DERS: VERİ ANALİZİ

KONU: DÖNEM PROJESİ - TITANIC VERİ SETİ ANALİZİ

ÖĞRENCİ: BERAT YILDIZ – EFE IŞIK

Numarası: 23430070072 - 23430070074

TARİH: 12.01.2026

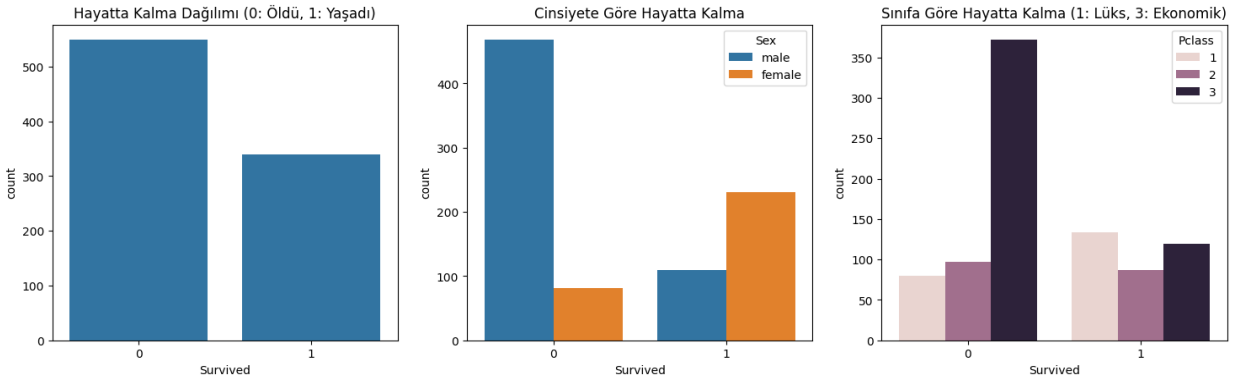
1. GİRİŞ

Bu projenin amacı, tarihin en bilinen deniz kazalarından biri olan Titanic faciasına ait yolcu verilerini analiz etmek ve makine öğrenmesi yöntemleri kullanarak yolcuların hayatta kalıp kalmayacağını tahmin eden bir model geliştirmektir. Projede Kaggle platformundan temin edilen "Titanic - Machine Learning from Disaster" veri seti kullanılmıştır. Bu veri seti, yolcuların yaş, cinsiyet, bilet sınıfı gibi demografik özelliklerini içermektedir.

2. VERİ ANALİZİ VE GÖRSELLEŞTİRME (EDA)

Veri seti toplam 891 yolcuya ait bilgilerden oluşmaktadır. Analiz sürecinde verinin dağılımı incelenmiş ve şu içgörüler elde edilmiştir:

- Genel Durum: Yolcuların büyük çoğunluğu hayatını kaybetmiştir.
- Cinsiyet Faktörü: Kadın yolcuların hayatta kalma oranının, erkek yolculara göre çok daha yüksek olduğu görülmüştür.
- Sınıf Faktörü: 1. Sınıf (lüks) bilet sahibi yolcuların hayatta kalma şansı, 3. sınıf (ekonomik) yolculara göre belirgin şekilde fazladır.



Şekil 1: Hayatta kalma durumunun cinsiyet ve bilet sınıfına göre dağılımı.

3. YÖNTEM

Proje kapsamında Python programlama dili ve Scikit-Learn kütüphanesi kullanılmıştır. Uygulanan adımlar şunlardır:

A. Veri Ön İşleme:

- Eksik Veriler: 'Age' (Yaş) sütunundaki eksik değerler, veri setinin yaş ortalaması ile doldurulmuştur. 'Cabin' sütunu çok fazla eksik içerdiği için analizden çıkarılmıştır.
- Kategorik Dönüşüm: Makine öğrenmesi modellerinin çalışabilmesi için 'Sex' (Cinsiyet) ve 'Embarked' (Liman) verileri sayısal değerlere (0 ve 1) dönüştürülmüştür (Encoding).

B. Kullanılan Algoritmalar:

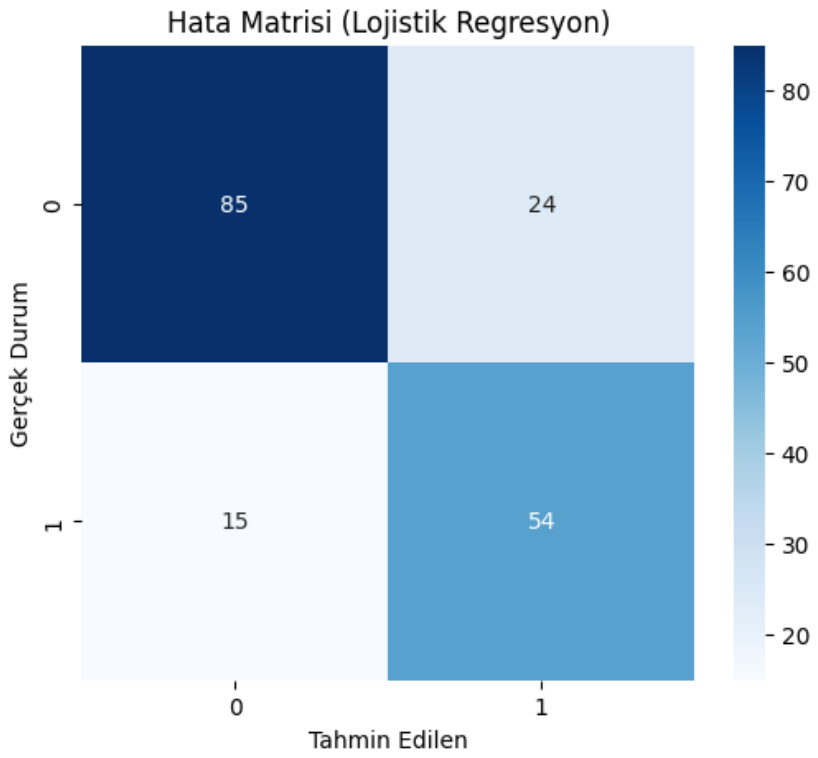
Problemin bir sınıflandırma (Classification) problemi olması nedeniyle iki farklı algoritma seçilmiştir:

1. Lojistik Regresyon (Logistic Regression): Temel sınıflandırma problemlerindeki başarısı nedeniyle seçilmiştir.
2. Rastgele Orman (Random Forest): Daha karmaşık yapıları modelleyebilmesi nedeniyle karşılaştırma amacıyla kullanılmıştır.

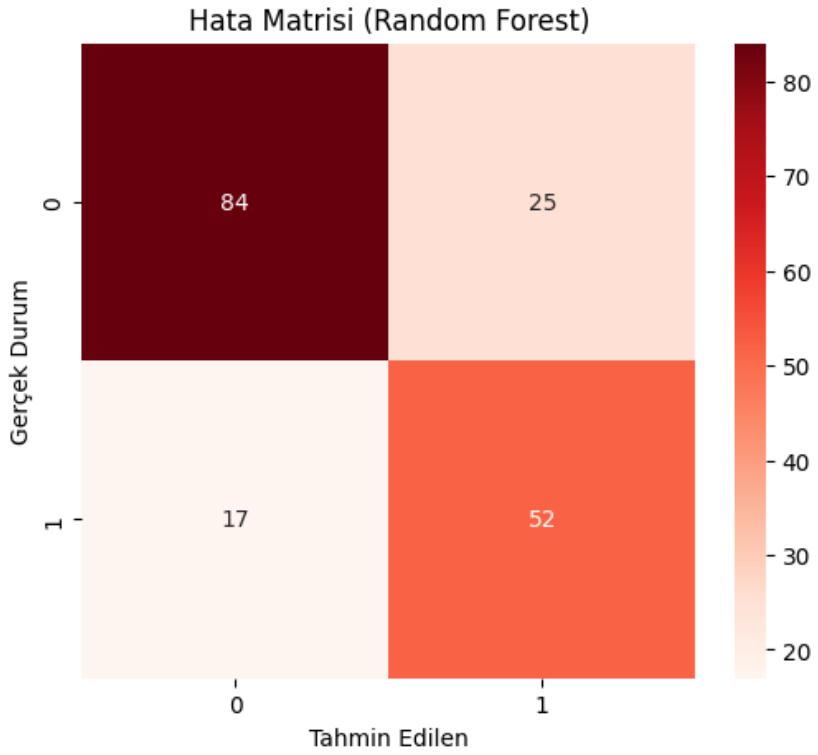
4. SONUÇLAR VE DEĞERLENDİRME

Veri setinin %20'si test verisi olarak ayrılmış ve modeller bu veri üzerinde test edilmiştir. Elde edilen başarı oranları (Accuracy) şöyledir:

- Lojistik Regresyon Başarısı: %78.09
- Random Forest Başarısı: %76.40



Şekil 2: Lojistik Regresyon Hata Matrisi



Şekil 3: Random Forest Hata Matrisi

"Karmaşıklık matrisleri incelendiğinde, ***Lojistik Regresyon*** modelinin hem hayatta kalanları hem de hayatını kaybedenleri tespit etmede **Random Forest** modeline göre daha kararlı olduğu görülmüştür."

5. KAYNAK KODLAR

Github Linki : https://github.com/Brostez/Veri-Analizi-Final-Projesi