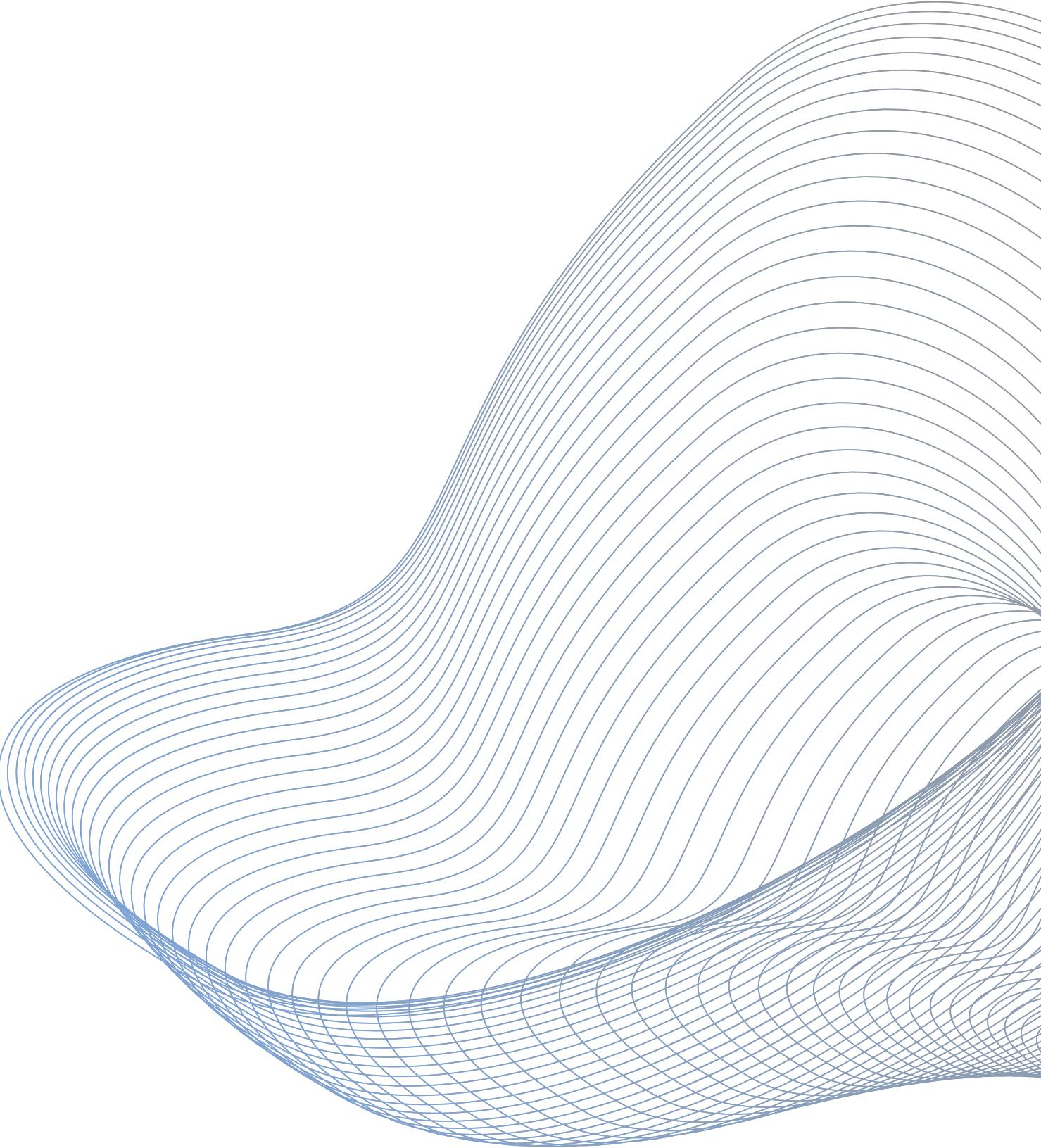


# **KALP HASTALIĞI RİSK TAHMİNİ PROJESİ**

**SİSTERSLAB-BİLİM VE TEKNOLOJİDE KADIN DERNEĞİ**

**Sema Nur Biçki**

- Giriş ve Problem Tanımı
- Veri Seti İncelemesi
- Keşifsel Veri Analizi (EDA)
- Veri Ön İşleme
- Uygulanan Modeller
- Model Karşılaştırması
- Sonuç ve Yorumlar



# Giriş ve Problem Tanımı

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

Neden Kullanılır?

Doktorlar çoğu zaman belirli testleri sadece hasta şikayetçi olduğunda uygulayabiliyor. Oysa bu model, kişinin temel sağlık verilerine bakarak henüz doktora gitmeden önce kalp hastalığı riski taşıyıp taşımadığını tahmin edebiliyor. Bu da bir tür erken uyarı sistemi gibi çalışıyor.

# Veri Seti İncelemesi

## Veri Seti Kaynağı

UCI Heart Disease veri kümesi (Kaggle üzerinden temin edilmiştir)

## Veri Seti Değişkenleri

918 gözlem, 12 özellik (feature), 6 tam sayı, 1 ondalıklı sayı, 5 kategorik veri olduğu gözlenmiştir.

## Kategorik Veriler

- Sex: Female (F) ve Male (M)
- ChestPainType (Göğüs Ağrısı Tipi): ATA → Atypical Angina (Atipik anjina), NAP → Non-Anginal Pain (Anjina dışı ağrı), ASY → Asymptomatic (Semptom yok), TA → Typical Angina (Tipik anjina)
- RestingECG (İstirahat EKG Sonucu): Normal, ST → ST-T dalga anomalilikleri, LVH → Sol ventriküler hipertrofi (LVH = Left Ventricular Hypertrophy)
- ExerciseAngina (Egzersize Bağlı Anjina): Y → Yes (Evet, egzersiz sırasında anjina var), N → No (Hayır, egzersiz sırasında anjina yok)
- ST\_Slope (ST Segmenti Eğimi): Up → Yukarı eğimli, Flat → Düz, Down → Aşağı eğimli

RangeIndex: 918 entries, 0 to 917

Data columns (total 12 columns):

| #  | Column         | Non-Null Count | Dtype   |
|----|----------------|----------------|---------|
| 0  | Age            | 918 non-null   | int64   |
| 1  | Sex            | 918 non-null   | object  |
| 2  | ChestPainType  | 918 non-null   | object  |
| 3  | RestingBP      | 918 non-null   | int64   |
| 4  | Cholesterol    | 918 non-null   | int64   |
| 5  | FastingBS      | 918 non-null   | int64   |
| 6  | RestingECG     | 918 non-null   | object  |
| 7  | MaxHR          | 918 non-null   | int64   |
| 8  | ExerciseAngina | 918 non-null   | object  |
| 9  | Oldpeak        | 918 non-null   | float64 |
| 10 | ST_Slope       | 918 non-null   | object  |
| 11 | HeartDisease   | 918 non-null   | int64   |

dtypes: float64(1), int64(6), object(5)

|   | Age | Sex | ChestPainType | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR | ExerciseAngina | Oldpeak | ST_Slope | HeartDisease |
|---|-----|-----|---------------|-----------|-------------|-----------|------------|-------|----------------|---------|----------|--------------|
| 0 | 40  | M   | ATA           | 140       | 289         | 0         | Normal     | 172   | N              | 0.0     | Up       | 0            |
| 1 | 49  | F   | NAP           | 160       | 180         | 0         | Normal     | 156   | N              | 1.0     | Flat     | 1            |
| 2 | 37  | M   | ATA           | 130       | 283         | 0         | ST         | 98    | N              | 0.0     | Up       | 0            |
| 3 | 48  | F   | ASY           | 138       | 214         | 0         | Normal     | 108   | Y              | 1.5     | Flat     | 1            |
| 4 | 54  | M   | NAP           | 150       | 195         | 0         | Normal     | 122   | N              | 0.0     | Up       | 0            |

|              | Age        | RestingBP  | Cholesterol | FastingBS  | MaxHR      | Oldpeak    | HeartDisease |
|--------------|------------|------------|-------------|------------|------------|------------|--------------|
| <b>count</b> | 918.000000 | 918.000000 | 918.000000  | 918.000000 | 918.000000 | 918.000000 | 918.000000   |
| <b>mean</b>  | 53.510893  | 132.396514 | 198.799564  | 0.233115   | 136.809368 | 0.887364   | 0.553377     |
| <b>std</b>   | 9.432617   | 18.514154  | 109.384145  | 0.423046   | 25.460334  | 1.066570   | 0.497414     |
| <b>min</b>   | 28.000000  | 0.000000   | 0.000000    | 0.000000   | 60.000000  | -2.600000  | 0.000000     |
| <b>25%</b>   | 47.000000  | 120.000000 | 173.250000  | 0.000000   | 120.000000 | 0.000000   | 0.000000     |
| <b>50%</b>   | 54.000000  | 130.000000 | 223.000000  | 0.000000   | 138.000000 | 0.600000   | 1.000000     |
| <b>75%</b>   | 60.000000  | 140.000000 | 267.000000  | 0.000000   | 156.000000 | 1.500000   | 1.000000     |
| <b>max</b>   | 77.000000  | 200.000000 | 603.000000  | 1.000000   | 202.000000 | 6.200000   | 1.000000     |

# Veri Seti İncelemesi

## 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 dengeli dir (%50'ye yakın dağılım), bu da modeli dengelemeye yardımcı olmuştur.

## Değişken Yorumları

- Yaş 44-60 aralığında dengeli dir.
- RestingBP (Dinlenme Kan Basıncı) 120-140 arası ancak 0 değerleri var, tıbbi olarak böyle bir durum imkansız, temizlenmesi gerekmektedir.
- Kolesterol ortalama 198 mg/dl, medyan 223 mg/dl yani ortalamanın üstünde, 603 mg/dl değeri de ortalamanın çok üstünde aykırı bir değer olabilir, ayrıca dinlenme kan basıncındaki gibi 0 değerleri var aynı şekilde bu tıbbi olark imkansız olduğundan temizlenmelidir.
- MaxHR (Maksimum Kalp Atım Hızı) medyan ve ortalama birbirine yakın, uyumlu bir dağılım söz konusudur.
- Oldpeak (ST Segmenti Depresyonu) dengeli görünse de 6.2 gibi uç değerler de mevcuttur.

# 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

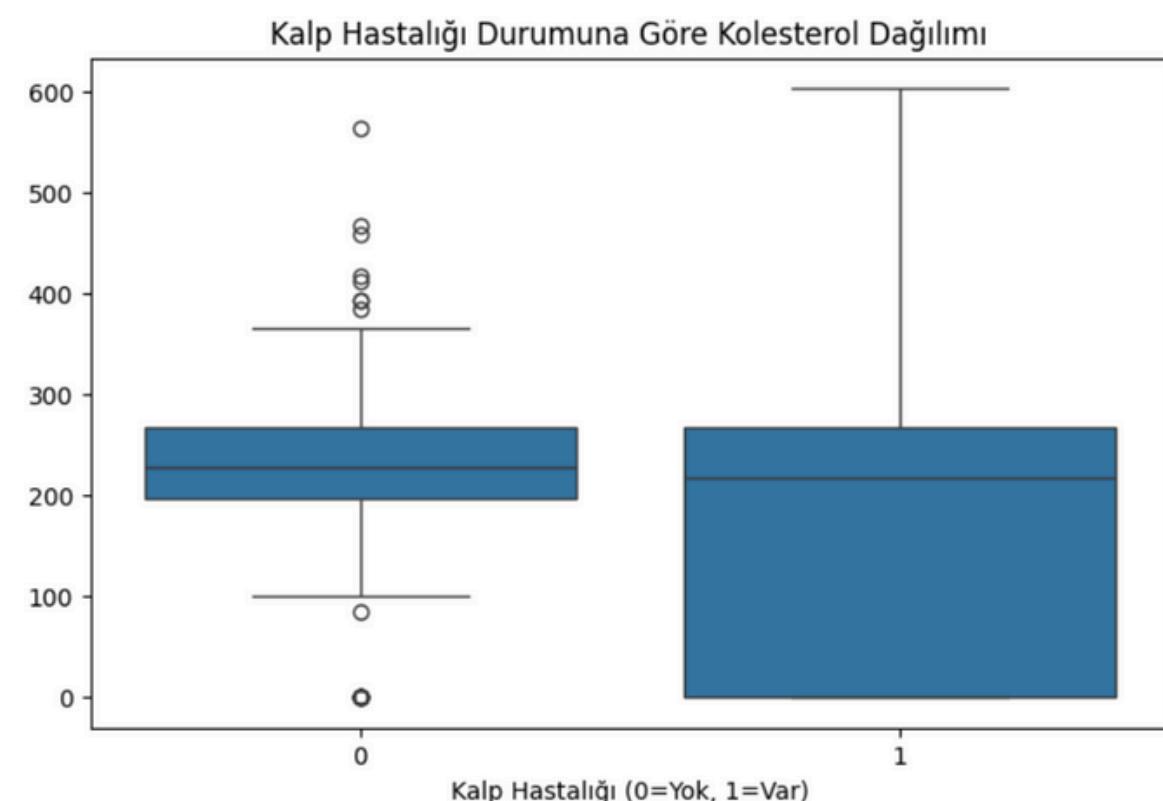
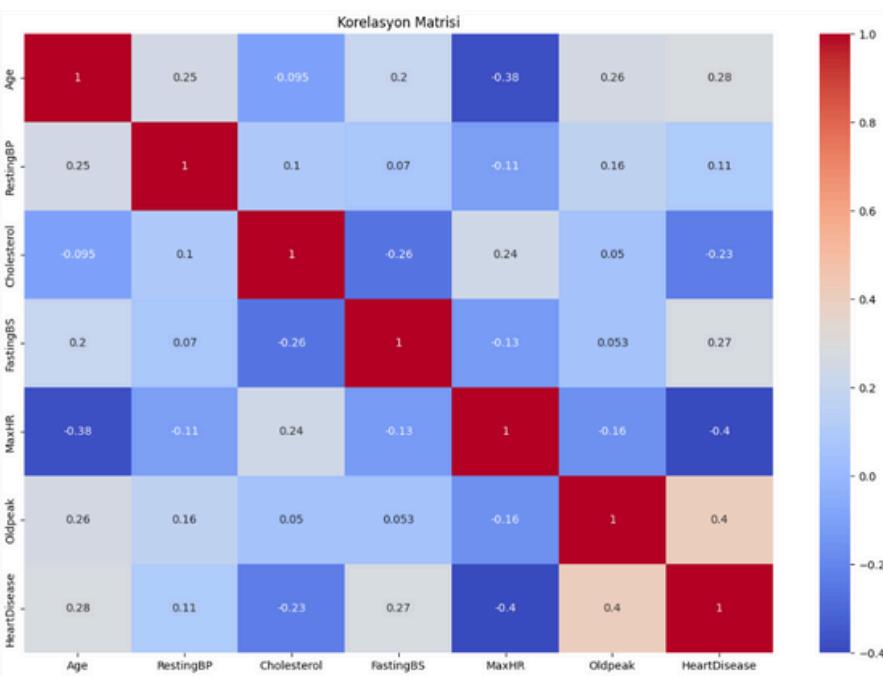
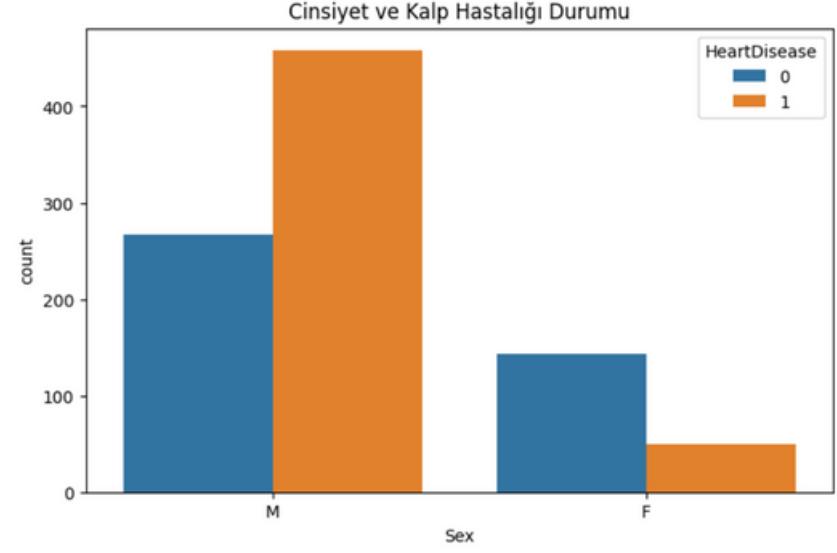
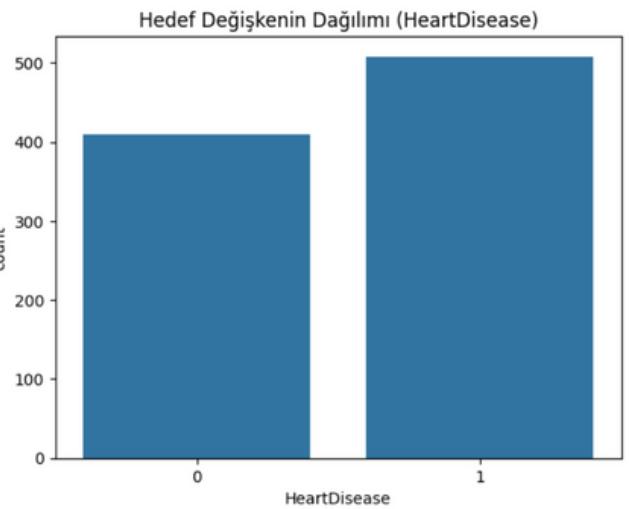
korelasyon(Maksimum kalp atım hızı ne kadar düşükse, kalp hastalığı riski o kadar yüksek. Kalbi zorlanmadan yüksek seviyeye çıkaramayan bireylerde hastalık ihtimali daha fazla.)

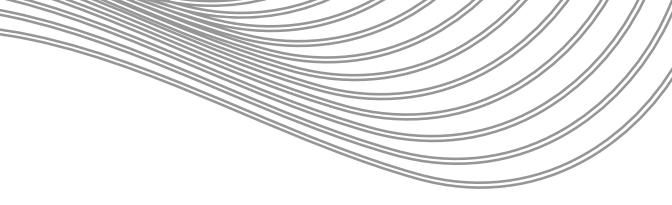
ExerciseAngina (Egzersize Bağlı Göğüs Ağrısı):Pozitif korelasyon

Age: Pozitif korelasyon

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

Kalp hastalığı bireylerin kolestrol seviyeleri daha yüksek görünse de, kalp hastalığı olmayan bireylerde bu kadar uç değerin çıkmış olması kolestrolün tek başına kalp hastalığı belirtisi olmadığını düşündürmektedir.





# Veri Ön İşleme

- IQR yöntemi ile 0 mg/dl tansiyon ya da 600 mg/dl kolesterol gibi tıbben imkânsız üç değerlerin varlığı tespit edilmiştir.
- Neden IQR?

**Gerçek dünya verileri genellikle bozulmuş, çarpık ya da gürültülüdür.** IQR yöntemi, verinin en uç %25'lik kısımlarına değil, ortadaki %50'lik kısmına odaklanır. Bu da onu uç değerlerden etkilenmeden çalışabilecek sağlam bir yöntem yapar.

- Uç değerleri tamamen silmek, özellikle **küçük veri setlerinde** bilgi kaybına neden olabilir. Bu yüzden **Winsorizing** ile verileri sadece sınıra kadar çekerek veri bütünlüğünü korunmuş ve modelin aşırı etkilenmesinin önüne geçilmiştir.
- Modeller metinle değil sayı ile çalışır. Bu sebeple **ikili 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.** Örneğin MaxHR değeri 180 iken Oldpeak sadece 0.8 olabilir ama bu onun daha önemsiz olduğu anlamına gelmez. Bu yüzden StandardScaler yöntemiyle tüm sayısal sütunları ortak bir ölçüye getirilmiştir.
- Ölçekleme işlemi, eğitim-test ayriminden sonra uygulanmıştır çünkü test verisi, modelin hiç “görmediği” bir veri olmalıdır. Eğer tüm veri setinde fit işlemi uygulanırsa, test verisinin ortalaması ve std'si modele sızar ve doğrulama güvenilir olmaz.

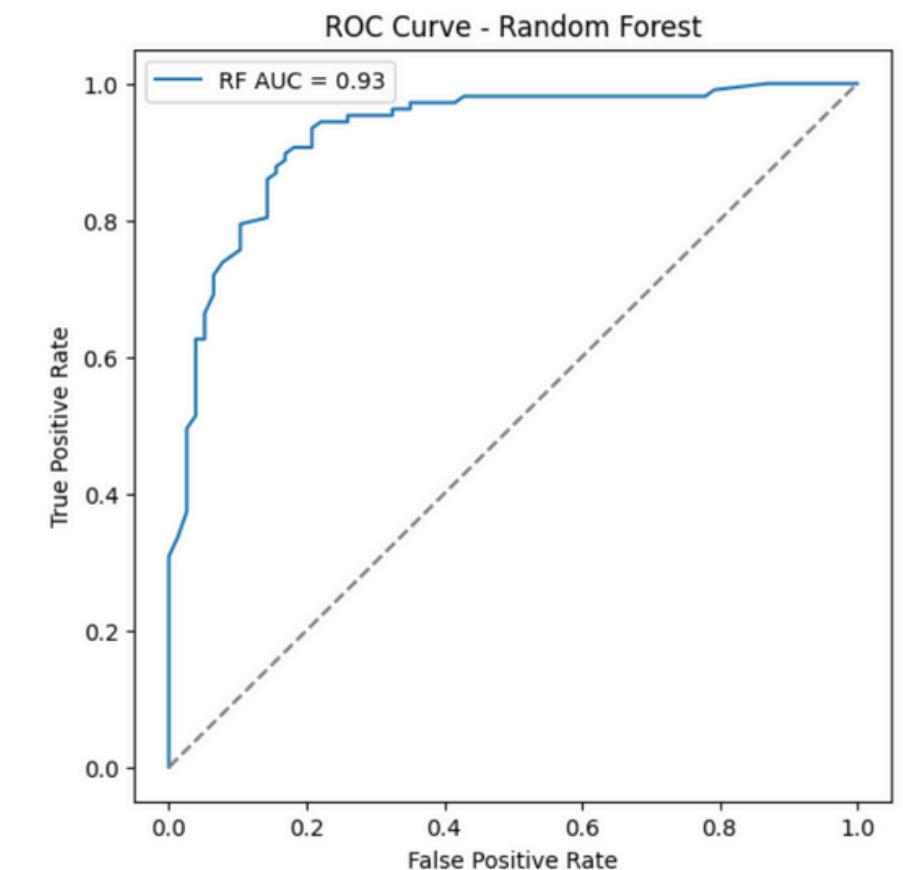
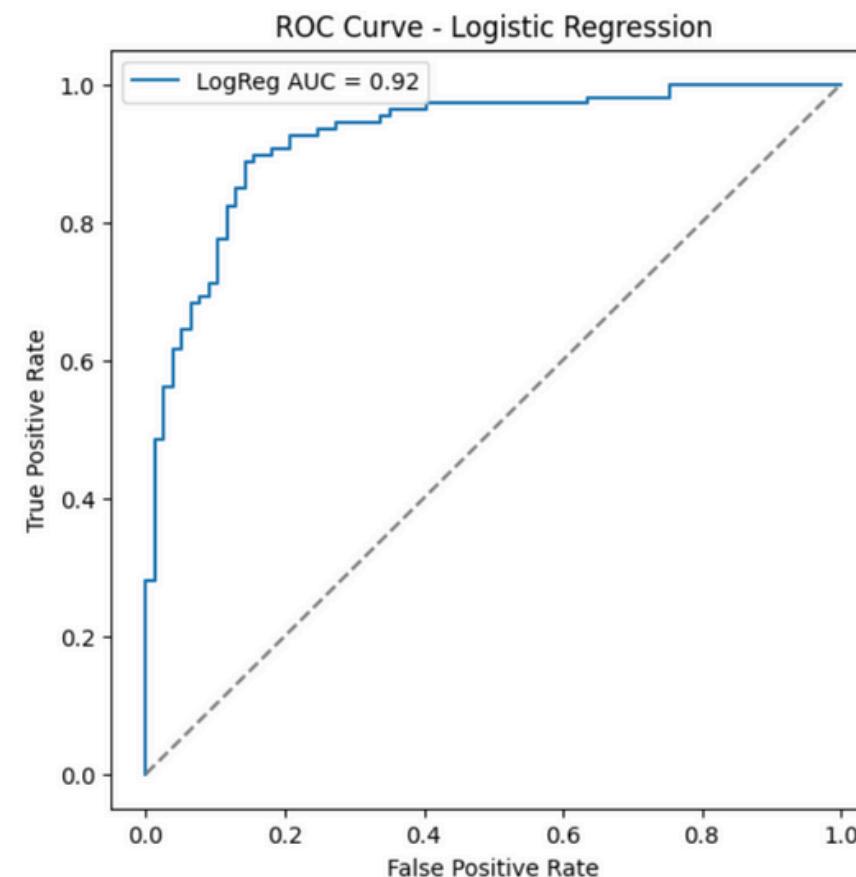
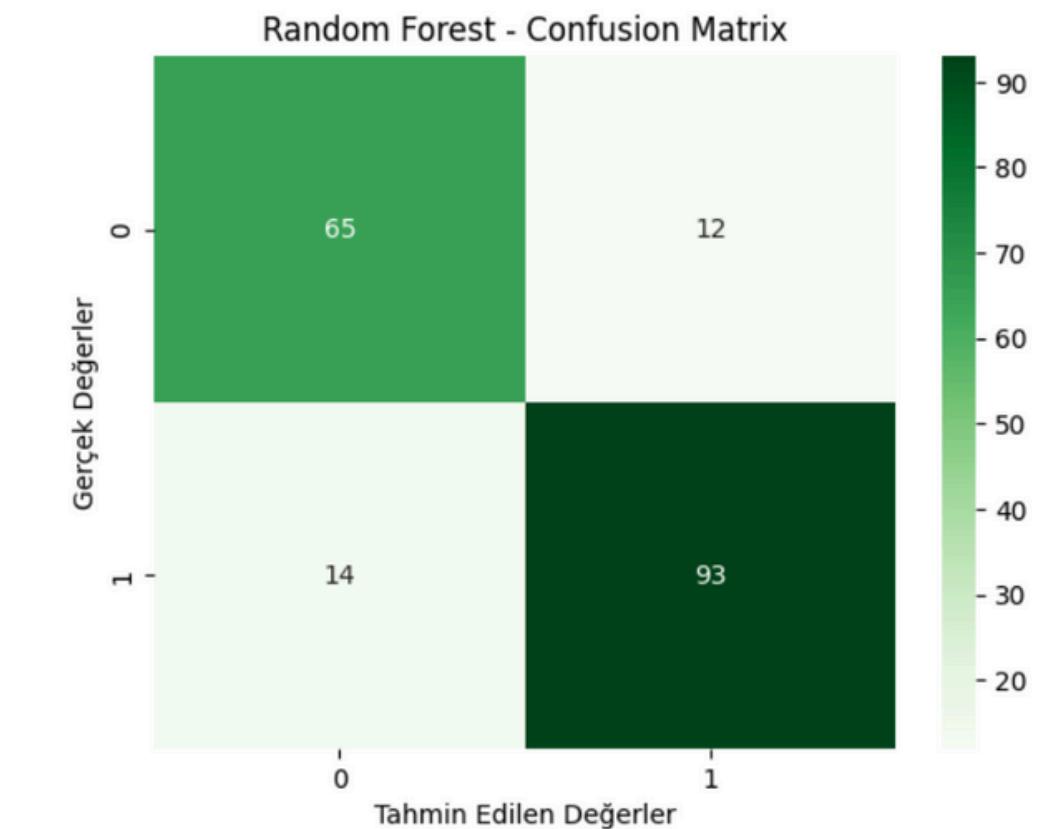
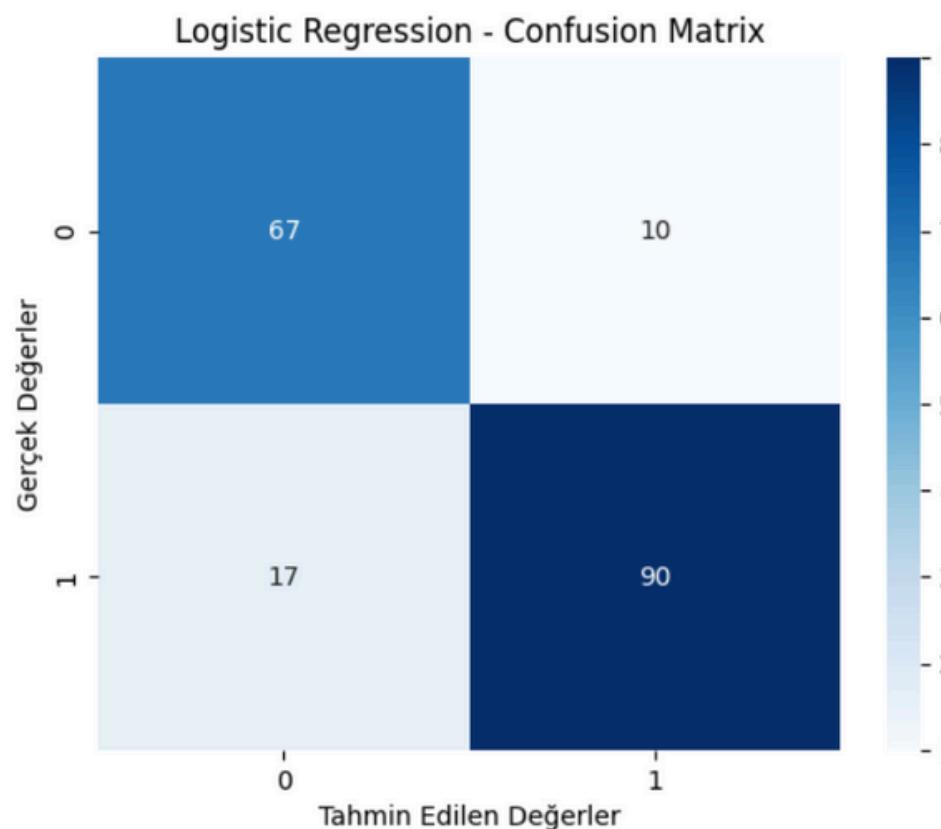
# Uygulanan Modeller ve Karşılaştırma

Confusion matrix bize modelin hangi tahminleri doğru yaptığı, hangilerinde hata yaptığı net bir şekilde gösteriyor.

Random Forest modeli, daha az hata yaparak hem sağlıklı bireyleri hem de hastaları daha doğru sınıflandırmıştır.

**Confusion matrix analizinde yanlış pozitif ve yanlış negatif oranlarının düşük olduğu görülmüştür.**

ROC (Receiver Operating Characteristic) eğrisi, modelin pozitif sınıfı ne kadar iyi tanıdığını gösteren bir grafiktir. Random forest için ROC-AUC skoru 0.93 olarak ölçülmüş, bu da modeli daha başarılı yapan diğer gerekçedir.



# 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    |

ROC AUC: Sınıflar arasındaki ayırt edebilme gücü → 1'e ne kadar yakınsa, o kadar iyi

Accuracy: Tüm doğru tahminlerin oranı (ama dengesiz veride yaniltıcı olabilir)

Precision: 'Hasta' dediğin kişilerin gerçekten hasta olma oranı

Recall: Gerçekten hasta olanların ne kadarını yakalayabildin? Sağlık alanında recall çok kritik çünkü hasta bireyi atlamamak gereklidir.

F1 Score: Precision ve Recall'un dengesi – en güvenilir genel başarı ölçüsüdür Random Forest, hem doğru hastaları bulmada, hem de yanlış alarmlardan kaçınmada daha dengeli bir model olduğunu gösteriyor.

# 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 örüntülerini 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.

# **SONUÇLAR**

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.

## **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.

## **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.



**DİNLEDİĞİNİZ İÇİN TEŞEKKÜR EDERİZ**