Titanik'te Hayatta Kalma Projesi: Dokümantasyon

1. Projenin Amacı

Projenin amacı, Titanik kazasında yolcuların hayatta kalma durumlarını tahmin etmek için makine öğrenimi modelleri geliştirmek ve değerlendirmektir. Bu doğrultuda, veri ön işleme, veri görselleştirme, modelleme ve hiperparametre optimizasyonu adımları izlenmiştir.

2. Veri Seti ve Özellikleri

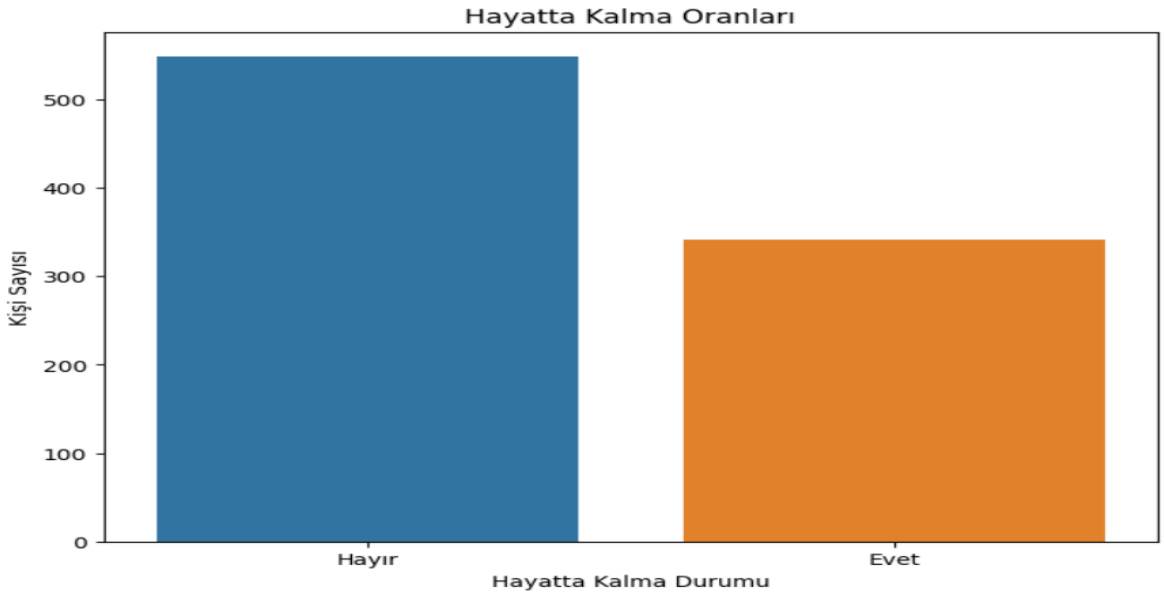
Veri seti, Titanik gemisindeki yolculara ait bilgileri içermektedir. Aşağıdaki özellikler bulunmaktadır:

- **Pclass**: Yolcu sınıfı (1 = Birinci Sınıf, 2 = İkinci Sınıf, 3 = Üçüncü Sınıf)
- **Sex**: Cinsiyet (male = Erkek, female = Kadın)
- **Age**: Yolcunun yaşı
- **SibSp**: Gemiye binen kardeş/eş sayısı
- **Parch**: Gemiye binen ebeveyn/çocuk sayısı
- **Ticket**: Bilet numarası
- **Fare**: Bilet ücreti
- **Cabin**: Kabin numarası
- **Embarked**: Bindiği liman (C = Cherbourg, Q = Queenstown, S = Southampton)
- **Survived**: Hayatta kalma durumu (1 = Hayatta kaldı, 0 = Hayatta kalmadı)

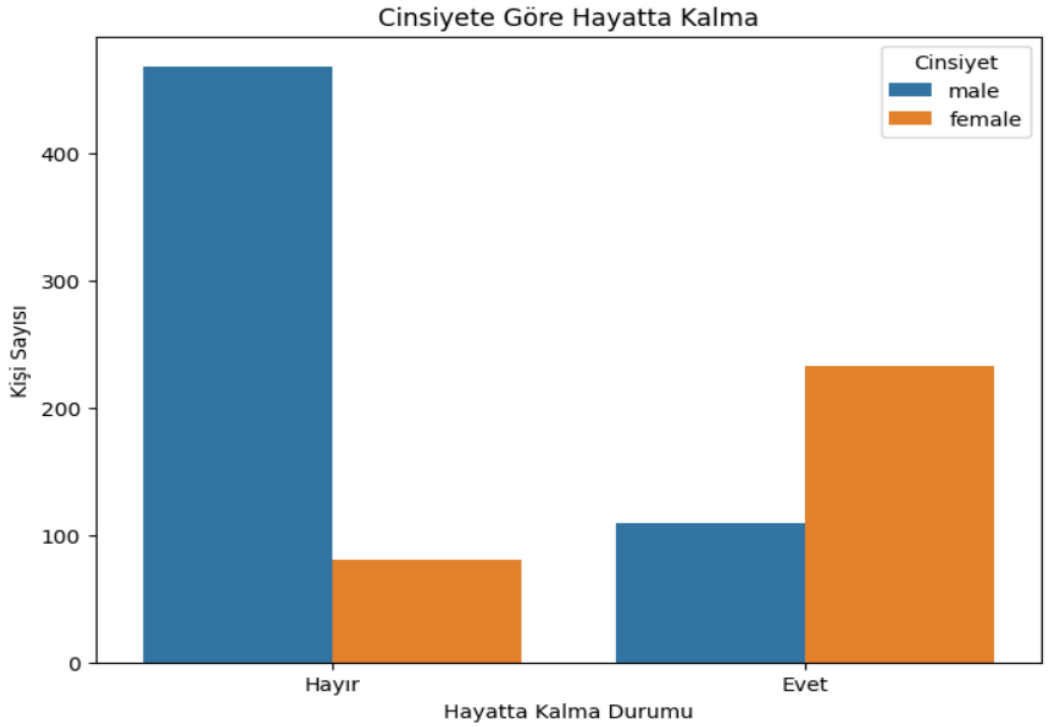
3. Veri Keşfi ve Görselleştirme

3.1. Genel Veri İncelemesi: Veri setinin ilk ve son beş satırı incelenmiştir. Veri setinin boyutları ve değişken tipleri belirlenmiştir. Ayrıca, veri setinin istatistiksel özetine bakılmıştır.

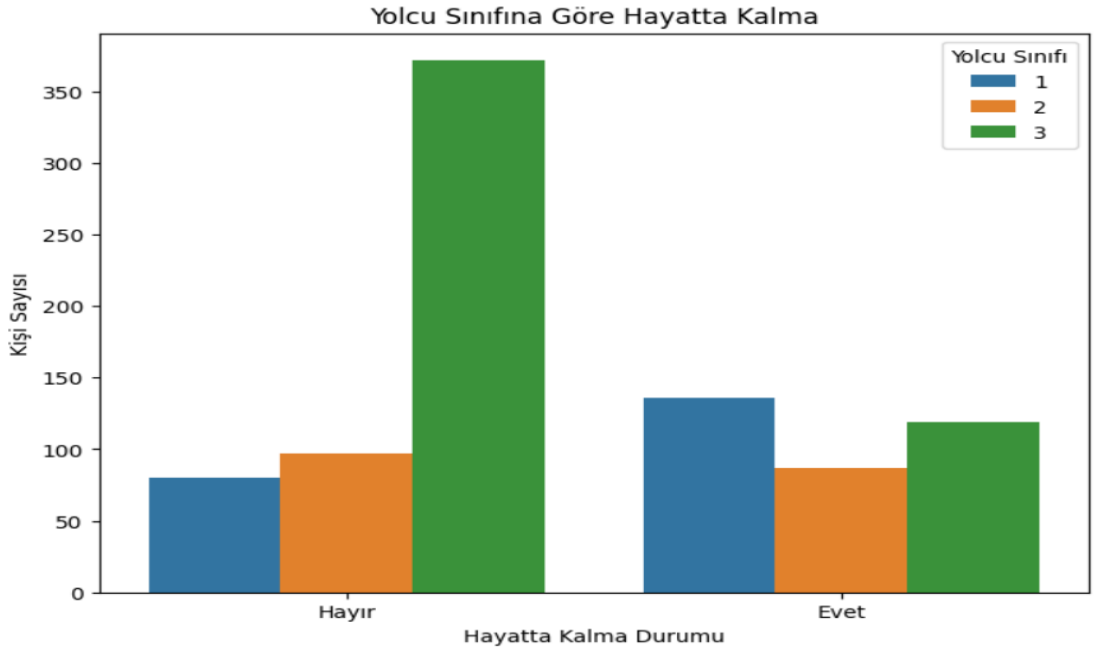
3.2. Hayatta Kalma Oranları: Hayatta kalan ve kalmayan yolcuların sayısal dağılımı ve yüzdesi hesaplanmıştır. Bar grafik kullanılarak bu oranlar görselleştirilmiştir. Bu grafikler, modelin genel performansını ve hayatta kalma oranlarını daha iyi anlamak için önemlidir.



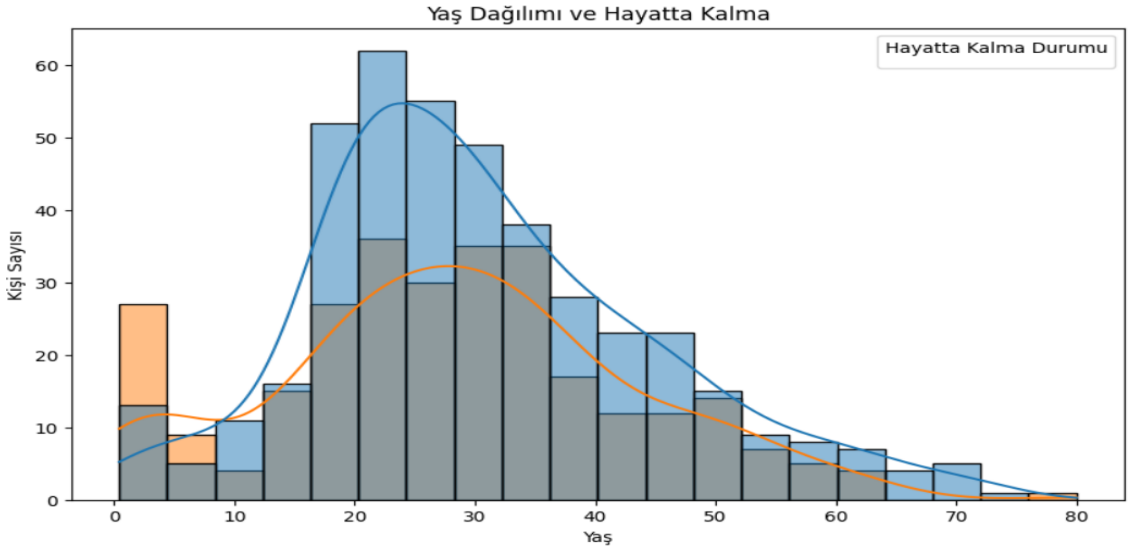
3.3. Cinsiyete Göre Hayatta Kalma: Kadın ve erkek yolcuların hayatta kalma oranları incelenmiştir. Kadınların hayatta kalma oranının erkeklere göre daha yüksek olduğu gözlemlenmiştir. Bu analiz, cinsiyetin hayatta kalma üzerinde önemli bir etkisi olduğunu göstermektedir.



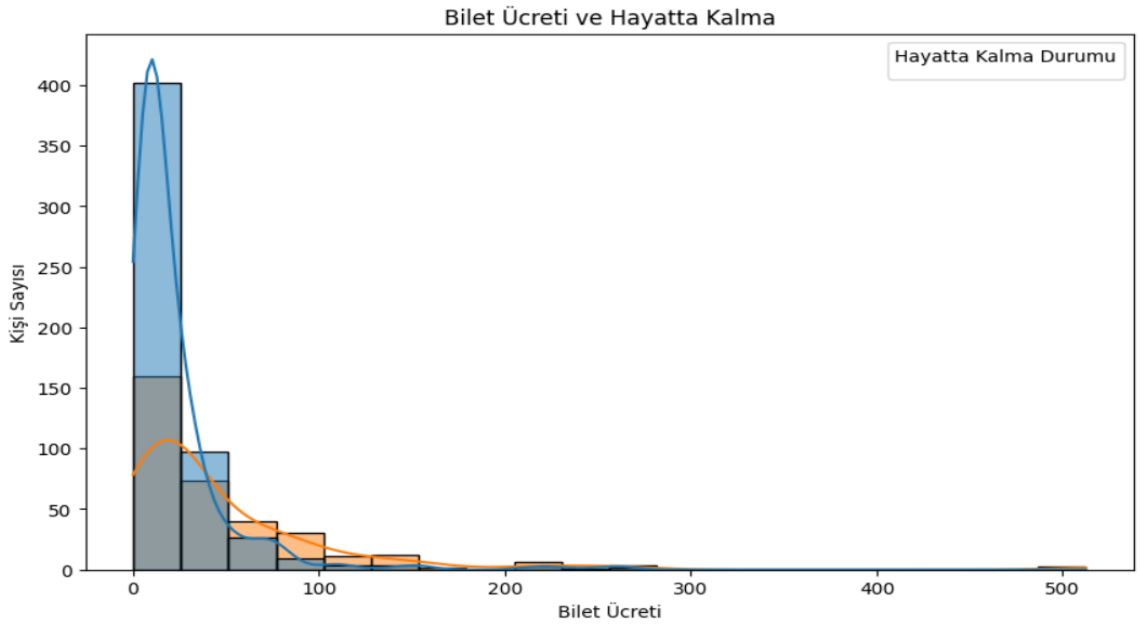
3.4. Yolcu Sınıfına Göre Hayatta Kalma: Birinci, ikinci ve üçüncü sınıftaki yolcuların hayatta kalma oranları analiz edilmiştir. Birinci sınıftaki yolcuların hayatta kalma oranının en yüksek olduğu, üçüncü sınıftaki yolcuların ise en düşük olduğu tespit edilmiştir. Bu bulgu, sosyoekonomik statünün hayatta kalma üzerinde önemli bir etkisi olduğunu göstermektedir.



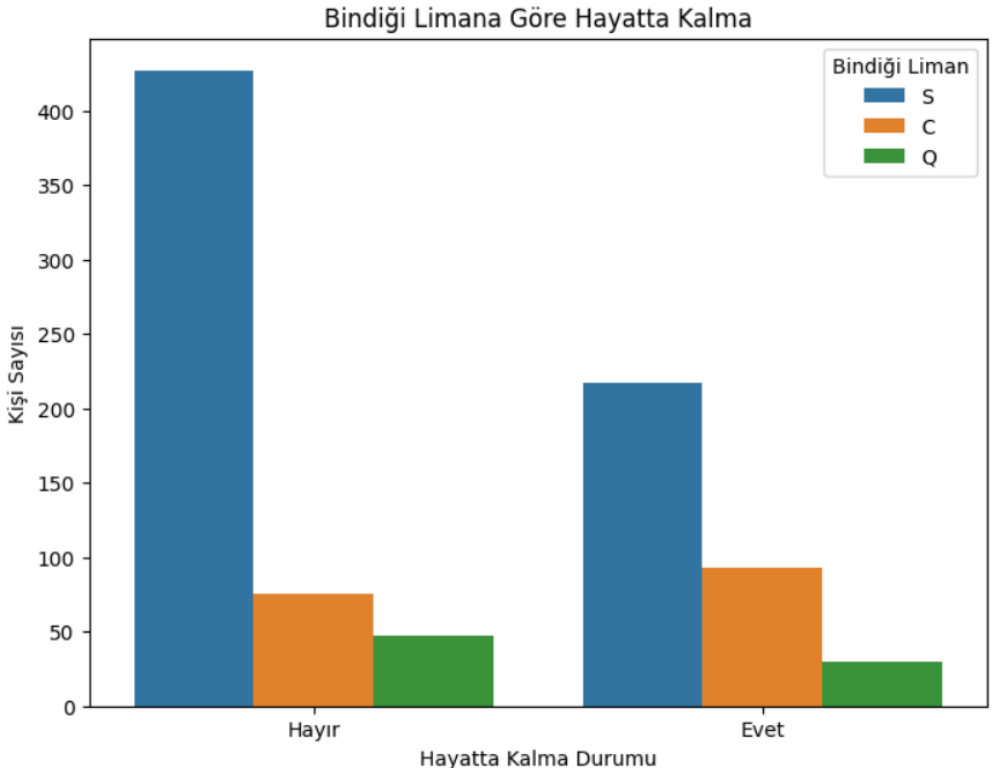
3.5. Yaş Dağılımı ve Hayatta Kalma: Yolcuların yaş dağılımı incelenmiş ve yaşa göre hayatta kalma oranları analiz edilmiştir. Genç yolcuların hayatta kalma oranının yaşlılara göre daha yüksek olduğu gözlemlenmiştir. Yaş dağılımı histogram ile görselleştirilmiştir.



3.6. Bilet Ücreti (Fare) ve Hayatta Kalma: Bilet ücretlerinin dağılımı incelenmiş ve log dönüşümü uygulanarak normalize edilmiştir. Bilet ücreti ile hayatta kalma arasında pozitif bir ilişki olduğu gözlemlenmiştir. Daha yüksek ücret ödeyen yolcuların hayatta kalma olasılığı daha yüksektir.



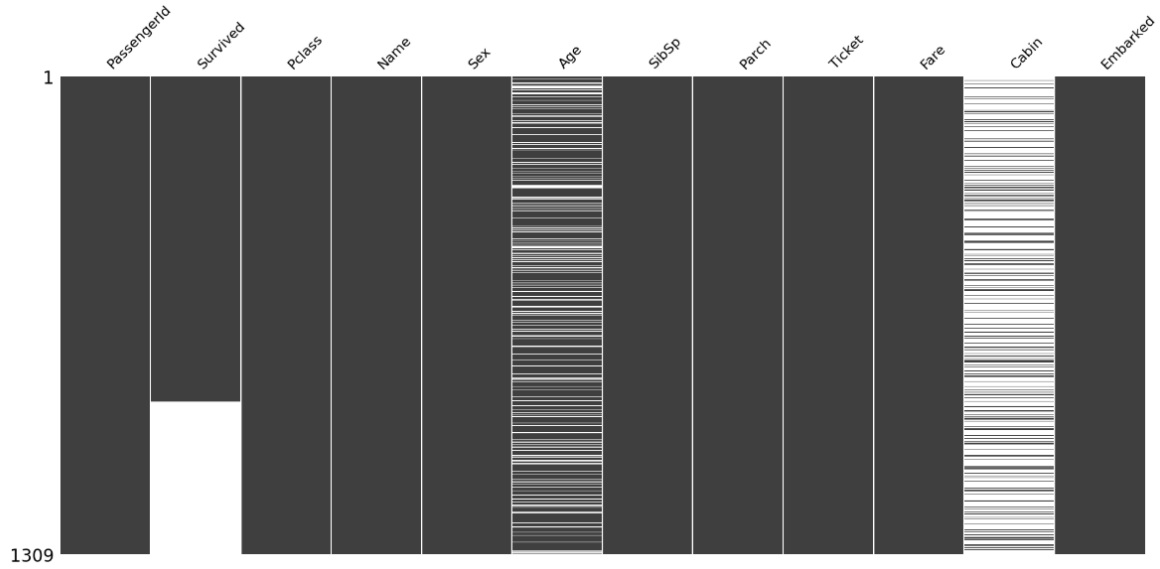
3.7. Bindiği Limana Göre Hayatta Kalma: Yolcuların bindiği limanlara göre hayatta kalma oranları incelenmiştir. Southampton'dan binen yolcuların en düşük hayatta kalma oranına sahip olduğu belirlenmiştir. Bu analiz, biniş limanının da hayatta kalma üzerinde bir etkisi olabileceğini göstermektedir.



4. Veri Ön İşleme

Veri setindeki eksik ve gereksiz bilgiler temizlenmiş, sayısal ve kategorik veriler düzenlenmiştir.

4.1. Eksik Değerlerin İncelenmesi: Eksik değerler analiz edilmiş ve veri setindeki eksik değerler uygun yöntemlerle doldurulmuştur. Cabin değişkenindeki eksik değerler nedeniyle bu sütun çıkarılmıştır. Yaş, fare ve bindiği liman değişkenlerindeki eksik değerler ise ortalama, mod veya uygun diğer yöntemlerle doldurulmuştur.



4.2. Sayısal Dönüşümler ve Normalizasyon: Fare değişkeni log dönüşüm ile normalize edilmiştir. Cinsiyet (Sex) ve bindiği liman (Embarked) değişkenleri sayısal değerlere dönüştürülmüştür. Bu dönüşümler, makine öğrenimi modellerinin daha iyi performans göstermesi için gereklidir.

4.3. Değişkenlerin Çıkarılması: Analiz için gerekli olmayan Ticket ve Name değişkenleri veri setinden çıkarılmıştır. Bu değişkenler modelin performansını etkilemeyecek bilgiler içerdiğinden çıkarılması uygun bulunmuştur.

5. Veri Setini Eğitim ve Test Olarak Ayırma

Veri seti, eğitim ve test seti olarak ikiye ayrılmıştır.

5.1. Bağımlı ve Bağımsız Değişkenler: Bağımlı değişken (y) hayatta kalma durumu (Survived) olup, bağımsız değişkenler (X) ise yolcu sınıfı, cinsiyet, yaş, kardeş/eş sayısı, ebeveyn/çocuk sayısı, bilet ücreti ve bindiği liman olarak belirlenmiştir.

5.2. Eğitim ve Test Setlerine Ayırma: Veri seti %80 eğitim ve %20 test olarak ikiye ayrılmıştır. Bu ayrım, modelin genel performansını değerlendirmek ve overfitting/underfitting problemlerini tespit etmek için önemlidir.

6. Model Eğitimi ve Değerlendirme

Bu aşamada, üç farklı model (Random Forest, XGBoost ve CatBoost) kullanılarak veri seti üzerinde eğitim yapılmış ve değerlendirilmiştir.

6.1. Random Forest Modeli: Random Forest sınıflandırıcısı kullanılarak model eğitilmiştir. Eğitim sırasında belirli hiperparametreler kullanılmıştır: **n_estimators=100**, **max_depth=None**, **min_samples_split=2**, **random_state=42**. Bu modelin performansı doğruluk (accuracy), F1 skoru, precision ve recall değerleri ile değerlendirilmiştir. Ayrıca, çapraz doğrulama (cross-validation) ile modelin güvenilirliği test edilmiştir.

6.2. XGBoost Modeli: XGBoost (Extreme Gradient Boosting) algoritması, karar ağaçlarını kullanarak ardışık bir şekilde eğitim yapar ve her adımda hataları düzeltmeye çalışır. XGBoost'un avantajları arasında yüksek performans, hız ve overfitting'i önleme yetenekleri bulunmaktadır. Modelin eğitimi sırasında belirli hiperparametreler kullanılmıştır: **n_estimators=100**, **learning_rate=0.1**, **max_depth=3**, **random_state=42**. Performans metrikleri ve çapraz doğrulama kullanılarak model değerlendirilmiştir.

6.3. CatBoost Modeli: CatBoost (Categorical Boosting), özellikle kategorik verilerle çalışırken yüksek performans gösteren bir gradient boosting algoritmasıdır. CatBoost, kategorik verileri otomatik olarak işler ve bu tür verilerle çalışmayı kolaylaştırır. Modelin eğitimi sırasında belirli hiperparametreler kullanılmıştır: **iterations=1000**, **learning_rate=0.1**, **depth=6**, **random_seed=42**. Modelin performansı doğruluk, F1 skoru, precision ve recall değerleri ile değerlendirilmiştir.

6.4. Model Performansının Karşılaştırılması: Her üç modelin performans metrikleri karşılaştırılmıştır. Genel olarak, model performansını değerlendirmek için kullanılan metrikler şunlardır:

- **Doğruluk (Accuracy)**: Modelin doğru tahmin ettiği örneklerin toplam örnek sayısına oranı.
- **F1 Skoru**: Precision ve recall'un harmonik ortalaması. Dengeli bir performans ölçütüdür.
- **Precision**: Doğru pozitif tahminlerin toplam pozitif tahminlere oranı.
- **Recall**: Doğru pozitif tahminlerin toplam gerçek pozitiflere oranı.
- **Çapraz Doğrulama (Cross-Validation)**: Modelin farklı veri alt kümelerinde performansını değerlendirerek genel performansı ve güvenilirliği test etme yöntemi.

7. Sonuçlar ve Değerlendirme

Modelin performansı ve elde edilen sonuçlar değerlendirilmiştir.

7.1. Performans Sonuçları:

- En iyi modelin doğruluk, F1 skoru, precision ve recall değerleri raporlandı.
- Eğitim ve test seti üzerindeki sonuçlar karşılaştırıldı.