# Kalp Hastalığı Riski Analizi - Proje Çıktısı

## 1. Projenin Amacı

Bu projenin temel amacı, Heart Disease UCI veri setini kullanarak bireylerin kalp hastalığı riski taşıyıp taşımadığını makine öğrenmesi modelleri aracılığıyla tahmin etmektir. Hedef değişken olan HeartDisease:

- 1: Kalp hastalığı var
- 0: Kalp hastalığı yok

Pratikteki önemine baktığımızda; kalp hastalıkları erken fark edilirse, oluşabilecek kötü sonuçlar engellenebilir. Model, kalp krizi risk taşıyan kişileri önceden tahmin ederek yardımcı olabilir.

## 2. Veri Seti İncelemesi

### Veri Seti Kaynağı:

- UCI Heart Disease veri kümesi (Kaggle üzerinden temin edilmiştir)

### Değişkenler:

Veri seti aşağıdaki özellikleri içermektedir:

- Age: Yaş (tam sayı)
- Sex: Cinsiyet ("M" veya "F")
- ChestPainType: Göğüs ağrısı tipi (kategorik veri)
- RestingBP: Dinlenme kan basıncı
- Cholesterol: Kolesterol seviyesi
- FastingBS: Açlık kan şekeri > 120 mg/dl (0 veya 1)
- RestingECG: EKG sonucu
- MaxHR: Maksimum kalp atış hızı
- ExerciseAngina: Egzersize bağlı anjina ("Y" veya "N")
- Oldpeak: ST segmenti depresyonu
- ST_Slope: ST segmentinin eğimi
- HeartDisease: Kalp hastalığı var mı? (0 = Yok, 1 = Var)

### İlk Analiz:

Eksik veri bulunmamaktadır.

Hedef değişken olan HeartDisease'in dağılımı:

- 1 (Kalp hastalığı var): %55.3
- 0 (Kalp hastalığı yok): %44.7
- Hedef değişken dengelidir (%50’ye yakın dağılım), bu da modeli dengelemeye yardımcı olmuştur.

Sayısal değişkenlerin temel istatistikleri:

- Yaş ortalaması: 53.5
- Dinlenme kan basıncı ortalaması: 132.4 mmHg
- Kolesterol ortalaması: 198.8 mg/dL
- Maksimum kalp atış hızı ortalaması: 136.8 bpm
- Sayısal ve kategorik değişkenlerin etki düzeyi EDA (Exploratory Data Analysis) ile görselleştirilmiştir.

### Keşifsel Veri Analizi (EDA):

Korelasyon matrisi incelendiğinde, hedef değişkenle en güçlü ilişkiye sahip özellikler:

- Oldpeak (ST segmenti depresyonu): Pozitif korelasyon
- MaxHR (Maksimum kalp atış hızı): Negatif korelasyon
- ExerciseAngina: Pozitif korelasyon
- Age: Pozitif korelasyon

Cinsiyet ve kalp hastalıkları arasındaki ilişki tespit edilmiş ve erkek bireylerde hastalık riskinin daha yüksek olabileceği tespit edilmiştir.

## 3. Veri Ön İşleme

- İkili kategorik veriler için label encoding, çoklu kategorik veriler için one-hot encoding işlemi yapılmıştır.
- Veri, %80 eğitim - %20 test şeklinde ayrılmıştır.
- Sayısal özellikler StandardScaler kullanılarak ölçeklendirilmiştir. Bu işlem, farklı ölçekteki değişkenlerin model performansını olumsuz etkilemesini engellemek için yapılmıştır.

## 4. Uygulanan Modeller ve Karşılaştırma

İki farklı model uygulanmıştır:

1. **Logistic Regression**
2. **Random Forest Classifier**

### Başarı Metrikleri:

| Model | Accuracy | Precision | Recall | F1 Score | ROC AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.85 | 0.82 | 0.88 | 0.85 | 0.90 |
| Random Forest | 0.88 | 0.86 | 0.91 | 0.88 | 0.93 |

### Model Karşılaştırması:

- Random Forest modeli tüm metriklerde daha üstün performans göstermiştir.
- ROC-AUC skoru 0.93, modelin sınıflar arasında güçlü bir ayırt edici güce sahip olduğunu göstermektedir.
- Confusion matrix analizinde yanlış pozitif ve yanlış negatif oranlarının düşük olduğu görülmüştür.

### Aşırı / Yetersiz Öğrenme Durumu:

- Eğitim ve test skorları benzer olduğundan aşırı öğrenme gözlemlenmemiştir.
- Logistic Regression modeli, veri setindeki bazı karmaşık örüntüleri öğrenmekte yetersiz kalmış olabilir.  Çünkü başarı metrikleri makuldur ama Random Forest'a göre belirgin şekilde düşüktür.Bu durum, çok büyük bir fark olmasa da, underfitting (yetersiz öğrenme) etkisinin hafif düzeyde olduğunu gösterir.

## 5. Sonuçlar ve Yorumlar

- Random Forest modeli, doğruluk, duyarlılık ve ROC-AUC açısından kalp hastalığı riskini tahmin etmede oldukça başarılı sonuçlar vermiştir.
- Aykırı verilerin temizlenmesi, modelin genelleme kapasitesini artırmıştır.

## 6. Projenin Önemi ve Katkısı

Bu proje, yapay zekanın sağlık alanındaki uygulamalarına güzel bir örnek sunmaktadır. Özellikle kalp hastalıklarının:

- Erken teşhisi
- Riskli bireylerin önceden belirlenmesi

gibi konularda makine öğrenmesi modellerinin uygulanabilirliğini göstermektedir.

## 7. Genel Yorumlar ve Geliştirme Önerileri

### Model Kullanımı:

- Random Forest modeli, pratik uygulamalar için uygundur.
- Yeni hastaların verisi ile kolayca tahminleme yapılabilir.

### Geliştirme Önerileri:

Daha iyi sonuçlar için;

- Veri Artırma: Daha fazla veri toplanarak modelin genelleme yeteneği artırılabilir.
- Ek Özellikler: Aile öyküsü, sigara kullanımı, stres seviyesi gibi ek özellikler model performansını artırabilir.