



Ankara Üniversitesi  
Fen Bilimleri Enstitüsü  
Bilgisayar Mühendisliği Bölümü



Makine Öğrenmesi Kullanarak  
Meme Kanseri Hastalarının Sağ Kalma Tahmini

Veri Bilimi

Prof. Dr. Semra Gündüç  
Dersin Öğretim Üyesi

22822606  
Oğuzhan Panatlı  
Doktora Öğrencisi

Aralık 2023

# İçindekiler

- Giriş
  - Veri Setine Genel Bakış
  - Veri Setinin İncelenmesi
  - Keşifsel Veri Analizi (Explatory Data Analysis - EDA)
- Yöntem
  - Veri Ön İşleme
  - Makine Öğrenmesi Modellerinin Uygulanması
    - Logistic Regression
    - K-Nearest Neighbor (KNN)
    - Support Vector Machine (SVM)
    - Decision Tree
    - Random Forest
  - Performans Metrikleri
- Sonuçların Değerlendirmesi
  - Modellerin Sonuçlarının Değerlendirilmesi ve Yorumlanması
- Sonuç

# Veri Setine Genel Bakış

≡ kaggle

+

 Create

🏠 Home

🏆 Competitions

📁 Datasets

👤 Models

🔗 Code

💬 Discussions

🎓 Learn

⌵ More

📋 Your Work

▶ VIEWED

🔍 Search

REIHANEH NAMDARI · UPDATED A YEAR AGO

232

New Notebook

Download (44 kB)

## Breast Cancer

Seer Breast Cancer Data - Labeled

Data Card Code (67) Discussion (1)


### About Dataset

This dataset of breast cancer patients was obtained from the 2017 November update of the SEER Program of the NCI, which provides information on population-based cancer statistics. The dataset involved female patients with infiltrating duct and lobular carcinoma breast cancer (SEER primary cites recode NOS histology codes 8522/3) diagnosed in 2006-2010. Patients with unknown tumour size, examined regional LNs, positive regional LNs, and patients whose survival months were less than 1 month were excluded; thus, 4024 patients were ultimately included.

**Usability** ⓘ  
10.00

**License**  
Attribution 4.0 International (CC ...)

**Expected update frequency**  
Never



## Veri Setine Genel Bakış

- Veri setinde **4024** hasta dahil edilmiştir. (gözlem sayısı)
- Değişkenlerin Açıklanması:
  - **Age** - hastanın yaşı (numerik)
  - **Race** – hastanın ırkı (kategorik)
  - **Marital Status** - medeni Durum (kategorik)
  - **T Stage** – primer tümörün büyüklüğünü ve kapsamını ifade ediyor. (kategorik)
  - **N Stage** - yakındaki lenf düğümlerinin tutulumunu ifade ediyor. (kategorik)
  - **6th Stage** – Kanserin kaç tane koltuk altı lenf düğümüne ve/veya iç meme lenf düğümlerine yayıldığına dair bilgi veriyor. Ayrıca tespit edilmişse tümör boyutu hakkında da bilgi veriyor. (kategorik)
  - **Differentiate** - Farklılaşma derecesi, kanser hücrelerinin yapı ve fonksiyon bakımından normal, sağlıklı hücrelere ne kadar benzediğini ifade ediyor. (kategorik)
  - **Grade** - Farklılaşma derecesi (kategorik)

## Veri Setine Genel Bakış

- Veri setinde **4024** hasta dahil edilmiştir. (gözlem sayısı)
- Değişkenlerin Açıklanması:
  - **A Stage** – Kanserin bölgesel mi uzak yerlere mi yayılmış bilgisini içeriyor. (kategorik)
  - **Tumor Size** - milimetre cinsinden tam boyutu gösteriyor. (numerik)
  - **Estrogen Status** - östrojen durumu: pozitif–negatif (kategorik)
  - **Progesterone Status** - progesteron durumu: pozitif–negatif (kategorik)
  - **Regional Node Examined** - tanı sürecinde incelenen bölgesel lenf düğümlerinin sayısını ifade ediyor. (numerik)
  - **Regional Node Positive** - bölgesel lenf düğümlerinde kanser hücrelerinin varlığını ifade ediyor. (numerik)
  - **Survival Months** – hastanın hayatta kaldığı ay sayısı (numerik)
  - **Status** - hastanın durumu: yaşıyor – yaşamıyor (kategorik)

## Veri Setinin İncelenmesi

### #Genel Bilgiler

```
df.info()
```

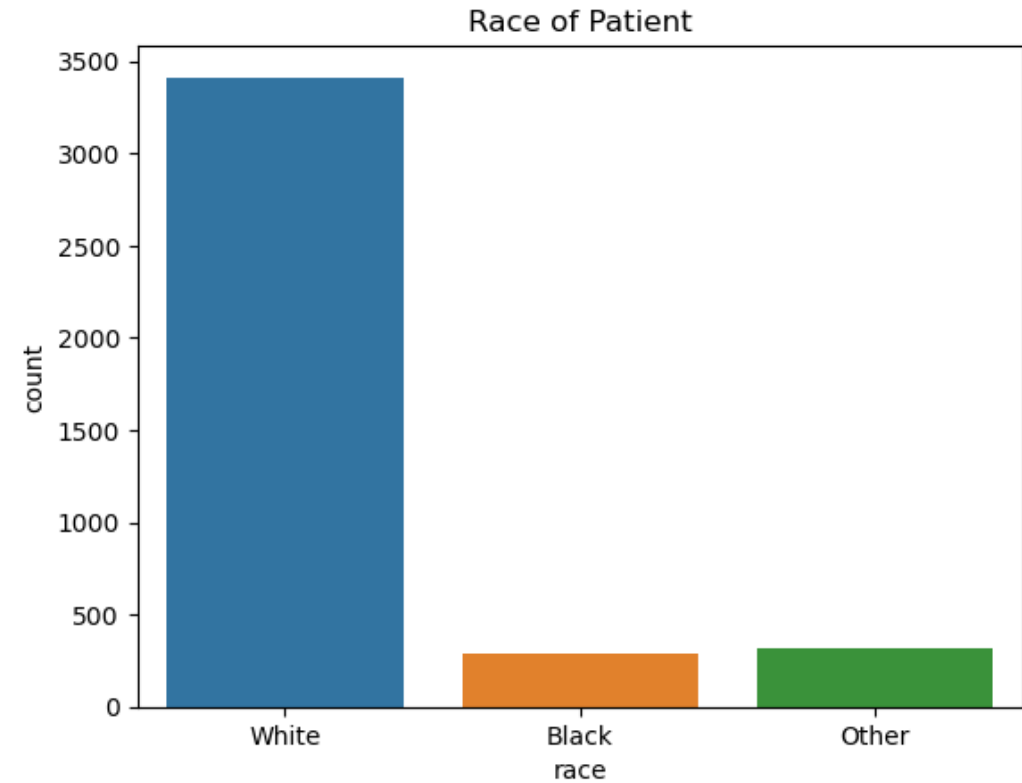
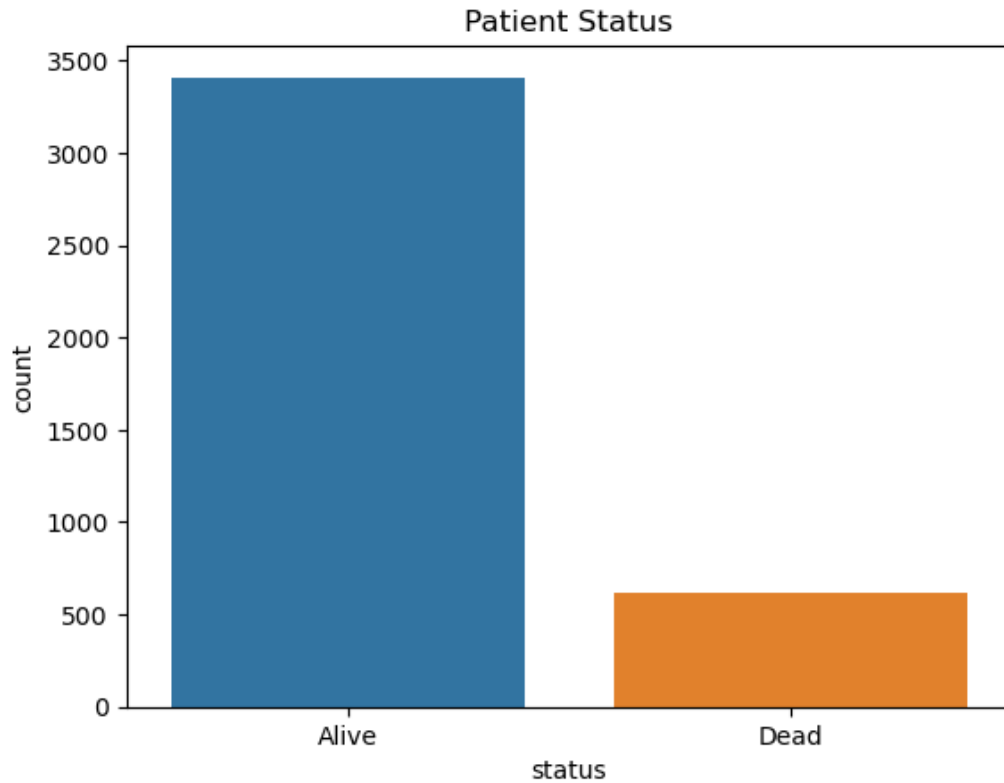
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4024 entries, 0 to 4023
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype
---  -
0   age                   4024 non-null   int64
1   race                  4024 non-null   object
2   marital_status        4024 non-null   object
3   t_stage               4024 non-null   object
4   n_stage               4024 non-null   object
5   6th_stage             4024 non-null   object
6   differentiate         4024 non-null   object
7   grade                 4024 non-null   object
8   a_stage               4024 non-null   object
9   tumor_size            4024 non-null   int64
10  estrogen_status       4024 non-null   object
11  progesterone_status   4024 non-null   object
12  regional_node_examined 4024 non-null   int64
13  regional_node_positive 4024 non-null   int64
14  survival_months       4024 non-null   int64
15  status                 4024 non-null   object
dtypes: int64(5), object(11)
memory usage: 503.1+ KB
```

### #Temel İstatistikler

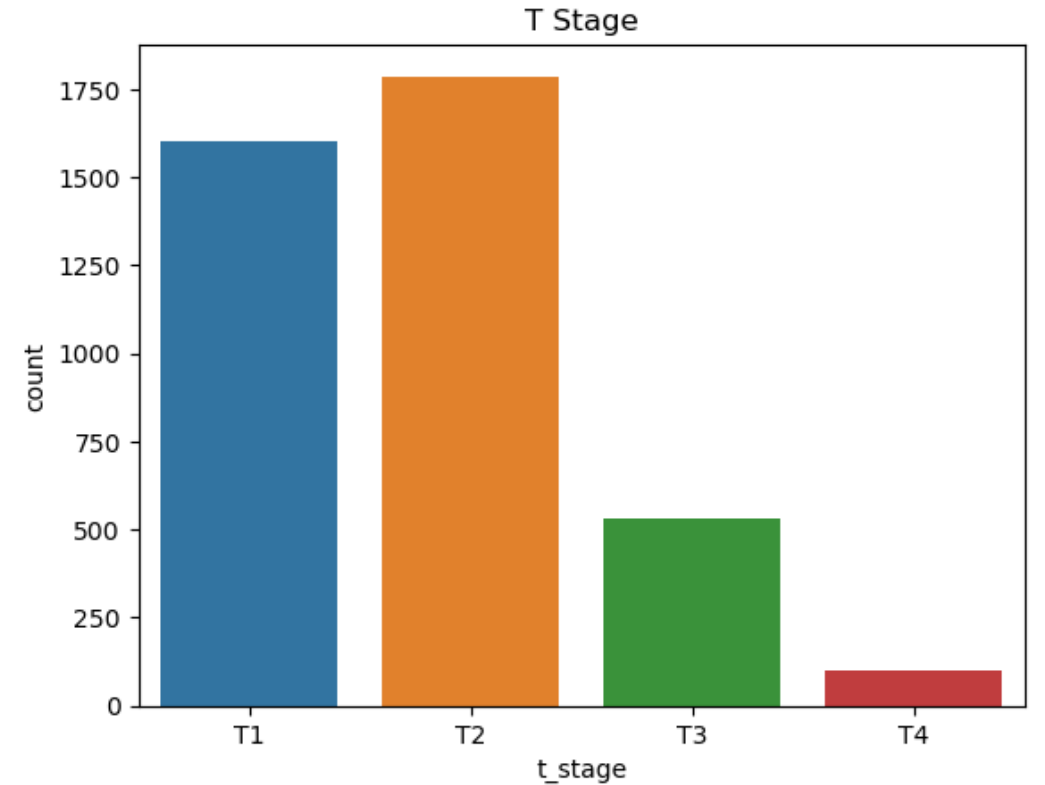
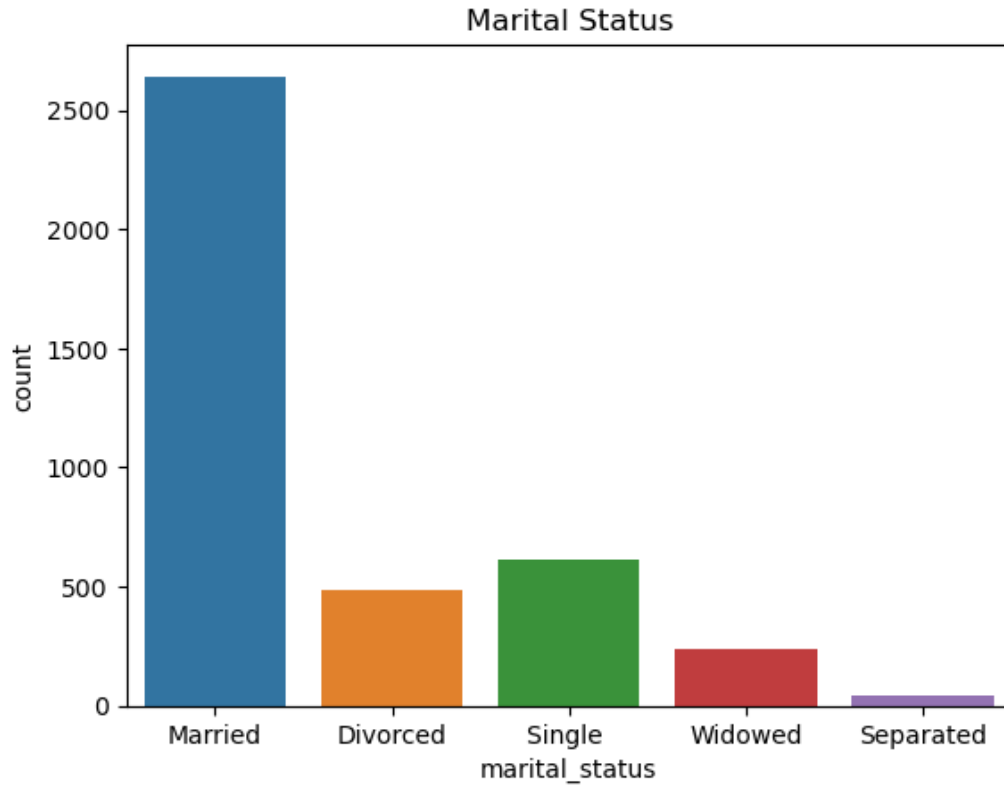
```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
age	4024.0	53.972167	8.963134	30.0	47.0	54.0	61.0	69.0
tumor_size	4024.0	30.473658	21.119696	1.0	16.0	25.0	38.0	140.0
regional_node_examined	4024.0	14.357107	8.099675	1.0	9.0	14.0	19.0	61.0
regional_node_positive	4024.0	4.158052	5.109331	1.0	1.0	2.0	5.0	46.0
survival_months	4024.0	71.297962	22.921430	1.0	56.0	73.0	90.0	107.0

## Keşifsel Veri Analizi (Explatory Data Analysis - EDA)



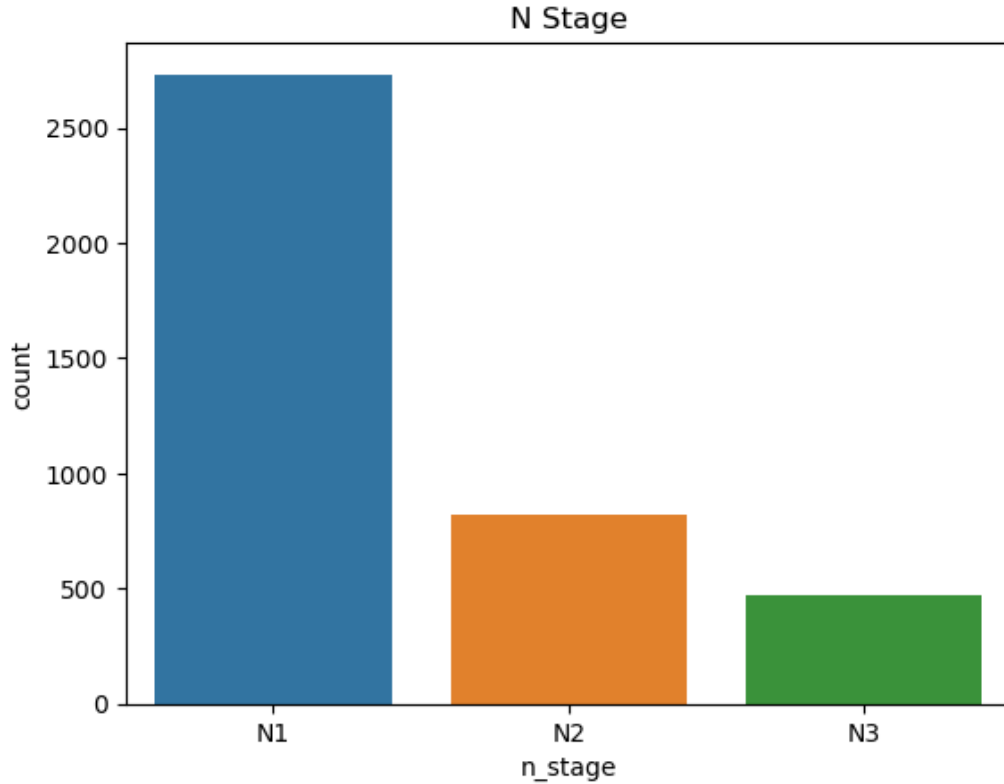
## Keşifsel Veri Analizi (Explatory Data Analysis - EDA)



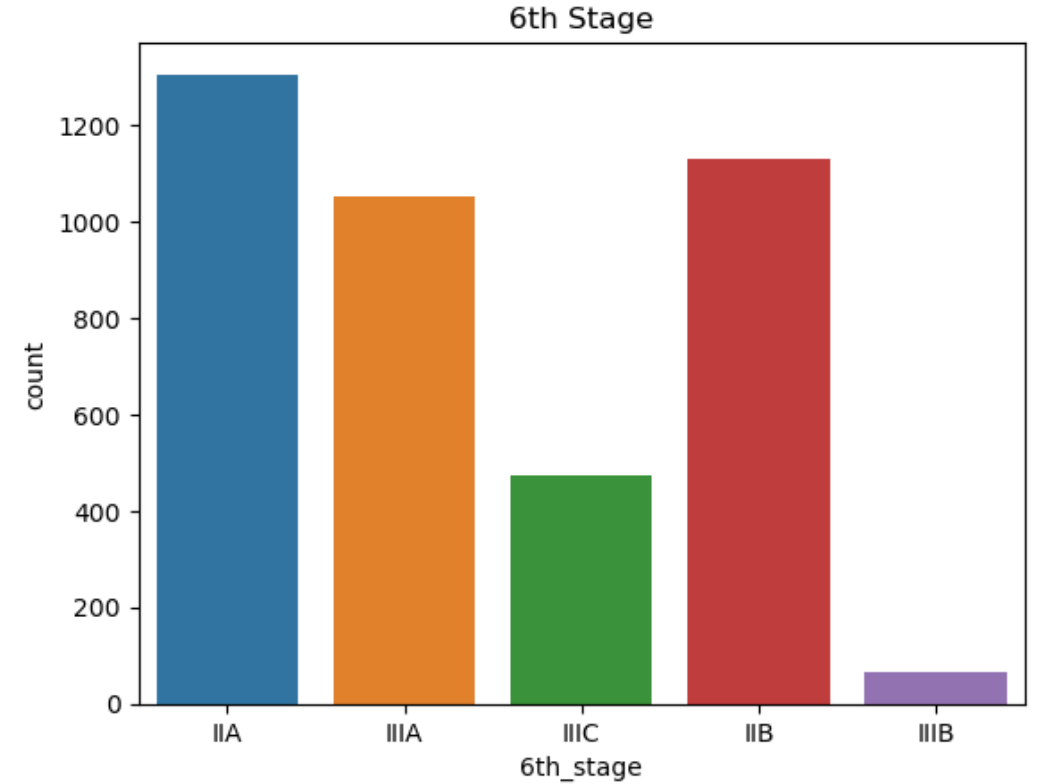
\* **T Stage** – primer tümörün büyüklüğünü ve kapsamını ifade ediyor.



## Keşifsel Veri Analizi (Explatory Data Analysis - EDA)

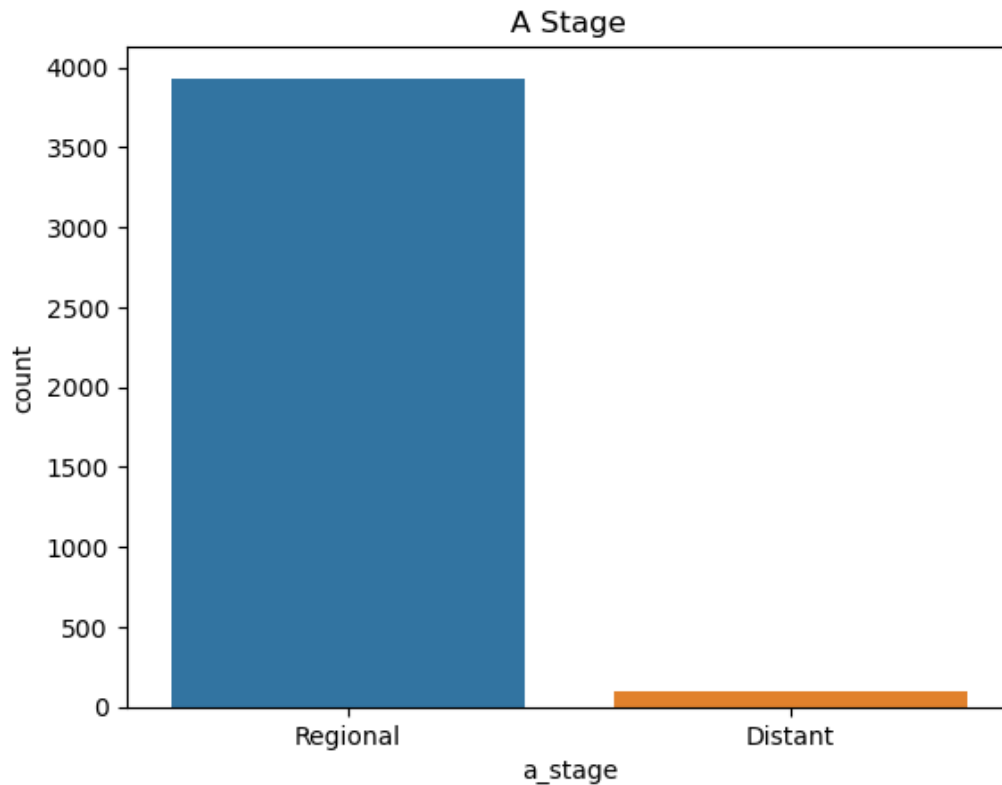


\* **N Stage** – yakındaki lenf düğümlerinin tutulumunu ifade ediyor.

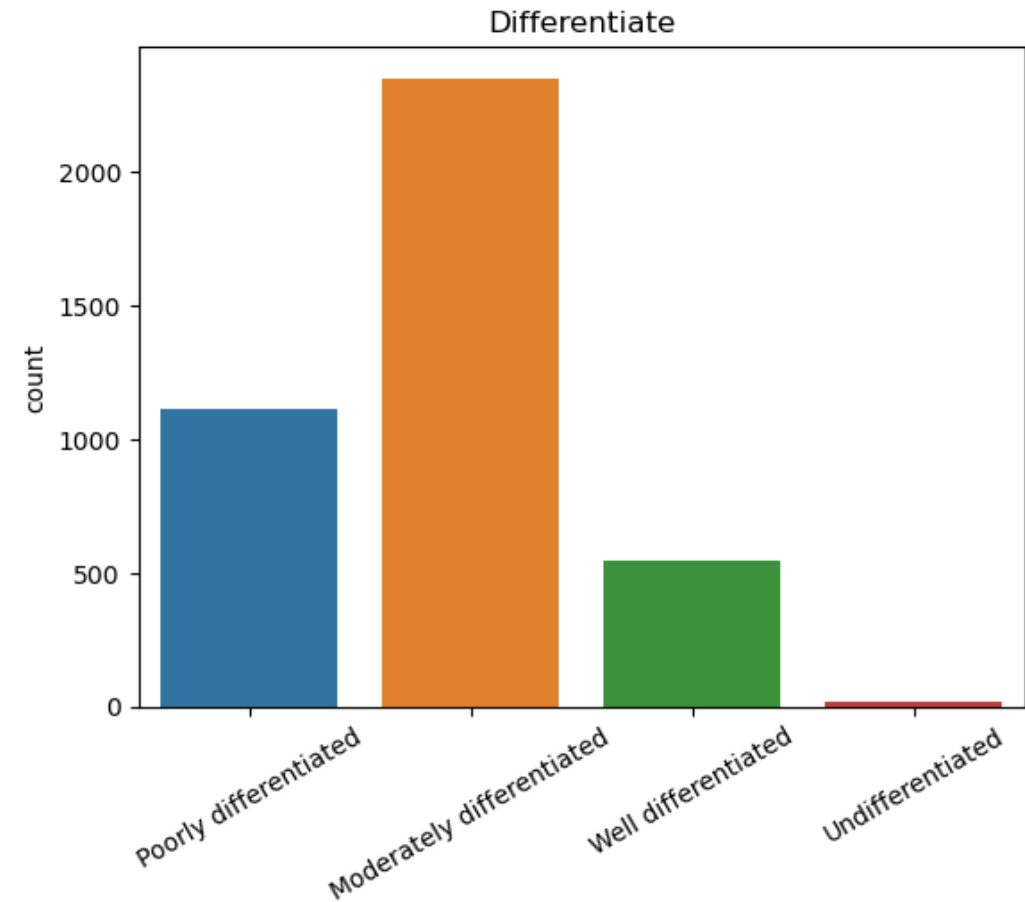


**6th Stage** – Kanserin kaç tane koltuk altı lenf düğümüne ve/veya iç meme lenf düğümlerine yayıldığına dair bilgi veriyor. Ayrıca tespit edilmişse tümör boyutu hakkında bilgi veriyor.

## Keşifsel Veri Analizi (Explatory Data Analysis - EDA)

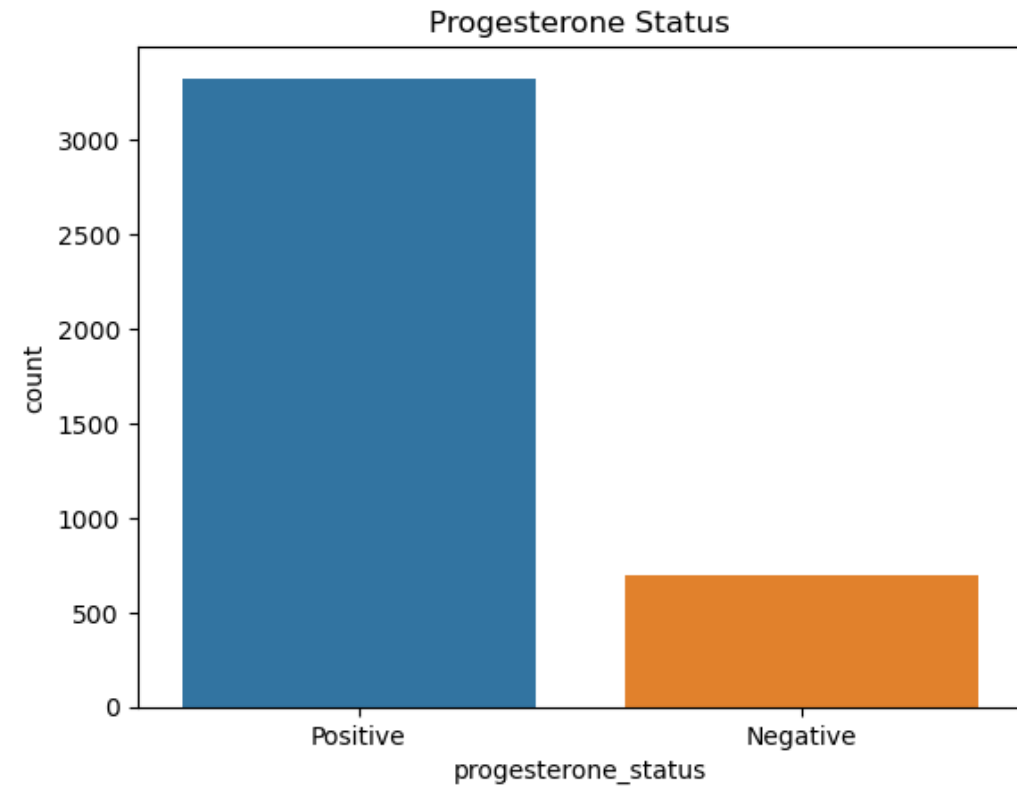
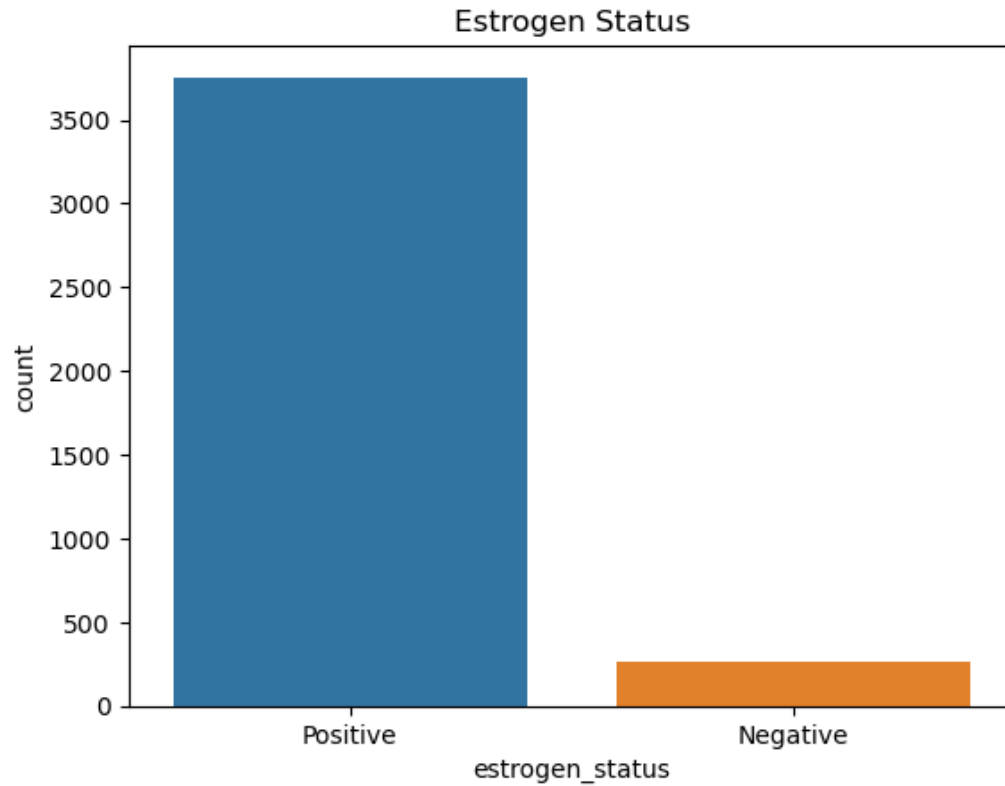


**A Stage** – Kanserin bölgesel mi uzak yerlere mi yayılmış bilgisini içeriyor.

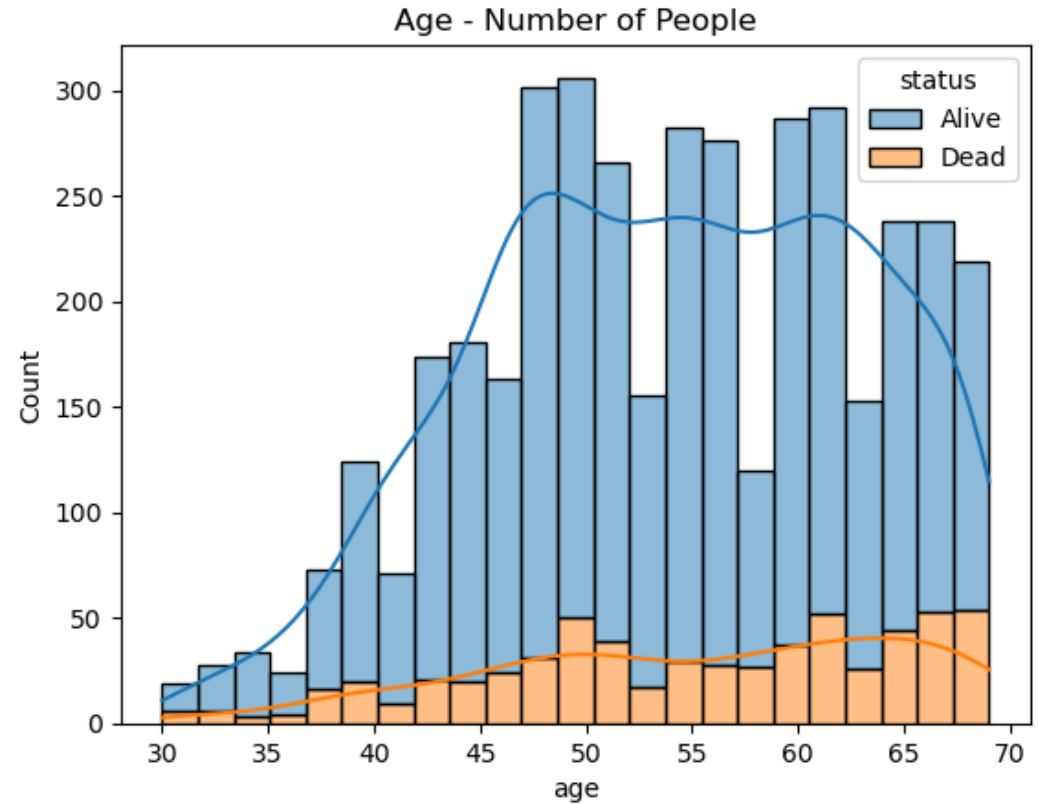
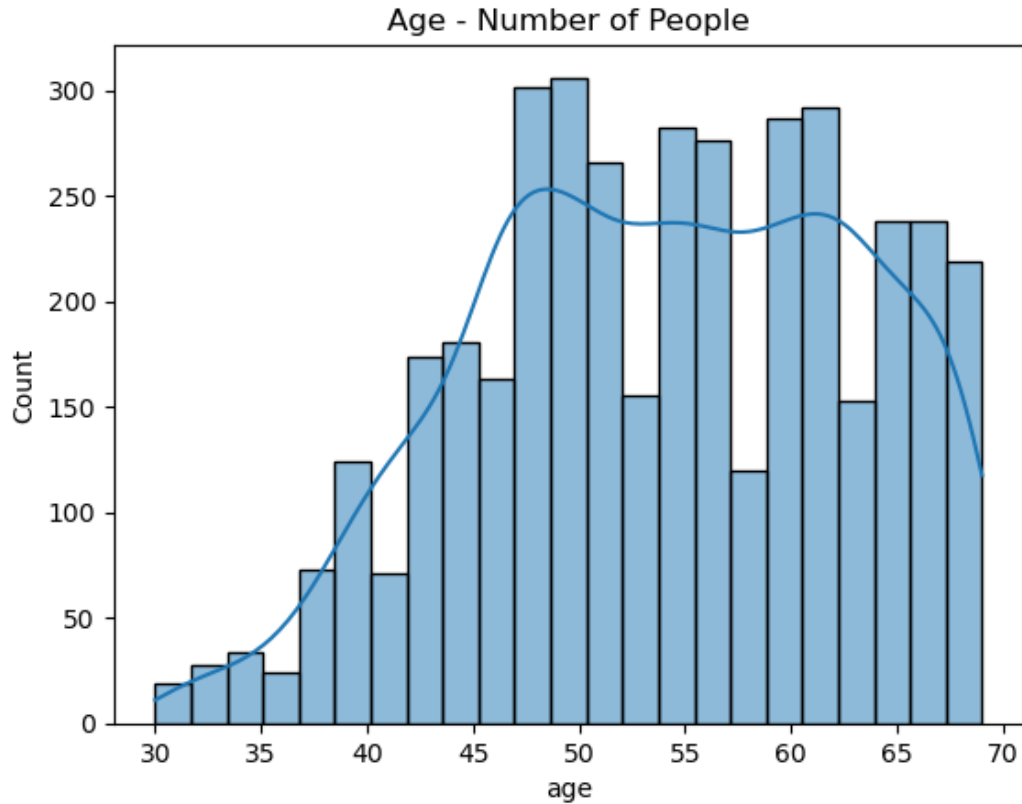


**Differentiate** - Farklılaşma derecesi, kanser hücrelerinin yapı ve fonksiyon bakımından normal, sağlıklı hücrelere ne kadar benzediğini ifade ediyor

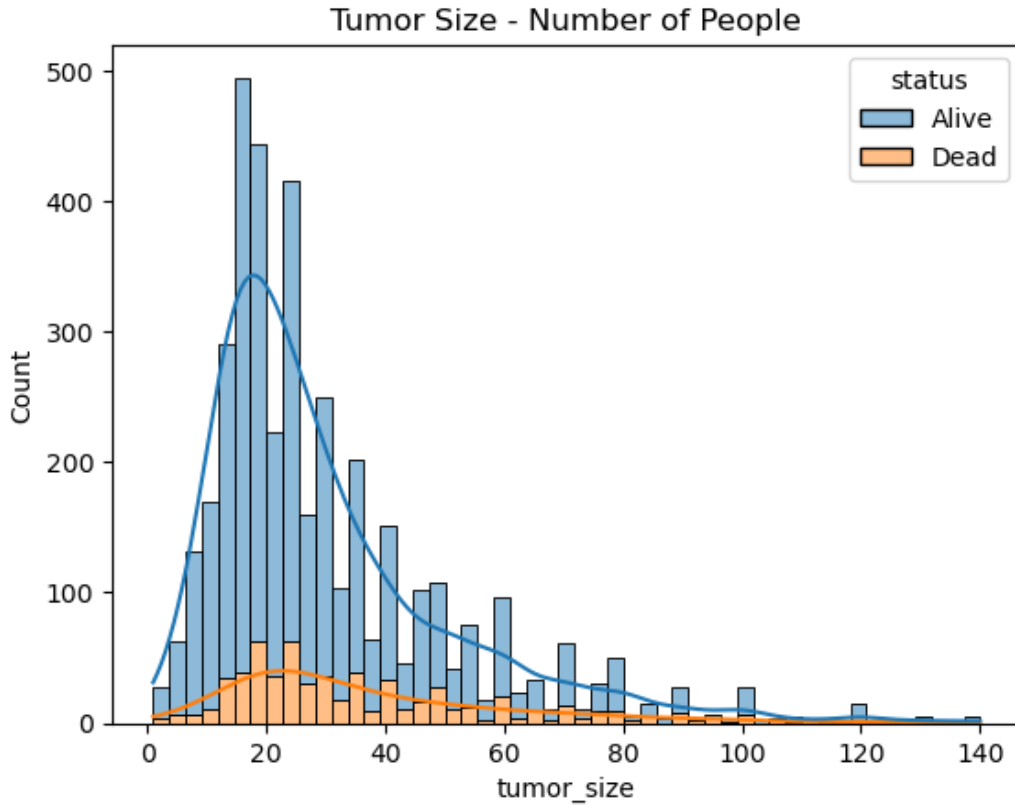
## Keşifsel Veri Analizi (Explatory Data Analysis - EDA)



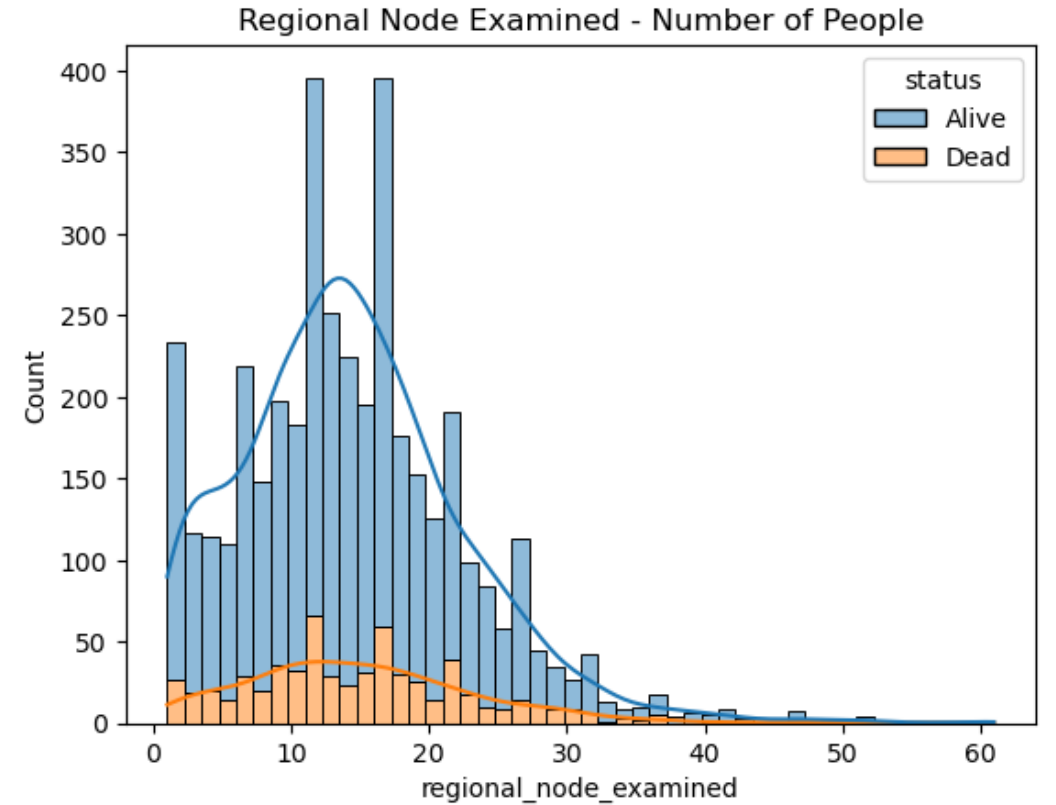
## Keşifsel Veri Analizi (Explatory Data Analysis - EDA)



## Keşifsel Veri Analizi (Explatory Data Analysis - EDA)

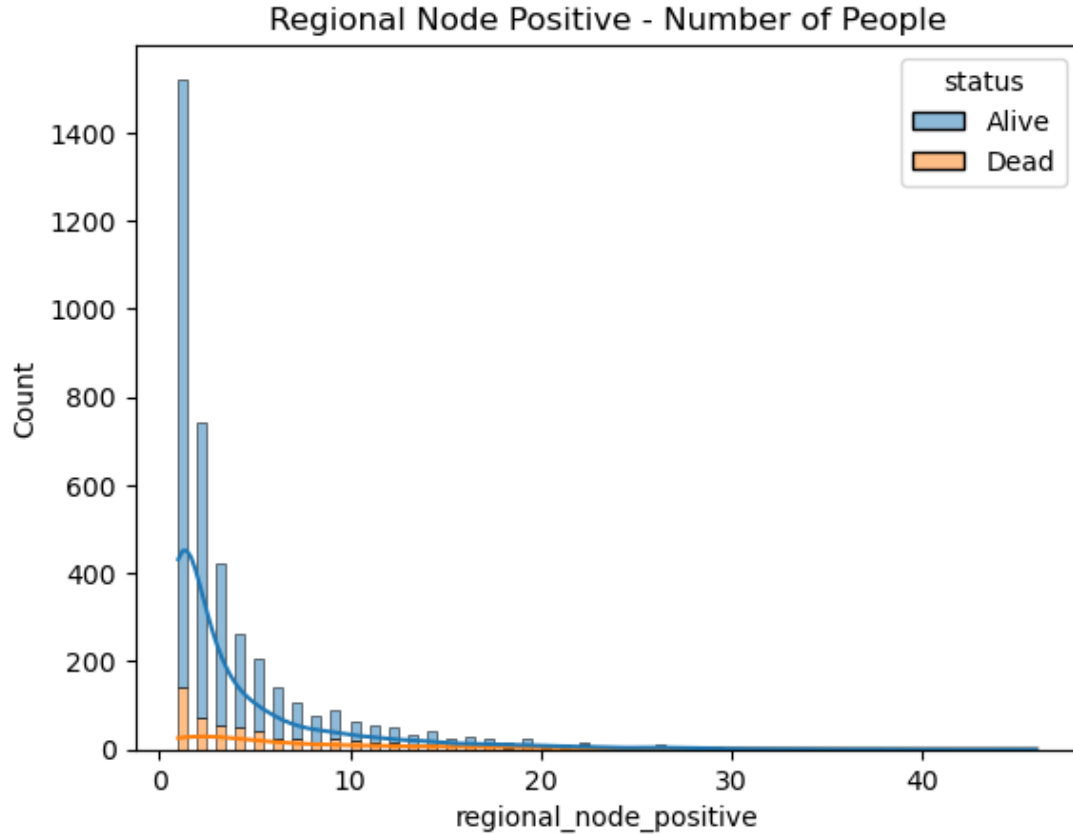


**Tumor Size** - milimetre cinsinden tam boyutu gösteriyor.

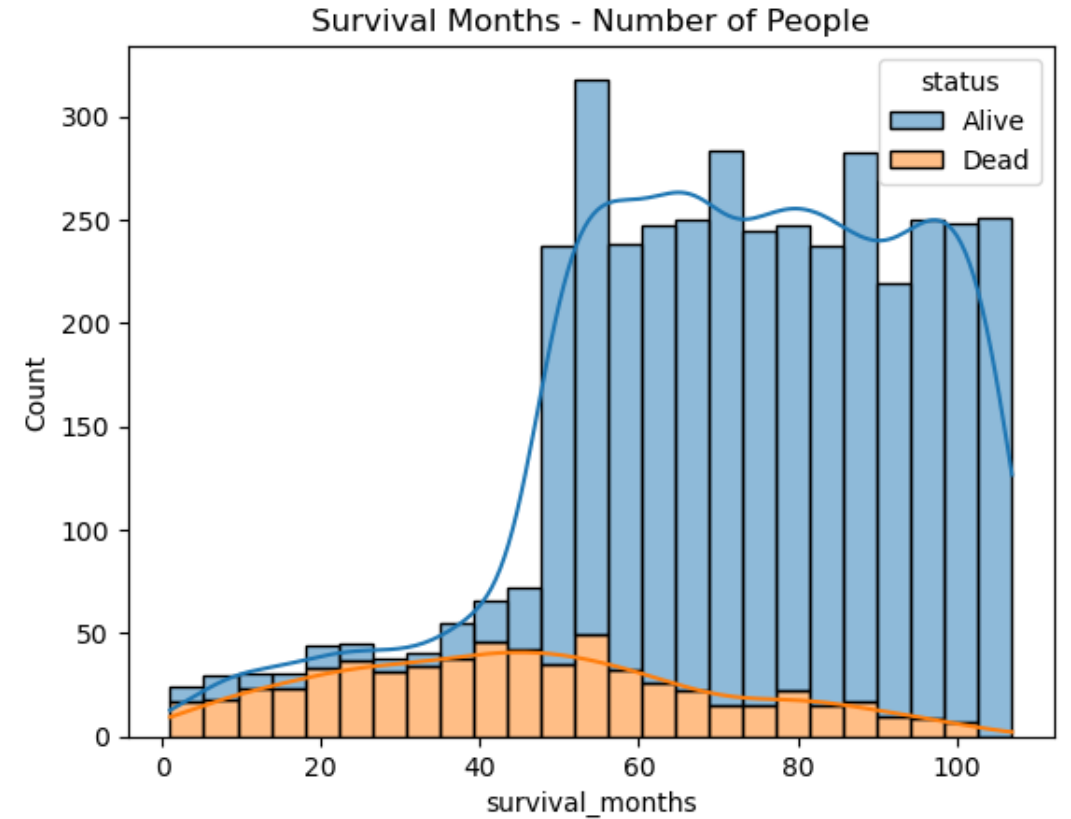


**Regional Node Examined** - tanı sürecinde incelenen bölgesel lenf düğümlerinin sayısını ifade ediyor.

## Keşifsel Veri Analizi (Explatory Data Analysis - EDA)

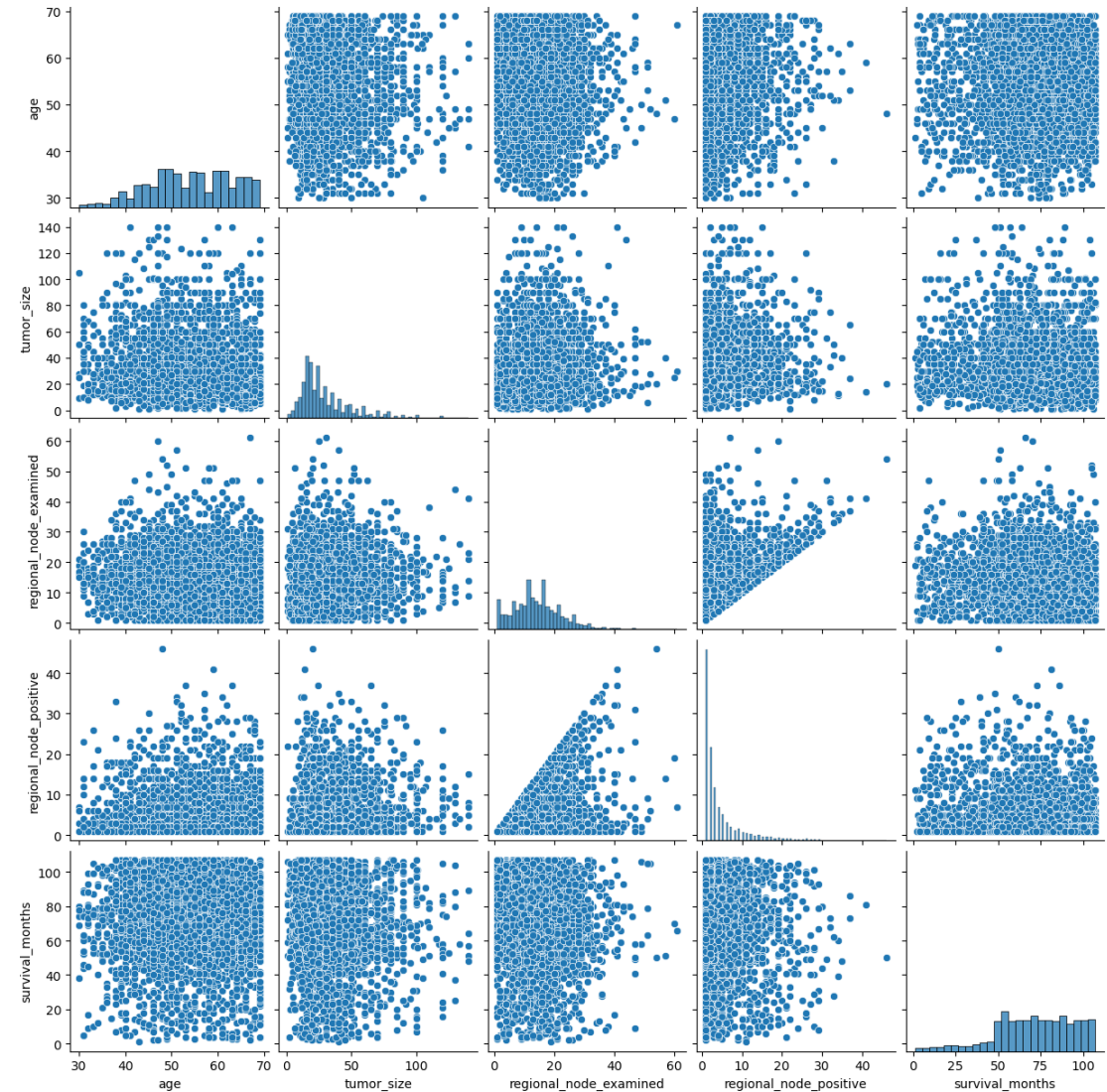
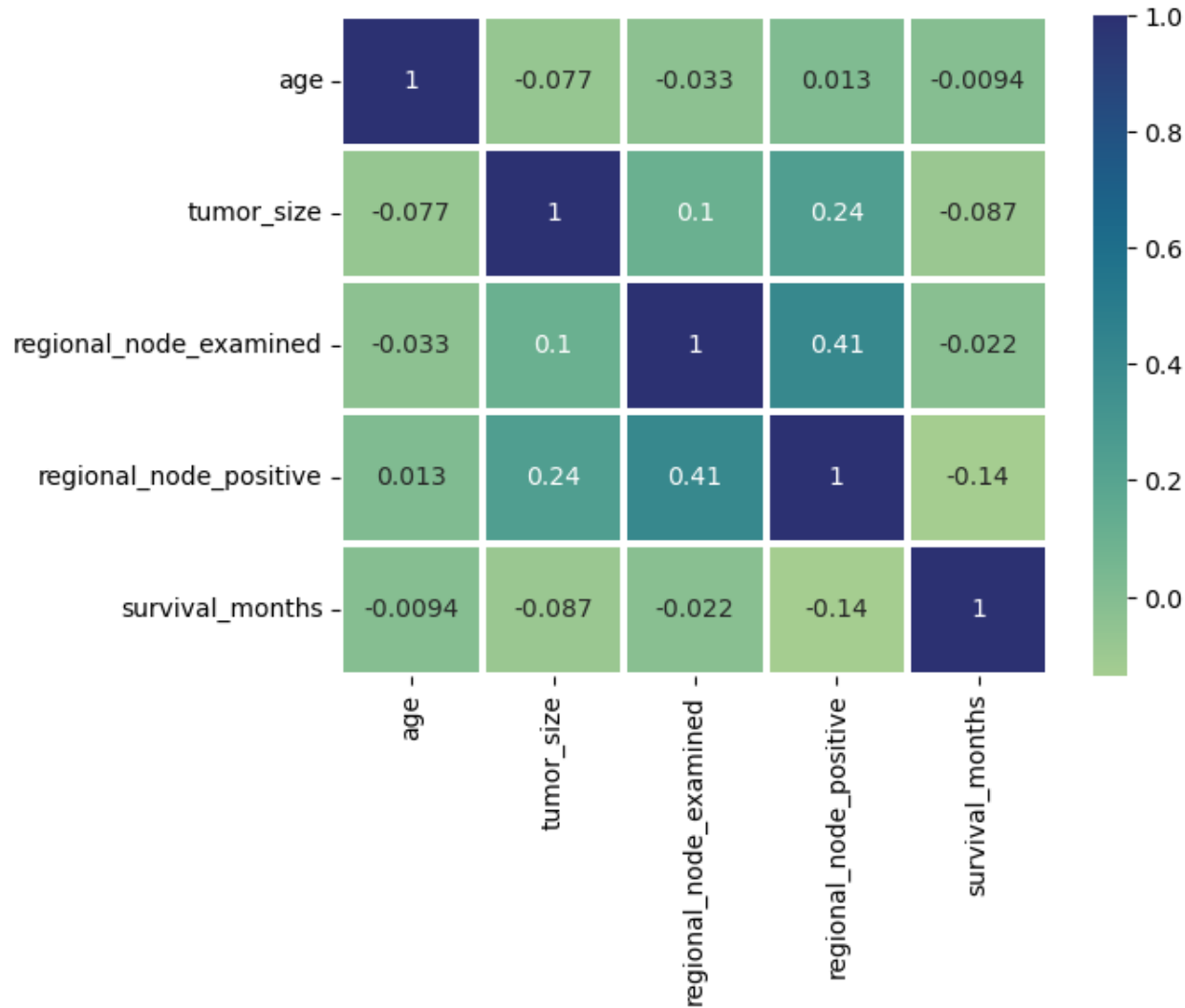


**Regional Node Positive** - bölgesel lenf düğümlerinde kanser hücrelerinin varlığını ifade ediyor.

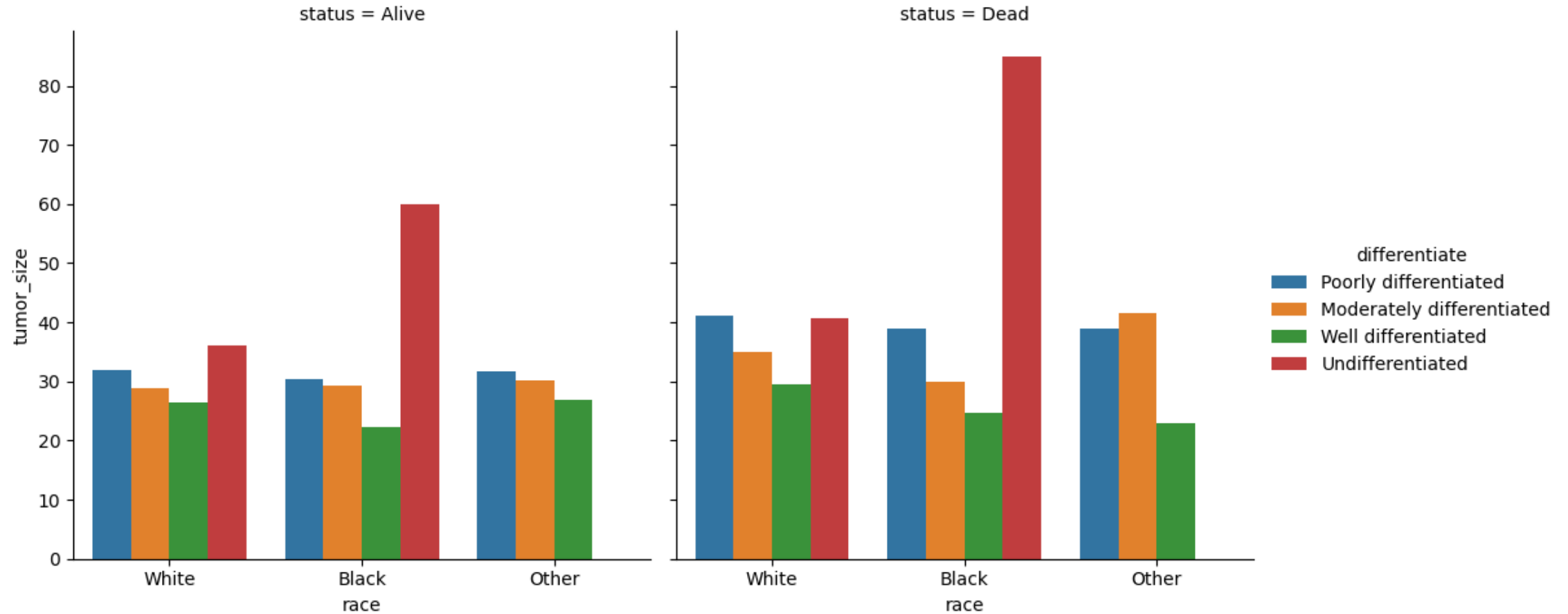


**Survival Months** – hastanın hayatta kaldığı ay sayısı

# Keşifsel Veri Analizi (Explatory Data Analysis - EDA)



## Keşifsel Veri Analizi (Explatory Data Analysis - EDA)

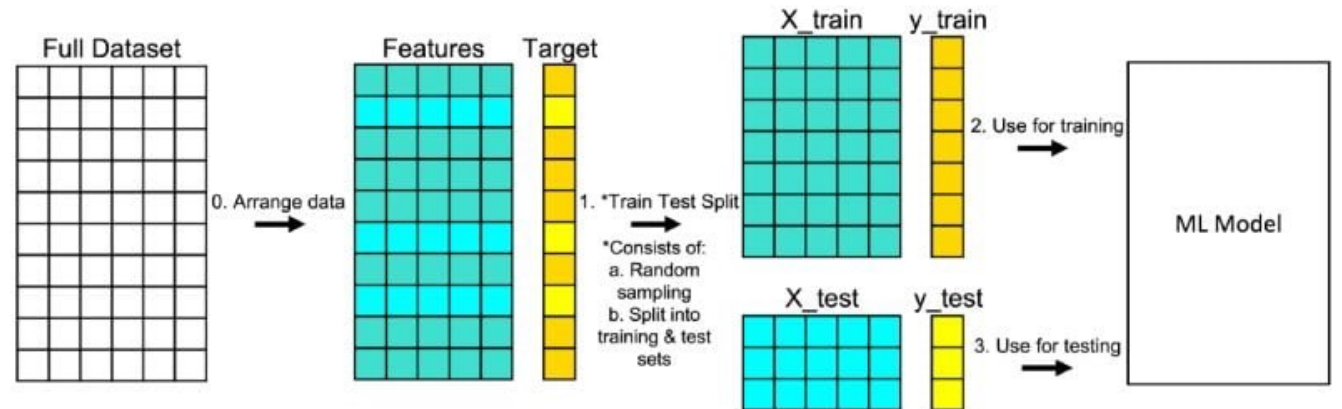


**Differentiate** - Farklılaşma derecesi, kanser hücrelerinin yapı ve fonksiyon bakımından normal, sağlıklı hücrelere ne kadar benzediğini ifade ediyor



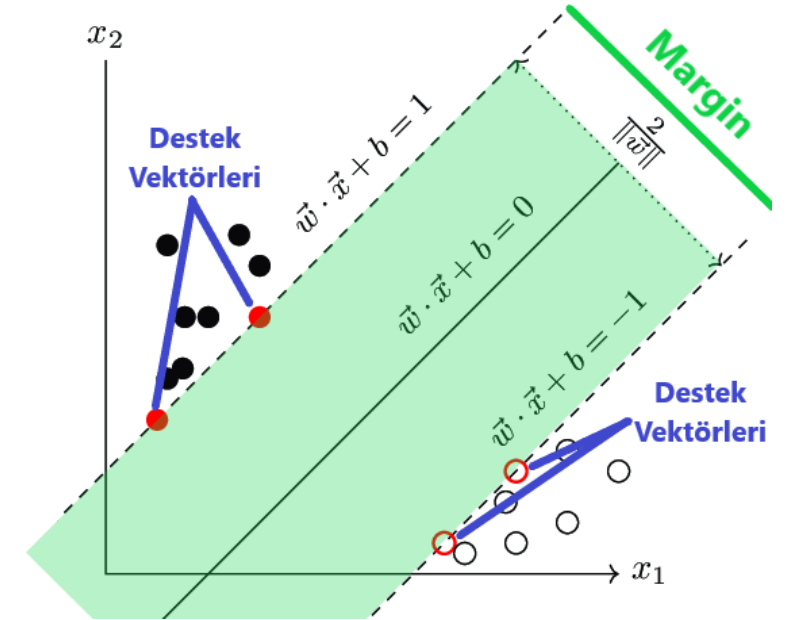
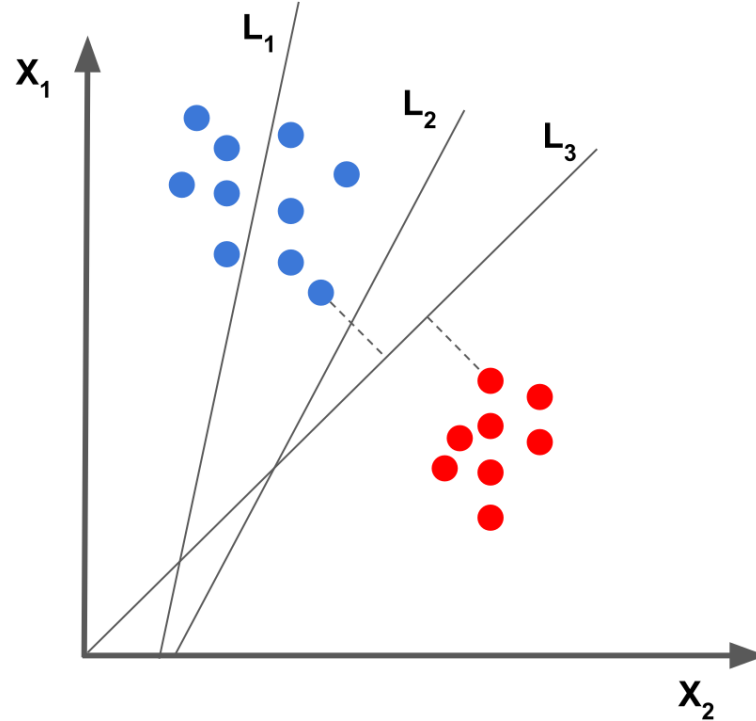
# Veri Ön İşleme

- Kategorik verilerin sayısal verilere dönüştürülmesi (label and one-hot encoding)
  - Race , Marital Status , A Stage , Estrogen Status, Progesterone Status ve Status -> **nominal**
  - Differentiate, 6th Stage, T Stage, ve N Stage verileri ---> **ordinal**
- Bağımsız değişkenlerin ve bağımlı değişkenin belirlenmesi (feature vector and target variable)
  - Status -> **Bağımlı Değişken**
  - Diğer değişkenler -> **Bağımsız Değişken**
- Veri setinin eğitim ve test olarak ayrılması (train-test split)
  - Eğitim Veri seti **%80** – Test Veri seti **%20**
- Ölçeklendirme (Feature Scaling)
  - **Standard scaler** ile ölçeklendirme yapıldı.



# Makine Öğrenmesi Modellerinin Uygulanması

- Logistic Regression
- K-Nearest Neighbor (KNN)
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest

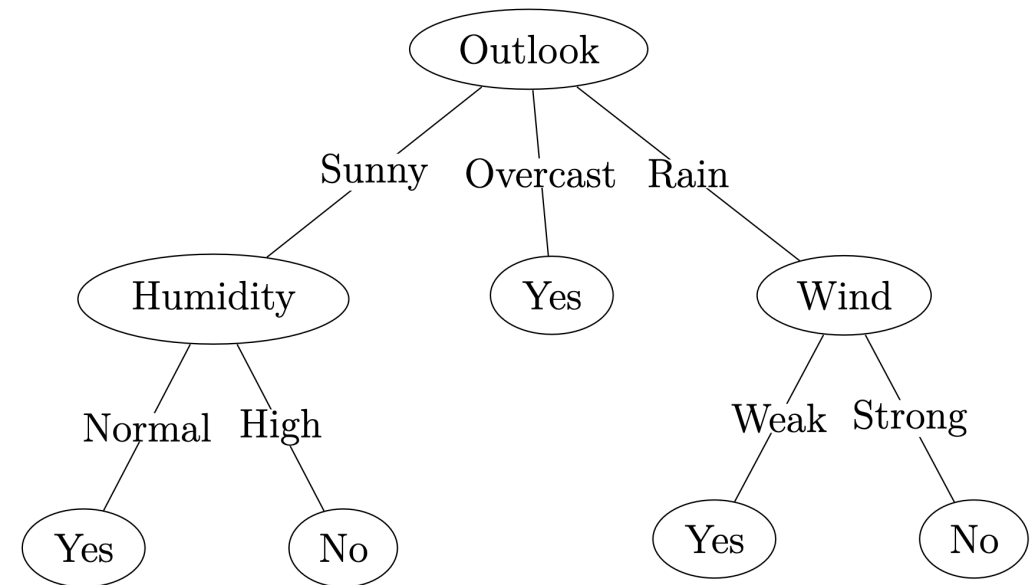


# Makine Öğrenmesi Modellerinin Uygulanması

- Logistic Regression
- K-Nearest Neighbor (KNN)
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest

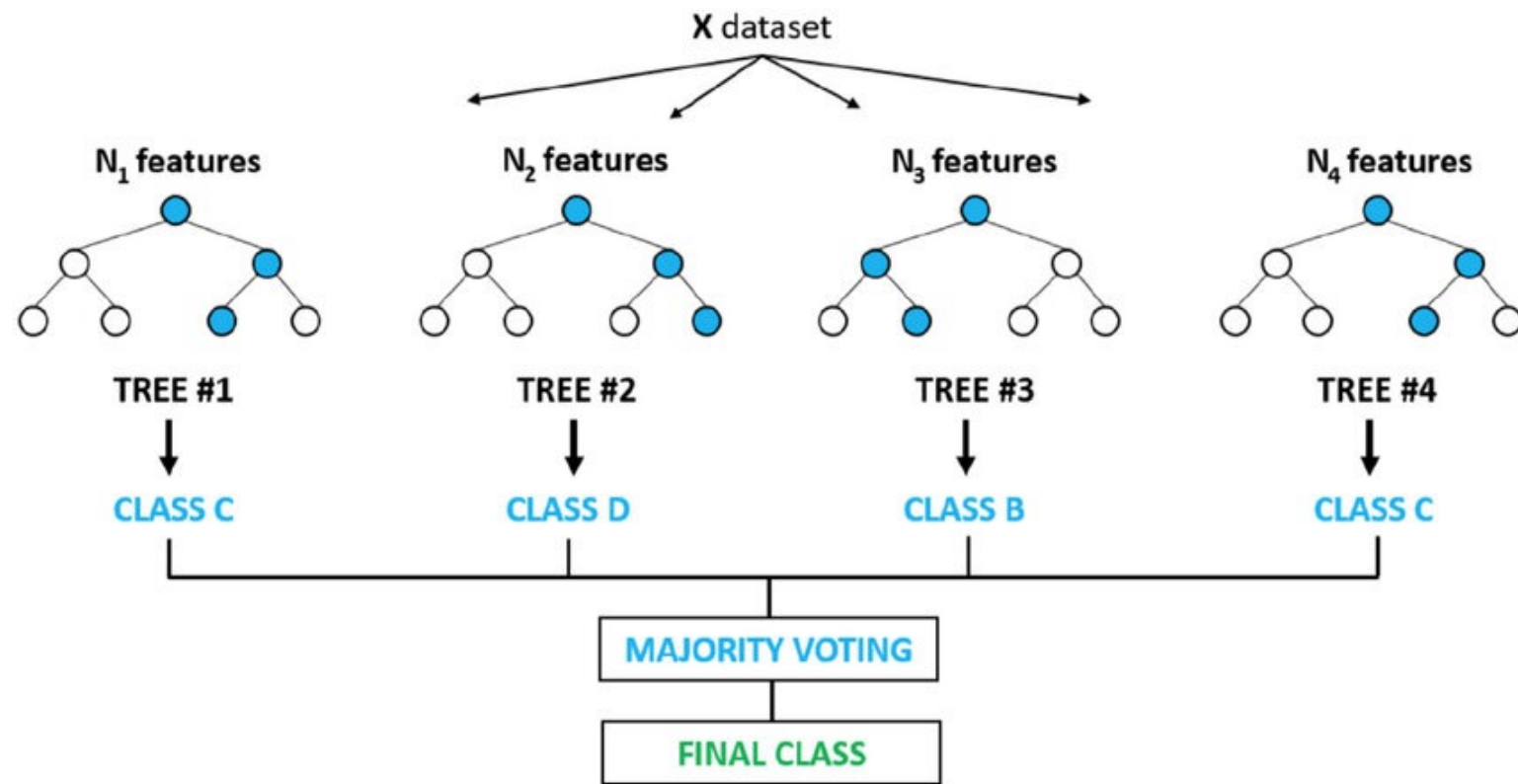
Belirli bir sorunu çözmek için **ağaç benzeri** bir yapı ve bunların olası kombinasyonlarını kullanır.

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Mild	High	Strong	No
2	Rain	Hot	Normal	Strong	No
3	Rain	Cool	High	Strong	No
4	Overcast	Hot	High	Strong	Yes
5	Overcast	Cool	Normal	Weak	Yes
6	Rain	Hot	High	Weak	Yes
7	Overcast	Mild	Normal	Weak	Yes
8	Overcast	Cool	High	Weak	Yes
9	Rain	Cool	High	Weak	Yes
10	Rain	Mild	Normal	Strong	No
11	Overcast	Mild	High	Weak	Yes
12	Sunny	Mild	Normal	Weak	Yes
13	Sunny	Cool	High	Strong	No
14	Sunny	Cool	High	Weak	No



# Makine Öğrenmesi Modellerinin Uygulanması

- Logistic Regression
- K-Nearest Neighbor (KNN)
- Support Vector Machine (SVM)
- Decision Tree
- **Random Forest**



## Performans Metrikleri

- Model sonuçlarının değerlendirilmesinde kullanılan performans metrikleri
  - Accuracy
  - Precision
  - Recall (Sensitivity)
  - F1 Score
  - AUC

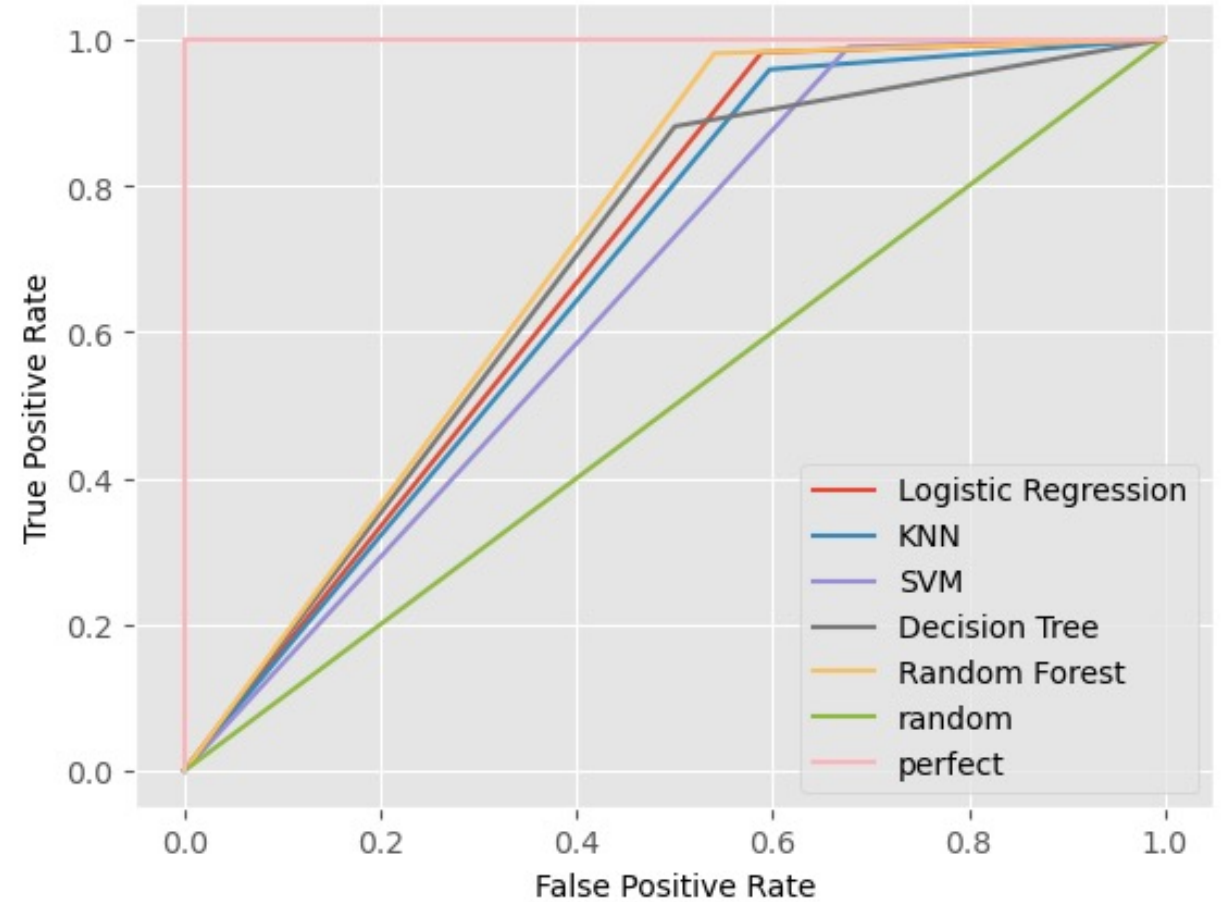
		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) <b>Type II Error</b>	<b>Sensitivity</b> $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) <b>Type I Error</b>	True Negative (TN)	<b>Specificity</b> $\frac{TN}{(TN + FP)}$
		<b>Precision</b> $\frac{TP}{(TP + FP)}$	<b>Negative Predictive Value</b> $\frac{TN}{(TN + FN)}$	<b>Accuracy</b> $\frac{TP + TN}{(TP + TN + FP + FN)}$

## Makine Öğrenmesi Modellerinin Sonuçlarının Değerlendirilmesi

Method	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.89	0.90	<b>0.98</b>	<b>0.94</b>	0.69
KNN	0.87	0.90	0.96	0.93	0.68
SVM	0.89	0.89	0.99	<b>0.94</b>	0.66
Decision Tree	0.82	<b>0.91</b>	0.88	0.89	0.69
Random Forest	<b>0.90</b>	0.90	<b>0.98</b>	<b>0.94</b>	<b>0.72</b>

# Makine Öğrenmesi Modellerinin Sonuçlarının Değerlendirilmesi

Method	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.89	0.90	0.98	0.94	0.69
KNN	0.87	0.90	0.96	0.93	0.68
SVM	0.89	0.89	0.99	0.94	0.66
Decision Tree	0.82	0.91	0.88	0.89	0.69
Random Forest	0.90	0.90	0.98	0.94	0.72



# Sonuç

- Breast Cancer veri setinde 5 farklı makine öğrenmesi modeli uygulanarak sınıflandırma yapılmıştır.
- Bu veri setinde en yüksek doğruluğa ve AUC değerine sırası ile %90 ve 0.72 değerleri ile Random Forest algoritması kullanarak ulaşılmıştır.
- Daha yüksek doğrulukta bir model elde edebilmek adına ileri çalışmalarda;
  - Veri sayısı artırılarak tekrar makine öğrenmesi yöntemleri uygulanabilir.
  - Hiperparametre optimizasyonu yapılarak tekrar modeller oluşturabilir.
  - Adaboost, XGBoost, LightGBM, ANN gibi farklı modeller kullanılabilir.



**DİNLEDİĞİNİZ İÇİN  
TEŞEKKÜRLER**