Makine Öğrenimi ile Kalp Rahatsızlığı Tahmini

Şevval MERTOĞLU

İÇİNDEKİLER

Giriş

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.

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

Proje için Flask Framework'ü kullanılmış ve proje Python diliyle yazılmıştır. Verimizi Kaggle'dan aldım. Verimiz 270 satır 13 kolondan oluşmaktadır.

Kullanılan Yöntemler

- *Kategorik ve Sayısal Değişkenlerin Belirlenmesi*
- *Kategorik değerler için benzersiz değer sayısı*
- *Kategorik Değişken Analizi*
- *Sayısal Değişken Analizi*
- *Hedef Değişken Analizi (Kategorik&Sayısal)*
- *Aykiri Değerlerin (Outliers) Analizi ve IQR Yöntemi ile Tespiti*
- *Eksik Değerlerin (Missing Values) Analizi*
- *Encoding (Label Encoding)*
- *RobustScaler*
- *Kategorik Değerler için One-Hot Encoding*
- *Random Forest Classifier*
- *Confusion Matrix*
- *Sınıflandırma Parametreleri*

İYİLEŞTİRMELER

- *Dubliceted (yinelenen) data kısmının çıkarılması*
- *Age'in kategorik veriye dönüştürülmesi*
- *St Depression'un kategorik veriye dönüştürülmesi*
- *Numerik değerlerin istatistiksel dağılımı*
- *Boxplot'ta sadece numerik değerlerin gösterimi*
- *RobutsScaler kullanımı*
- *Kategorik değerlerde One-hot encoding kullanımı*
- *Benzersiz değer sayılarının sadece kategorik verilerde gösterimi*

Flask Projesi

CSV Dosyasını Yükle

Şevval Mertoğlu

Dosya Seç

Yükle

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:

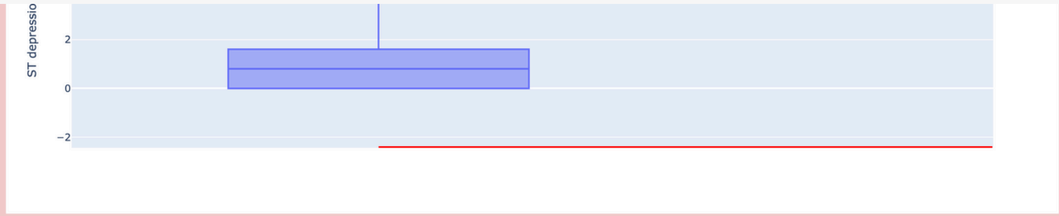
Sayısal Görünümlü Kategorik Özellikler: Sex, Chest pain type, FBS over 120, EKG results, Exercise angina, Slope of ST, Number of vessels fluro, Thallium, Heart Disease

First 5 Rows

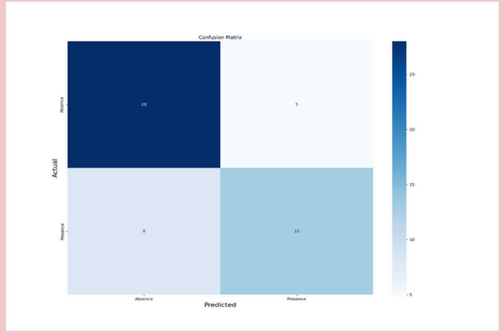
	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart Disease
0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	1
1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	0
2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	1
3	64	1	4	128	263	0	0	105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



Confusion Matrix



KATEGORİK VE NUMERİK DEĞİŞKENLERİN BELİRLENMESİ

```
# Yaş aralıklarını tanımlama
bins = [0, 18, 30, 50, 60, 100]
labels = ['0-18', '18-30', '30-50', '50-60','60+',]
```

```
# Yaş sütununu kategorik hale getirir
df['Age'] = pd.cut(df['Age'], bins=bins, labels=labels, right=False)
```

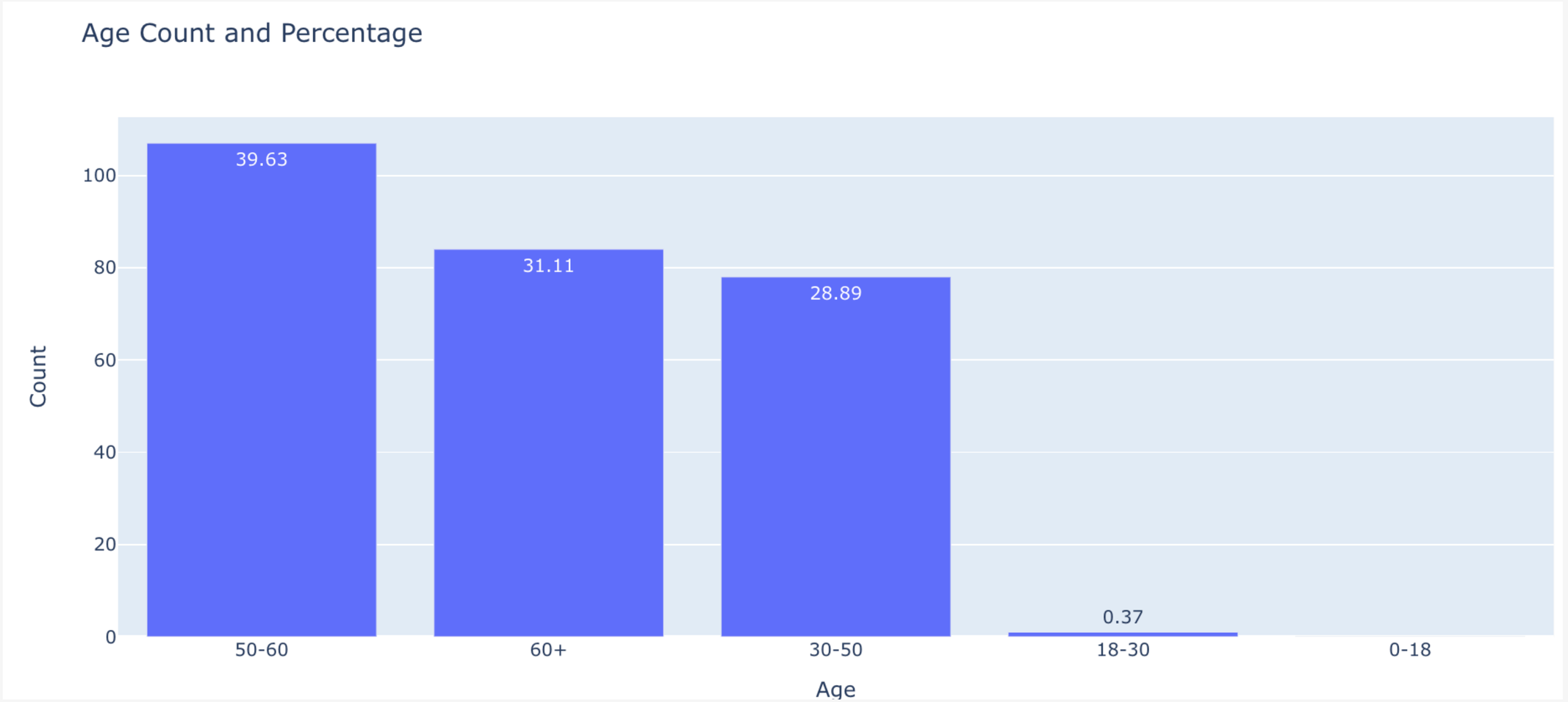
```
st_depression_bins = [0, 0.5, 1, 2, 100]  # 0'dan 100'e kadar olan değerler için
st_depression_labels = ['Normal (0.0-0.5 mm)','Orta (0.5-1 mm)', 'Yüksek (1-2 mm)', 'Ciddi (>2 mm)']
```

```
df['ST depression'] = pd.cut(df['ST depression'], bins=st_depression_bins, labels=st_depression_labels, right=False)
```

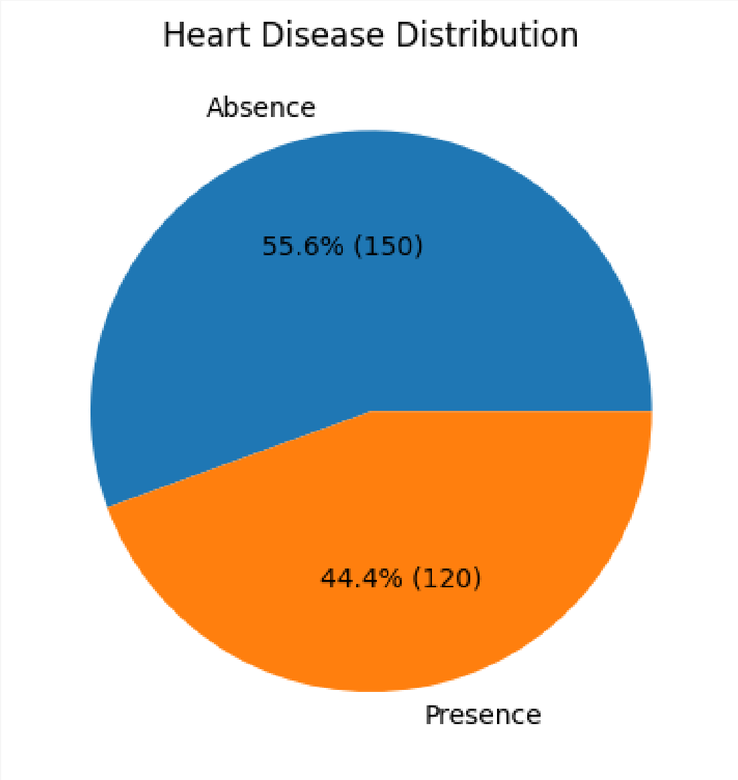
Yaş ve St depression değerlerini numerikten kategorik değer olarak değiştirdim.

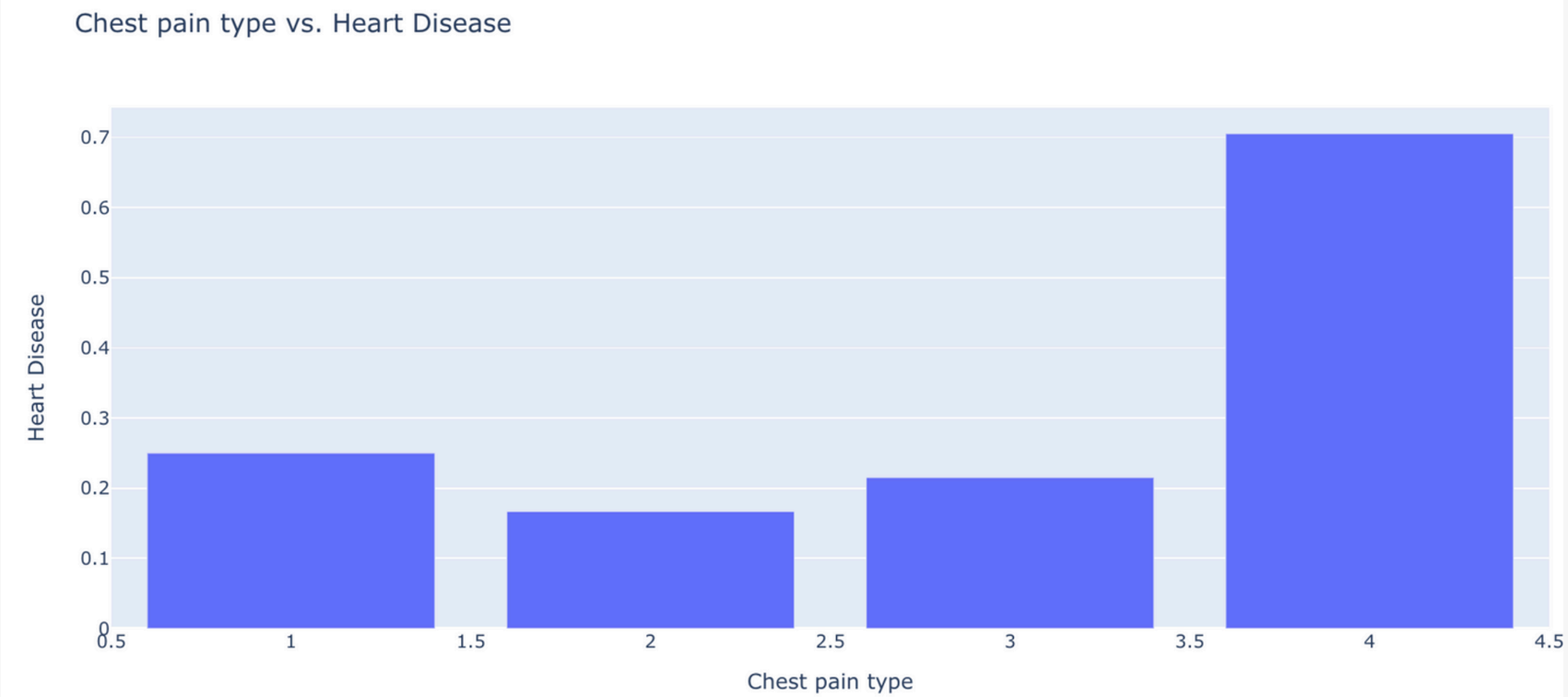
```
cat_cols: 11
num_cols: 3
cat_but_car: 0
num_but_cat: 10
Kategorik Değişkenler: ['Heart Disease', 'Age', 'Sex', 'Chest pain type', 'FBS over 120', 'EKG results', 'Exercise angina', 'ST d
epression', 'Slope of ST', 'Number of vessels fluro', 'Thallium']
Sayısal Değişkenler: ['BP', 'Cholesterol', 'Max HR']
```

• Kategorik Değişken Analizi

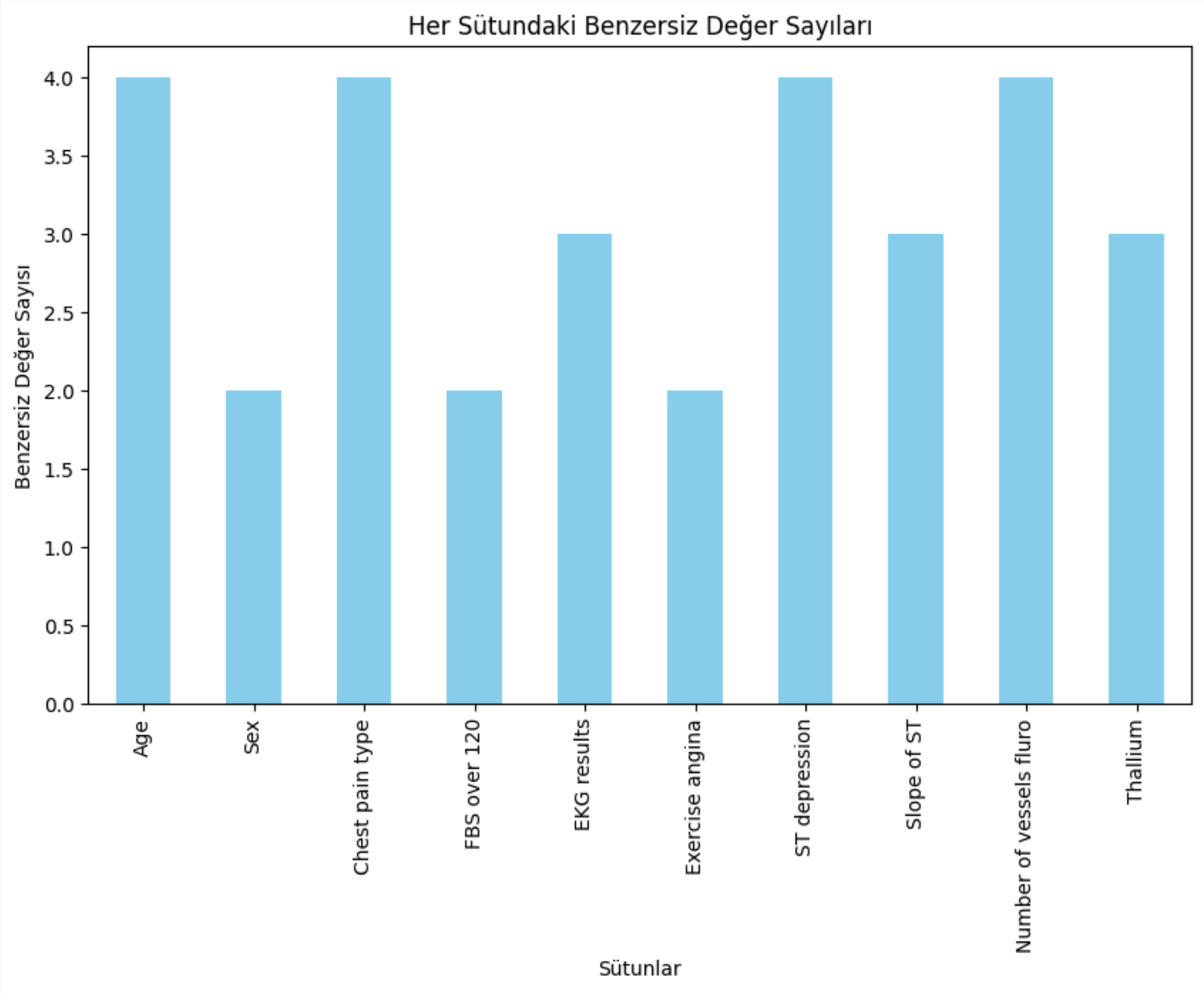


Hedef Değişken Analizi (Kategorik)

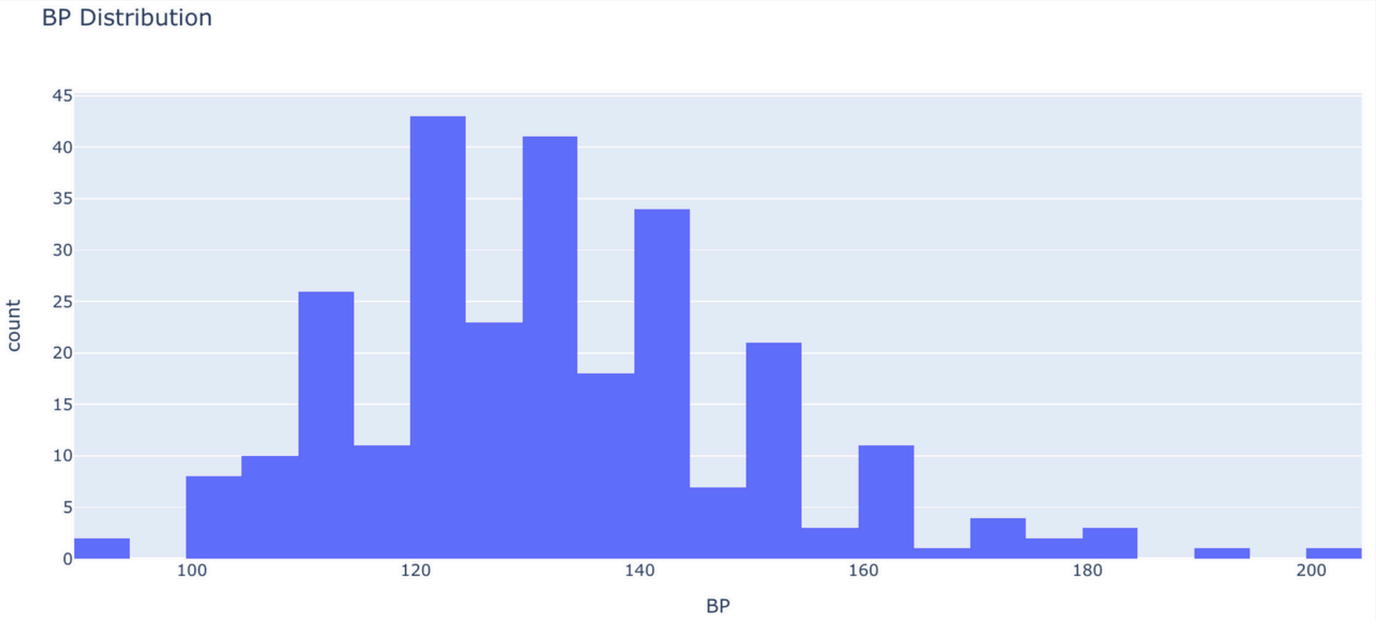


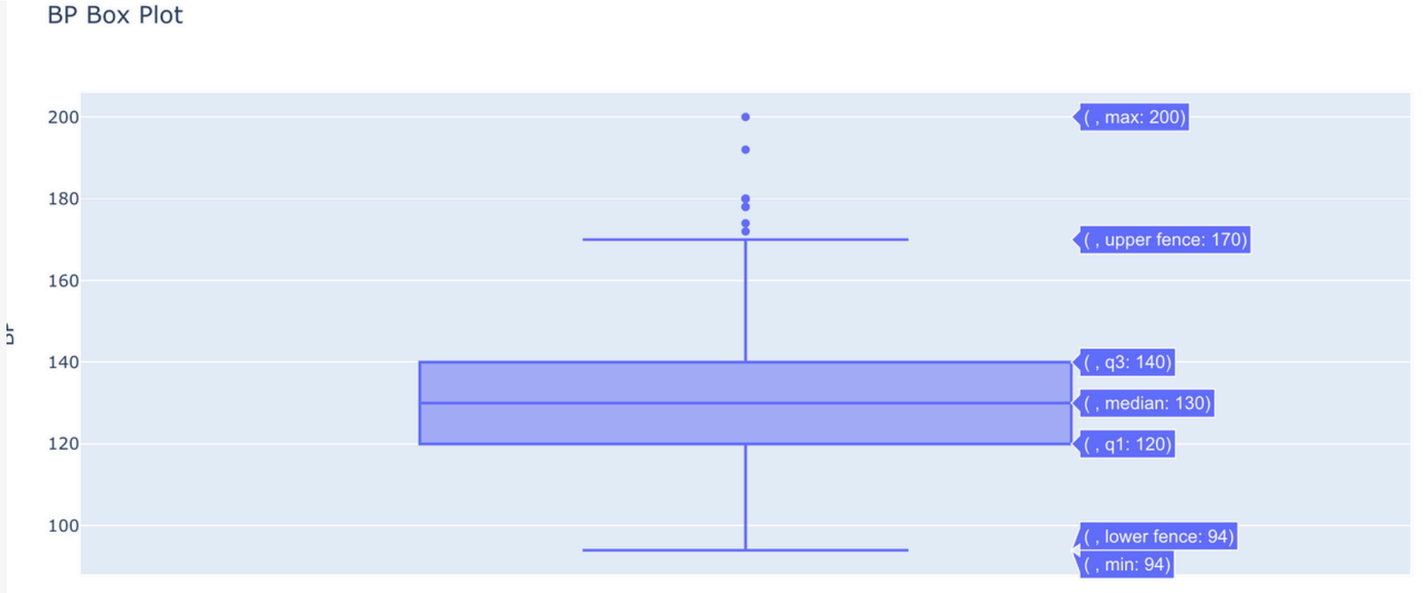


Kategorik değerlerdeki benzersiz değer sayısı

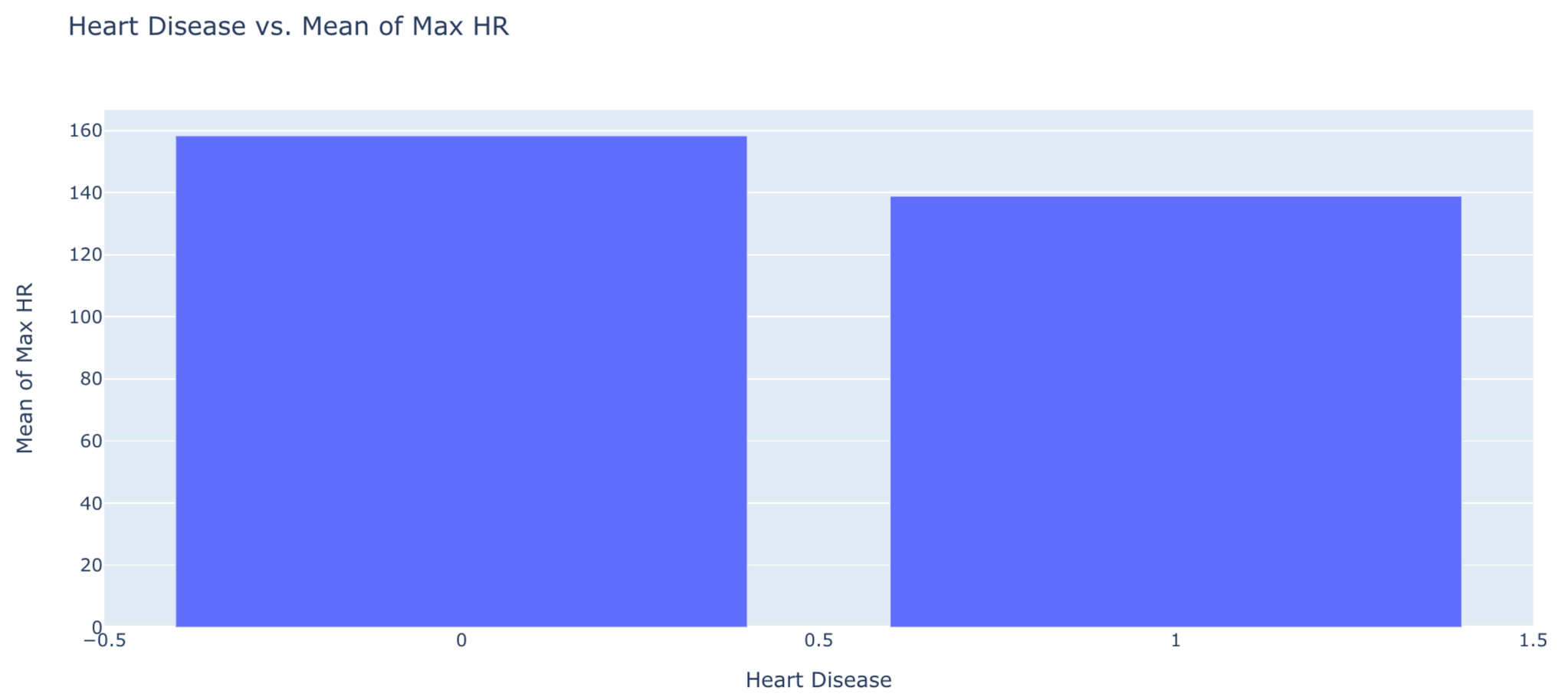


• Sayısal Değişken Analizi





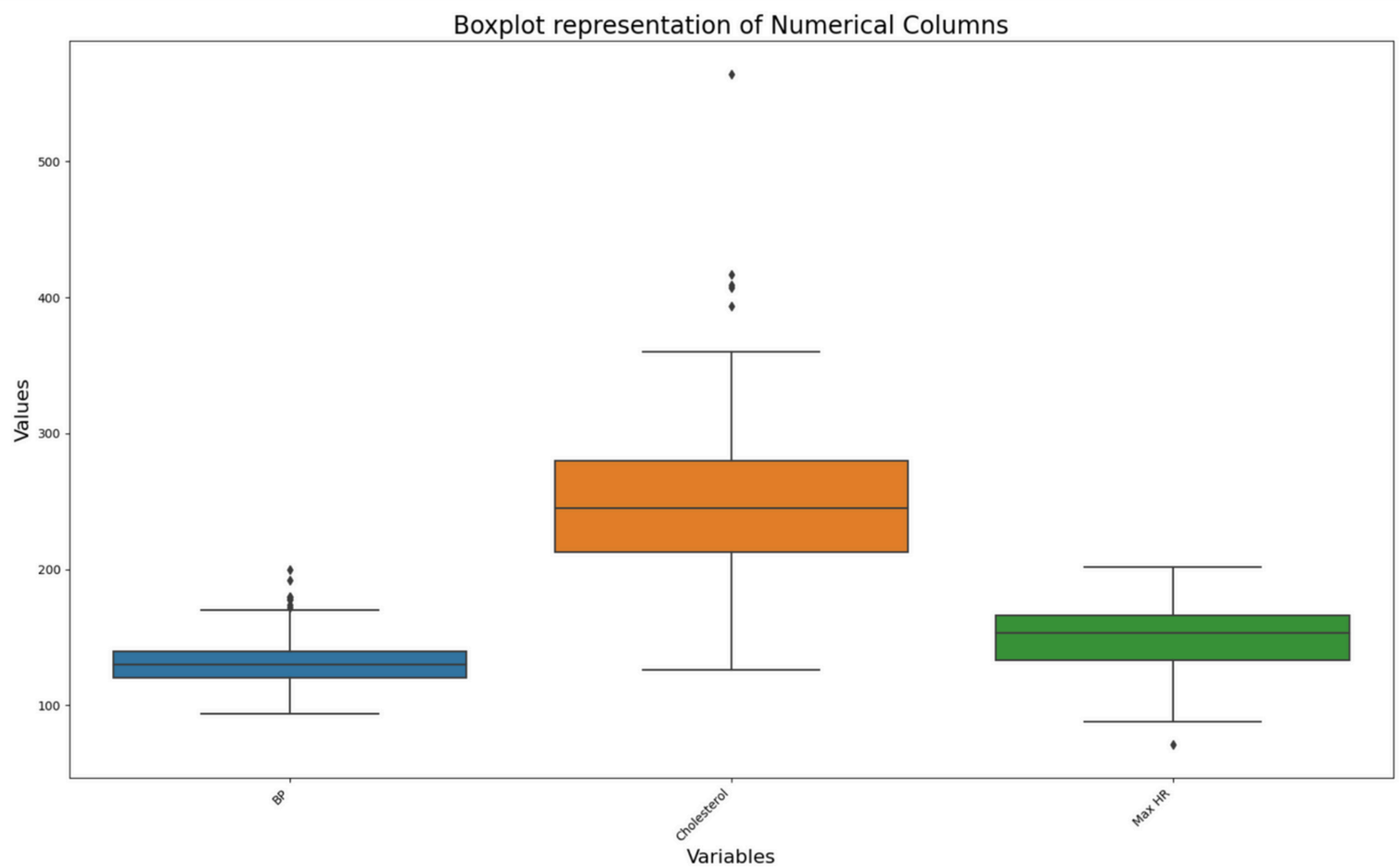
Hedef Değişken Analizi (Sayısal)



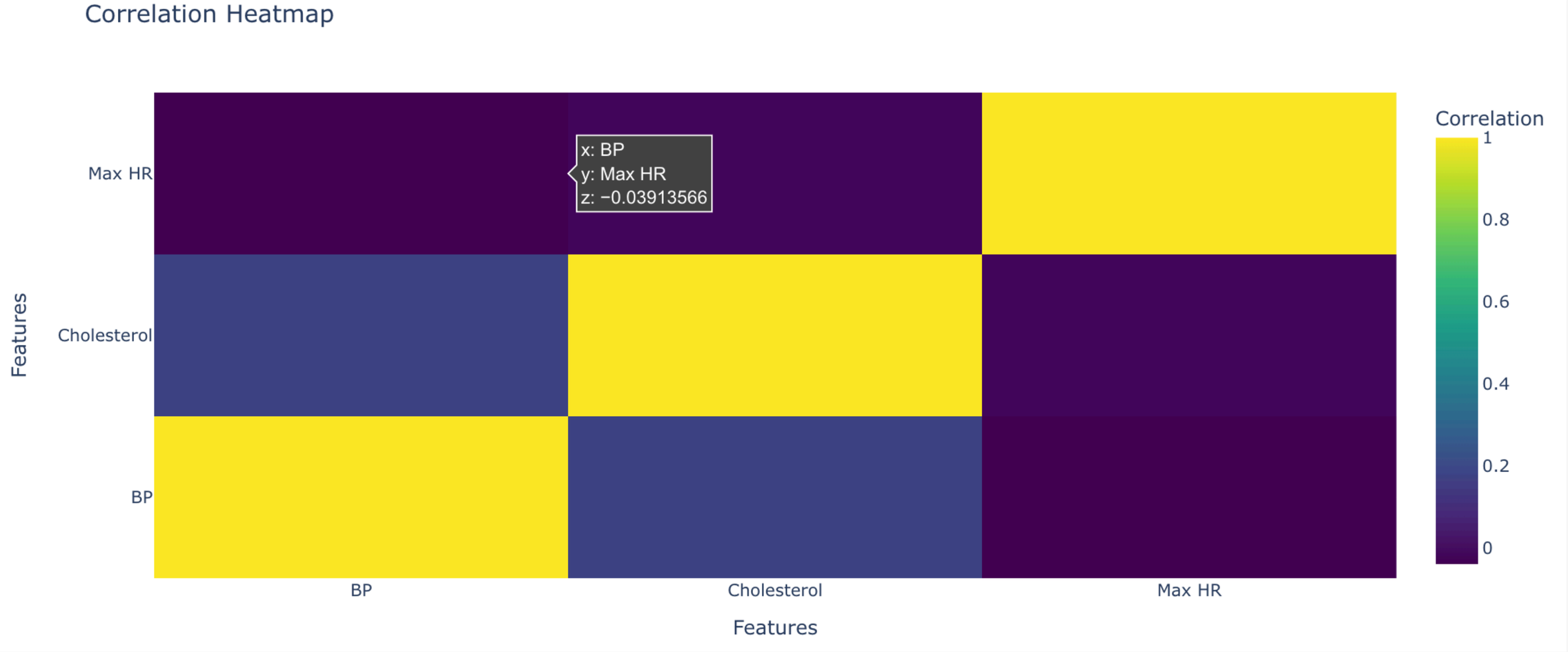
Sayısal Kolonların İstatistiksel Dağılımı

	BP	Cholesterol	Max HR
count	270.000000	270.000000	270.000000
mean	131.344444	249.659259	149.677778
std	17.861608	51.686237	23.165717
min	94.000000	126.000000	71.000000
25%	120.000000	213.000000	133.000000
50%	130.000000	245.000000	153.500000
75%	140.000000	280.000000	166.000000
max	200.000000	564.000000	202.000000

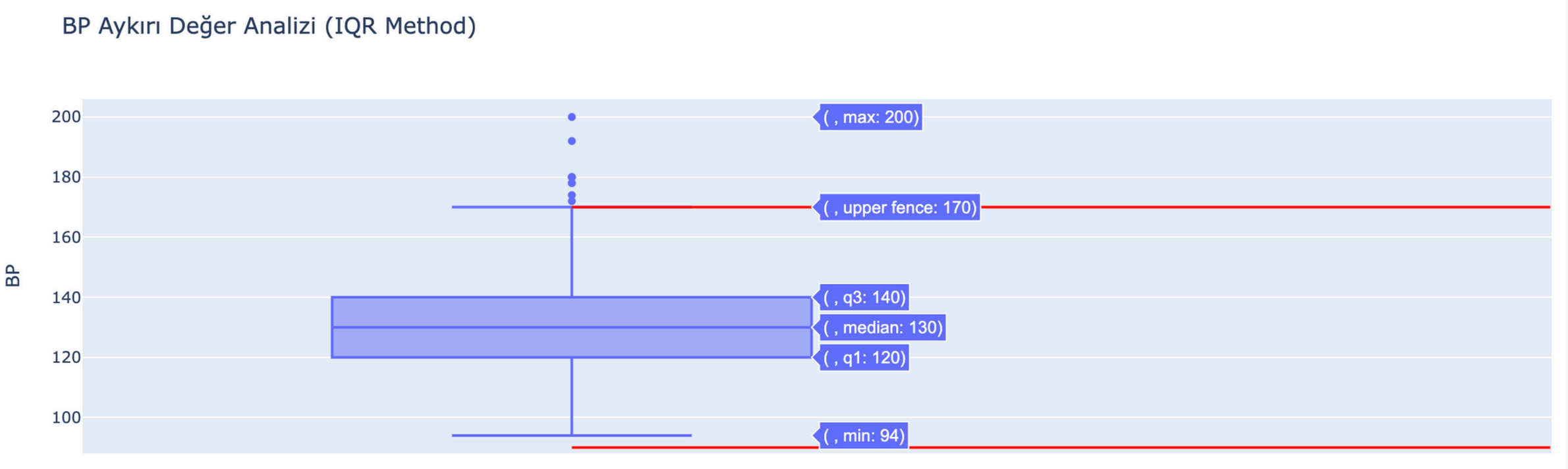
Sayısal Kolonların Boxplot gösterimi



Sayısal Verilerde Korelasyon Analizi



Aykırı Değer Analizi



RobutsScaler ve One-Hot Encoding Kullanımı

```
# RobustScaler kullanarak veri ölçeklendirme
scaler = RobustScaler()
df[num_cols] = scaler.fit_transform(df[num_cols])
```

```
# Kategorik değişkenleri one-hot encoding yap
df = pd.get_dummies(df, columns=cat_cols, drop_first=True)
```

RandomForestClassifier modelini daha az karmaşık hale getirmek için parametreler:

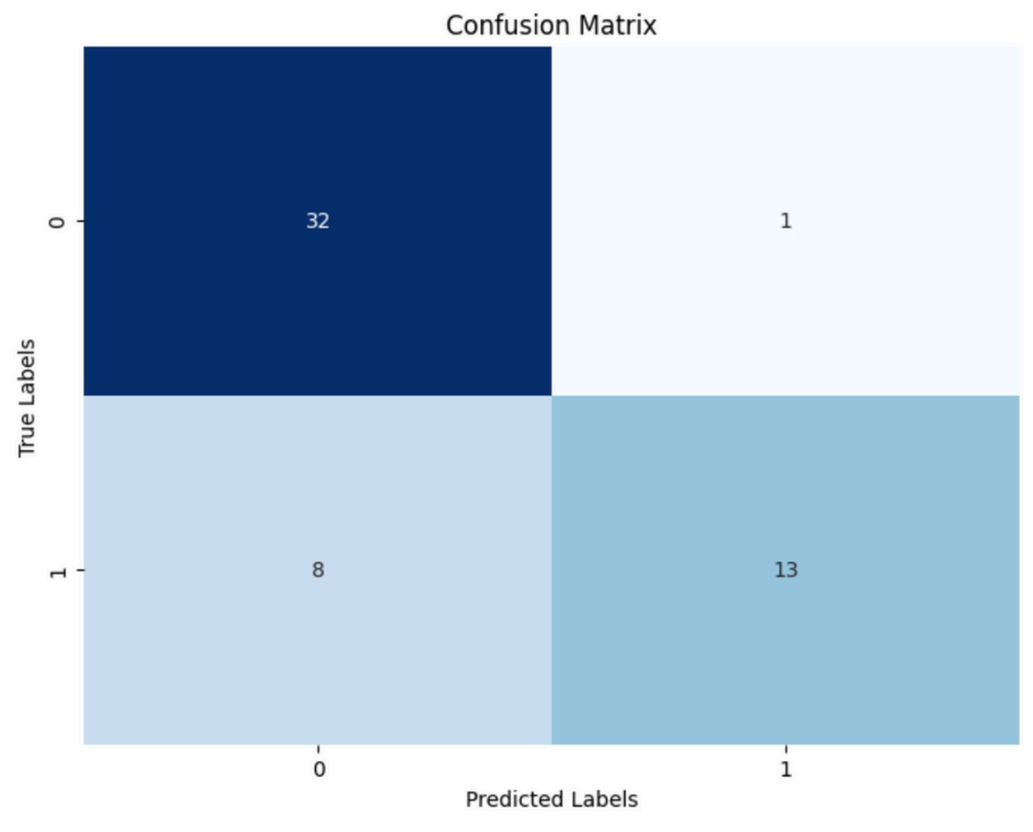
```
model = RandomForestClassifier(max_depth=10, min_samples_split=10, min_samples_leaf=4, n_estimators=100, random_state=42)
model.fit(X_train, y_train)

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

```
Model Accuracy: 0.8333333333333334
```

- max_depth=10: Bu parametre, her bir ağacın maksimum derinliğini belirler. Daha küçük bir değer, modelin aşırı öğrenmesini (overfitting) önleyebilir.
- min_samples_split=10: Bu parametre, bir düğümün bölünebilmesi için gerekli olan minimum örnek sayısını belirtir.
- min_samples_leaf=4: Bu parametre, yaprak düğümde bulunması gereken minimum örnek sayısını belirler. Her yaprakta en az 4 örnek bulunmasını sağlamak, çok az örnek içeren yapraklar oluşmasını önler ve bu da aşırı öğrenmeyi azaltır.
- n_estimators=100: Bu parametre, ormandaki ağaç sayısını tanımlar. Daha fazla ağaç genellikle daha iyi performans sağlar.
- random_state=42: Bu parametre, rastgele sayı üreteci için tohum değerini ayarlar.

Confusion Matrix



Sınıflandırma Parametreleri

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.97	0.88	33
1	0.93	0.62	0.74	21
accuracy			0.83	54
macro avg	0.86	0.79	0.81	54
weighted avg	0.85	0.83	0.82	54

Eğitim Doğruluk Oranı:

```
Training Set Accuracy: 0.8935185185185185
```

Test Doğruluk Oranı:

```
Test Set Accuracy: 0.8333333333333334
```

Teşekkürler!