
# Gölhisar Uygulamalı Bilimler Yüksekokulu

## NESNE TABANLI PROGRAMLAMA II DERSİ

## PROJE KONUSU: Kanser Teşhisi ve Sınıflandırması

**Öğrenci Ad-Soyad/Numara**
**1-Nazmi Yücel Çan 2212903018**
**2-Emir can Manici 2212903044**
**3-Ahmet Arif Güneri 2212903034**
**4-Aslan Karaca 2212903016**
**5-Hacer Maltar 2212903059**

MAYIS 2024
BURDUR

.

# 1. GİRİŞ

Meme kanseri, meme dokusunda başlayan ve meme hücrelerinin kontrolsüz ve anormal bir şekilde büyümesiyle karakterize edilen bir kanser türüdür. Meme kanseri genellikle kadınlarda görülse de, nadiren erkeklerde de görülebilir. Her yaşta ortaya çıkabilir, ancak genellikle menopoz sonrası dönemde daha sık görülür. Meme kanseri çeşitli alt tiplere ayrılabilir ve bu alt tipler farklı hücresel özelliklere ve tedavi yöntemlerine sahip olabilir. Tedavi seçenekleri meme kanserinin tipine, evresine ve diğer faktörlere bağlı olarak değişebilir, ancak cerrahi, kemoterapi, radyoterapi, hormonsal terapi ve hedefe yönelik ilaç tedavisi gibi yöntemler genellikle kullanılır.

Meme kanseri teşhisinde makine öğrenimi tekniklerinin kullanımını incelemekte ve bu tekniklerin başarı oranlarını değerlendirmektedir. Meme kanseri teşhisinde kullanılan yaygın özellikleri ve sınıflandırma algoritmalarını ele alarak, özellikle Yığın Kalınlığı, Hücre Boyutunun Eş biçimliliği ve diğerleri gibi özelliklerin kanser teşhisindeki önemini vurgulamaktadır. Ayrıca, farklı makine öğrenimi yöntemlerinin bu veri seti üzerindeki performansını karşılaştırmak için yapılan çalışmalara da odaklanmaktadır.

Proje konumuz Gaussian Naive Bayes (GNB), sınıflandırma problemlerinde kullanılan, istatistiksel bir makine öğrenme algoritmasıdır. Naive Bayes sınıflandırıcılarının bir türü olan GNB, özellikle sürekli veriyle çalışmak için uygundur. Temel olarak, Naive Bayes algoritması, Bayes teoremi ve "naive" (saf) varsayımı üzerine kuruludur. Bu varsayım, tüm özelliklerin birbirinden bağımsız olduğunu kabul eder. Diğer algoritmalar arasında Gaussian Naive Bayes algoritması ise diğer algoritmalara göre daha iyi sonuçlar ortaya koymaktadır.

Projemizdeki veri kümemiz 684 satır 11 sütun 9 adet özellik bulunmaktadır.

Veri özellikleri:

1. Clump Thickness (Yığın kalınlığı)
2. Uniformity of Cell Size (Hücre Boyutunun Tekdüzeliği)
3. Uniformity of Cell Shape (Hücre Şeklinin Tekdüzeliği)
4. Marginal Adhesion (Marjinal adezyon)
5. Single Epithelial Cell Size (Tek epitel hücre boyutu)
6. Bare Nuclei (Çıplak Çekirdekler)
7. Bland Chromatin (Yumuşak Kromatin )
8. Normal Nucleoli (Normal Nükleoller)
9. Mitoses ( Mitozlar)

# 2. GEREÇ VE YÖNTEM

## 2.1 Veri Seti

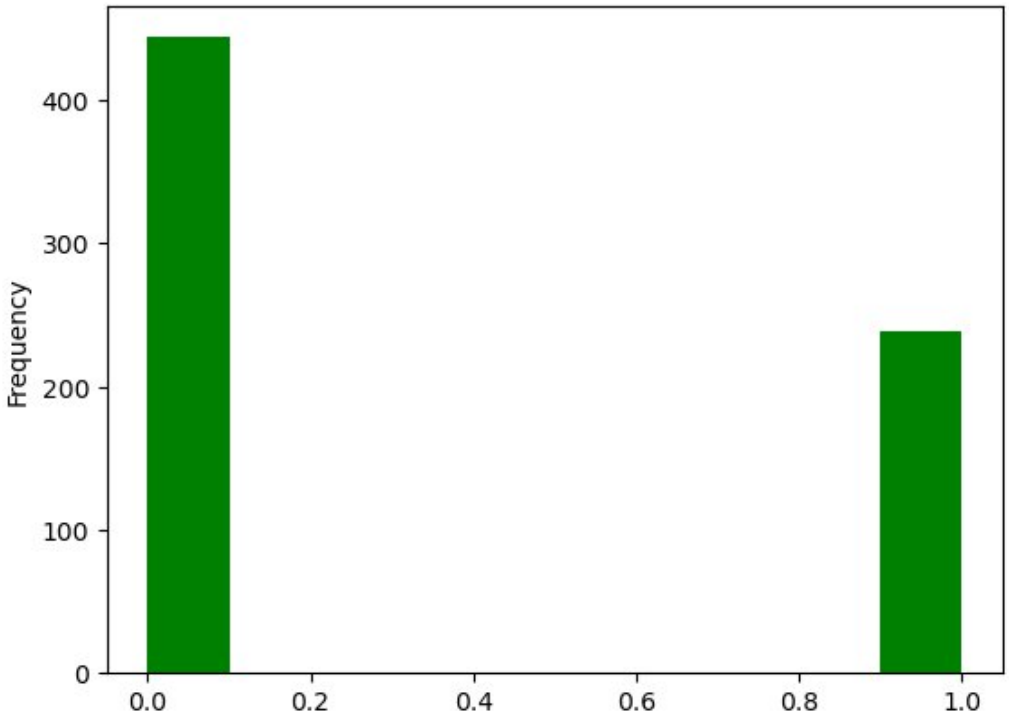
Projemizdeki veri kümemiz son temizlenmiş halinde 683 satır 11 sütun 9 adet özellik bulunmaktadır.

Veri özellikleri:

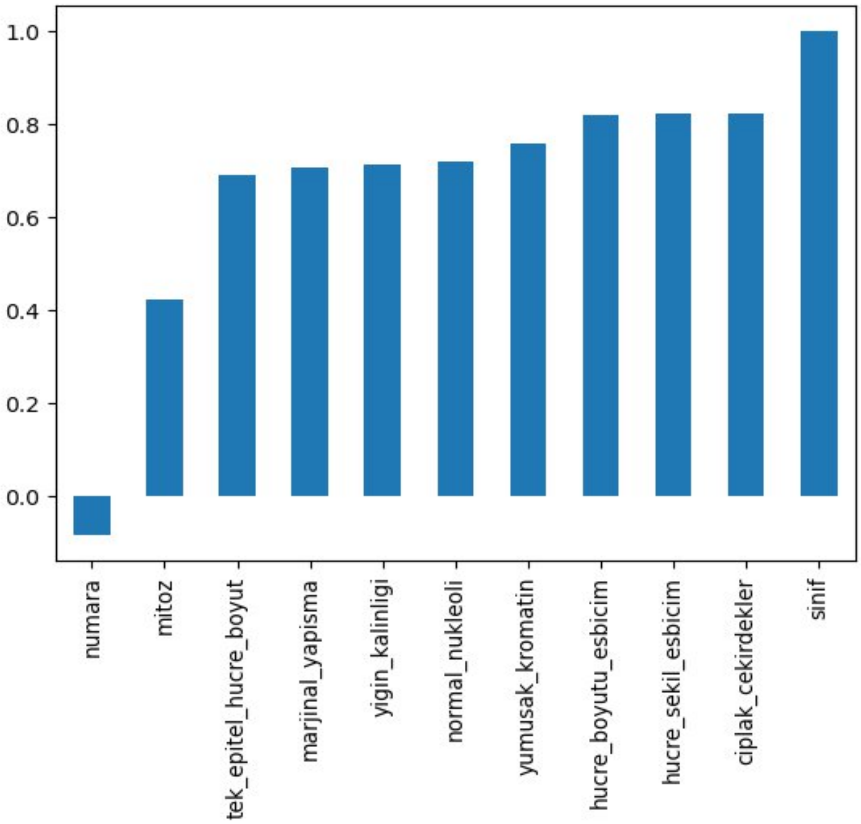
1. Clump Thickness (Yığın kalınlığı)
2. Uniformity of Cell Size (Hücre Boyutunun Tekdüzeliği)
3. Uniformity of Cell Shape (Hücre Şeklinin Tekdüzeliği)
4. Marginal Adhesion (Marjinal adezyon)
5. Single Epithelial Cell Size (Tek epitel hücre boyutu)
6. Bare Nuclei (Çıplak Çekirdekler)

7. Bland Chromatin (Yumuşak Kromatin )
8. Normal Nucleoli (Normal Nükleoller)
9. Mitoses ( Mitozlar)

Veri setinde 444 adet iyi huylu örnek, 239 adet ise kötü huylu örnek vardır.




Grafik 1. İyi huylu örnekler 0 ile gösterilmiş, kötü huylu örnekler ise 1 ile gösterilmiştir.




Grafik 2. Burda veri setindeki özelliklerin korelasyon değerlerine göre sıralaması

gösterilmektedir. Tablodan anlaşıldığı üzere sınıf sütununa en yakın korelasyon değerine sahip olan ciplak_cekirdekler özelliğidir.

## 2.2 GAUSSIAN NAIVE BAYES

### 2.2.1 Nedir?

Gaussian Naive Bayes, Bayes ‘in teorisini temel alan olasılık tabanlı bir makine öğrenme algoritmasıdır. Her parametrenin çıkış değişkenini tahmin etme konusunda bağımsız bir kapasiteye sahip olduğu varsayılır. Bağımlı bir değişkenin her grupta sınıflandırılma olasılığını tahmin edebilir.

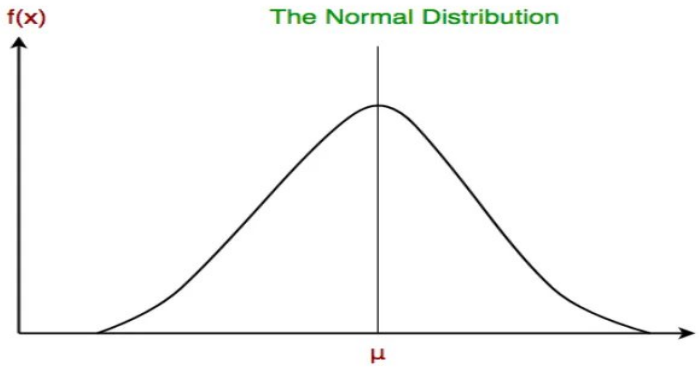
**Gaussian Dağılım**

Gaussian dağılımı, aşağıdaki formülle tanımlanır:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}\right)$$

Gauss Naive Bayes'de, her bir özellik ile ilişkili sürekli değerlerin, Normal dağılım olarak da bilinen bir Gauss dağılımını takip ettiği varsayımı yapılır. Görselleştirildiğinde bu dağılım, özellik değerlerinin ortalamasına göre simetrik olan çan şeklinde bir eğri oluşturur. Normal dağılım düzgün, sürekli bir eğri ile karakterize edilir ve çeşitli istatistiksel ve makine öğrenimi uygulamalarında sürekli değişkenlerin dağılımını modellemek için yaygın olarak kullanılır.

Çan şeklindeki eğri, özellik değerlerinin çoğunun ortalama etrafında kümelendiğini ve ortalamadan her iki yönde uzaklaştıkça oluşum sayısının simetrik olarak azaldığını gösterir. Bu varsayım, modelleme sürecini basitleştirir ve sınıf etiketi göz önüne alındığında özellik bağımsızlığının varsayıldığı Naive Bayes sınıflandırması kapsamında verimli olasılıksal hesaplamalara izin verir.




Gaussian Naive Bayes, basitlik, hızlı eğitim süresi ve düşük bellek gereksinimleri gibi avantajlara sahiptir. Bu nedenle, özellikle büyük veri setlerinde ve hızlı çözüm gerektiren durumlarda tercih edilir. Ancak, verilerin normal dağılıma sahip olduğu varsayımını karşılamadığı durumlarda performansı düşebilir. Bu nedenle, veri setinizin özelliklerini ve gereksinimlerinizi dikkate alarak Gaussian Naive Bayes'in kullanımını değerlendirmeniz önemlidir.

.

### 2.2.2 TARİHSEL GELİŞİM

Naive Bayes algoritmasının kökenleri 18. yüzyıla kadar uzanır, ancak modern makine öğrenimi literatüründeki kullanımı daha yakın bir tarihe, özellikle 1960'lara dayanır. Gaussian Naive Bayes'in tam olarak ne zaman ortaya çıktığına dair net bir tarih vermek zordur, çünkü sürekli verilerle çalışan algoritmalara olan ihtiyaç, veri bilimi ve istatistik alanındaki gelişmelere bağlı olarak zamanla değişmiştir.

Bayes Teoremi şu şekilde ifade edilir:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Gaussian Naive Bayes'in yaygın olarak kullanılmaya başlaması, makine öğrenimi ve istatistik alanlarının popülerleştiği son yıllara, yani 20. yüzyılın sonlarına ve 21. yüzyılın başlarına dayanır. Özellikle, Gaussian Naive Bayes'in metin sınıflandırma ve doğal dil işleme gibi alanlarda etkili olduğu keşfedildikten sonra popülerliği artmıştır.

### 2.2.3 GAUSSİAN NAIVE BAYES KULLANIM ALANLARI

Gaussian Naive Bayes'in kullanım alanlarından bazıları:

- Metin Sınıflandırma: Gaussian Naive Bayes, metin verilerini sınıflandırmak için yaygın olarak kullanılır. Özellikle spam filtreleme, duygu analizi ve belge sınıflandırma gibi metinle ilgili problemlerde etkilidir.

- Tıbbi Teşhis: Tıbbi teşhislerde, özellikle laboratuvar test sonuçları gibi sürekli verilerin olduğu durumlarda Gaussian Naive Bayes sıkça kullanılır. Örneğin, bir hastalığın varlığını veya yokluğunu belirlemek için laboratuvar test sonuçlarını kullanabilir.

- Ekonomik Tahminler: Finansal piyasalardaki verilerin analizinde Gaussian Naive Bayes sıklıkla kullanılır. Hisse senedi fiyat tahminleri, kredi riski değerlendirmeleri ve müşteri segmentasyonu gibi ekonomik tahminlerde bu algoritma kullanılabilir.

- Mühendislik ve Endüstri: Üretim süreçleri, kalite kontrolü, hata tespiti gibi endüstriyel uygulamalarda Gaussian Naive Bayes kullanılabilir. Sensör verileri gibi sürekli ölçümlerin analizi için uygundur.

- Biyoinformatik: Biyolojik verilerin analizinde Gaussian Naive Bayes sıklıkla kullanılır. Örneğin, gen ifadesi verilerini sınıflandırmak veya protein fonksiyonlarını tahmin etmek için kullanılabilir.

- Makine Öğrenimi Modellerinde Baseline Olarak: Makine öğrenimi modellerinin başlangıç noktası olarak kullanılabilir. Özellikle daha karmaşık modelleri oluşturmadan önce, veri setinin genel yapısını anlamak için basit bir Gaussian Naive Bayes modeli oluşturulabilir.

## 2.3 Tekrarlanan K Katlı Çapraz Doğrulama (Repeated K-Fold Cross-Validation)

.

### 2.3.1 Nedir?

Tekrarlanan K Katlı Çapraz Doğrulama (TKCD), bir makine öğrenimi modelinin performansını değerlendirmek için kullanılan bir yöntemdir. Bu yöntem, modelin aşırı uyum riskini azaltmak ve genellenebilirliğini doğrulamak için kullanılır.

### 2.3.2 Çalışma Prensibi

- Veri Bölünmesi: Veri seti, K eşit büyüklükte alt kümeye (kat) bölünür.
- Tekrarlama: Aşağıdaki adımlar, belirli bir tekrar sayısı (L) kadar tekrarlanır:
    - Eğitim ve Test Katmanı: K katmandan biri test katmanı olarak seçilir, kalan K-1 katman ise eğitim katmanı olarak kullanılır.
    - Model Eğitimi: Model, eğitim katmanı üzerinde eğitilir.
    - Performans Değerlendirmesi: Modelin performansı, test katmanı üzerinde değerlendirilir.
- Genel Performans Hesaplama: Her tekrar için elde edilen performans değerleri ortalama alınarak modelin genel performansı hesaplanır.

### 2.3.3 Avantajlar

- Aşırı uyum riskini azaltır.
- Modelin genellenebilirliğini daha iyi tahmin eder.
- Farklı model parametrelerinin ve mimarilerinin karşılaştırılmasında kullanılabilir.

### 2.3.4 Dezavantajlar

- Hesaplama açısından pahalı olabilir.
- Veri setinin büyüklüğünden etkilenir.
- K ve L değerlerinin seçimi modelin performansını etkileyebilir.

### 2.3.5 Tarihsel Gelişimi

Tekrarlanan K Katlı Çapraz Doğrulama (TKCD), nispeten yeni bir yöntem olmasına rağmen, istatistik ve makine öğrenimi alanlarında köklü bir geçmişe sahiptir. Gelişimi, birden fazla araştırmacının katkılarından oluşmuştur:

**1950'ler:**

- Leo Breiman: "Çapraz Doğrulama" terimini ilk kez 1952 yılında "Çapraz Doğrulama Kullanarak Çeşitli Tahmin Prosedürlerinin Karşılaştırılması" adlı makalesinde kullandı.
- George Stone: 1954 yılında "Çapraz Doğrulama Seçimi" adlı makalesinde K Katlı Çapraz Doğrulama yöntemini ilk kez açıkladı.

**1960'lar:**

- Marvin Minsky: 1963 yılında "Beyin Modelleri" adlı kitabında TKCD'yi nöron ağları bağlamında tartıştı.

**1970'ler:**

- Donald Rubin: 1978 yılında "Çapraz Doğrulama ve Tahmin Seçimi" adlı makalesinde TKCD'nin istatistiksel temellerini sağlamlaştırdı.

**1980'ler:**

- Reginald N. Mantell: 1980 yılında "Çapraz Doğrulama Kullanarak Tahmin Performansının Değerlendirilmesi" adlı makalesinde TKCD'nin uygulamalarını

genişletti.

**1990'lar ve Sonrası:**

- Makine öğreniminin gelişmesiyle birlikte TKCD, model seçiminde ve hiperparametre optimizasyonunda popüler bir araç haline geldi.
- Birçok varyasyon ve geliştirme TKCD üzerine yapıldı.

**Günümüzde:**

- TKCD, makine öğrenimi ve istatistiksel modellemede yaygın olarak kullanılan standart bir yöntemdir.
- Farklı araştırma alanlarında, özellikle biyoloji, tıp ve mühendislikte de kullanılmaktadır.

## 3. BULGULAR

Bu araştırmada temel olarak Gaussian Naive Bayes üzerinde durulmuş ayrıca ek olarak farklı sınıflandırma metodları ve Tekrarlanan K Katlı Çapraz Doğrulama ile diğer araştırmaların sonuçları karşılaştırılmıştır. Bu araştırma makalelerini belirtirken kullanılacak olan numaralandırmalar kaynaklar kısmında belirtilmiştir.

| Sahiplik | Metodlar | Temel/R-KF | Parametre Değerleri | Doğruluk(%) |
|---|---|---|---|---|
| Şuanki Araştırma | Gaussian Naive Bayes | Temel | train_size=0.60 random_state=5 | 95.6 |
| | | | train_size=0.60 random_state=9 | 98.1 |
| | | | train_size=0.75 random_state=9 | 99.4 |
| | | | train_size=0.75 random_state=5 | 98.2 |
| | | R-KF | n_splits=5 n_repeats=4 random_state=5 | Ort. 96.2 |
| | AdaBoost | Temel | train_size=0.60 random_state=5 | 95.6 |
| | | | train_size=0.60 random_state=9 | 97.0 |
| | | | train_size=0.75 random_state=9 | 96.4 |
| | | | train_size=0.75 random_state=5 | 97.0 |
| | | R-KF | n_splits=5 n_repeats=4 random_state=5 | Ort. 95.8 |
| | Bernouli Naive Bayes | Temel | train_size=0.60 random_state=5 | 65.6 |
| | | | train_size=0.60 random_state=9 | 63.5 |

| | | | |
|---|---|---|---|
| | | train_size=0.75<br>random_state=9 | 61.4 |
| | | train_size=0.75<br>random_state=5 | 66.6 |
| | R-KF | n_splits=5<br>n_repeats=4<br>random_state=5 | 97.0 |
| Multinomial Naive Bayes | Temel | train_size=0.60<br>random_state=5 | 90.8 |
| | | train_size=0.60<br>random_state=9 | 90.1 |
| | | train_size=0.75<br>random_state=9 | 92.9 |
| | | train_size=0.75<br>random_state=5 | 90.6 |
| Decision Tree | Temel | train_size=0.60<br>random_state=5 | 92.3 |
| | | train_size=0.60<br>random_state=9 | 95.9 |
| | | train_size=0.75<br>random_state=9 | 95.9 |
| | | train_size=0.75<br>random_state=5 | 94.1 |
| | R-KF | n_splits=5<br>n_repeats=4<br>random_state=5 | 94.1 |
| K-Means | Temel | train_size=0.60<br>random_state=5 | 94.8 |
| | | train_size=0.60<br>random_state=9 | 94.8 |
| | | train_size=0.75<br>random_state=9 | 97.0 |
| | | train_size=0.75<br>random_state=5 | 01.7 |
| K-Neighbor | Temel | train_size=0.60<br>random_state=5 | 95.6 |
| | | train_size=0.60<br>random_state=9 | 98.1 |
| | | train_size=0.75<br>random_state=9 | 98.8 |
| | | train_size=0.75<br>random_state=5 | 98.2 |
| | R-KF | n_splits=5<br>n_repeats=4<br>random_state=5 | 96.8 |
| Logistic Regression | Temel | train_size=0.60<br>random_state=5 | 95.9 |
| | | train_size=0.60<br>random_state=9 | 98.5 |

| | | | | |
|---|---|---|---|---|
| | | | train_size=0.75<br>random_state=9 | 98.8 |
| | | | train_size=0.75<br>random_state=5 | 98.2 |
| | | R-KF | n_splits=5<br>n_repeats=4<br>random_state=5 | 96.5 |
| | SVM | Temel | train_size=0.60<br>random_state=5 | 95.9 |
| | | | train_size=0.60<br>random_state=9 | 97.8 |
| | | | train_size=0.75<br>random_state=9 | 98.8 |
| | | | train_size=0.75<br>random_state=5 | 97.6 |
| | Random Forest | Temel | train_size=0.60<br>random_state=5 | 95.9 |
| | | | train_size=0.60<br>random_state=9 | 98.5 |
| | | | train_size=0.75<br>random_state=9 | 98.2 |
| | | | train_size=0.75<br>random_state=5 | 98.2 |
| | | R-KF | n_splits=5<br>n_repeats=4<br>random_state=5 | 97.1 |
| | Xgboost | Temel | train_size=0.60<br>random_state=5 | 97.4 |
| | | | train_size=0.60<br>random_state=9 | 96.3 |
| | | | train_size=0.75<br>random_state=9 | 98.2 |
| | | | train_size=0.75<br>random_state=5 | 95.9 |
| | | R-KF | n_splits=5<br>n_repeats=4<br>random_state=5 | 96.4 |
| Verisetinin bulunduğu sitedeki Temel Model Performansı (Max Değer) | Xgboost | Temel | - | 98.2 |
| | SVM | | - | 98.8 |
| | Random Forest | | - | 98.8 |
| | NNC | | - | 98.8 |
| | Logistic Regression | | - | 98.8 |
| Makale-1 | Naive Bayes | - | - | 97.5 |

| | | | | |
|---|---|---|---|---|
| Makale-2 | SVM | - | - | Ort. 66.06 |

**Tablo 1. Farklı araştırma ve farklı metodların doğruluk oranlarının karşılaştırılması**

Tablo 1 üzerinde yapılan farklı araştımalar ve uygulanan farklı sınıflandırma metodlarının doğruluk oranları karşılaştırılmıştır. Yapılan şunaki araştırmada farklı parametre değerleri denenmiş olup Gaussian Naive Bayes Metodunun temel hali üzerinde 99.4% oranına varan sonuçlar elde edilmiştir. Fakat model üzerinde ezberleme durumuna engel olmak amacıyla Tekrarlanan K Katlı Çapraz Doğrulama(R-KF) tekniği kullanılmıştır. Bu teknikle tabloda belirtilen parametreler ile ortalama olarak 96.2% sonucuna ulaşılmıştır.

Aynı parametreler diğer metodlar üzerinde kullanıldığında ulaşılan en yüksek doğruluk oranına yetişememiştir. Fakat yine de daha farklı parametreler ile daha iyi sonuçlar çıkarılabileceği unutulmamalıdır.

Verisetinin bulunduğu sitede bazı metodların Temel Model Performansları belirtilmiştir. Bu doğruluk performansı değerlerinin yine hiç biri ulaşılan en yüksek doğruluk oranına yetişememiştir.

1 numaralı makale üzerinde Naive Bayes metodu ile ulaşılan doğruluk oranı 97.5 olduğu belirtilmiştir. Parametre değerleri ise bilinmemektedir.

2 numaralı makale üzerinde ise SVM metodu kullanılmış ve ulaşılan doğruluk değerinin ortalama olamk üzere 66.06 olduğu belirtilmiştir. Ayıca ulaşılan bu sonucun metot üzerinde bir optimizasyon yapılmadan bu sonuca ulaşıldığı belirtilmiştir.

.

## KAYNAKLAR

(1) Rathi, M., & Singh, A. K. (2012). Breast cancer prediction using naïve bayes classifier. International Journal of Information Technology & Systems, 1(2), 77-80.

(2) Indraswari, R., & Arifin, A.Z. (2017). RBF KERNEL OPTIMIZATION METHOD WITH PARTICLE SWARM OPTIMIZATION ON SVM USING THE ANALYSIS OF INPUT DATA'S MOVEMENT.

https://medium.com/@thommaskevin/tinyml-gaussian-naive-bayes-classifier-31f8d241c67c

https://builtin.com/artificial-intelligence/gaussian-naive-bayes

Ontivero-Ortega, M., Lage-Castellanos, A., Valente, G., Goebel, R., & Valdes-Sosa, M. (2017). Fast Gaussian Naïve Bayes for searchlight classification analysis. Neuroimage, 163, 471-479.

## KAYNAK KOD ADRESİ

**Nazmi Yücel Çan: https://github.com/N4YuC4/Breast-Cancer-Classification**

**Hacer Maltar. https://github.com/Hacermaltar/python-projeleri.git**

**Emir Can Manici: https://github.com/ideaswamp/gogus-kanser--s-n-fland-rmas**