Mipower Women Empowerment
Program Group Project

DATE
31/05/2024

HEART DISEASE PREDICTION

EDA + INSIGHTS

PRESENTED BY
Betül Altınel
Şevval Mertoğlu
Rana Gül Serpil
Buse Salı

İÇERİK

1 GİRİŞ

Her yıl dünya genelinde yaklaşık 17.9 milyon can kaybına neden olan kardiyovasküler hastalıklar, diğer kanser türlerinden daha yüksek ölüm oranlarına sahiptir. Bu sebeple özellikle kalp yetmezliği yaşayan bireylerde, sağ kalım oranlarını tahmin etmek büyük önem taşır. Sağ kalım tahmini, hastalığın erken teşhisi ve temel risk faktörlerinin belirlenmesine önemli bir katkı sağlar.

Son 25 yılda bilgisayar bilimlerindeki hızlı ilerleme sayesinde, tıp alanındaki araştırmalar makine öğrenimi ve yapay zeka teknikleriyle birleştirilerek daha etkili hale gelmiştir.

Bu çalışmada, kalp hastalıklarının tespiti için bir veri seti üzerinde çeşitli veri analizi ve görselleştirme teknikleri uygulanmıştır ve bir web projesi oluşturulmuştur.

Proje için Flask Framework'ü kullanılmış ve proje Python diliyle yazılmıştır.

Anahtar Kelimeler: Veri Analizi, Kalp hastalığı tespiti, Medikal veri analizi.

2 VERİ SETİ

Son yıllarda bilişim teknolojilerinin her sektörde yaygın bir şekilde kullanıldığı bilinmektedir. Özellikle sağlık sektöründe hastalıkların belirlenmesinde veri analizi tekniklerinin kullanımı her geçen gün artmaktadır. Erken teşhis aşamasında bilişim teknolojileri faydalı ve başarılı sonuçlar vermektedir. Verimizi Kaggle'dan aldık.

Özellikler:

Yaş, Cinsiyet, Göğüs Ağrısı Tipi, BP (kan basıncı), Kolesterol seviyesi, FBS'nin 120'nin üzerinde olması (açlık kan şekeri), EKG Sonuçları (elektrokardiyogram sonuçları), Max HR (maksimum kalp atış hızı), Egzersiz Anjina durumu, ST Depresyonu (EKG'de ST segmentinin depresyonu), ST'nin Eğimi (EKG'de ST segmentinin eğimi), Damar Sayısı Fluroskopi (floroskopide görülen damar sayısı) ve Talyum Stres testi sonuçlarıdır.

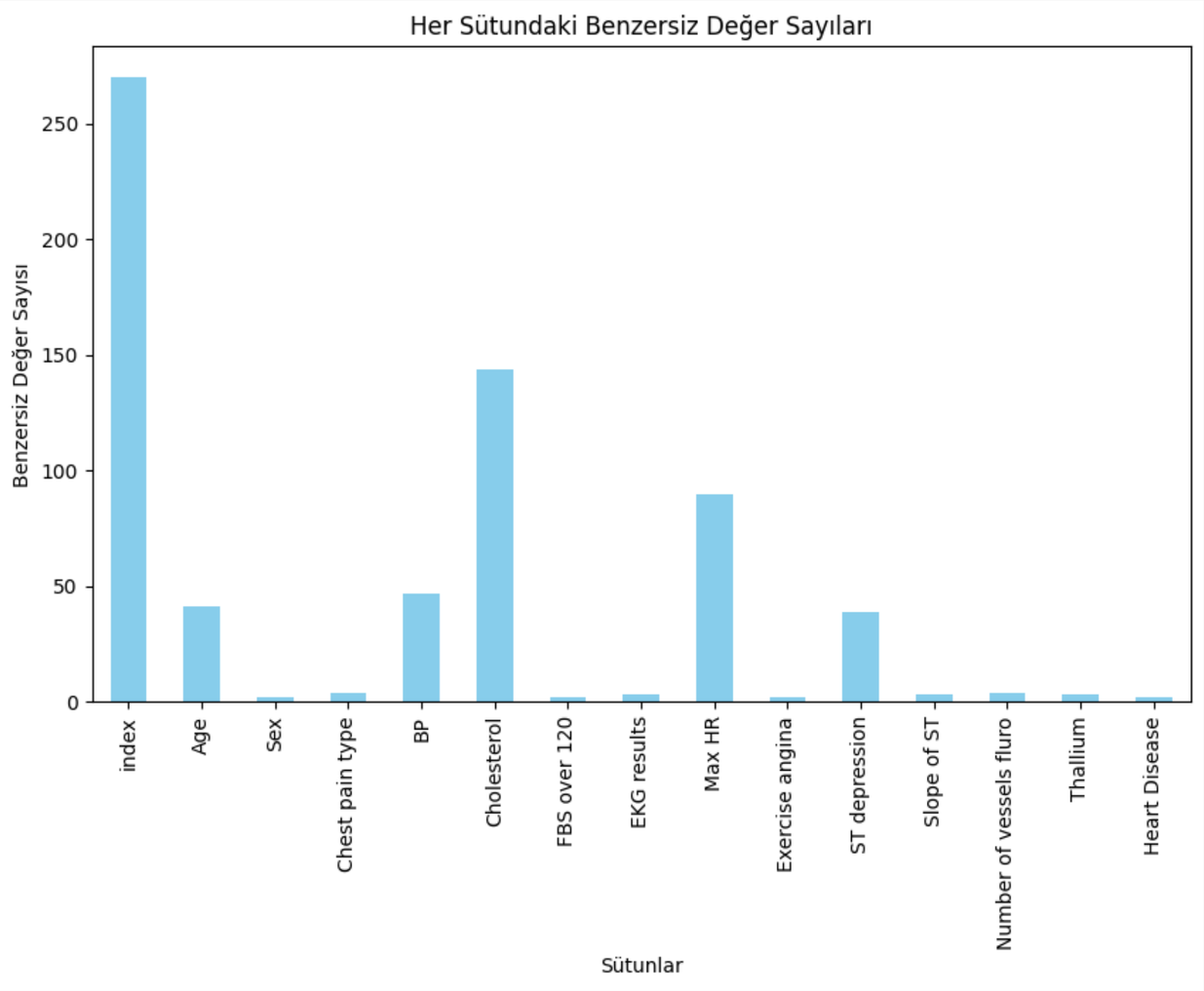
```
data=pd.read_csv('Heart_Disease_Prediction.csv')
data.head()
```

	index	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart_Disease
0	0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
1	1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
2	2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
3	3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
4	4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence

3 YÖNTEMLER

- Kategorik ve Sayısal Değişkenlerin Belirlenmesi
- Kategorik Değişken Analizi
- Sayısal Değişken Analizi
- Hedef Değişken Analizi (Kategorik&Sayısal)
- Aykiri Değerlerin (Outliers) Analizi ve IQR Yöntemi ile Tespiti
- Yinelenen Değerlerin Analizi (Duplicated data)
- Eksik Değerlerin (Missing Values) Analizi
- Encoding (Label Encoding)
- Özellik Ölçeklendirme (Feature Scaling-Standard Scaler)
- Random Forest Classifier
- Confusion Matrix
- Sınıflandırma Parametreleri

KATEGORİK VE NUMERİK DEĞİŞKENLERİN BELİRLENMESİ



Bu grafikte, benzersiz değer sayısına göre sütunlar ayrılmıştır: yüksek benzersiz değer sayısına sahip sütunlar genellikle numerik veri içerir (örneğin, yaş veya kolesterol), düşük benzersiz değer sayısına sahip sütunlar ise genellikle kategorik veri içerir (örneğin, cinsiyet veya EKG sonuçları).

Bu sayede kategorik ve numerik değişkenler belirlenmiştir.

KATEGORİK VE NUMERİK DEĞİŞKENLERİN BELİRLENMESİ

KATEGORİK DEĞİŞKENLER:

- Sex (Cinsiyet)
- Chest pain type (Göğüs ağrısı tipi)
- FBS over 120 (FBS değerinin 120 üzerinde olması)
- EKG Results (EKG sonuçları)
- Exercise Angina (Egzersiz Anjina)
- Slope of ST (ST Eğimi)

NUMERİK DEĞİŞKENLER

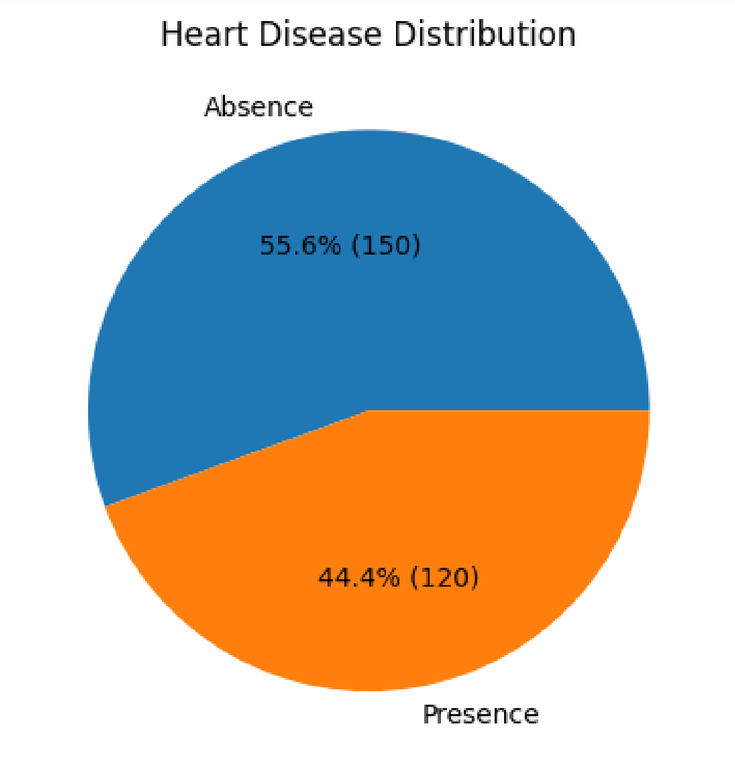
- Age (Yaş)
- BP (Kan basıncı)
- Cholesterol (Kolestrol)
- Max HR (Maksimum kalp atış hızı)
- ST Depression (ST depresyonu)
- Number of vessels furo (Flüroskopide görülen damar sayısı)
- Thallium (Talyum)

DEĞİŞKENLERİN İSTATİKSEL ANALİZİ

	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR
count	270.000000	270.000000	270.000000	270.000000	270.000000	270.000000	270.000000	270.000000
mean	54.433333	0.677778	3.174074	131.344444	249.659259	0.148148	1.022222	149.677778
std	9.109067	0.468195	0.950090	17.861608	51.686237	0.355906	0.997891	23.165717
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000
25%	48.000000	0.000000	3.000000	120.000000	213.000000	0.000000	0.000000	133.000000
50%	55.000000	1.000000	3.000000	130.000000	245.000000	0.000000	2.000000	153.500000
75%	61.000000	1.000000	4.000000	140.000000	280.000000	0.000000	2.000000	166.000000
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000

Bu tablo, veri kümesindeki yaş, cinsiyet, göğüs ağrısı tipi, kan basıncı (BP), kolesterol, açlık kan şekeri (FBS), EKG sonuçları ve maksimum kalp hızı (Max HR) gibi değişkenlerin merkezi eğilim ve yayılım ölçütlerini özetlemektedir.

HEDEF DEĞİŞKEN ANALİZİ

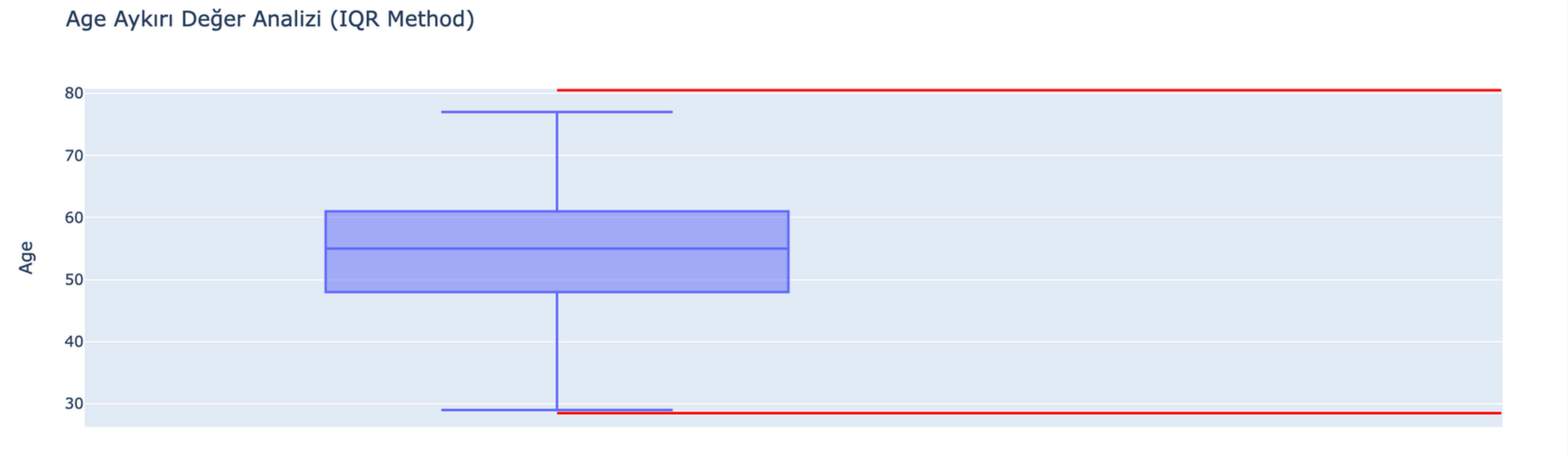


Bu grafikten hedef değişkenin dengeli bir dağılıma sahip olduğu söylenebilir.
Target yani etiket sütunundaki (Heart Disease) değerler ise Absence ve Presence olmak üzere 2'ye ayrılır. Absence; hastalık yok yani sağlıklı anlamına geliyor, Presence ise hastalık var yani hastalıklı anlamına geliyor. Dağılımları pasta dilimi grafiğinde gösterilmiştir.

Hedef Değişken: Kalp hastalığı varlığı (presence: 1) veya yokluğu (absence: 0).

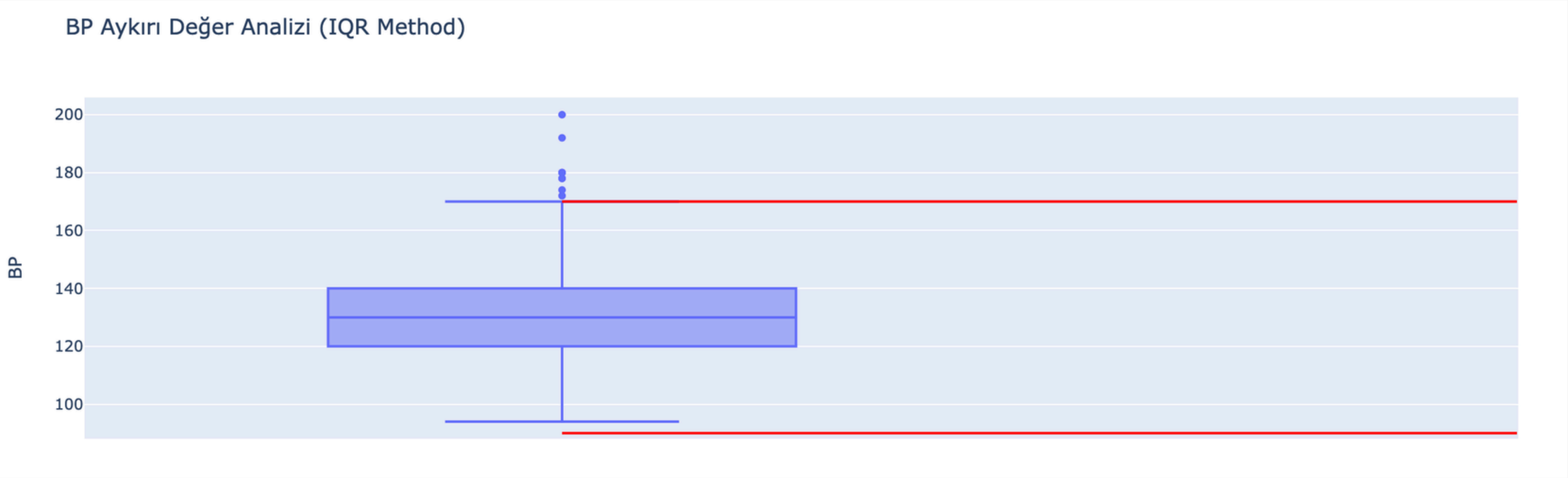
AYKIRI DEĞER ANALİZİ

3.1.AGE AYKIRI DEĞER ANALİZİ (IQR METODU)



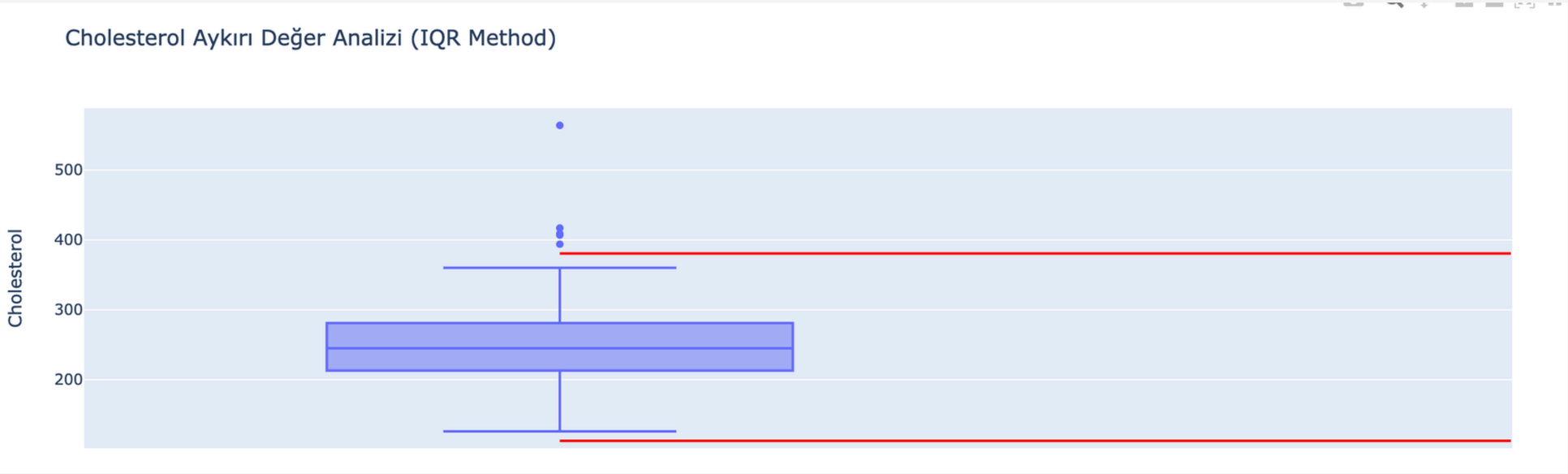
Yaş verisinin çoğunluğunu 45-55 yaş arasında yoğunlaşmış ve 30-70 yaş aralığında yayılım göstermektedir. Ancak 20 ve 80 yaşlarındaki değerlerin aykırı olduğunu göstermektedir.

3.2. BP AYKIRI DEĞER ANALİZİ (IQR METODU)



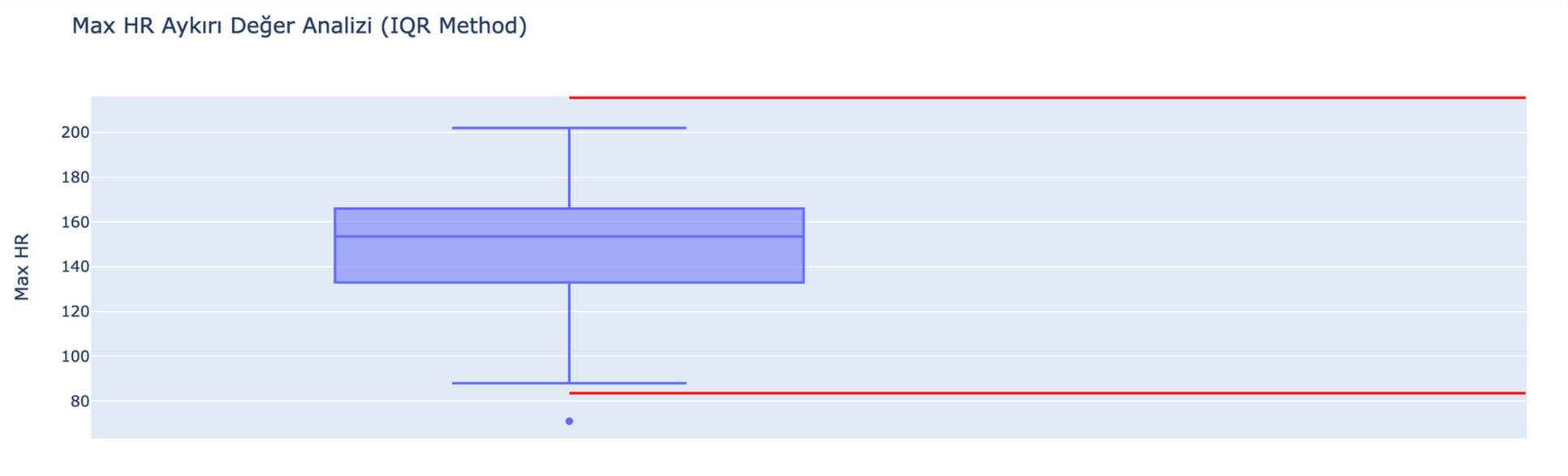
Kan basıncı verilerinin çoğunluğunu 120-140 aralığında yoğunlaşmış ve 160'ın üzerindeki değerler aykırıdır..

3.3. CHOLESTEROL AYKIRI DEĞER ANALİZİ (IQR METODU)



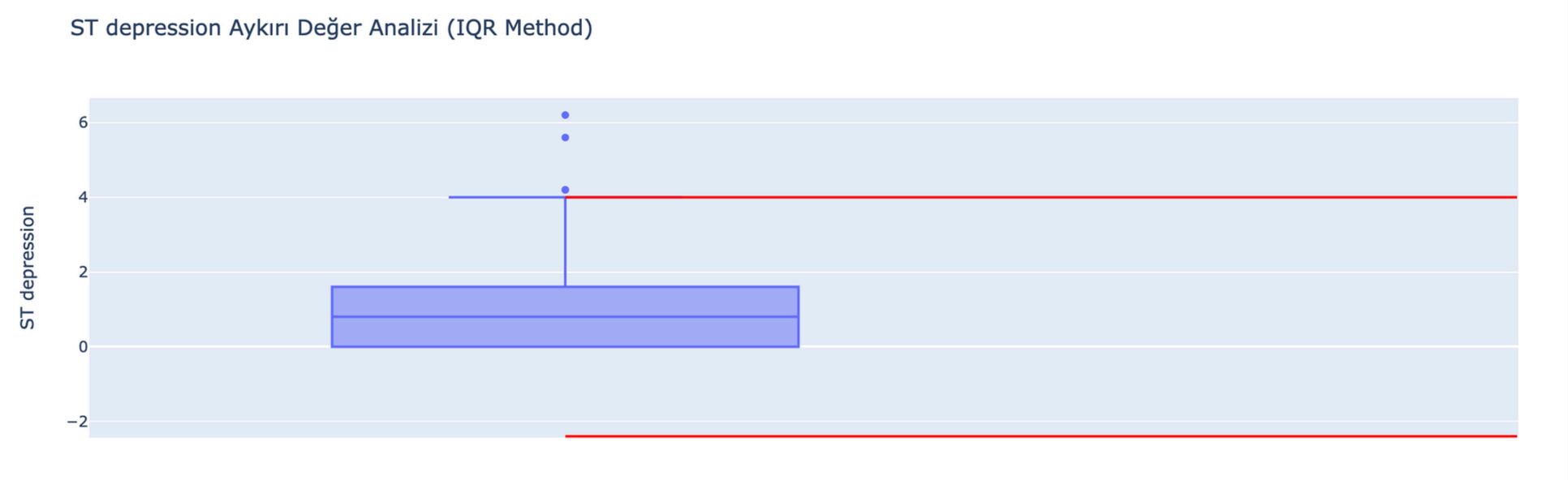
Kolesterol verilerinin çoğunluğu 200-300 aralığında yoğunlaşmıştır ve 400'ün üzerindeki değerler aykırıdır.

3.4. MAX HR AYKIRI DEĞER ANALİZİ (IQR METODU)



Maksimum kalp hızı (Max HR) verilerinin çoğunluğunu 140-160 aralığında yoğunlaşmıştır ve 80'in altındaki değerlerin aykırı olduğunu gözükmektedir.

3.5. ST DEPRESSION AYKIRI DEĞER ANALİZİ (IQR METODU)



ST depresyon verilerinin çoğunluğunun 0-1 aralığında yoğunlaşmaktadır ve 2'nin üzerindeki değerlerin aykırı olduğu görünmektedir.

YİNELENEN DEĞERLERİN ANALİZİ

In [10]:

```
# Duplicated data
duplicated_rows = df[df.duplicated()]

if not duplicated_rows.empty:
    print(duplicated_rows)
else:
    print("There are no duplicated rows in dataframe!")
```

```
There are no duplicated rows in dataframe!
```

Veri setinde yinelenen değer bulunmamıştır.

EKSİK DEĞERLERİN ANALİZİ

```
# Columns that are missing data
missing_values_count = df.isnull().sum()
missing_values_count
```

index	0
Age	0
Sex	0
Chest pain type	0
BP	0
Cholesterol	0
FBS over 120	0
EKG results	0
Max HR	0
Exercise angina	0
ST depression	0
Slope of ST	0
Number of vessels fluro	0
Thallium	0
Heart Disease	0

Veri setinde eksik değer bulunmamıştır.

ENCODING (LABEL ENCODING)

```python
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
df['Heart Disease'] = label_encoder.fit_transform(df['Heart Disease'])

df.head()
```

	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart Disease
0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	1
1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	0
2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	1
3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	0
4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	0

Bu kod ile "Heart Disease" sütunundaki kategorik verileri sayısal değerlere dönüştürmek için LabelEncoder kullanarak veri ön işleme yapılmıştır.

ÖZELLİK ÖLÇEKLENDİRME (FEATURE SCALING - STANDART SCALER)

```
# Feature scaling yapmadan önce hedef değişkeni ayırdık
X = df.drop('Heart Disease', axis=1)  # Özellikler
y = df['Heart Disease']  # Hedef değişken-Target

# StandardScaler'ı kullanarak feature scaling işlemi
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

'Heart Disease' hedef değişkenini veri setinden ayırarak kalan özellikleri StandardScaler kullanarak ölçeklendirmiştir.

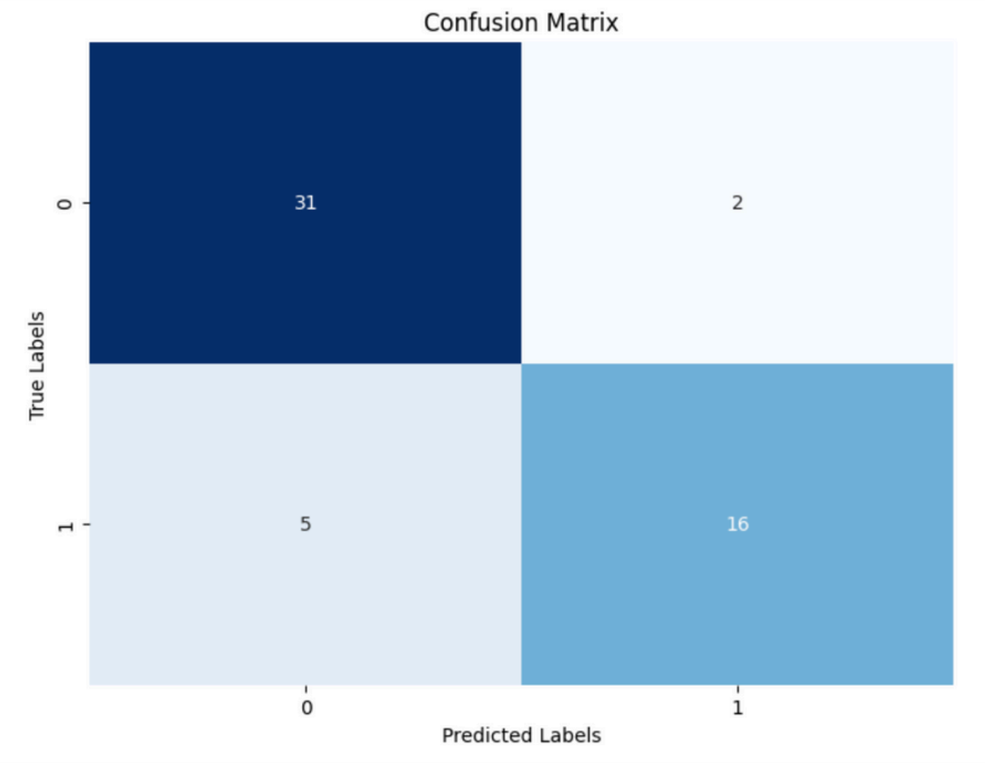
RANDOM FOREST CLASSIFIER

```python
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split

# Eğitim ve test setlerini oluştur
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

# RandomForestClassifier modelini eğit
model = RandomForestClassifier()
model.fit(X_train, y_train)

# Modelin performansını değerlendir
accuracy = model.score(X_test, y_test)
print("Model Accuracy:", accuracy)
```

```
Model Accuracy: 0.8518518518518519
```

Random forest accuracy (doğruluk) değerini %85 bulduk.

CONFUSION MATRIX



- **Doğru Pozitif (TP)**: Sağ alt hücredeki **16** değeri. Bu, gerçek sınıfın 1 olduğu ve tahmin edilen sınıfın da 1 olduğu durumlardır.
- **Doğru Negatif (TN)**: Sol üst hücredeki **31** değeri. Bu, gerçek sınıfın 0 olduğu ve tahmin edilen sınıfın da 0 olduğu durumlardır.
- **Yanlış Pozitif (FP)**: Sağ üst hücredeki **2** değeri. Bu, gerçek sınıfın 0 olduğu ancak tahmin edilen sınıfın 1 olduğu durumlardır.
- **Yanlış Negatif (FN)**: Sol alt hücredeki **5** değeri. Bu, gerçek sınıfın 1 olduğu ancak tahmin edilen sınıfın 0 olduğu durumlardır.

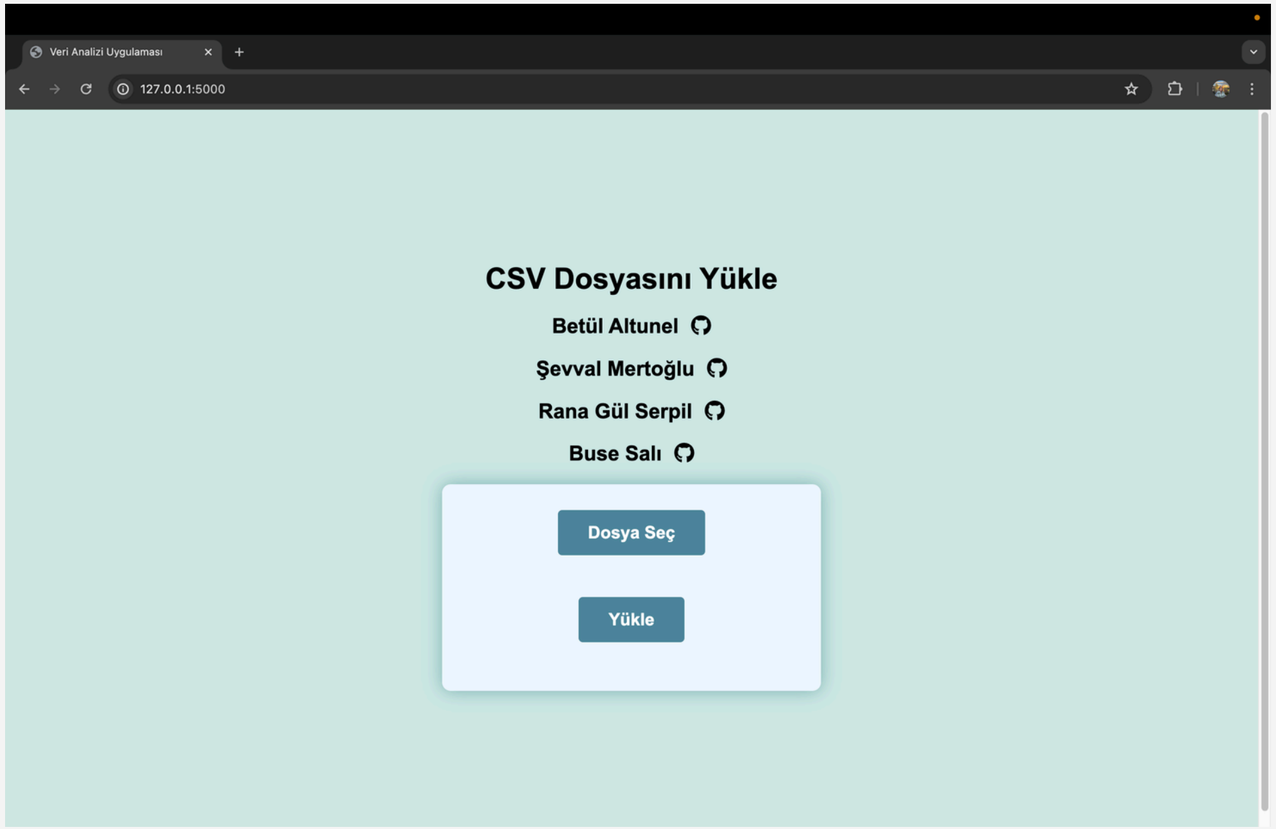
SINIFLANDIRMA PARAMETRELERİ

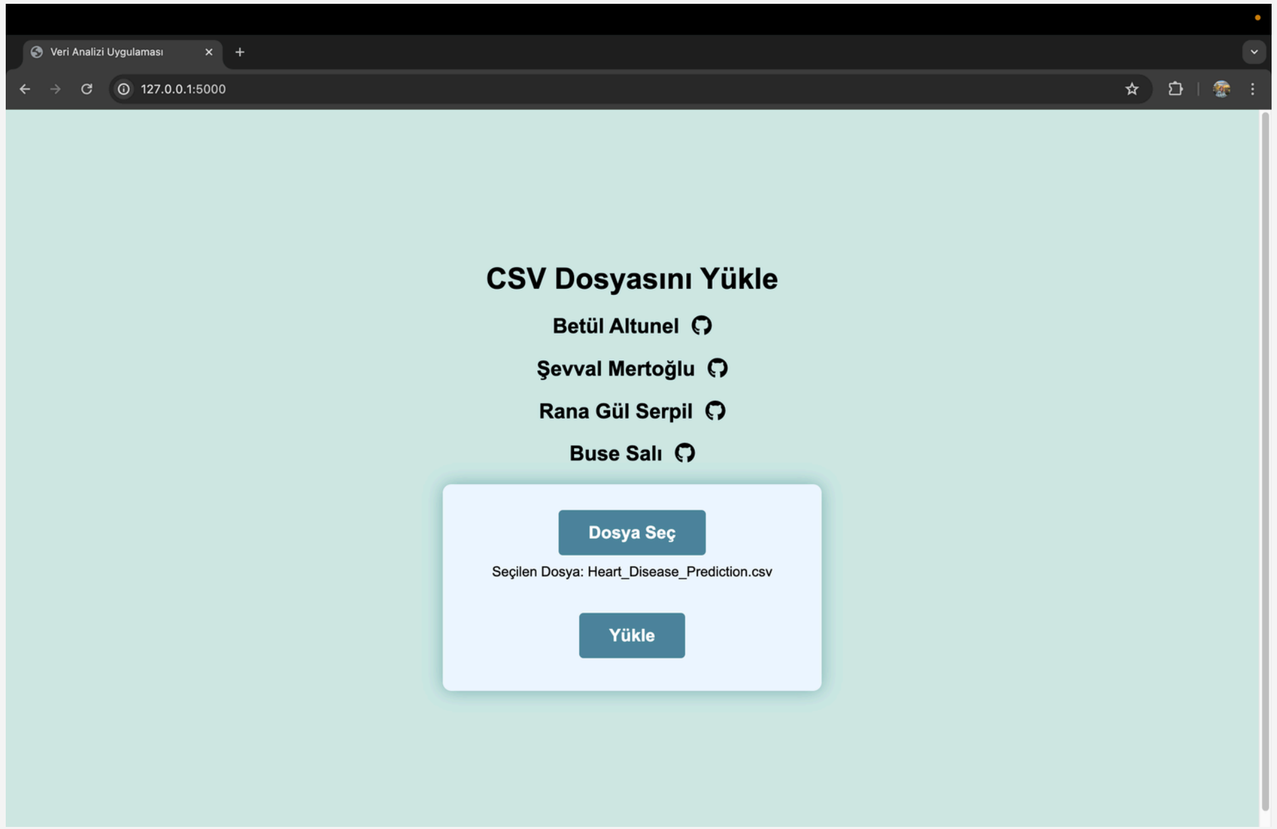
Classification Report:

	precision	recall	f1-score	support
0	0.84	0.94	0.89	33
1	0.88	0.71	0.79	21
accuracy			0.85	54
macro avg	0.86	0.83	0.84	54
weighted avg	0.86	0.85	0.85	54

Sınıflandırma Raporunda, modelin iki sınıf ('Absence' ve 'Presence') için precision, recall, f1-score ve support metriklerini içerir.
Ayrıca, modelin genel doğruluğu (%85), makro ortalama ve ağırlıklı ortalama değerleri de rapordadır.
Bu rapor sayesinde modelin her iki sınıf için ne kadar iyi performans gösterdiğini ayrıntılı bir şekilde değerlendiririz.

4 PROJE ARAYÜZÜ





127.0.0.1:5000/results/Heart_Disease_Prediction.csv

Veri Seti Analiz Sonuçları

Örnek Sayısı: 270

Özellik Sayısı: 14

Kategorik Özellikler: Sex, Chest pain type, FBS over 120, EKG results, Exercise angina, Slope of ST, Number of vessels fluro, Thallium

Sayısal Özellikler: Age, BP, Cholesterol, Max HR, ST depression

Kategorik Görünümlü Kardinal Özellikler:

Sayısal Görünümlü Kategorik Özellikler: Sex, Chest pain type, FBS over 120, EKG results, Exercise angina, Slope of ST, Number of vessels fluro, Thallium, Heart Disease

First 5 Rows

	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart Disease
0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	1
1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	0
2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	1
3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	0
4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	0

Last 5 Rows

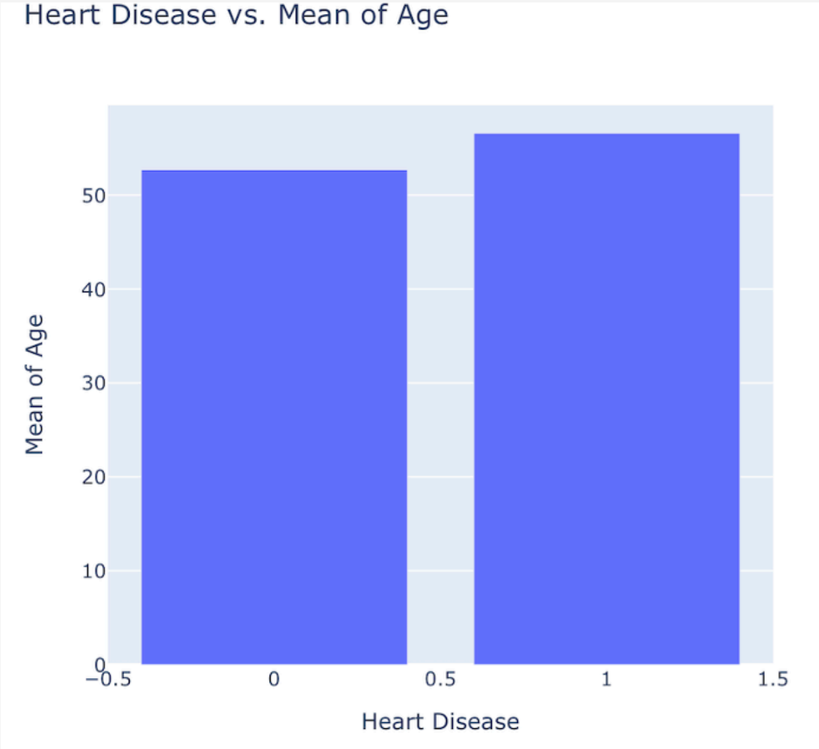
	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart Disease
265	52	1	3	172	199	1	0	162	0	0.5	1	0	7	0

5 INSIGHTS

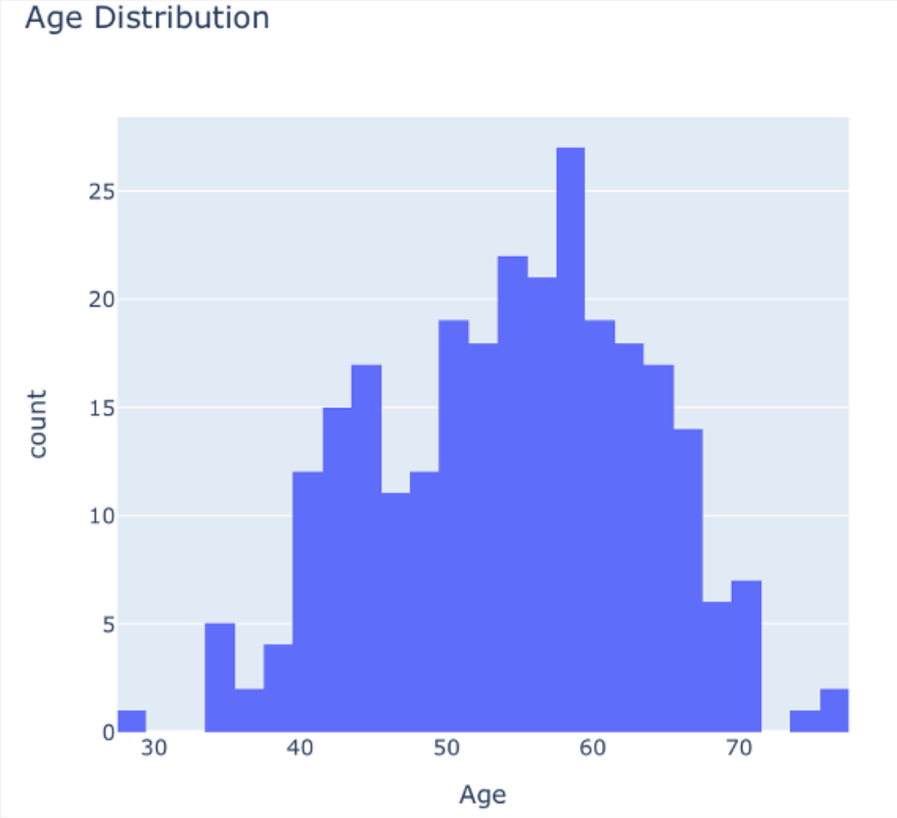
1-AGE (YAŞ)

- **Yaş ve Kalp Hastalığı İlişkisi:** Grafikte görüldüğü üzere, kalp hastalığı olan kişilerin ortalama yaşı, kalp hastalığı olmayanlara göre daha yüksektir. Bu durum, yaşın kalp hastalığı riskini artıran önemli bir faktör olduğunu göstermektedir.

- **Risk Grupları:** Kalp hastalığı olan grubun ortalama yaşının daha yüksek olması, yaşın ilerlemesiyle kalp hastalığı riskinin arttığını söyleyebilir. Bu nedenle, yaşlı bireylerin kalp hastalığı açısından daha yüksek risk altında olduğu söylenebilir.



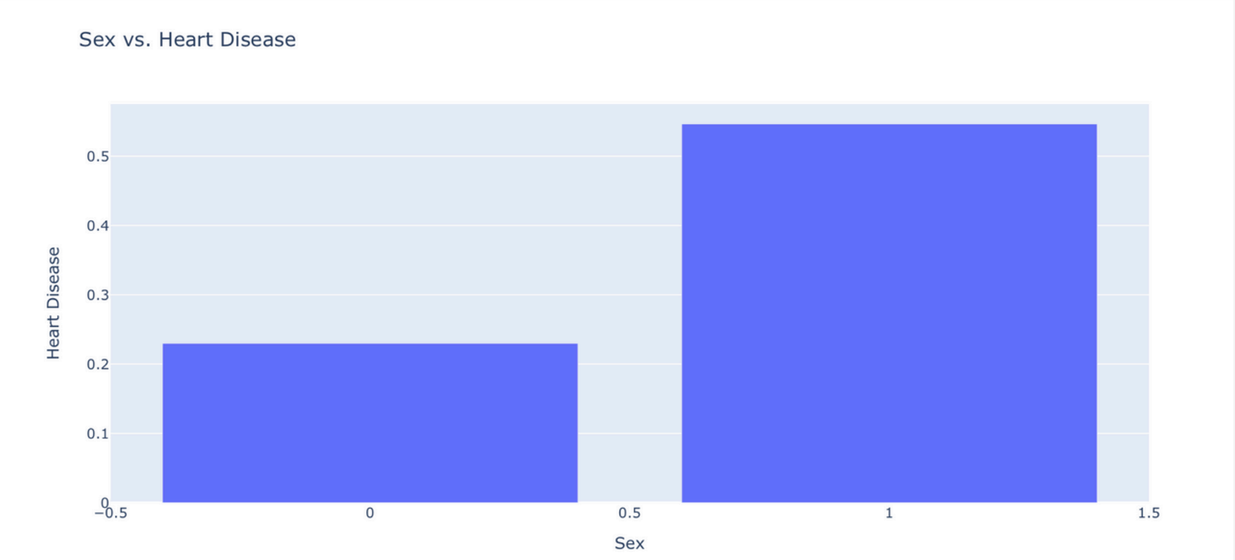
- Kalp hastalığı olmayanlar (0)'ın yaş ortalaması 52-53, kalp hastalığı olanlar (1)'ın yaş ortalaması 56-57'dir.
- Verideki en yoğun yaş grubu, 55-60 yaş aralığında yer almakta olup, bu yaş grubundaki yüksek kişi sayısı, kalp hastalığı görülme sıklığını artırmaktadır.

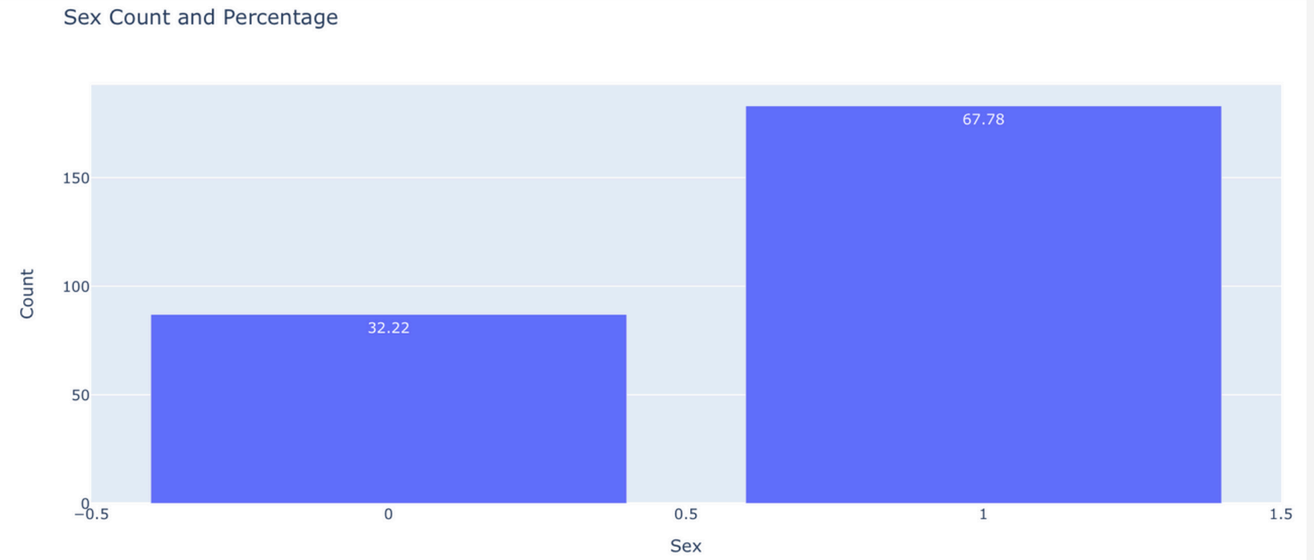


2-SEX (CINSIYET)

Kalp Hastalığı Prevalansı (Görülme Sıklığı):

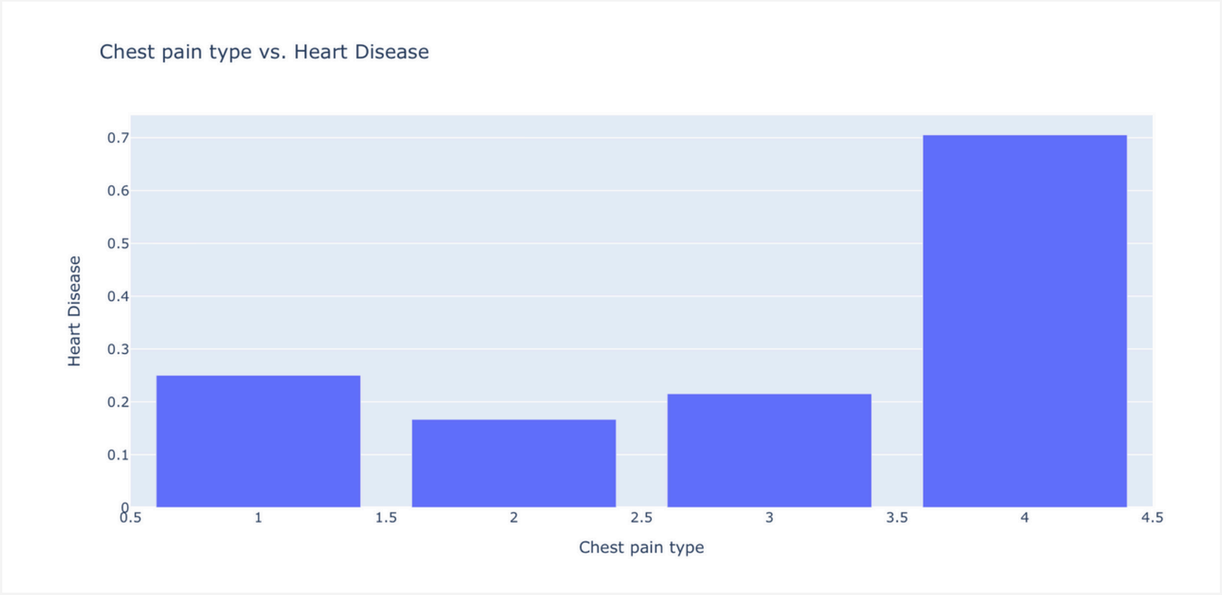
- Kadınlarda (0), kalp hastalığı prevalansı yaklaşık 0.2 (veya %20) civarındadır.
- Erkeklerde (1), kalp hastalığı prevalansı yaklaşık 0.5 (veya %50) civarındadır.

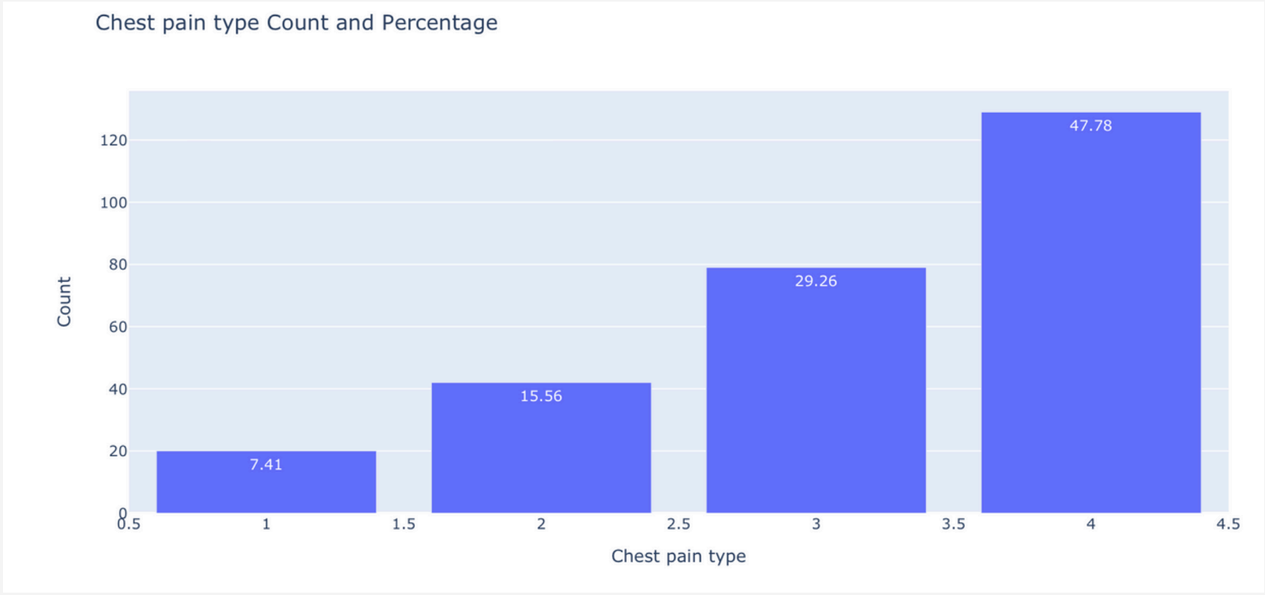




1-CHEST PAIN TYPE (GÖĞÜS AĞRISI TİPİ)

Özellikle göğüs ağrısı türü 4 olan kişilerde kalp hastalığı riski oldukça yüksektir. Bu tür bir göğüs ağrısı ciddiye alınmalı ve tıbbi yardım alınmalıdır. Göğüs ağrısı türü 2 ise en düşük kalp hastalığı oranına sahiptir, bu da bu tür ağrının diğer türlere göre daha az risk taşıdığını göstermektedir.



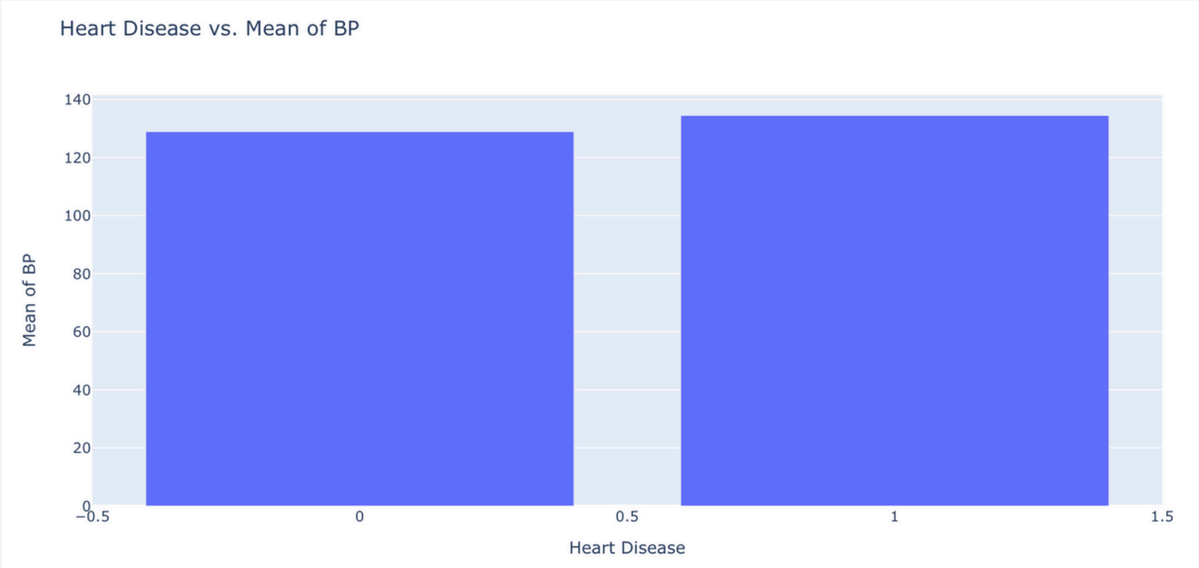


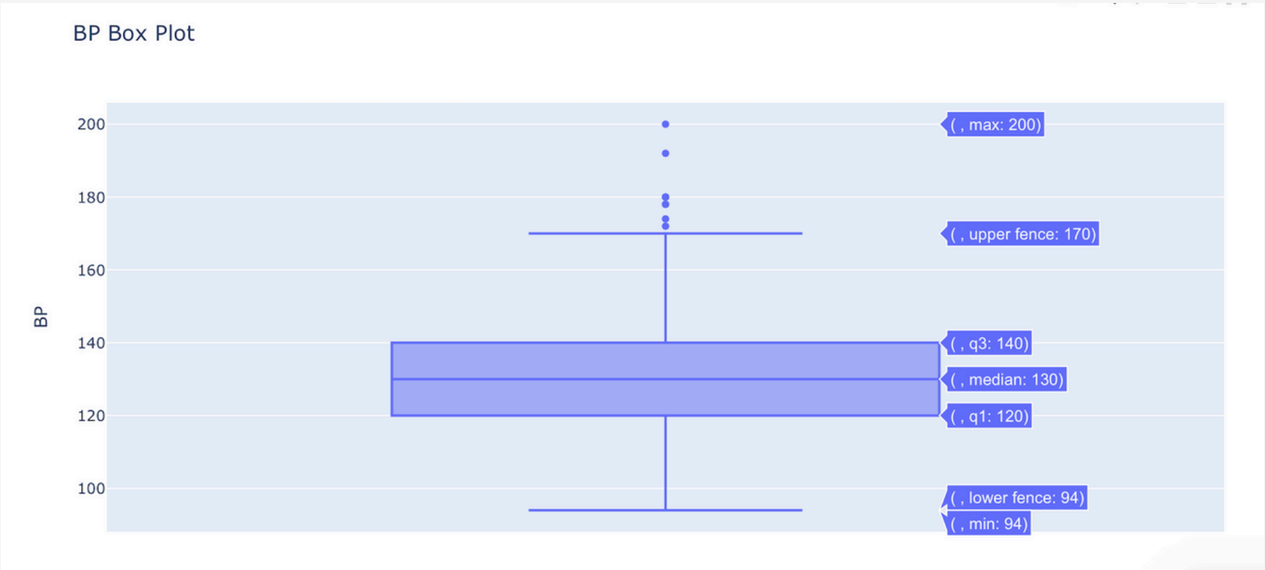
- Göğüs ağrısı türü 4, hem yaygın hem de kalp hastalığı riski açısından en tehlikeli türdür.
- Göğüs ağrısı türü 2, daha az yaygın ve kalp hastalığı riski açısından en az tehlikeli türdür.

Bu bulgular, göğüs ağrısı türlerinin kalp hastalığı riskini ve yaygınlığını anlamada önemli bir rol oynayabilir

4-BP (KAN BASINCI)

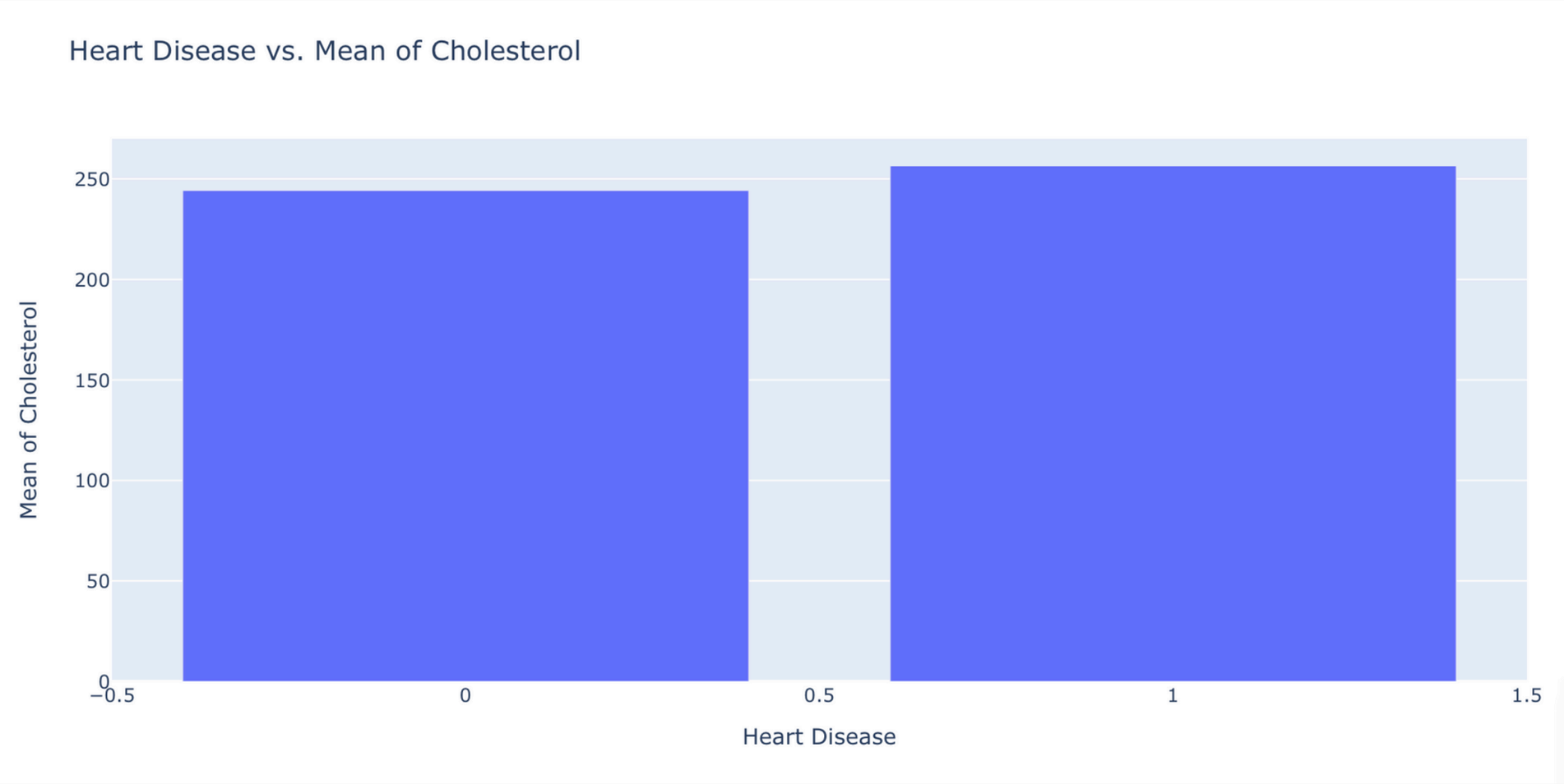
Soldaki grafiğe göre, kalp hastalığı olup olmaması ortalama kan basıncı üzerinde belirgin bir fark yaratmamaktadır. Her iki grup için de ortalama kan basıncı değeri hemen hemen aynıdır, yani 130 civarındadır.





5-CHOLESTEROL (KOLESTROL)

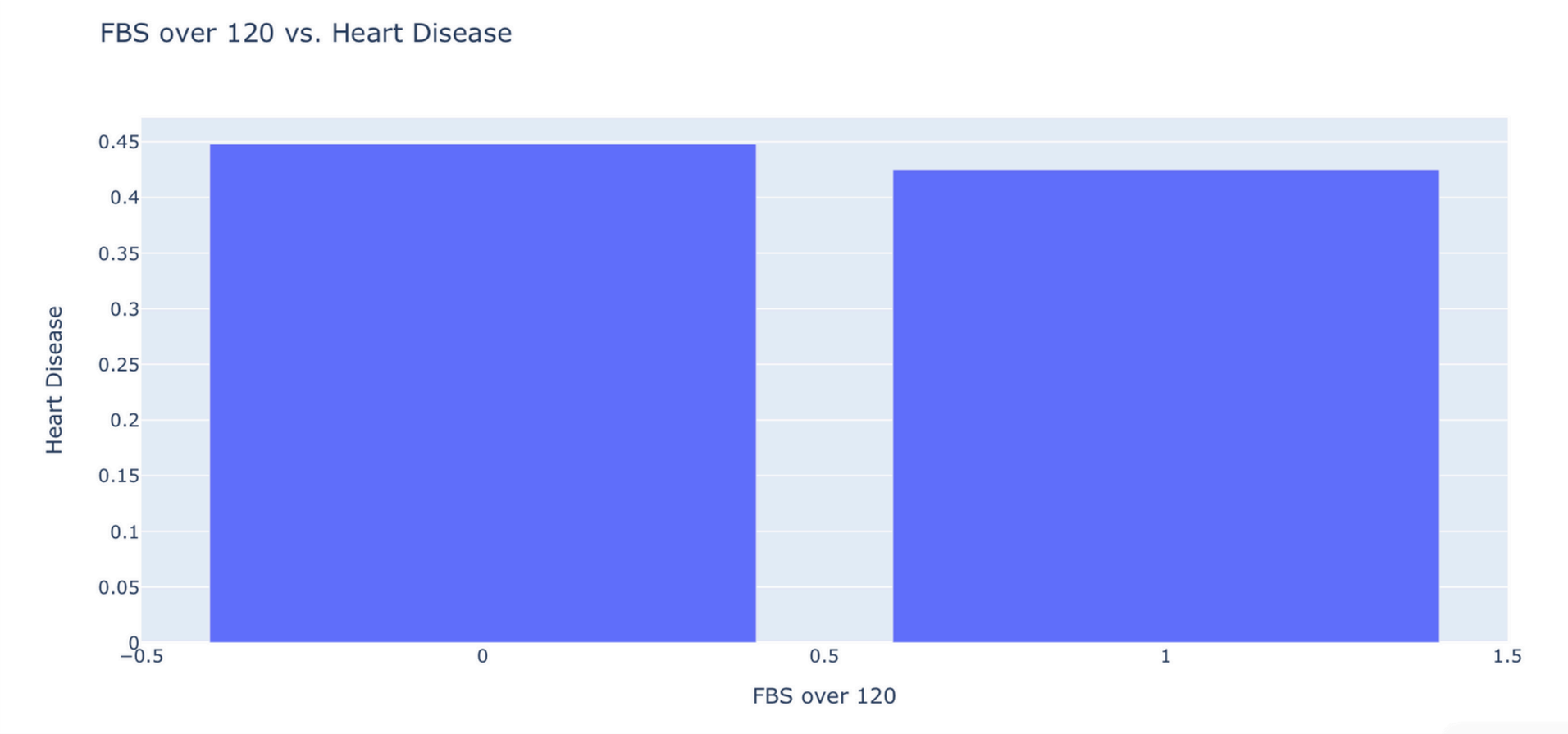
Grafiğe göre, kalp hastalığı olup olmaması ortalama kolesterol üzerinde belirgin bir fark yaratmamaktadır. Her iki grup için de ortalama kolesterol değeri hemen hemen aynıdır, yani 240 civarındadır.



6-FBS (AÇLIK KAN ŞEKERI)

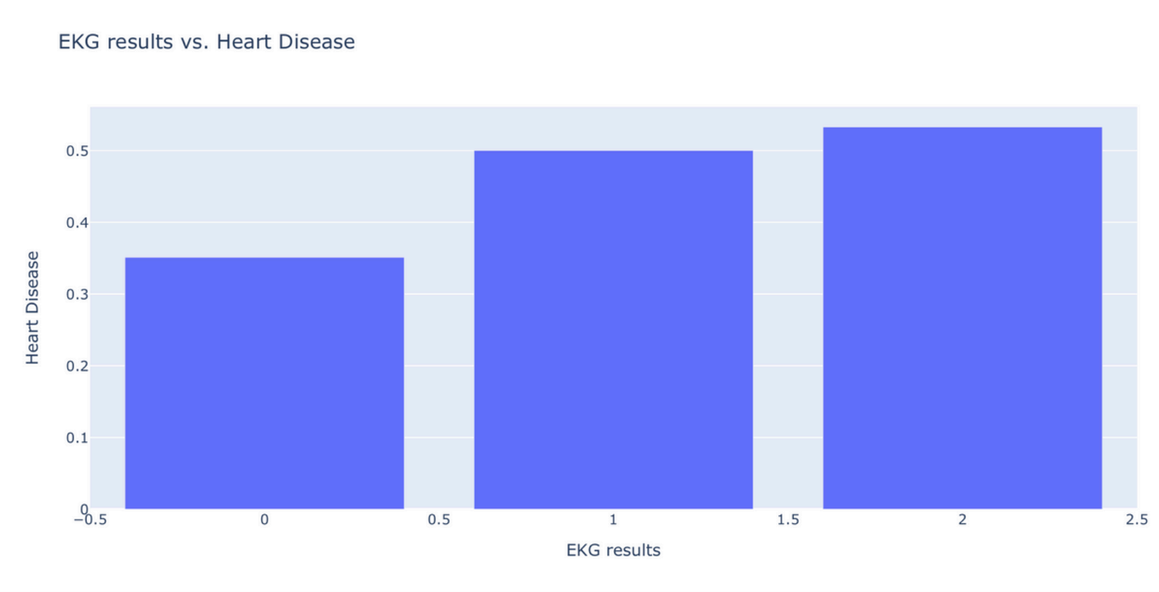
- FBS ≤ 120 (0): Kalp hastalığı oranı yaklaşık %45'tir.
- FBS > 120 (1): Kalp hastalığı oranı yaklaşık %40'tır.

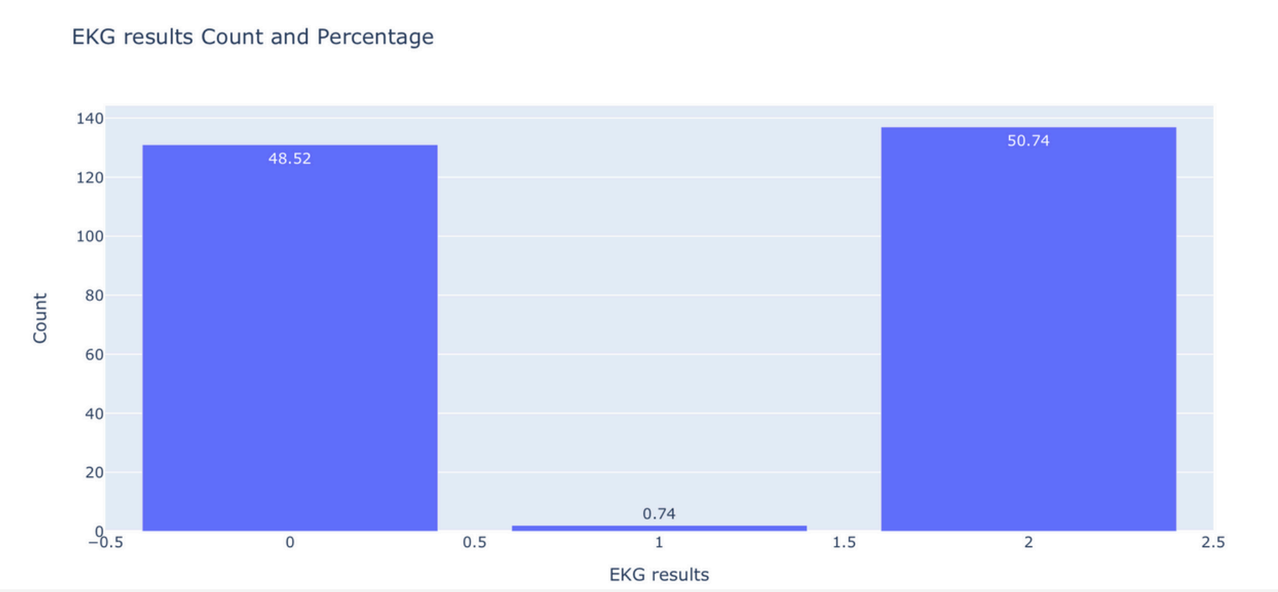
FBS'nin 120'nin üzerinde olup olmaması kalp hastalığı oranında belirgin bir fark yaratmamaktadır. Her iki grup için de kalp hastalığı oranı birbirine yakındır, %40 ile %45 arasında değişmektedir.



7- EKG

Soldaki grafiğe göre, EKG sonuçlarının kalp hastalığı oranı üzerinde belirgin bir etkisi vardır. EKG sonucu 1 ve 2 olan kişilerde kalp hastalığı oranı %50 civarındayken, EKG sonucu 0 olan kişilerde bu oran %30'dur.



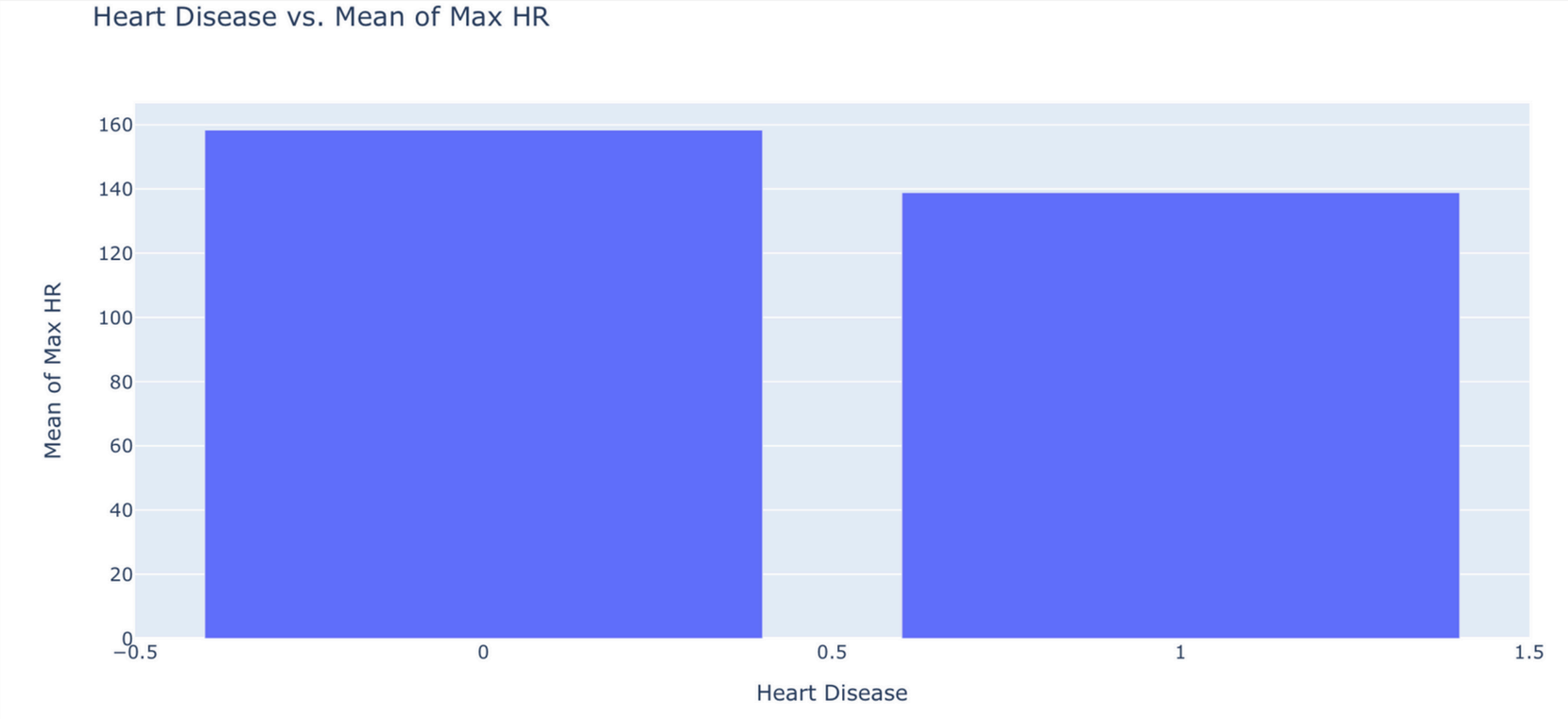


Bu, EKG sonuçlarının çoğunlukla normal (0) veya belirli bir anormalliği (2) gösterdiğini ve orta düzey anormalliğin (1) nadir olduğunu göstermektedir

8- MAX HR (MAKSIMUM KALP ATIŞ HIZI)

Kalp Hastalığı Olmayanlar (0): Ortalama maksimum kalp atış hızı yaklaşık 160 civarındadır.
Kalp Hastalığı Olanlar (1): Ortalama maksimum kalp atış hızı yaklaşık 140 civarındadır.

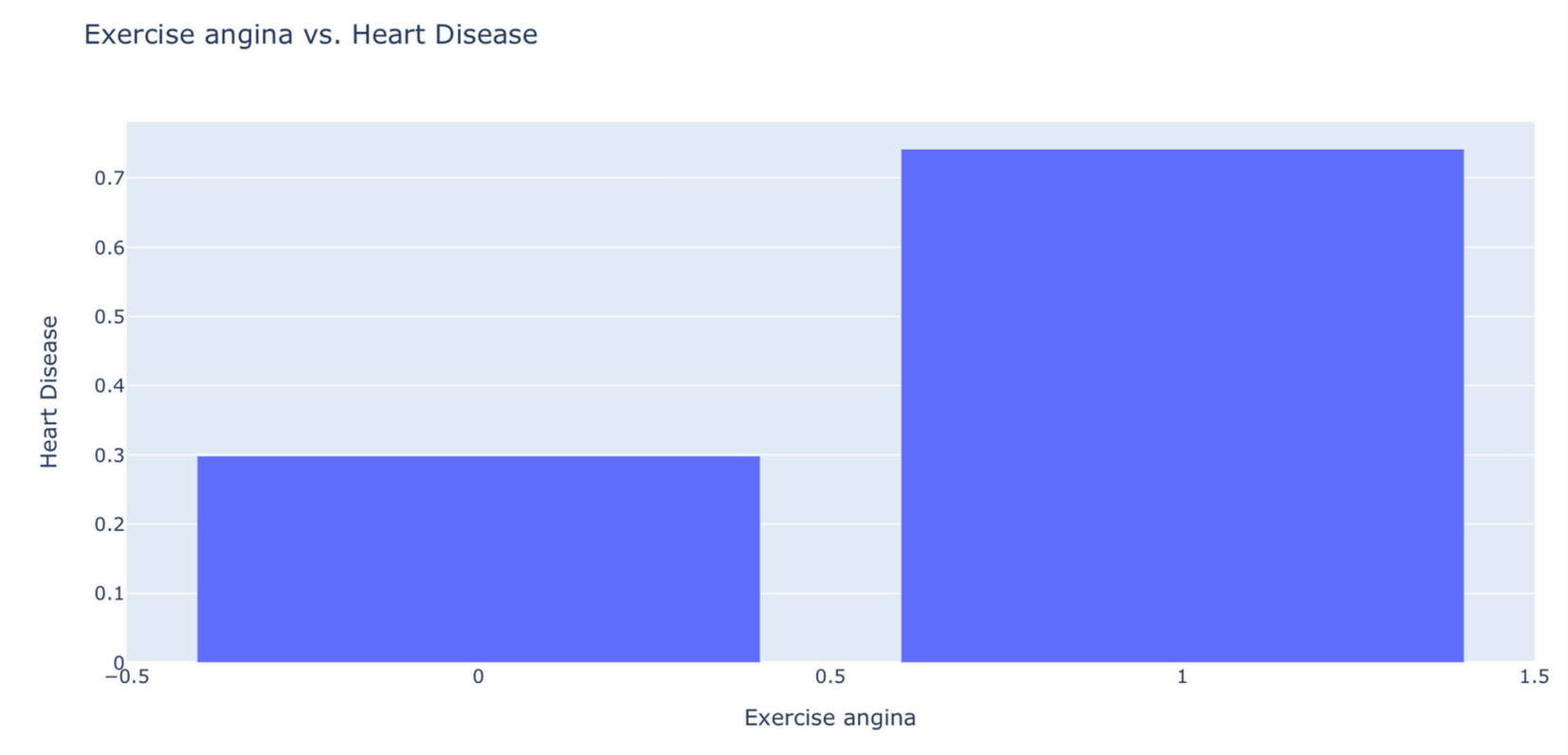
Grafik, kalp hastalığı olan bireylerin ortalama maksimum kalp atış hızının daha düşük olduğunu göstermektedir. Bu durum, kalp hastalığının maksimum kalp atış hızını etkileyebileceğini ve bu bireylerde kalbin daha az performans gösterdiğini işaret edebilir.



9- EXERCISE ANGINA (EGZERSIZ ANJINA)

Egzersiz Sırasında Anjina Yok (0): Kalp hastalığı oranı yaklaşık %30'dur.
Egzersiz Sırasında Anjina Var (1): Kalp hastalığı oranı yaklaşık %70'tir.

Grafik, egzersiz sırasında anjina yaşayan bireylerin kalp hastalığı oranının çok daha yüksek olduğunu göstermektedir. Bu durum, egzersiz sırasında ortaya çıkan anjinanın kalp hastalığı riskinin önemli bir göstergesi olduğunu işaret ediyor olabilir.

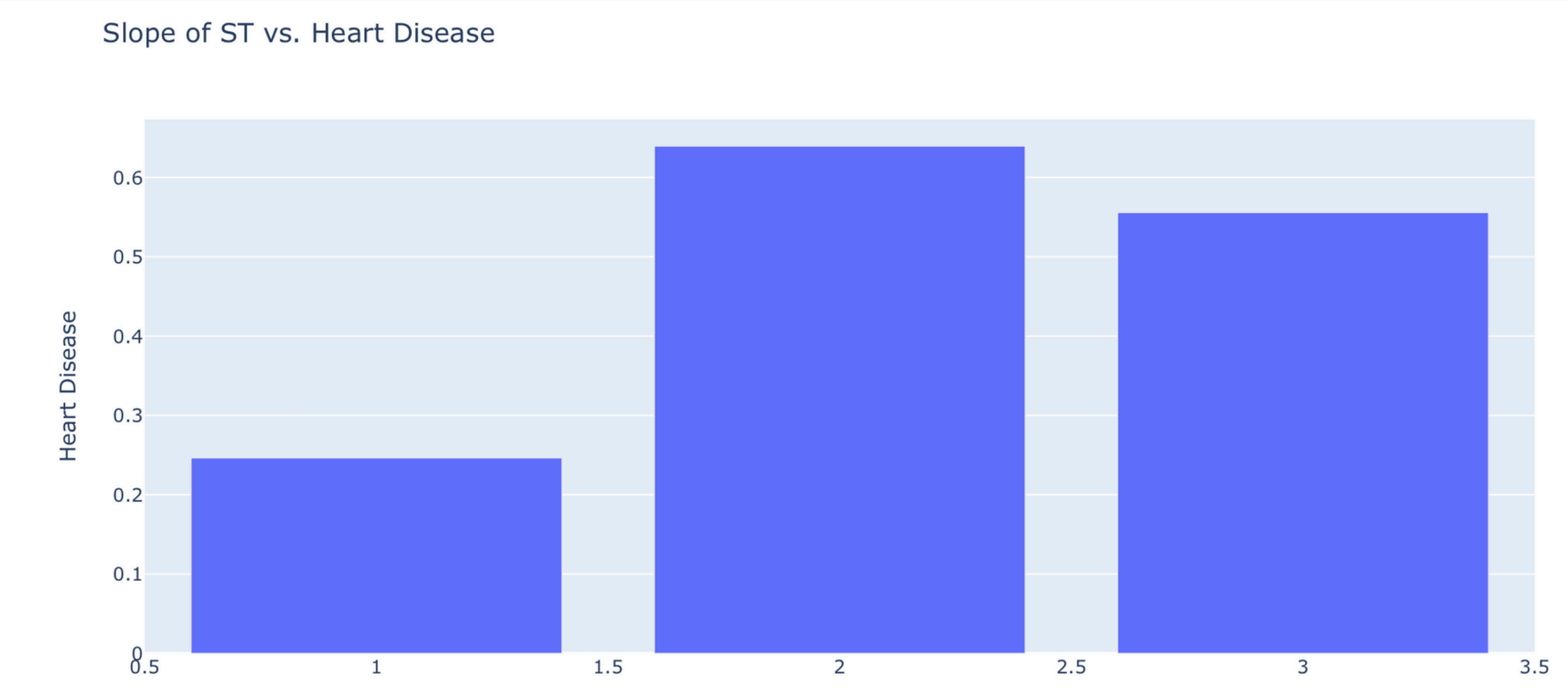


10- ST DEPRESSION(ST DEPRESYONU)

ST depresyonu değeri arttıkça, genel olarak kalp hastalığı oranının da arttığı görülmektedir.
Özellikle ST depresyonu 2 ve üzeri olan değerlerde kalp hastalığı oranı oldukça yüksek olup %60 ila %100 arasında değişmektedir.

11- SLOPE OF ST (ST EĞIMI)

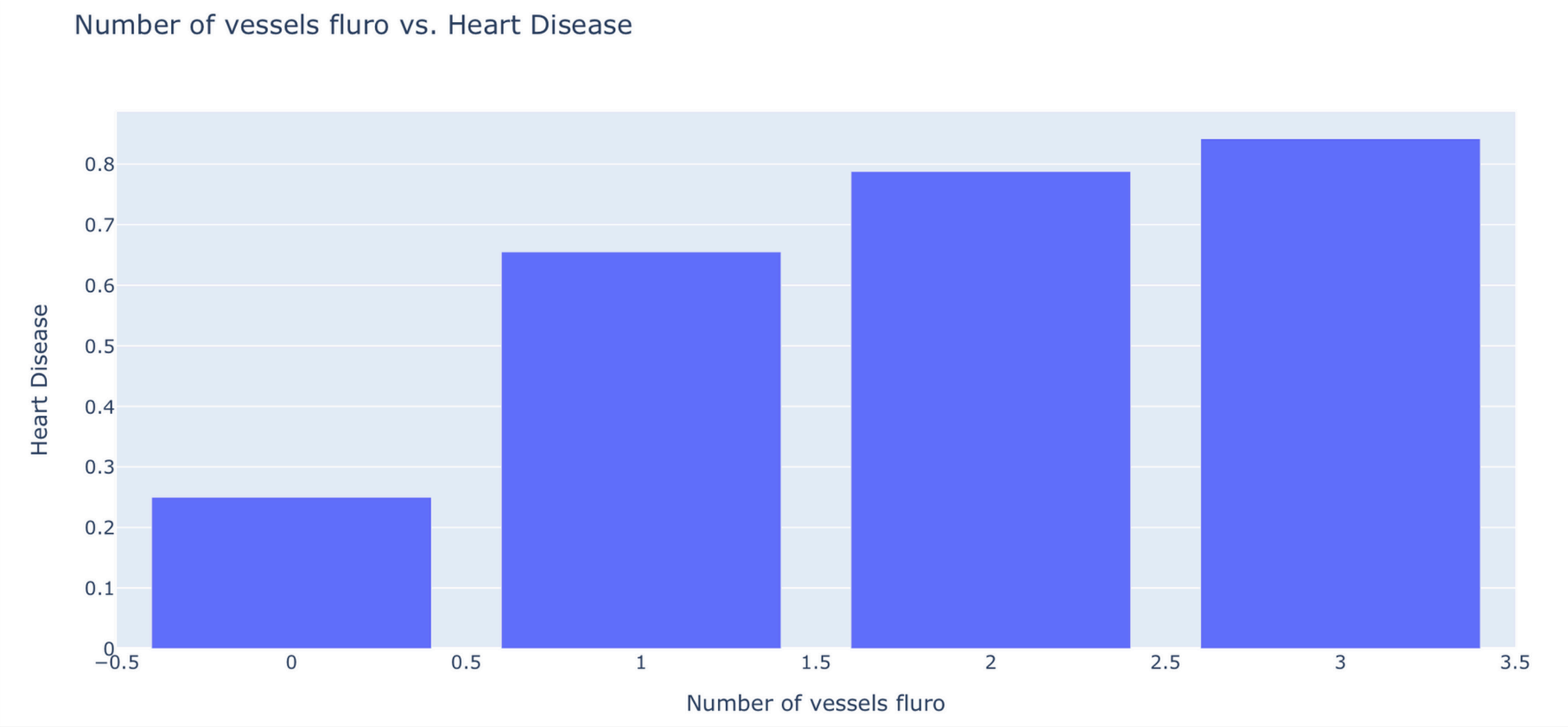
Grafik, ST segmenti eğiminin kalp hastalığı ile ilişkili olduğunu göstermektedir. ST segmenti düz olan bireylerde kalp hastalığı riski en yüksektir, bunu aşağı eğimli olanlar takip eder. ST segmenti yukarı eğimli olanlarda ise kalp hastalığı riski en düşüktür.



12- NUMBER OF VESSELS (DAMAR SAYISI)

- 0 Damar: Kalp hastalığı oranı yaklaşık %25'tir.
- 1 Damar: Kalp hastalığı oranı yaklaşık %60'tır.
- 2 Damar: Kalp hastalığı oranı yaklaşık %70'tir.
- 3 Damar: Kalp hastalığı oranı yaklaşık %80'dir.

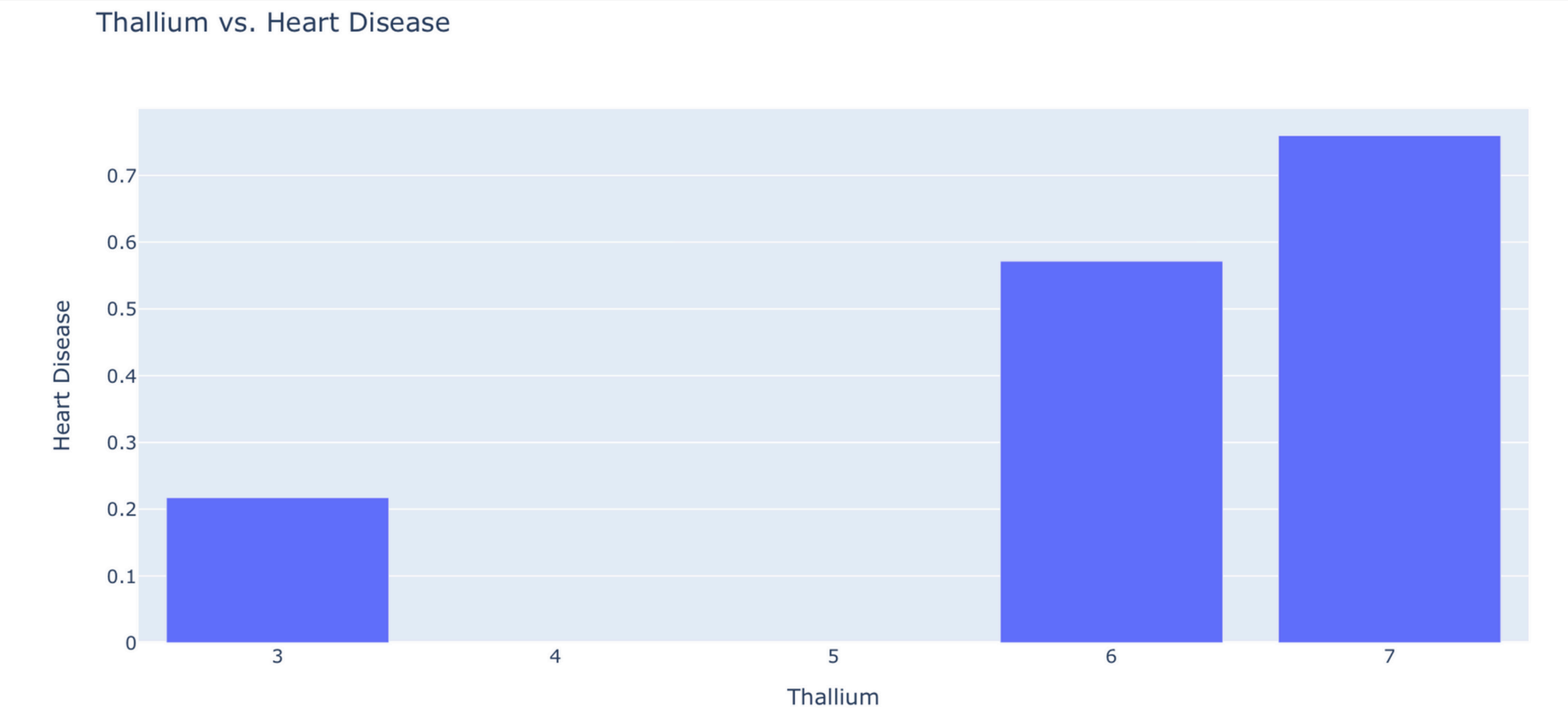
Flüoroskopi ile görülebilen damar sayısı arttıkça kalp hastalığı oranı da artmaktadır.



13- THALLIUM (TALYUM)

- Talyum Testi Sonucu 3: Kalp hastalığı oranı yaklaşık %20'dir.
- Talyum Testi Sonucu 6: Kalp hastalığı oranı yaklaşık %55'tir.
- Talyum Testi Sonucu 7: Kalp hastalığı oranı yaklaşık %70'tir.

Grafiğe göre Talyum testi sonuçlarının kalp hastalığı riski ile ilişkilidir. Talyum testi sonuçları 6 ve 7 olan bireylerde kalp hastalığı oranı belirgin şekilde yüksektir.



6 SONUÇLAR

Giriş:
Bilişim teknolojilerinin sağlık sektöründe kullanımı artmakta olup, veri analizi teknikleri hastalıkların erken teşhisinde başarılı sonuçlar vermektedir.

Veri ve Özellikler:
Veri Kaggle'dan alınmış olup, yaş, cinsiyet, göğüs ağrısı tipi, kan basıncı, kolesterol seviyesi, EKG sonuçları gibi özellikler içerir. Hedef değişken, kalp hastalığı olup olmadığını belirten "Heart Disease" sütunudur.

Analiz ve İşlemler:
Kategorik ve sayısal değişkenler belirlendi.
Değişken analizi, aykırı değerler, yinelenen ve eksik değerler analiz edildi.
Label Encoding ve Standart Scaler kullanılarak veri ön işlendi.
Random Forest modeli kullanılarak %85 doğruluk elde edildi.
Confusion Matrix ve sınıflandırma raporu ile modelin performansı değerlendirildi.

Bulgular:
Yaş, cinsiyet, göğüs ağrısı tipi gibi özellikler kalp hastalığı riskini etkileyen önemli faktörlerdir.
Kalp hastalığı riski yaş, EKG sonuçları, maksimum kalp atış hızı, egzersiz sırasında anjina, ST depresyonu, floroskopi ile görülen damar sayısı ve Talyum testi sonuçları ile ilişkilidir.
Erkeklerde kalp hastalığı görülme olasılığı daha yüksektir.

Sonuç:
Model, kalp hastalığı tahmininde başarılı olup, belirli demografik ve klinik özellikler kalp hastalığı riskini anlamada önemlidir. Bu bulgular, erken teşhis ve tedavi süreçlerine katkı sağlayabilir.

DİNLEDİĞİNİZ İÇİN TEŞEKKÜR EDERİZ.