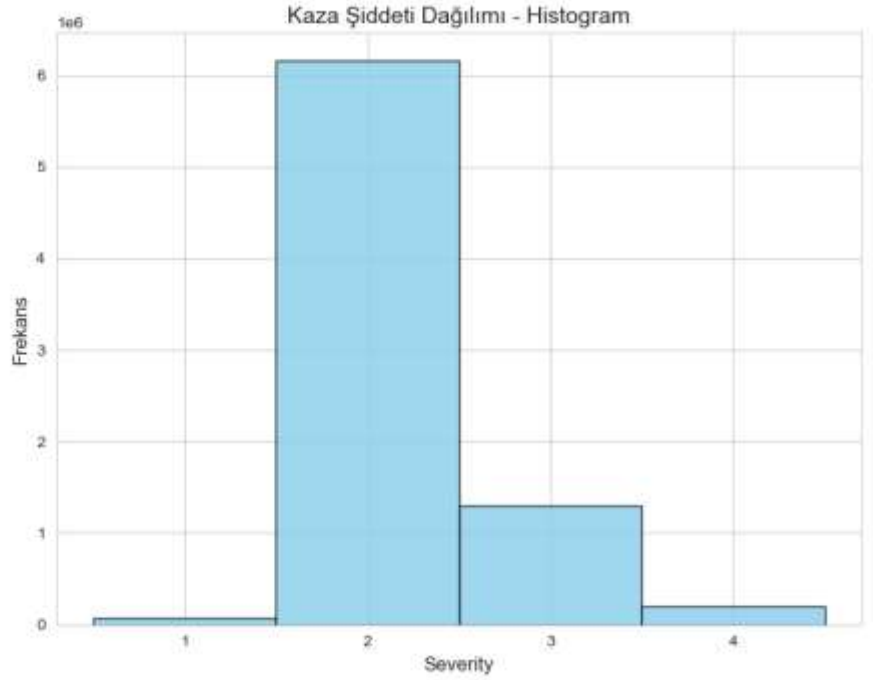
Linki : https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

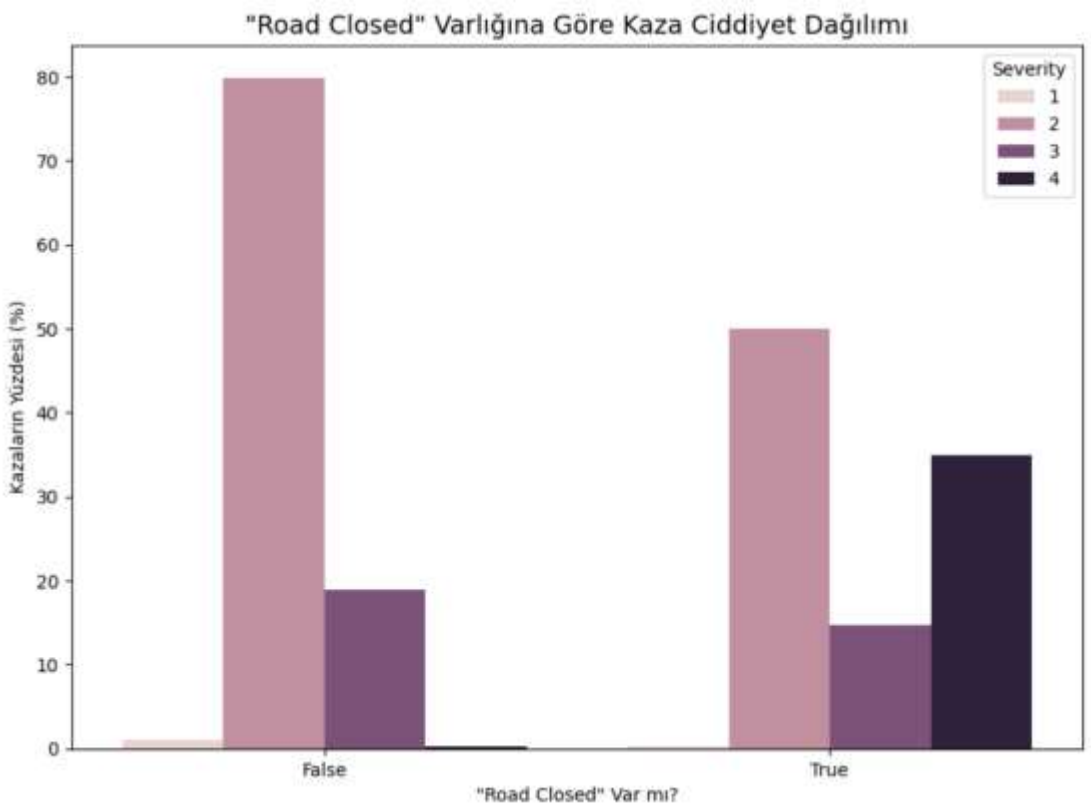
7728394 Satırdan oluşmaktadır, 46 sütun (özellik vardır) numerik ve kategorik veriler vardır.

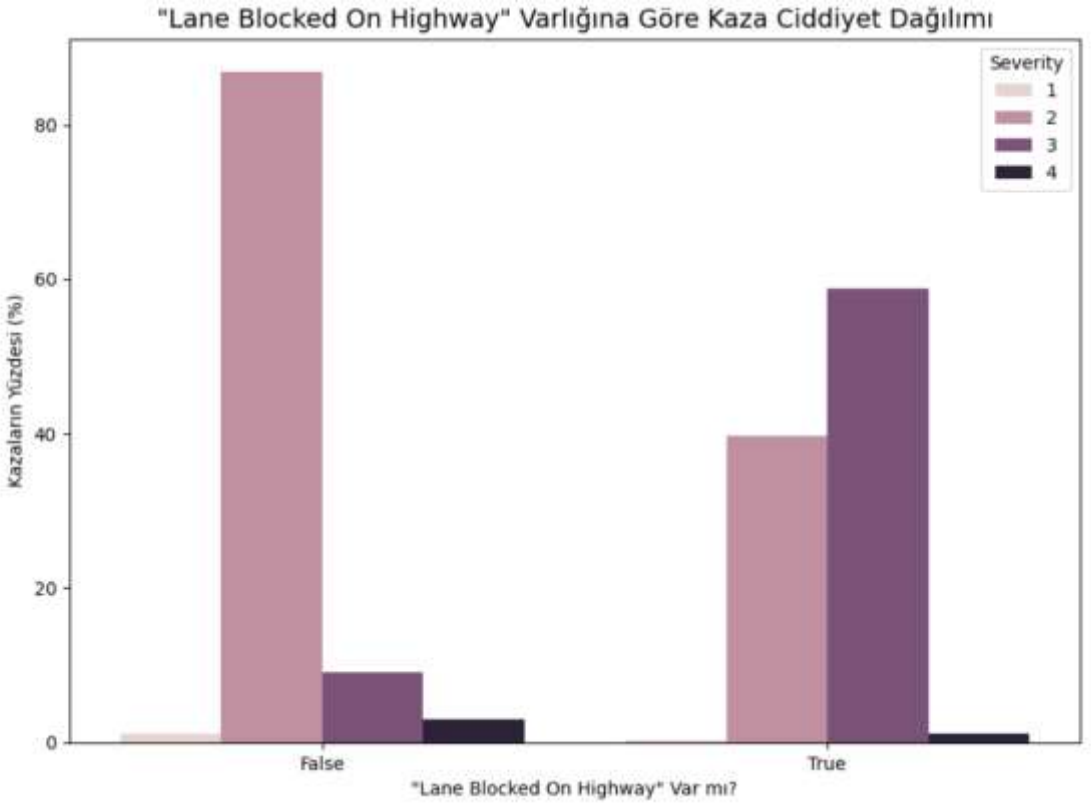
Veri seti Şiddet(severity) dağılımı

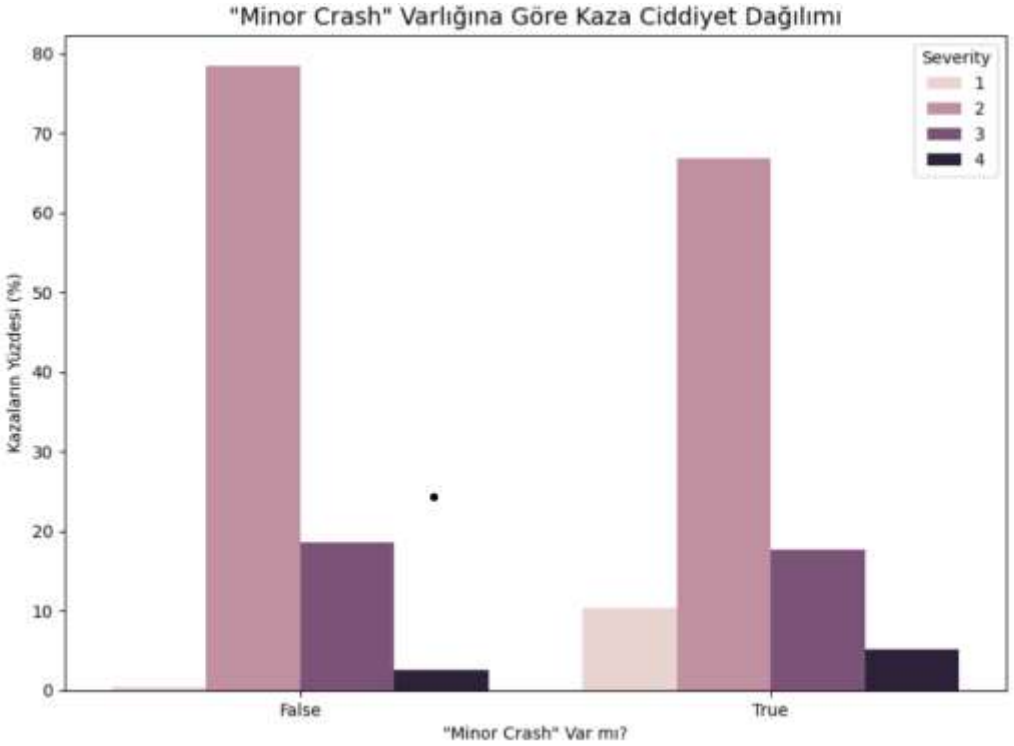


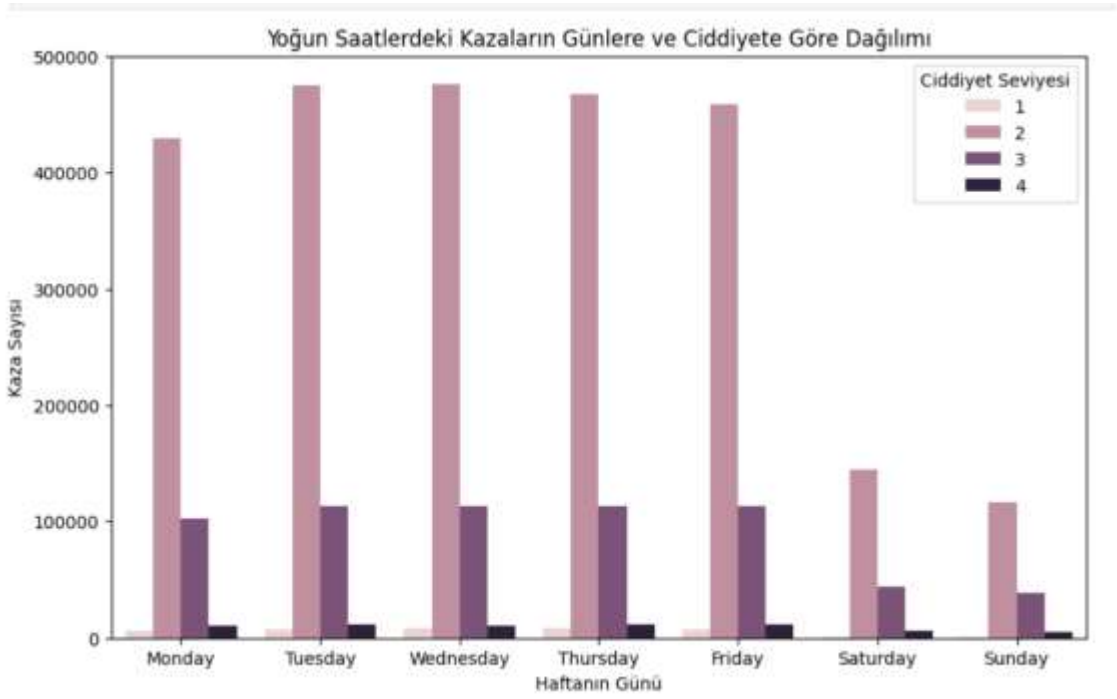





"Road Closed" Varlığına Göre Kaza Ciddiyet Dağılımı
Severity
1
2
3
4
Kazaların Yüzdesi (%)
0
10
20
30
40
50
60
70
80
False
True
"Road Closed" Var mı?


"Lane Blocked On Highway" Varlığına Göre Kaza Ciddiyet Dağılımı
Severity
1
2
3
4
Kazaların Yüzdesi (%)
0
20
40
60
80
False
True
"Lane Blocked On Highway" Var mı?


"Minor Crash" Varlığına Göre Kaza Ciddiyet Dağılımı
Severity
1
2
3
4
Kazaların Yüzdesi (%)
0
10
20
30
40
50
60
70
80
False
True
"Minor Crash" Var mı?


Yoğun Saatlerdeki Kazaların Günlere ve Ciddiyete Göre Dağılımı
Ciddiyet Seviyesi
1
2
3
4
Kaza Sayısı
0
100000
200000
300000
400000
500000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Haftanın Günü

ABD Trafik kazalarında Şiddet Düzeyi Öngörüsü

Train test split oranını her grup üyesi %80-%20 seçmiştir ve data leakage olmaması için dikkat edilmiştir.

F1 ana metrik olarak seçilmiştir diğer metrikler;

- F1-Score
- Precision
- Recall
- Accuarcy

En iyi model Catboost 2m örneklem

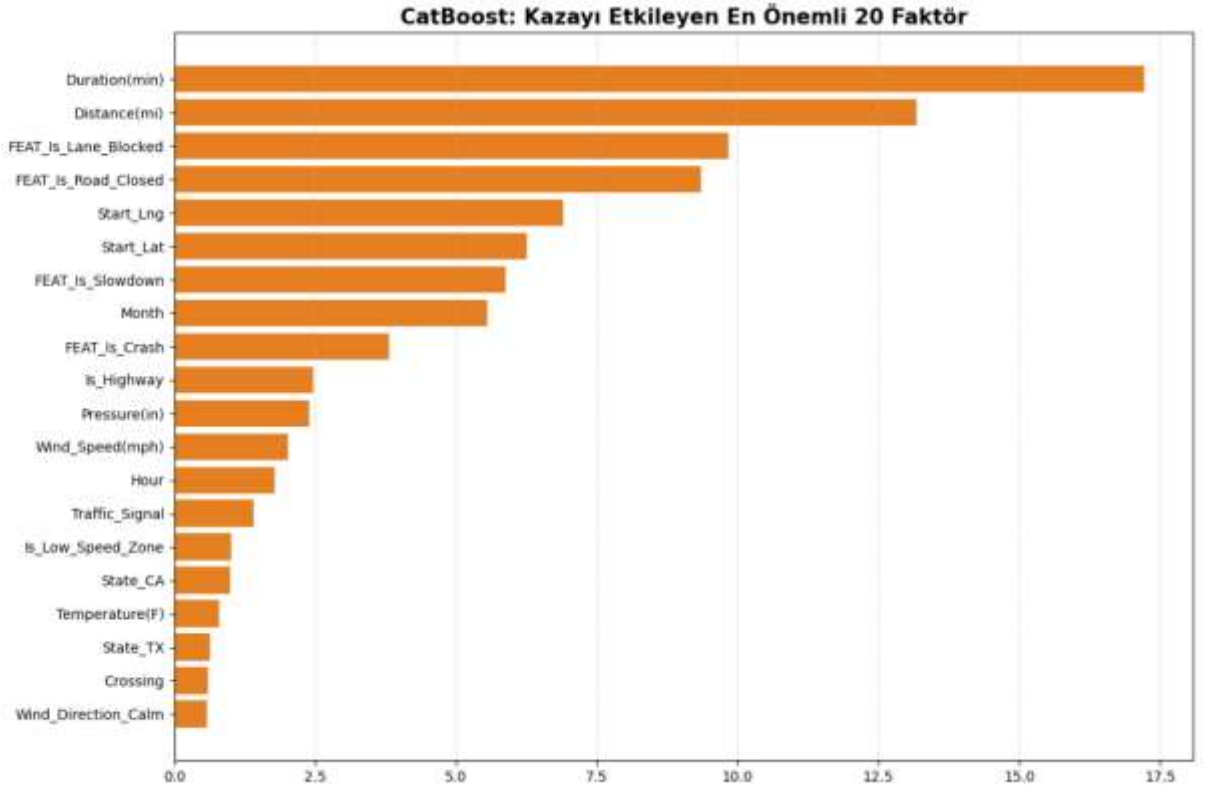
```
############################################################
### CATBOOST MODEL DEĞERLENDİRMESİ ###
############################################################

🚀 CatBoost F1-Macro Score: 0.8344

📋 Classification Report:
              precision    recall  f1-score   support

  Severity 1       0.84      0.76      0.80     13473
  Severity 2       0.89      0.85      0.87    320000
  Severity 3       0.83      0.89      0.86    259868
  Severity 4       0.80      0.82      0.81     40942

    accuracy                           0.86    634283
   macro avg       0.84      0.83      0.83    634283
weighted avg       0.86      0.86      0.86    634283
```



1. Model Geliştirme Yaklaşımları

Büyük ölçekli veri setiyle çalışırken performansı artırmak ve işlem yükünü dengelemek için şu stratejiler uygulanmıştır:

- **Stratejik Alt Örnekleme:** Veri setinin %80'ini oluşturan *Severity 2* sınıfı 1.6 Milyon örneğe indirilerek dengesizlik azaltılmış, diğer sınıflar (*Severity 1, 3, 4*) korunmuştur.
- **Gelişmiş Özellik Mühendisliği:**
  - **Metin Madenciliği:** 'Description' sütunundan kaza nedenini belirten (Closed, Blocked vb.) yeni özellikler türetilmiştir.
  - **Yol ve Zaman Analizi:** Otoban/sokak ayrımı (*Is_Highway*) ve iş çıkış saati/hafta sonu (*Is_Rush_Hour*, *Is_Weekend*) gibi davranışsal özellikler oluşturulmuştur.
- **Kategorik Veri Yönetimi:** CatBoost'un yerleşik yetenekleri kullanılarak One-Hot Encoding ihtiyacı minimize edilmiştir.

2. Hiperparametreler

En iyi pratiklere dayalı manuel yapılandırma ile şu parametreler seçilmiştir:

iterations: 500 (Dengeli öğrenme döngüsü).

learning_rate: 0.1 (Kararlı yakınsama hızı).

depth: 8 (Karmaşık ilişkiler için orta-derin ağaç yapısı).

loss_function: 'MultiClass'

eval_metric: 'TotalF1'

early_stopping_rounds: 30

3. Feature Set

Modelin eğitiminde kullanılan ve ham veriden türetilen temel öznitelikler şunlardır:

1.) Türetilmiş Yol Durumu: FEAT_Is_Road_Closed (Yol kapalı mı?), FEAT_Is_Lane_Blocked (Şerit tıkalı mı?), FEAT_Is_Crash (Kaza mı?), FEAT_Is_Slowdown (Yavaşlama var mı?).

2.) Konum ve Yol Tipi: Is_Highway (Otoban/Ana Yol), Is_Low_Speed_Zone (Sokak/Cadde), Start_Lat, Start_Lng, Distance(mi).

3.) Zaman Bilgisi: Hour (Saat), Month (Ay), Is_Rush_Hour (Trafik saati mi?), Is_Weekend (Hafta sonu mu?), Duration(min) (Kaza süresi).

4.) Hava Koşulları: Temperature(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Speed(mph), Weather_Condition.

5.)Trafik İşaretçileri: Traffic_Signal, Junction (Kavşak), Crossing (Geçit)

Modellerin Karşılaşılaştırılması

Model	Accuracy	F1-Macro	Recall	Precision
LightGBM(Ahmet)	0.88	0.7137	0.88	0.64
CatBoost (Ahmet)	0.87	0.6991	0.87	0.63
AdaBoost(İlkay)	0.79	0.7420	0.76	0.74
CatBoost (İlkay)	**0.86**	**0.8344**	**0.83**	**0.84**
RandomForest(İrem)	0.54	0.36	0.69	0.38
GradientBoostin(İrem)	0.74	0.46	0.61	0.43
ExtraTreesClass(Rabia)	0.55	0.34	0.50	0.34
HistGradientB(Rabia)	0.56	0.34	0.51	0.34
XGboost(Sıla)	0.78	0.62	0.57	0.73
Bagging Classifier(Sıla)	0.83	0.72	0.68	0.77

Sonuç ve değerlendirme

1. Temel Strateji & Metodoloji

Veri Yönetimi: Aşırı sınıf dengesizliği, "Stratejik Alt Örnekleme" ile çözüldü. Çoğunluk sınıfı (Sev 2) optimize edilirken, kritik azınlık sınıfların (%100) tamamı korundu.

Özellik Mühendisliği: Metin madenciliği (Description) ve zamansal analizlerle (Rush Hour, Weekend) veriye derinlik kazandırıldı.

2. Şampiyon Model: CatBoost

Performans: 0.86 Accuracy ve 0.83 F1-Macro skoru ile en başarılı model oldu.

Güvenilirlik: Nadir görülen ciddi kazaları yakalama (Recall) başarısı %80'in üzerine çıktı.

3. Kazayı Etkileyen Kritik Faktörler

Süre ve Mesafe: Kazanın etki süresi (Duration) ve uzunluğu (Distance).

Yol Durumu: Yolun kapanması (Road_Closed) veya şeridin tıkanması.

Yapısal Faktörler: Olayın otobanda (Highway) veya şehir içinde olması.

Genel Yargı: Geliştirilen model, trafik kazalarının ciddiyetini yüksek doğrulukla tahmin ederek, acil müdahale ve trafik yönetimi için güvenilir bir karar destek sistemi sunmaktadır.