

**Gölhisar Uygulamalı Bilimler Yüksekokulu**

**NESNE TABANLI PROGRAMLAMA II DERSİ**

**PROJE KONUSU: Kanser Teşhisi ve Sınıflandırması**

**Öğrenci Ad-Soyad/Numara**

1. **2212903060 Muhammet Emir HAKSEVER**
2. **2212903046 Mustafa KARAKILIÇ**
3. **2212903036 Mehmet İLCİ**
4. **2212903074 Mustafa KARADAŞ**

MAYIS 2024 BURDUR

# GİRİŞ

## Veri kümesi ile ilgili detaylı bilgiler

Veri kümesi meme kanseri teşhisi için kullanılan hücresel özelliklere ait çeşitli verileri içermektedir. Aşağıda bu veri kümesi hakkında detaylı bilgiler yer almaktadır:

**Genel Bilgiler**

* Veri Kayıt Sayısı (Satır Sayısı): 683
* Özellik Sayısı (Sütun Sayısı): 11

**Sütunlar ve Özellikler**

* Sample code number: Örnek kod numarası (int)
* Clump Thickness: Hücre yığını kalınlığı (int 1-10)
* Uniformity of Cell Size: Hücre boyutunun düzenliliği (int 1-10)
* Uniformity of Cell Shape: Hücre şeklinin düzenliliği (int 1-10)
* Marginal Adhesion: Hücrelerin birbirine yapışma derecesi (int 1-10)
* Single Epithelial Cell Size: Tek epitelyal hücre boyutu (int 1-10)
* Bare Nuclei: Çıplak çekirdekler (int 1-10)
* Bland Chromatin: Kromatin dokusu (int 1-10)
* Normal Nucleoli: Normal çekirdekçikler (int 1-10)
* Mitoses: Mitoz bölünme sayısı (int 1-10)
* Class: Teşhis sınıfı (int 2 = benign (iyi huylu) 4 = malignant (kötü huylu))

**İstatistiksel Özet**

* Clump Thickness:
* Ortalama: 4.44
* Standart Sapma: 2.82
* Min: 1
* Max: 10
* Uniformity of Cell Size:
* Ortalama: 3.15
* Standart Sapma: 3.07
* Min: 1
* Max: 10
* Bare Nuclei:
* Ortalama: 3.54
* Standart Sapma: 3.65
* Min: 1
* Max: 10
* Class:
* Ortalama: 2.70
* Standart Sapma: 0.95
* Min: 2
* Max: 4

**İlk Beş Kayıt**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample code number | Clump Thickness | Uniformity of Cell Size | Uniformity of Cell Shape | Marginal Adhesion | Single Epithelial Cell Size | Bare Nuclei | | Bland Chromatin | Normal Nucleoli | Mitoses | Class |
| 1000025 |  |  | 1 |  |  | | 1 | 3 | 1 | 1 | 2 |
| 1002945 |  |  | 4 |  |  | | 1 | 3 | 2 | 1 | 2 |
| 1015425 |  |  | 1 |  |  | | 2 | 3 | 1 | 1 | 2 |
| 1016277 |  |  | 8 |  |  | | 4 | 3 | 7 | 1 | 2 |
| 1017023 |  |  | 1 |  |  | | 1 | 3 | 1 | 1 | 2 |

**Özet**

Bu veri kümesi meme kanseri teşhisi için kullanılabilecek 10 farklı hücresel özellikten oluşmaktadır. Bu özellikler kanserli ve kanserli olmayan hücreleri ayırt etmek için kullanılır. Veri kümesi sınıflandırma algoritmaları geliştirmek ve test etmek için oldukça uygun bir yapıdadır.

## 

## LOJİSTİK REGRESYON

Lojistik regresyon, sınıflandırma problemlerinde kullanılan istatistiksel bir modeldir. Temel amacı, bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi modelleyerek bağımlı değişkenin belirli sınıflara ait olma olasılığını tahmin etmektir. Lojistik regresyon, makine öğrenimi ve veri bilimi alanlarında yaygın olarak kullanılan güçlü bir sınıflandırma algoritmasıdır ve birçok farklı alanda başarılı bir şekilde uygulanabilir.

# GEREÇ VE YÖNTEM

## Tarihsel Geçmiş

Lojistik regresyon, 19. yüzyılın ortalarında Belçikalı matematikçi Pierre François Verhulst tarafından popülasyon büyümesini modellemek amacıyla geliştirilmiştir. Verhulst, 1838'de lojistik fonksiyonun temelini atmış ve bu fonksiyonun ayrıntılı bir versiyonunu 1845'te yayınlamıştır. Bu çalışmanın ardından Chester Ittner Bliss, 1934'te probit modelini geliştirmiş ve Ronald Fisher, 1935'te maksimum olabilirlik tahminini ortaya koymuştur. Joseph Berkson ise lojistik modeli "logit" olarak adlandırmış ve bu terminoloji yıllarca kullanılmıştır. 1966'da Cox tarafından çoklu logit modeli tanıtılmış ve 1973'te McFadden bu modeli ayrık seçim teorisi ile ilişkilendirerek önemli bir ilerleme kaydetmiştir. McFadden, bu çalışmaları nedeniyle 2000 yılında Nobel Ödülü almıştır.

## Kullanım Alanları

Lojistik regresyon, basit ve etkili olması nedeniyle pek çok alanda yaygın olarak kullanılmaktadır. Bazı önemli kullanım alanları şunlardır:

1. **Kredi Skorlama**: Finansal kurumlar, kredi başvurularını değerlendirirken lojistik regresyon kullanarak modeller oluşturur. Bu modeller, kararların kolayca anlaşılabilir olmasını sağlar ve düzenleyici taleplere hızlı yanıt verilmesini mümkün kılar.
2. **Tıp**: Genetik hastalıkların tespiti ve kan testleri gibi biyomedikal araştırmalarda lojistik regresyon kullanılır. Miroculus gibi şirketler, mikroRNA ve genler arasındaki ilişkileri belirlemek için lojistik regresyon modelleri kullanarak hızlı ve doğru sonuçlar elde etmektedir.
3. **Doğal Dil İşleme (NLP)**: Metin sınıflandırma, toksik konuşma tespiti ve e-posta sıralama gibi alanlarda lojistik regresyon yaygın olarak kullanılmaktadır. Metin verilerini işlemek ve sınıflandırmak için kullanılan bu yöntem, basit ve etkili çözümler sunar.
4. **Otel Rezervasyonu**: Booking.com gibi hizmetler, kullanıcıların niyetlerini ve tercihlerini tahmin etmek için lojistik regresyon kullanır. Bu modeller, kullanıcı davranışlarını analiz ederek daha iyi hizmet sunmayı hedefler.

## Görseller ve Şekiller

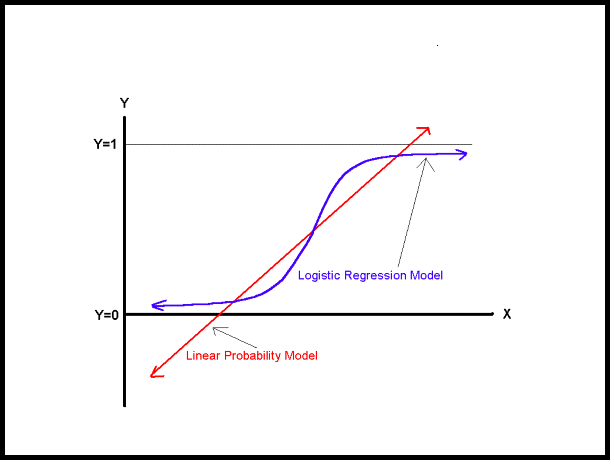
Lojistik regresyon modeli, S-eğrisi şeklinde bir grafikle görselleştirilir. Bu grafik, bağımsız değişkenlerin olasılıkla ilişkilendirildiği bir sigmoid fonksiyonunu temsil eder. Aşağıdaki görseller, lojistik regresyonun nasıl çalıştığını ve uygulandığı alanları göstermektedir:

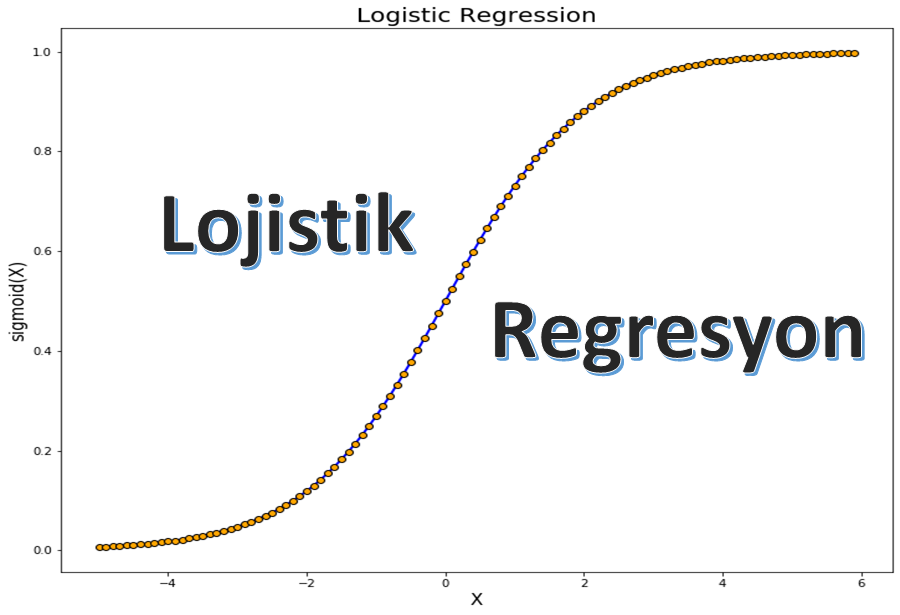
1. **Lojistik Fonksiyon Grafiği**: Bu grafik, -6 ile +6 arasında değişen x değerleri için lojistik fonksiyonun nasıl davrandığını gösterir.
2. **Uygulama Örnekleri**: Kredi skorlama, tıbbi teşhis ve doğal dil işleme gibi alanlarda lojistik regresyonun nasıl kullanıldığını gösteren diyagramlar ve akış şemaları.

Lojistik regresyonun avantajları arasında basitlik, hız ve şeffaflık bulunmaktadır. Bu özellikleri, özellikle düzenleyici gerekliliklerin yüksek olduğu sektörlerde çok değerlidir.

Daha fazla bilgi için bu konularla ilgili detaylı kaynaklara ve görsellere ulaşabilirsiniz:

* [HolyPython](https://holypython.com/logistic-regression-history/)
* [MIT OpenCourseWare](https://ocw.mit.edu/courses/sloan-school-of-management/15-071-the-analytics-edge-spring-2017/logistic-regression/)
* [ActiveWizards](https://activewizards.com/blog/5-real-world-examples-of-logistic-regression-application/)





## LOJİSTİK REGRESYON

Lojistik regresyon, sınıflandırma problemlerinde kullanılan istatistiksel bir modeldir. Temel amacı, bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi modelleyerek bağımlı değişkenin belirli sınıflara ait olma olasılığını tahmin etmektir. İşte lojistik regresyon hakkında bazı temel bilgiler:

Temel Kavramlar

1. Bağımlı Değişken: İkili (binary) ya da çoklu (multinomial) kategorik bir değişken. Genellikle 0 ve 1 şeklinde kodlanır.

2. Bağımsız Değişkenler: Bağımlı değişkeni tahmin etmek için kullanılan bir veya daha fazla sürekli ya da kategorik değişken.

Lojistik Fonksiyonu

Lojistik regresyon modeli, lojistik fonksiyon (sigmoid fonksiyon) kullanılarak tahmin edilir:

\[ P(Y=1|X) = \frac{1}{1 + e^{-(\beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2 + ... + \beta\_n X\_n)}} \]

Burada:

- \( P(Y=1|X) \): Y'nin 1 olma olasılığı

- \( \beta\_0 \): Modelin sabit terimi (intercept)

- \( \beta\_1, \beta\_2, ..., \beta\_n \): Bağımsız değişkenlerin katsayıları

- \( X\_1, X\_2, ..., X\_n \): Bağımsız değişkenler

Modelin Amacı

Model, bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi belirler ve bağımlı değişkenin belirli bir sınıfa ait olma olasılığını tahmin eder. Örneğin, bir hastanın belirli bir hastalığa sahip olup olmadığını, bir e-posta'nın spam olup olmadığını tahmin edebilir.

Yorumlama

- Katsayılar (\(\beta\)): Katsayılar, bağımsız değişkenlerdeki bir birimlik değişimin, bağımlı değişkenin logaritmik olasılığı üzerindeki etkisini gösterir.

- Odds ve Odds Ratio: Odds, bir olayın olma olasılığının olmama olasılığına oranıdır. Odds ratio ise, iki farklı durumda odds'ların oranıdır.

Avantajları ve Dezavantajları

- Avantajları:

- Basit ve yorumlaması kolaydır.

- Modelin katsayıları, değişkenlerin önemini ve yönünü gösterir.

- İkili ve çoklu sınıflandırma problemlerinde etkilidir.

- Dezavantajları:

- Doğrusal olmayan ilişkileri iyi yakalayamaz.

- Bağımsız değişkenler arasında yüksek korelasyon varsa (multikolineerlik), model performansı düşebilir.

- Veri setinin dengeli olmasını gerektirir; dengesiz veri setlerinde performansı düşebilir.

Lojistik regresyon, makine öğrenimi ve veri bilimi alanlarında yaygın olarak kullanılan güçlü bir sınıflandırma algoritmasıdır ve birçok farklı alanda başarılı bir şekilde uygulanabilir.

# BULGULAR

**Kütüphanelerin Yüklenmesi**

* import pandas as pd
* import numpy as np
* from sklearn.linear\_model import LogisticRegression
* from sklearn.model\_selection import train\_test\_split
* from sklearn.metrics import accuracy\_score, confusion\_matrix, classification\_report

**Veri Setinin Yüklenmesi ve Ön İşleme**

* data = pd.read\_csv("breast-cancer-wisconsin.csv")
* data.replace('?', -99999, inplace=True) # Eksik verileri uygun bir değere değiştirme
* data.drop(['id'], axis=1, inplace=True) # ID sütununu kaldırma

**Eğitim ve Test Verilerinin Ayrılması**

* X = data.drop(['class'], axis=1)
* y = data['class']
* X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2)

**Modelin Eğitilmesi ve Başarı Oranı**

* model = LogisticRegression(max\_iter=1000)
* model.fit(X\_train, y\_train)
* predictions = model.predict(X\_test)
* accuracy = accuracy\_score(y\_test, predictions)
* print(f"Doğruluk Oranı: {accuracy \* 100:.2f}%")

**Sonuçların Değerlendirilmesi**

* conf\_matrix = confusion\_matrix(y\_test, predictions)
* class\_report = classification\_report(y\_test, predictions)
* print(conf\_matrix)
* print(class\_report)

**Grafik Analizler**

Bu kısımda elde edilen sonuçları grafiksel olarak axis=1, inplace=True) # ID sütununu kaldırma analiz edip sunabilirsiniz. Örneğin, doğru ve yanlış sınıflandırmaları gösteren bir çubuk grafik ekleyebilirsiniz

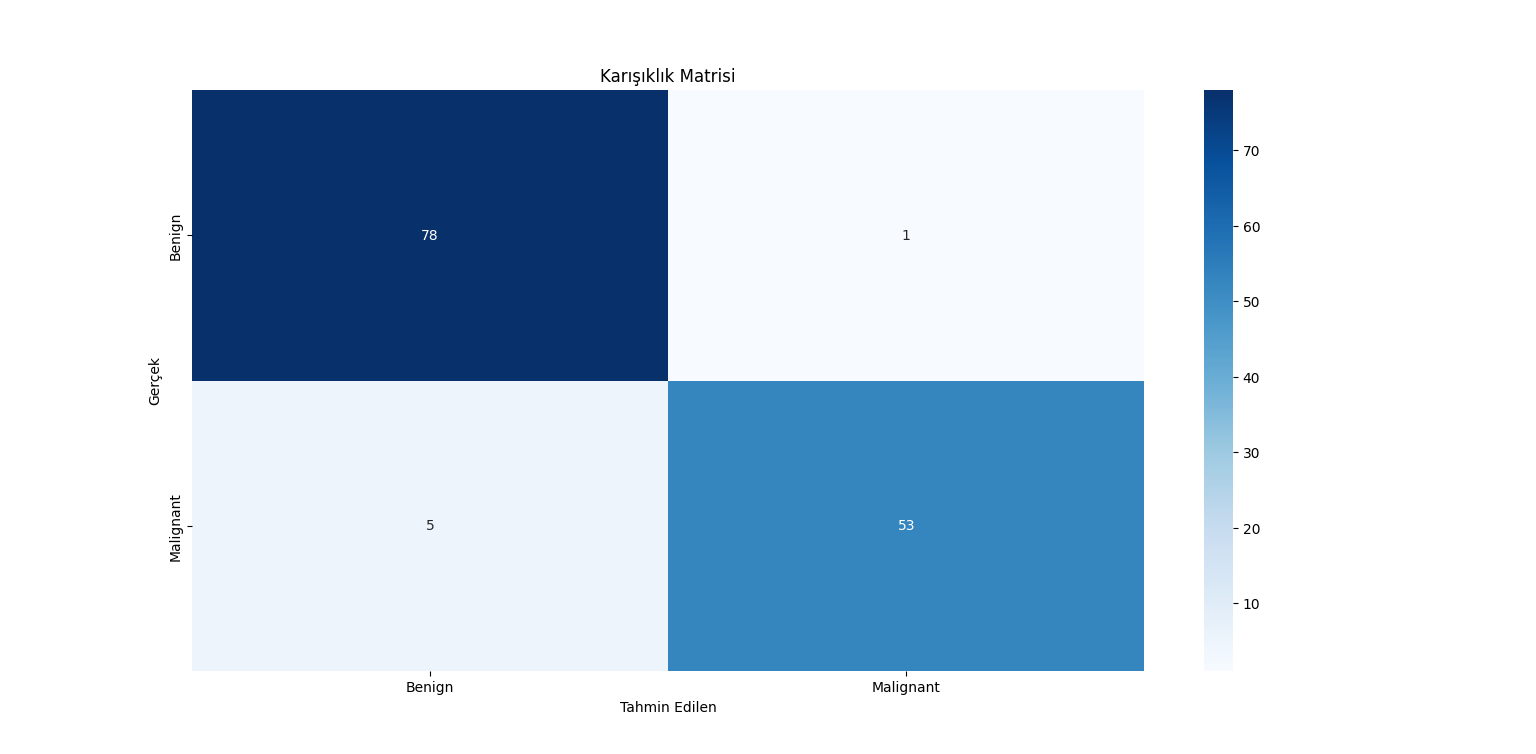
* import matplotlib.pyplot as plt
* import seaborn as sns
* sns.heatmap(conf\_matrix, annot=True, fmt="d")
* plt.title("Confusion Matrix")
* plt.xlabel("Tahmin Edilen")
* plt.ylabel("Gerçek")
* plt.show()

ekran görüntüsü, dikdörtgen, metin, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

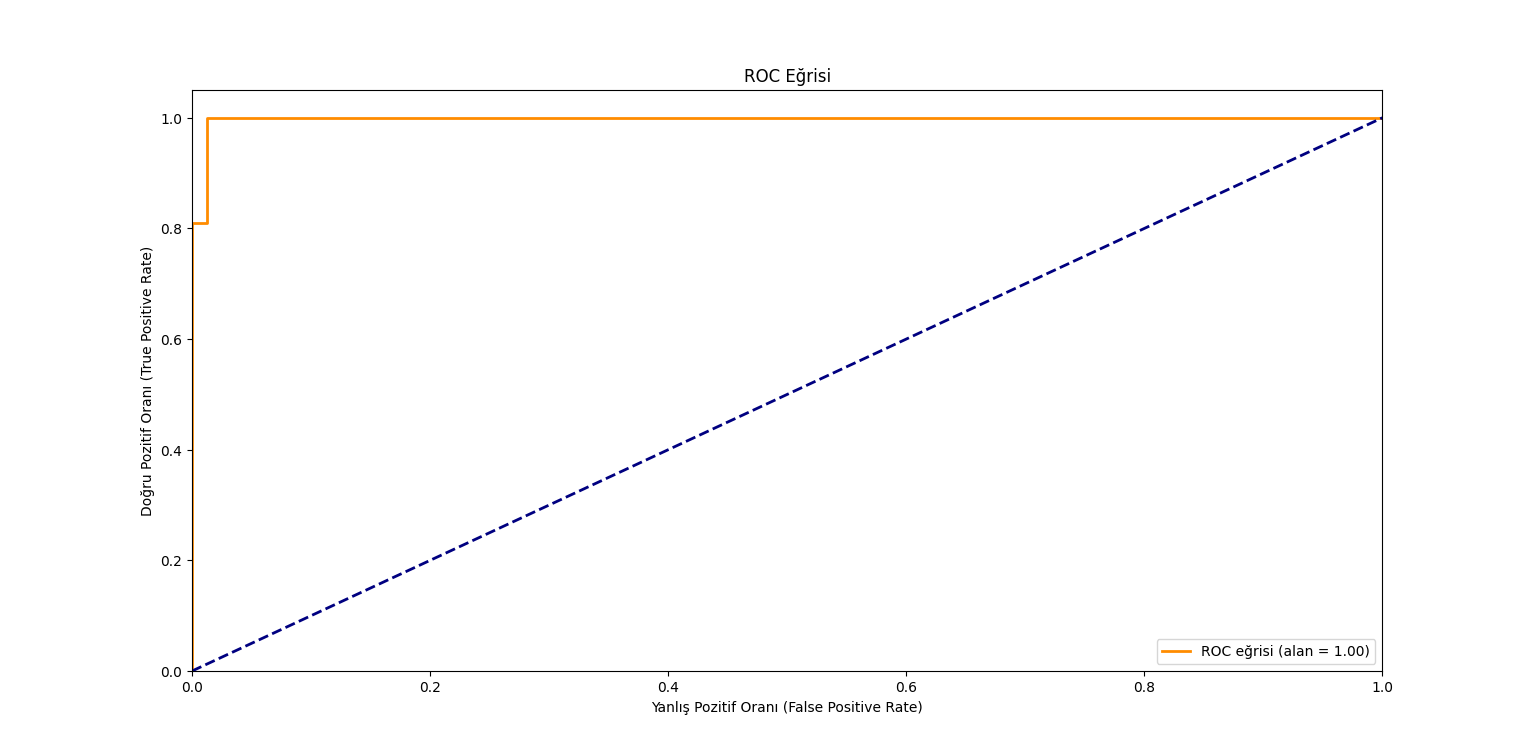
Şekil 1

1. **Model Performans Metrikleri Grafiği:**
   * Açıklama: Bu çubuk grafik, modelin performansını değerlendirmek için kullanılan dört farklı metriğin değerlerini gösterir.
   * Metri (Metric):
     + Accuracy (Doğruluk): Modelin tüm doğru tahminlerinin oranı. Yani, modelin doğru tahmin ettiği örneklerin toplam örneklere oranı.
     + F1 Score (F1 Skoru): Doğruluk (precision) ve geri çağırma (recall) arasındaki dengeyi ölçen bir metri. Özellikle sınıf dağılımı dengesiz olan durumlarda kullanışlıdır.
     + Precision (Doğruluk): Modelin pozitif sınıf için yaptığı tahminlerin doğruluk oranı. Yani, modelin pozitif sınıf olarak tahmin ettiği örneklerin gerçekten pozitif olma oranı.
     + Recall (Geri Çağırma): Modelin gerçek pozitif örnekleri ne kadar iyi bulduğunu gösterir. Yani, gerçek pozitif örneklerin model tarafından doğru bir şekilde pozitif olarak tahmin edilme oranı.
   * Grafikteki Değerler: Her bir metriğin değeri, 0 ile 1 arasında bir skala üzerinde gösterilmiştir. Değerler, performansın yüksek olduğunu göstermektedir.



Şekil 2

1. **Karışıklık Matrisi (Confusion Matrix):**
   * Açıklama: Bu ısı haritası, modelin doğru ve yanlış sınıflandırmalarını ayrıntılı olarak gösterir.
   * Ekseni Etiketleri:
     + X Ekseni (Tahmin Edilen): Modelin tahmin ettiği sınıflar (Benign veya Malignant).
     + Y Ekseni (Gerçek): Gerçek sınıflar (Benign veya Malignant).
   * Matris Hücreleri:
     + True Negatives (TN): Modelin negatif (benign) olarak doğru tahmin ettiği örneklerin sayısı.
     + False Positives (FP): Modelin yanlışlıkla pozitif (malignant) olarak tahmin ettiği negatif (benign) örneklerin sayısı.
     + False Negatives (FN): Modelin yanlışlıkla negatif (benign) olarak tahmin ettiği pozitif (malignant) örneklerin sayısı.
     + True Positives (TP): Modelin pozitif (malignant) olarak doğru tahmin ettiği örneklerin sayısı.
   * Renk Skalası: Renkler, hücrelerdeki sayıların büyüklüğüne göre değişir; daha koyu renkler daha yüksek sayıları temsil eder.



Şekil 3

1. **ROC Eğrisi ve AUC (Area Under Curve):**
   * Açıklama: Bu çizim, modelin sınıflandırma performansını değerlendirmek için ROC (Receiver Operating Characteristic) eğrisini ve AUC (Area Under the Curve) değerini gösterir.
   * Eksenler:
     + X Ekseni (Yanlış Pozitif Oranı - False Positive Rate): Yanlış pozitif tahminlerin oranı (yanlış alarmlar).
     + Y Ekseni (Doğru Pozitif Oranı - True Positive Rate): Doğru pozitif tahminlerin oranı (hassasiyet veya geri çağırma).
   * Çizgiler:
     + ROC Eğrisi: Doğru pozitif oranı ile yanlış pozitif oranı arasındaki ilişkiyi gösterir. Eğri ne kadar yukarıda ve sola yakınsa, modelin performansı o kadar iyidir.
     + Diyagonal Çizgi (Random Guessing): Rastgele tahmin eden bir modelin ROC eğrisidir. Eğrinin bu çizginin üstünde olması, modelin rastgele tahminlerden daha iyi olduğunu gösterir.
   * AUC Değeri: ROC eğrisinin altındaki alan. 1'e ne kadar yakınsa, modelin performansı o kadar iyidir. Bu değer, modelin genel olarak pozitif ve negatif sınıfları ayırt etme yeteneğini özetler.

# KAYNAKLAR

* - Abramowitz, M., & Stegun, I. A. (1964). \*Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables\*. US Government Printing Office.
* - ActiveWizards. (2024). \*5 Real-world Examples of Logistic Regression Application\*. ActiveWizards.
* - ActiveWizards. (n.d.). \*5 Real-World Examples of Logistic Regression Application\*. Retrieved from https://activewizards.com/blog/5-real-world-examples-of-logistic-regression-application/
* - Cox, D. R. (1966). \*Multinomial Extensions of Logistic Regression\*. Biometrika.
* - McFadden, D. (1973). \*Conditional Logit Analysis of Qualitative Choice Behavior\*.
* - MIT OpenCourseWare. (n.d.). \*Logistic Regression\*. Retrieved from https://ocw.mit.edu/courses/sloan-school-of-management/15-071-the-analytics-edge-spring-2017/logistic-regression/

**KAYNAK KOD ADRESİ**

**Projenizde kullanılan kaynak kodlarını eklediğiniz GITHUB sayfalarınızın bağlantı adresini buraya eklemeniz gerekmektedir.**

**https://github.com/EmirHaksever/Python/blob/main/README.md**