Heart Failure Prediction Dataseti ile Veri Madenciliği Uygulaması

1. Projenin Amaç ve Yöntemi

Bu projenin amacı, kalp yetmezliği riski taşıyan bireyleri tahmin edebilmek için bir makine öğrenmesi modeli geliştirerek veriden anlamlı çıkarımlar yapmak, sağlık alanında karar destek sistemlerine katkı sağlamak ve riskli bireylerin önceden belirlenmesine yardımcı olmaktır. Projede, Random Forest sınıflandırma algoritması kullanılarak model oluşturulmuş ve farklı değerlendirme yöntemleriyle modelin başarısı ölçülmüştür. Veri görselleştirme ve ön işleme tekniklerinden faydalanılmıştır.

2. Veri Seti Hakkında

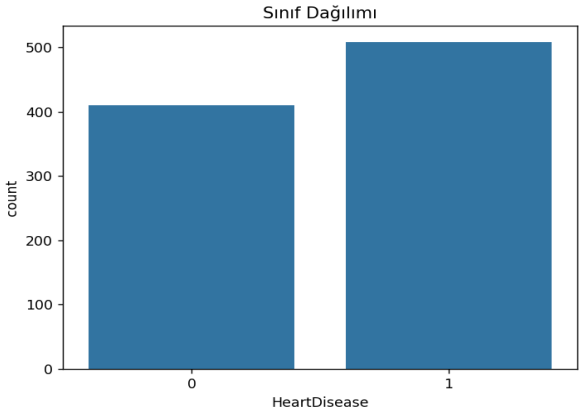
Veri seti Kaggle'dan alınmış olup, kalp yetmezliği tahmini için kullanılabilecek çeşitli özellikler içermektedir.

Veri Seti Özellikleri:

- **Age**: Hastanın yaşı (yıl)
- **Sex**: Hastanın cinsiyeti (M: Erkek, F: Kadın)
- **ChestPainType**: Göğüs ağrısı tipi (TA: Tipik Anjina, ATA: Atipik Anjina, NAP: Anjinal Olmayan Ağrı, ASY: Asemptomatik)
- **RestingBP**: Dinlenme kan basıncı (mm Hg)
- **Cholesterol**: Serum kolesterol (mg/dl)
- **FastingBS**: Açlık kan şekeri (1: 120 mg/dl'den büyük, 0: 120 mg/dl'den küçük)
- **RestingECG**: Dinlenme elektrokardiyografi sonuçları (Normal, ST: ST-T dalgası anormalliği, LVH: Sol ventrikül hipertrofisi)
- **MaxHR**: Ulaşılan maksimum kalp atış hızı
- **ExerciseAngina**: Egzersizle tetiklenen anjina (Y: Evet, N: Hayır)
- **Oldpeak**: ST depresyonu (egzersizle dinlenme arasındaki fark)
- **ST_Slope**: Egzersiz ST segmentinin eğimi (Up: Yukarı, Flat: Düz, Down: Aşağı)
- **HeartDisease**: Çıktı sınıfı (1: Kalp hastalığı, 0: Normal)

Veri Seti Boyutu:

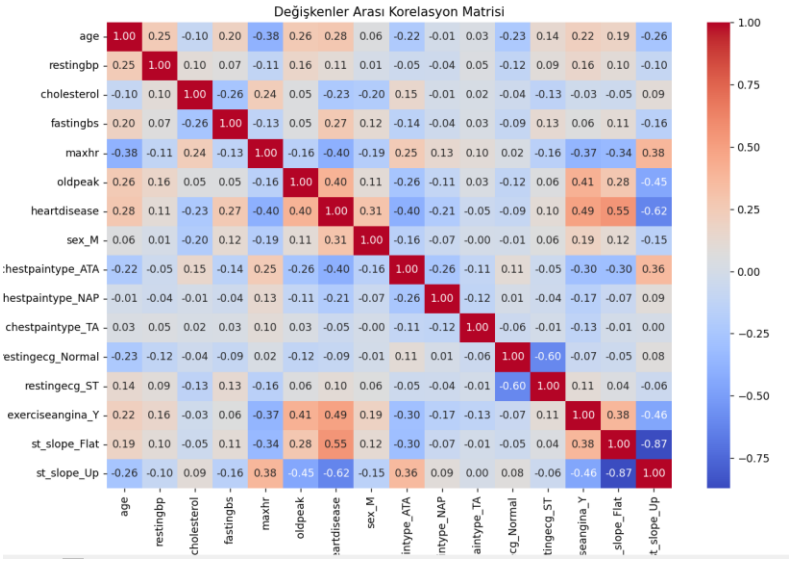
- Toplam örnek sayısı: 918
- Toplam özellik sayısı: 11
- Sınıf dağılımı:
    - Kalp hastalığı olan (1): 508 örnek (%55.3)
    - Kalp hastalığı olmayan (0): 410 örnek (%44.7)



3. Görselleştirilmiş Veriler

3.1. Korelasyon Matrisi

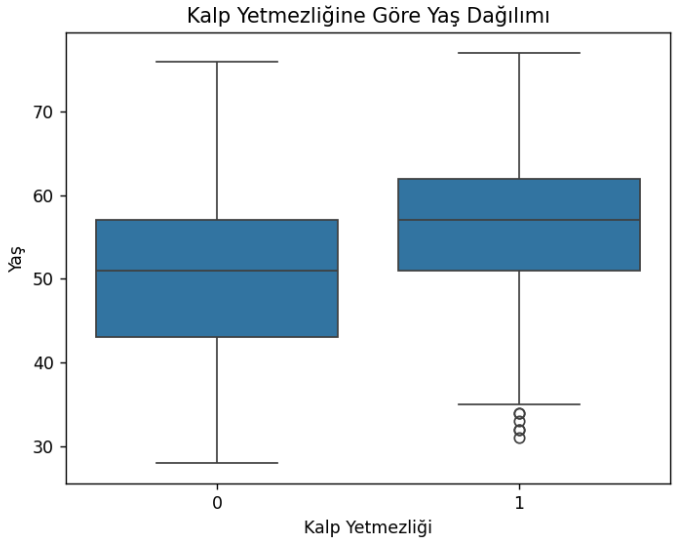
Korelasyon matrisi, özellikler arasındaki ilişkileri göstermektedir. Özellikle Oldpeak, MaxHR ve Age özelliklerinin HeartDisease ile yüksek korelasyona sahip olduğu görülmektedir.



Şekil 1: Özellikler arasındaki korelasyon matrisi

3.2. Yaş Dağılımı - Kalp Hastalığına Göre

Yaş dağılımı grafiği, kalp hastalığı olan ve olmayan hastaların yaş dağılımlarını göstermektedir. Kalp hastalığı olan hastaların yaş ortalamasının daha yüksek olduğu gözlemlenmiştir.

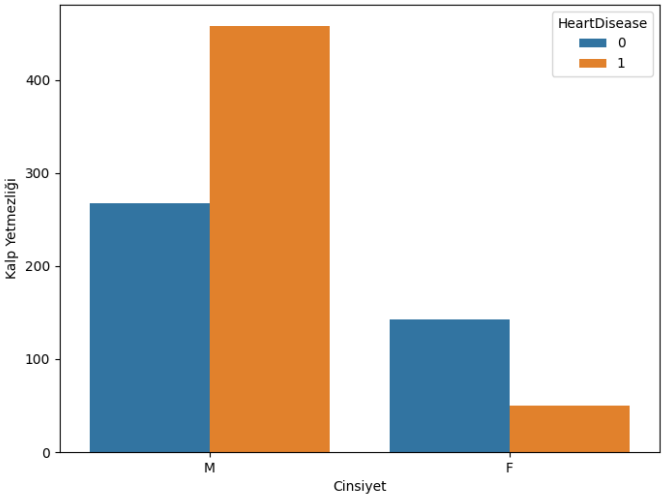


Şekil 2: Kalp hastalığına göre yaş dağılımı

Grafiğe göre, 50-65 yaş aralığında kalp hastalığı riskinin belirgin şekilde arttığı görülmektedir.

3.3. Cinsiyet Dağılımı - Kalp Hastalığına Göre

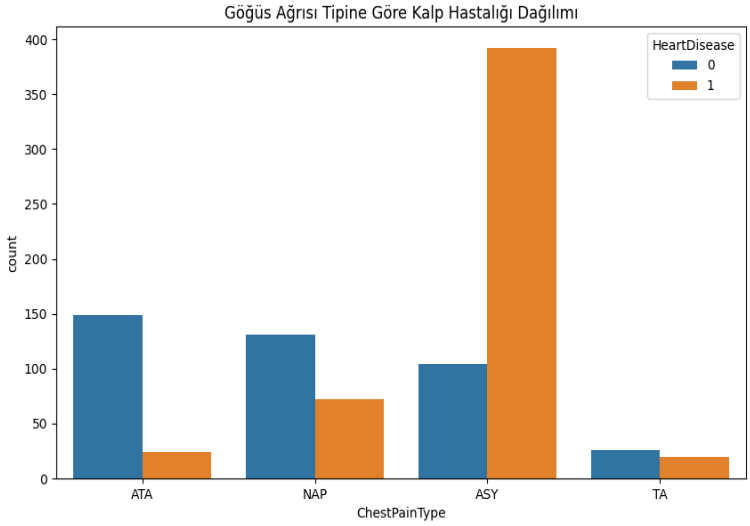
Cinsiyet dağılımı grafiği, erkeklerin kadınlara göre kalp hastalığına daha yatkın olduğunu göstermektedir.



Şekil 3: Cinsiyete göre kalp hastalığı dağılımı

3.4. Göğüs Ağrısı Tipine Göre Kalp Hastalığı Dağılımı

Göğüs ağrısı tipine göre kalp hastalığı dağılımı, Asemptomatik (ASY) göğüs ağrısı tipinin kalp hastalığı ile en yüksek ilişkiye sahip olduğunu göstermektedir.



Şekil 4: Göğüs ağrısı tipine göre kalp hastalığı dağılımı

4. Veri Ön İşleme

Veri seti üzerinde aşağıdaki ön işleme adımları uygulanmıştır:

1. **Sütun İsimleri Düzenleme:** Sütun isimleri küçük harfe dönüştürülmüş, boşluklar alt çizgi ile değiştirilmiştir.

2. **Kategorik Değişkenlerin Dönüştürülmesi**: One-hot encoding yöntemi ile kategorik değişkenler sayısal forma dönüştürülmüştür.

3. **Özellik Standardizasyonu:** StandardScaler kullanılarak veri seti ölçeklendirilmiştir.

4. **Özellik Seçimi:** Random Forest algoritması kullanılarak özellik önem düzeyleri (feature importance) belirlenmiş ve en önemli 5 özellik seçilmiştir. Karşılaştırma yapabilmek adına SelectKBest yöntemiyle de 5 özellik belirlenmiştir.

    - **K-Best'e Göre Seçilen En Önemli Özellikler:** ['oldpeak', 'chestpaintype_ATA', 'exerciseangina_Y', 'st_slope_Flat', 'st_slope_Up'], olmuştur.
    - **Random Forest'a Göre Seçilen En Önemli Özellikler**: ['st_slope_Up', 'maxhr', 'st_slope_Flat', 'oldpeak', 'cholesterol'] olmuştur.

5.Model Karşılaştırmaları

Random Forest sınıflandırma algoritması kullanılarak üç farklı model oluşturulmuştur:

1. Tüm özellikleri kullanan model
2. SelectKBest yöntemiyle seçilen özellikleri kullanan model
3. Random Forest özellik önemine göre seçilen özellikleri kullanan model

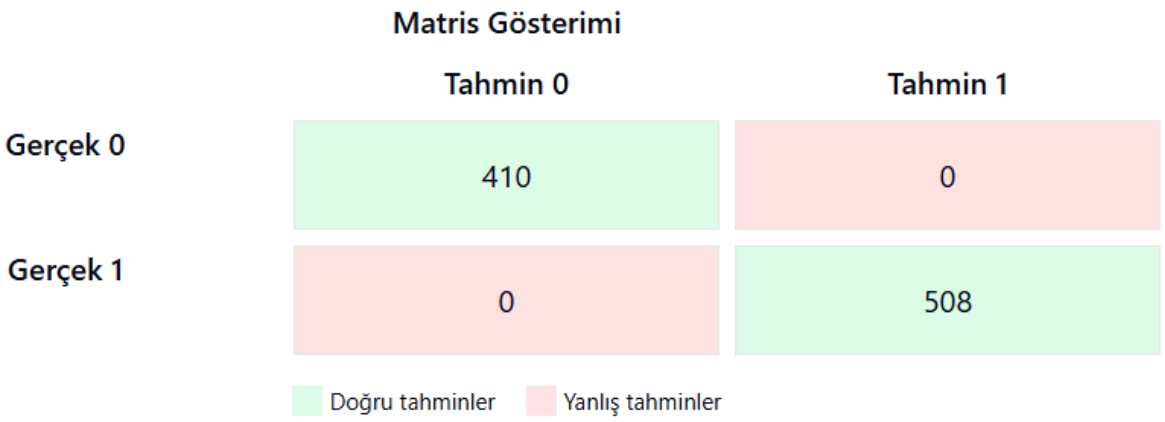
Her bir model için eğitim verisi ile test ve %70 eğitim-%30 test bölünmüş verileri kullanarak performans karşılaştırması yapılmıştır.

5.1. Tüm Özelliklerle Oluşturulan Model

5.1.1. Eğitim = Test (Overfitting riski yüksek)

- **Doğruluk**: %100
- **Sınıflandırma Doğruluğu**:
    - Sınıf 0 (Hastalık yok) - Precision: 1.00, Recall: 1.00, F1-score: 1.00
    - Sınıf 1 (Hastalık var) - Precision: 1.00, Recall: 1.00, F1-score: 1.00

- **Konfüzyon Matrisi**:

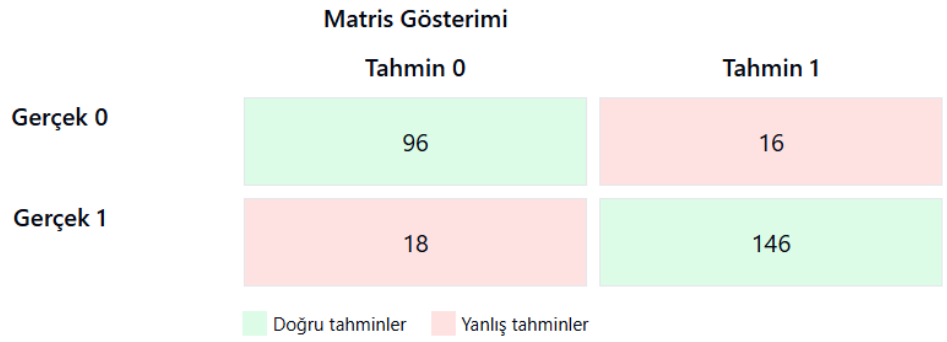


➢ Bu sonuçlar, modelin eğitim verisi üzerinde mükemmel bir şekilde çalıştığını gösteriyor, ancak bu durum aşırı öğrenme (overfitting) göstergesidir.

5.1.2. %70 Eğitim - %30 Test

- **Doğruluk**: %87,68
- **Sınıflandırma Doğruluğu**:
    - Sınıf 0 (Hastalık yok) - Precision: 0.84, Recall: 0.86, F1-score: 0.85
    - Sınıf 1 (Hastalık var) - Precision: 0.90, Recall: 0.89, F1-score: 0.90

- **Konfüzyon Matrisi**:

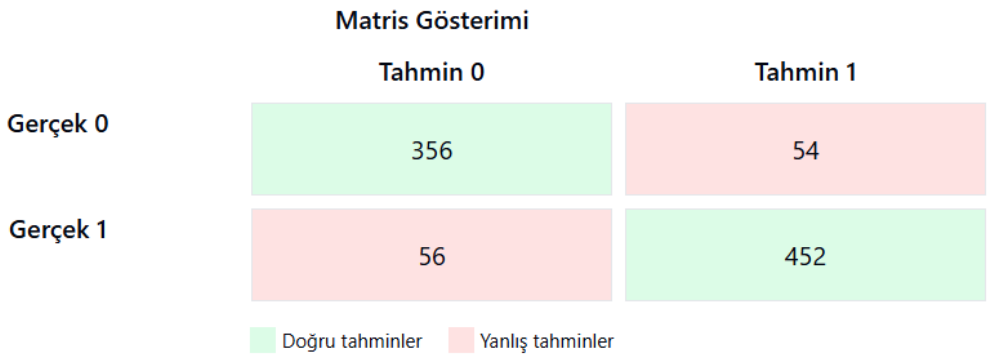


➢ Bu sonuçlar, modelin daha önce görmediği veriler üzerinde iyi bir performans gösterdiğini belirtiyor.

5.2. KBest ile Seçilen Özelliklerle Oluşturulan Model

5.2.1. Eğitim = Test (Overfitting riski yüksek)

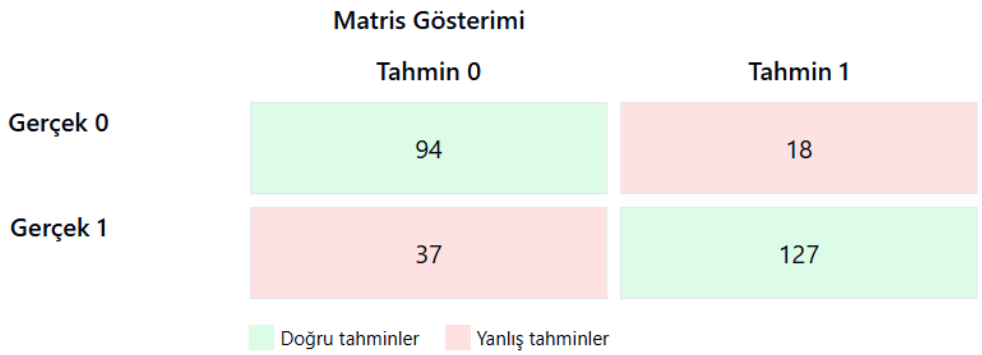
- **Doğruluk**: %88.02
- **Sınıflandırma Doğruluğu**:
    - Sınıf 0 (Hastalık yok) - Precision: 0.86, Recall: 0.87, F1-score: 0.87
    - Sınıf 1 (Hastalık var) - Precision: 0.89, Recall: 0.89, F1-score: 0.89



- KBest yöntemi ile seçilen özelliklerle eğitim verisi üzerinde %100 doğruluk yerine %88.02 doğruluk elde edilmesi, modelin daha az özellikle daha az kompleks olduğunu ve aşırı öğrenme riskinin azaldığını gösteriyor.

5.2.2. %70 Eğitim - %30 Test

- **Doğruluk**: %80.07
- **Sınıflandırma Doğruluğu**:
  - Sınıf 0 (Hastalık yok) - Precision: 0.72, Recall: 0.84, F1-score: 0.77
  - Sınıf 1 (Hastalık var) - Precision: 0.88, Recall: 0.77, F1-score: 0.82
- **Konfüzyon Matrisi**:

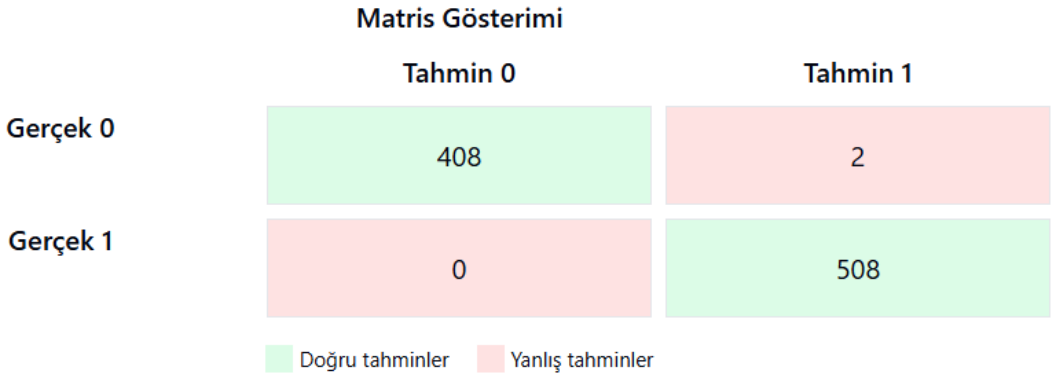


- Test verisi üzerindeki performansın tüm özelliklerle oluşturulan modele göre düşük olması, KBest yöntemiyle seçilen özelliklerin tek başına yeterli olmadığını gösteriyor.

5.3. Random Forest Özellik Önemine Göre Seçilen Özelliklerle Oluşturulan Model

5.3.1 Eğitim = Test (Overfitting riski yüksek)

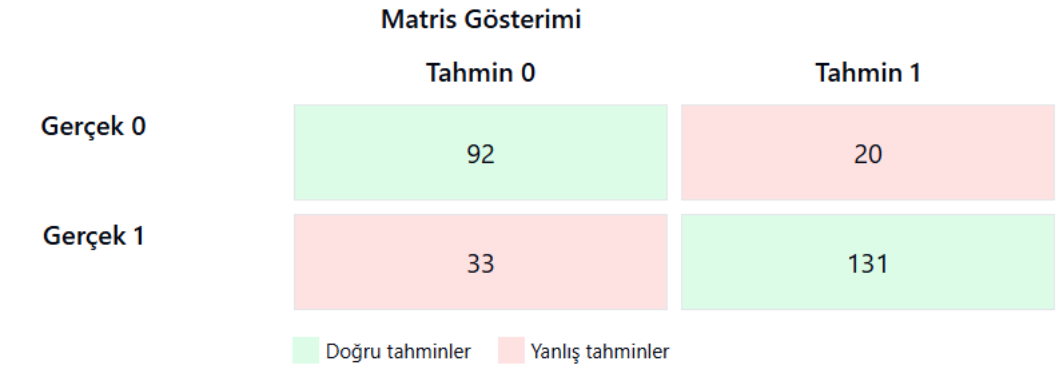
- **Doğruluk**: %99.78
- **Sınıflandırma Doğruluğu**:
  - Sınıf 0 (Hastalık yok) - Precision: 1.00, Recall: 1.00, F1-score: 1.00
  - Sınıf 1 (Hastalık var) - Precision: 1.00, Recall: 1.00, F1-score: 1.00



- Random Forest'ın özellik önemine göre seçilen özelliklerle eğitim verisi üzerinde neredeyse mükemmel sonuçlar elde edilmiştir, bu da yine aşırı öğrenme riskini göstermektedir.

5.3.2. %70 Eğitim - %30 Test

- **Doğruluk**: %80.80
- **Sınıflandırma Doğruluğu**:
  - Sınıf 0 (Hastalık yok) - Precision: 0.74, Recall: 0.82, F1-score: 0.78
  - Sınıf 1 (Hastalık var) - Precision: 0.87, Recall: 0.80, F1-score: 0.83
- **Konfüzyon Matrisi**:



➢ Bu model, KBest modeline göre biraz daha iyi performans göstermiştir, ancak tüm özellikleri kullanan modelin gerisinde kalmıştır.

6. Çapraz Doğrulama Sonuçları

- **Ortalama Doğruluk**: %84.62
- **Standart Sapma**: %6.61

10-fold çapraz doğrulama sonuçları, modelin genelleştirme yeteneği hakkında daha güvenilir bir değerlendirme sunmaktadır. %84.62'lik ortalama doğruluk oranı, modelin farklı veri bölümlerinde makul bir performans gösterdiğini işaret ediyor.

KAYNAKÇA

- Kaggle- HEART FAİLURE PREDİCTİON DATASET
  https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data
- www.medium.com