# FET445 VERİ MADENCİLİĞİ

## Trafik Kazalarından Yaralanma Şiddetinin Tahmini

## Grup: Luminia

**Video Link :** https://youtu.be/el7oMJSitf0?si=HYJjJK01avN6Cc5_

Tarih: 25.12.2025

# Trafik Kazalarından Yaralanma Şiddetinin Tahmini

Trafik kazaları; hava koşulları, yol yapısı, hız ve zaman gibi birçok faktörün etkileşimiyle meydana gelmektedir. Bu projede, bu değişkenler kullanılarak trafik kazalarındaki yaralanma şiddetini tahmin eden bir makine öğrenmesi modeli geliştirilmiştir. Amaç yaralanma şiddetinin önceden öngörülmesi ve kazalara yol açan riskli koşulların daha iyi anlaşılması hedeflenmektedir.

# VERİ SETİ

**Veri Seti Kaynağı:**

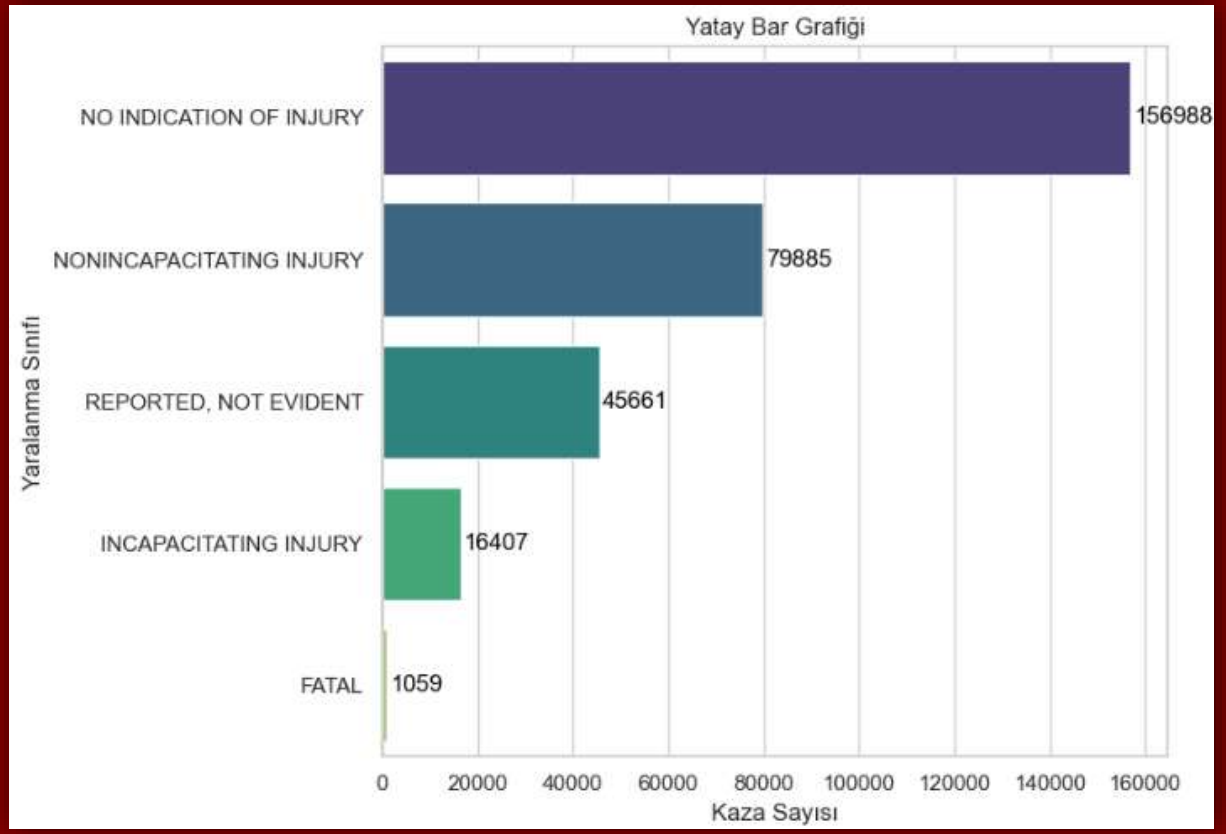
Bu projede Chicago Traffic Crashes veri seti kullanılmıştır. Veri seti, Chicago şehrinde gerçekleşen trafik kazalarına ait ayrıntılı kayıtları içermektedir. Veri seti Chicago Open Data Portal üzerinden temin edilmiştir.

**Kaynak Linki:**
https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if

**Boyut:**

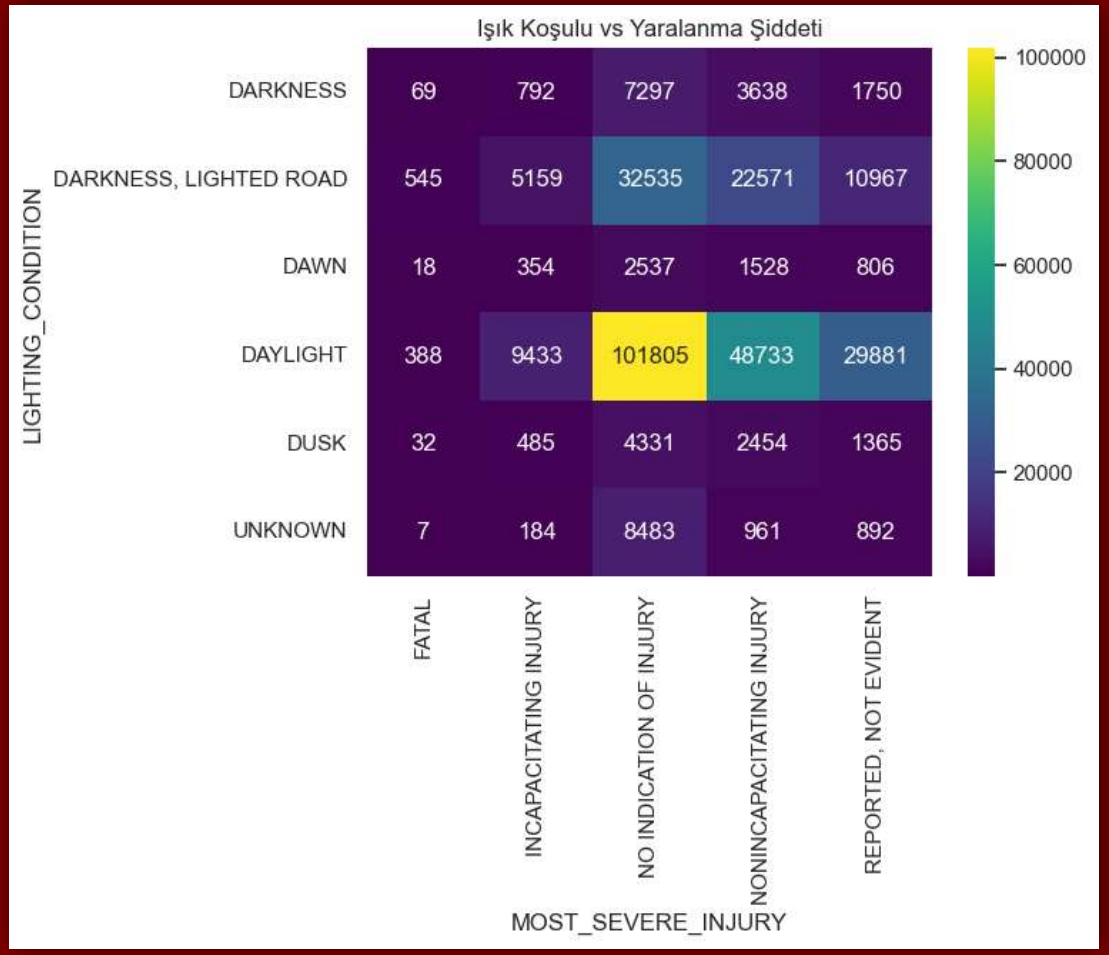
Veri seti, 300.000 trafik kazası kaydını içermekte olup toplam 48 özellikten (sütundan) oluşmaktadır. Bu özelliklerin 31 tanesi kategorik 17 tanesi sayısaldır.

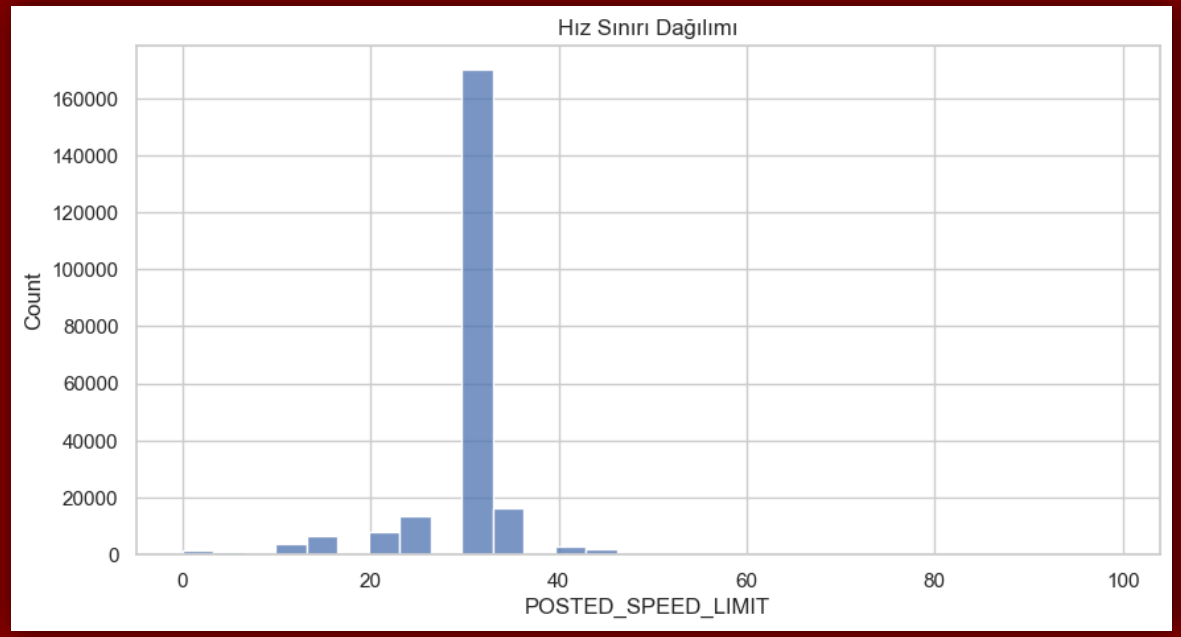
- **Hedef Değişkenin Sınıf Dağılımı:**

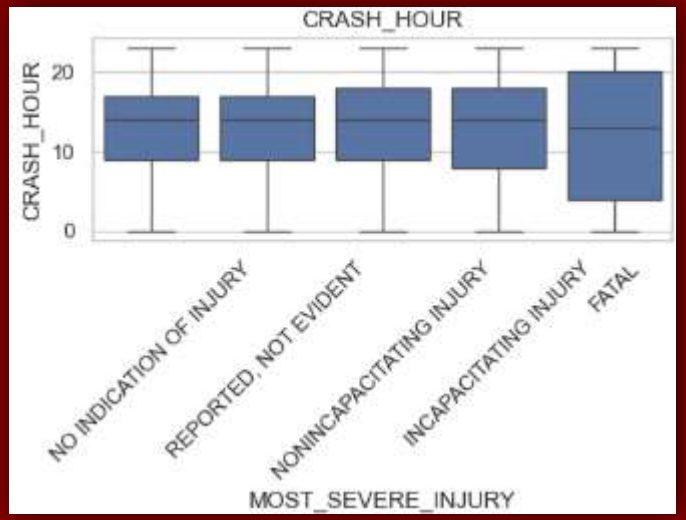


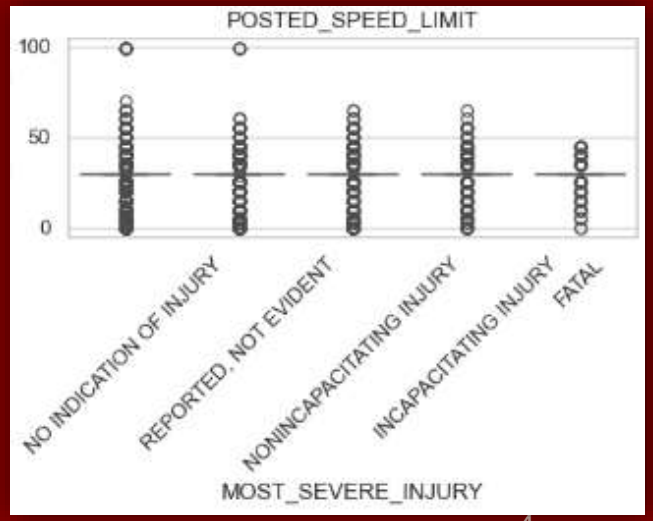

- Veri seti sınıf dağılımı açısından oldukça dengesizdir; “NO INDICATION OF INJURY” sınıfı baskınken “FATAL” ve benzeri kritik sınıflar oldukça az gözleme sahiptir.

# Veri Setindeki Özelliklerden Bazı Örnekler:



# Modelleme Yaklaşımı ve Deneysel Tasarım

- **Modelleme Stratejisi:** Temel olarak Logistic Regression, CatBoost, Decision Tree, XGBoost, LightGBM ,MLP, AdaBoost, ExtraTrees, GradientBoosting gibi modeller kullanılmıştır.
- **Veri Dengesizliği:** Dengesiz sınıf dağılımı SMOTE yöntemiyle ele alınmıştır.
- **Feature Engineering:** Zaman (gündüz/gece, yoğun saatler), hız grupları, mevsim ve etkileşim (hız × saat) özellikleri üretilmiştir.
- **Dönüşümler:** Sayısal değişkenler ölçeklendirilmiş, kategorik değişkenler One-Hot Encoding ile kodlanmıştır.
- **Feature Selection:** En bilgilendirici değişkenleri seçmek için Mutual Information(MI), SelectKBest gibi yöntemler uygulanmıştır.
- **Veri Bölme:** Veri seti %80 eğitim – %10 doğrulama – %10 test olarak ayrılmıştır.
- **Performans Metrikleri:** Model başarımı Accuracy, Macro F1-score, Recall ve Precision metrikleriyle değerlendirilmiştir.

# MODELLER

## BEST MODEL 1 : XGBoost

**Model geliştirilirken kullanılan yaklaşımlar:** Veri seti üzerinde öncelikle eksik değer analizi yapılmıştır. Kategorik değişkenler OneHotEncoder ile sayısal forma dönüştürülmüş, sayısal veriler ise StandardScaler ile ölçeklendirilmiştir. Veri sızıntısını önlemek amacıyla yaralanma şiddetiyle doğrudan ilişkili olan "INJURIES_TOTAL", "INJURIES_FATAL" gibi sütunlar eğitim öncesinde veri setinden çıkarılmıştır.

**Kullanılan Feature Engineering Yöntemleri:**

1. Kategorik Değişkenlerin Sayısallaştırılması
2. Ölçeklendirme (Scaling)
3. Gürültü ve Aykırı Değerlerin Temizlenmesi

**Hyper parametreler:** Optimizasyon işlemi için RandomizedSearchCV tekniği kullanılmıştır.

## BEST MODEL 2 :LightGBM

**Model geliştirilirken kullanılan yaklaşımlar:** Büyük veri setlerinde hızlı çalışması ve karmaşık ilişkileri yakalayabilmesi nedeniyle tercih edilmiştir.Hedef değişkende sınıf dengesizliği bulunduğu için Imbalanced-learn Pipeline kullanılmıştır. Model, dengesiz veriyle uyumlu olacak şekilde eğitilmiştir. Stratified K-Fold Cross Validation kullanılmıştır.

**Feature Engineering:**

- One-Hot Encoding
- StandardScaler
- Veri Temizleme
- Sınıf Dengesini Sağlama

**Hyper parametreler:** LightGBM modeli için Optuna kullanılarak manuel hyperparametre optimizasyonu yapılmıştır.

## BEST MODEL 3 :Extra Trees

Bu çalışmada sınıflandırma problemi için ensemble learning tabanlı bir yöntem olan Extra Trees Classifier kullanılmıştır. Extra Trees, çok sayıda karar ağacının rastgeleleştirilmiş bölünmelerle oluşturulması prensibine dayanır ve özellikle yüksek boyutlu veri setlerinde güçlü genelleme performansı sunar.

**Kullanılan Feature Engineering Yöntemleri:**

1. Eksik Değer İşleme
2. Ölçeklendirme (Scaling)
3. Gürültünün Azaltılması

**Feature Selection:** Extra Trees Classifier'ın kendi içinde yer alan feature_importances_ mekanizması kullanılmıştır.

**Hyper parametreler:** Modelin en iyi performansını bulmak için RandomizedSearchCV tekniği kullanılmıştır.

| | Model | Accuracy | F1 Score | Precision |
|---|---|---|---|---|
| 2 | XGBoost | 0.726203 | 0.702609 | 0.733375 |

| Model | Accuracy | F1 Macro | Recall Macro | Precision Macro |
|---|---|---|---|---|
| LightGBM | 0.659983 | 0.417735 | 0.457785 | 0.418773 |
| GradientBoosting | 0.726683 | 0.387411 | 0.401951 | 0.488495 |

| Model | Accuracy | F1 Macro | Recall Macro | Precision Macro |
|---|---|---|---|---|
| ExtraTrees | 0.66 | 0.42 | 0.44 | 0.42 |

**SONUÇLAR:**

- Bu veri seti ve problem için en uygun yaklaşımın XGBoost ve GradientBoosting gibi boosting temelli algoritmalar olduğunu gördük. GradientBoosting modeli en yüksek genel doğruluğu sağlarken, XGBoost F1 skoru ve hassasiyet dengesi açısından daha istikrarlı bir performans sergiledi; bu durum veri setindeki sınıfların karmaşıklığını yönetmede bu modellerin daha güçlü olduğunu kanıtladı.
- Veri seti sınıf dağılımı açısından oldukça dengesiz olduğu için çok yüksek doğruluk ve tahmin oranları elde edilemedi.

# TEŞEKKÜR EDERIZ