Naïve Bayes Sınıflandırma Yöntemi İle Meme Kanseri Teşhisi

Fatih ATEŞ
Bursa Teknik Üniversitesi, 19360859074@ogrenci.btu.edu.tr

***Özetçe* – Günümüzde artarak devam eden meme kanseri sorununun teşhis edilmesi aşamasında, bilime katkıda bulunması amacıyla, ABD'nin Wisconsin eyaletindeki meme kanseri hastalarına ait veriler baz alınarak, Naïve Bayes yöntemi ile bir sınıflandırıcı geliştirilmiştir. Önerilen yöntem, daha önce aynı veri seti ile yapılmış bir akademik makale ile karşılaştırılmış ve sonuçlarıyla birlikte sunulmuştur. Meme kanserinin erken teşhis edilmesi konusunda her geçen gün ilerlemeler kaydedilmektedir. Geliştirilen sınıflandırıcı %96 oranında doğruluğa sahip olmasıyla beraber erken teşhis aşamasında büyük bir rol oynaması beklenmektedir.**

Anahtar Kelimeler – Meme kanseri, Naïve Bayes, Sınıflandırma Yöntemi, Veri Analizi, Wisconsin

1. GİRİŞ

Araştırmalara göre dünyada her yıl 2.1 milyon kadın, ülkemizde ise 20 bin kadın meme kanserinden etkilenmektedir. Yaşam boyu her 8 kadından biri meme kanseri riski, her 38 kadından birisi ise meme kanserinden ölme riski ile karşı karşıyadır.[1]

İnsan vücudu, her biri kendine özgü işlevi olan milyonlarca hücreden oluşmaktadır. Vücudumuzdaki sağlıklı hücreler bölünebilme yeteneğine sahiptirler. Yaşamın ilk yıllarında hücreler daha hızlı bölünürken, erişkin yaşlarda bu hız yavaşlar. Fakat hücrelerin bu yetenekleri sınırlıdır, sonsuz bölünemezler. Her hücrenin hayatı boyunca belli bir bölünebilme sayısı vardır. Sağlıklı bir hücre ne kadar bölüneceğini bilir ve gerektiğinde ölmesini de bilir. Normalde vücudun sağlıklı ve düzgün çalışması için hücrelerin büyümesi, bölünmesi ve daha çok hücre üretmesine gereksinim vardır. Bazen buna rağmen süreç doğru yoldan sapar. Yeni hücrelere gerek olmadan hücreler bölünmeye devam eder. Bu hücrelerin düzensiz büyümesi kansere sebep olmaktadır. Bu sayede hücreler kontrolsüz olarak bölünür ve büyürler, tümör denilen anormal doku kütlesini oluştururlar. Her bir tümör kanserli olmamasına rağmen, vücudun normal işleyişini bozan sindirim, sinir ve dolaşım sistemlerini geliştirir ve istila eder.[2]

Tümörler iyi huylu veya kötü huylu olabilirler. İyi huylu tümörler kanser değildir. Bunlar sıklıkla alınırlar ve çoğu zaman tekrarlamazlar. İyi huylu tümörlerdeki hücreler vücudun diğer taraflarına yayılmazlar. En önemlisi iyi huylu tümörler nadiren hayatı tehdit ederler. Kötü huylu tümörler kanserdir. Bu tümörler normal dokuları sıkıştırabilirler, içine sızabilirler ya da tahrip edebilirler. Eğer kanser hücreleri oluştukları tümörden ayrılırsa, kan ya da lenf dolaşımı aracılığı ile vücudun diğer bölgelerine gidebilirler. Gittikleri yerlerde tümör kolonileri oluşturur ve büyümeye devam ederler. Kanserin bu şekilde vücudun diğer bölgelerine yayılması olayına metastaz adı verilir.[3]

Meme kanseri meme hücrelerinde başlayan kanser türüdür. Akciğer kanserinden sonar dünyada görülme sıklığı en yüksek olan kanser türüdür. Erkeklerde de görülmekler beraber kadın vakaları erkek vakalarından 100 kat fazladır. 1970'lerden bu yana meme kanserinin görülme sıklığında artış yaşanmaktadır ve bu artışa modern, Batılı yaşam tarsi sebep olarak gösterilmektedir. Kuzey Amerika ve Avrupa ülkelerinde görülme sıklığı, dünyanın diğer bölgelerinde görülme sıklığından daha fazladır.

Meme kanseri, yayılmadan önce, erken tespit edilirse, hasta %96 yaşam şansına sahiptir. Her yıl 44000'de bir kadın meme kanserinden ölmektedir. Meme kanserine karşı en iyi koruyucu yöntem erken teşhistir.

Memedeki iyi huylu veya kötü huylu olduğu kesin anlamanın tek yolu vardır. Biyopsi ile mikroskobik tetkik sonucu tanı koymak. Ama bazı özellikler var ki, o kitlenin daha çok neye benzediği konusunda muayene eden hekime ortalama bir fikir verebilir.[4]

2. METODOLOJİ

A. *BAYES TEOREMİ*

Bayes teoremi, olasılık kuramı içinde incelenen önemli bir konudur. Bu teorem bir rastgele değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi gösterir. Bu kavram için **Bayes Kuralı**, **Bayes Savı** veya **Bayes Kanunu** adları da kullanılır.

Bayes Teoreminde B olasılığının gerçekleşme durumu altında A olasılığının gerçekleşme durumu Denklem 2.1 ile açıklanabilir.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Denklem 2.1 Bayes Teoremi

P(A) ve P(B); A ve B olaylarının marjinal olasılıklarıdır.

Burada önsel olasılık Bayes teoremine öznellik katar. Diğer bir ifadeyle örneğin P(A) henüz elde veri toplanmadan A olayı hakkında sahip olunan bilgidir. Diğer taraftan P(B|A) ardıl olasılıktır çünkü veri toplandıktan sonra, A olayının gerçekleşmiş olduğu durumlarda B olayının gerçekleşme ihtimali hakkında bilgi verir.[5]

B. NAÏVE BAYES SINIFLANDIRICISI

Naïve Bayes sınıflandırıcıları, Bayes'in Teoremine dayalı bir sınıflandırma algoritmaları koleksiyonudur. Tek bir algoritma değil, hepsinin ortak bir ilkeyi paylaştığı bir algoritma ailesidir, yani sınıflandırılan her özellik çifti birbirinden bağımsızdır.

Her bir niteliği ve sınıf etiketini rastgele değişkenler olarak düşünün. Öznitelikleri kümesi verilen bir kayıt verildiğinde: amaç, elinizdeki verilere dayanarak Y sınıfını tahmin etmektir.

2.A içerisinde verilen Bayes Teoreminden yola çıkılarak elde edilen amaç fonksiyonu Denklem 2.2 ile ifade edilir.

$$P(Y|X_1, X_2, \dots, X_d) = \frac{P(X_1, X_2, \dots, X_d|Y)P(Y)}{P(X_1, X_2, \dots, X_d)}$$

Denklem 2.2 Amaç fonksiyonu

Burada X_i nitelikleri arasındaki bağımlılıkların olmadığı varsayılarak işlem yapılır (Denklem 2.3).

$$P(X_1, X_2, \dots, X_d|Y_j) = P(X_1|Y_j)P(X_2|Y_j) \dots P(X_d|Y_j)$$

Denklem 2.3 Bayes Açılımı

Sınıflandırma sütunu için olasılık değeri Denklem 2.4 ile hesaplanır.

$$P(X) = N_X/N$$

Denklem 2.4 Sınıflandırma sütunu olasılık hesabı formülü

Kategorik öznitelikler için olasılık değeri Denklem 2.5 ile hesaplanır.

$$P(X_i|Y_k) = X_{ik}/N_{Xk}$$

Denklem 2.4 Kategorik öznitelikle için olasılık hesabı formülü

C. MEME KANSERİ VERİ SETİ

Bu projede ABD'nin Wisconsin eyaletinde teşhis edilen meme kanseri hastalarının UCI(University of California, Irvine) üzerinde yayınlanan veri seti kullanılmıştır.

Ayrıca veri seti, California Üniversitesi-Irvine'de bulunan Makine Öğrenmesi Deposunda (Machine Learning Repository) çevrimiçi olarak mevcuttur. Yarıçap, doku, çevre uzunluğu, alan, pürüzsüzlük, kompaktlık, konkavlık, konkav noktalar, simetri ve fraktal boyut olmak üzere 10 gerçek değerli nitelik vardır.[6]

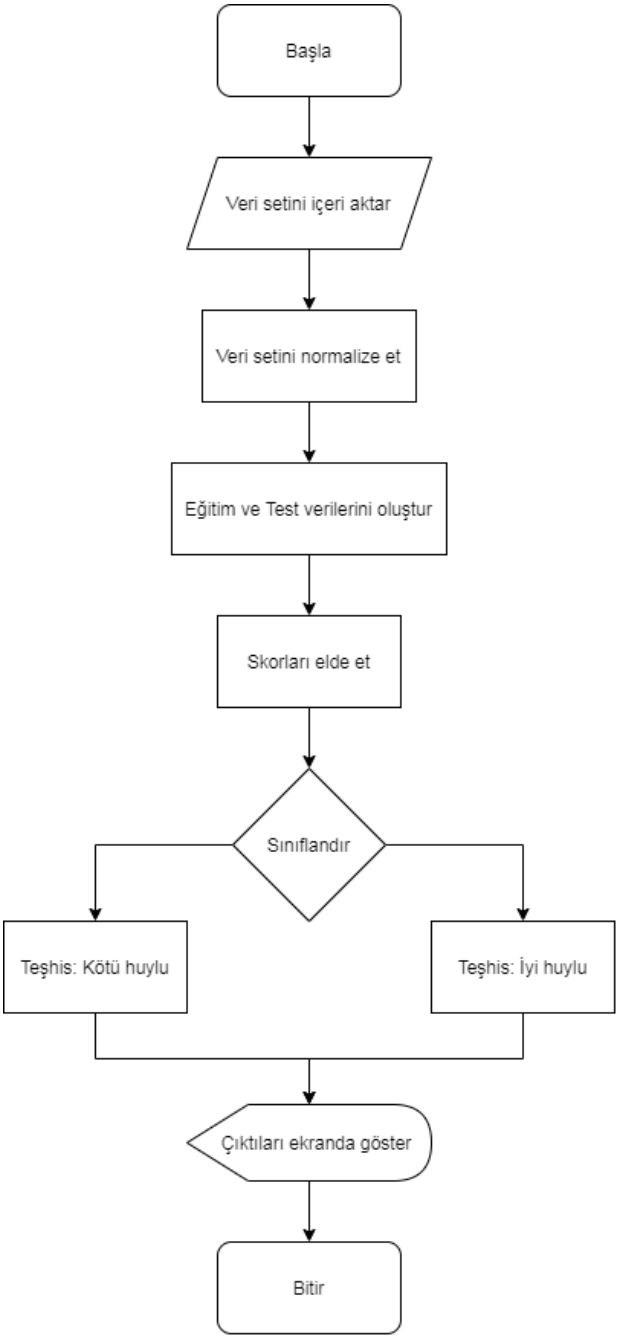
Öznitelik isimleri ve anlamları Çizelge 2.1 ile sunulmuştur.

Öznitelik Adı	**Öznitelik Anlamı**
id	Benzersiz ayraç
diagnosis	Tümör teşhisi (B= Kötü huylu, M= İyi huylu)
radius_mean	Merkezden çevre üzerindeki noktalara olan mesafelerin ortalaması
texture_mean	Gri-skala değerinin standart sapması
perimeter_mean	Çekirdek tümörün ortalama boyutu
area_mean	Tümörün yayıldığı alanın ortalaması
smoothness_mean	Yarıçap uzunluklarındaki yerel değişimin ortalaması
compactness_mean	Perimeter ve area değerlerine bağlı yoğunluk oranlarının ortalaması
concavity_mean	Konturun içbükey kısımlarının ortalama şiddeti
concave_points_mean	Konturun içbükey kısımlarının sayısının ortalaması
symmetry_mean	## Bilinmiyor
fractal_dimension_mean	"Kıyı şeridi yaklaşımı için ortalama" - 1
radius_se	Merkezden çevre üzerindeki noktalara olan mesafelerinin standart hatası
texture_se	Gri-skala değerinin standart sapmasının standart hatası
perimeter_se	## Bilinmiyor
area_se	## Bilinmiyor
smoothness_se	Yarıçap uzunluklarındaki yerel standart hatası
compactness_se	Perimeter ve area değerlerine bağlı yoğunluk oranlarının standart hatası
concavity_se	Konturun içbükey kısımlarının ciddiyeti için standart hata
concave_points_se	Konturun içbükey kısımlarının sayısı için standart hata
symmetry_se	## Bilinmiyor
fractal_dimension_se	"Kıyı şeridi yaklaşımı için standart hata" - 1
radius_worst	Merkezden çevre üzerindeki noktalara olan mesafeleri için "en kötü" veya en büyük ortalama değer
texture_worst	Gris kala değerinin standart sapması için "en kötü" veya en büyük ortalama değer
perimeter_worst	## Bilinmiyor
area_worst	## Bilinmiyor
smoothness_worst	Yarıçap uzunluklarındaki yerel değişim için "en kötü" veya en büyük ortalama değer
compactness_worst	Perimeter ve area değerlerine bağlı "en kötü" veya en büyük ortalama değer
concavity_worst	Konturun içbükey kısımlarının ciddiyeti için "en kötü" veya en büyük ortalama değer
concave_points_worst	Konturun içbükey kısımlarının sayısı için "en kötü" veya en büyük ortalama değer
symmetry_worst	## Bilinmiyor
fractal_dimension_worst	"Kıyı şeridi yaklaşımı için en kötü veya en büyük ortalama değer" - 1

Veri seti bilgileri:

- Veri seti karakteristiği çok değişkenlidir
- Öznitelikler gerçek sayılardan oluşmuştur
- 569 kayıt bulunmaktadır (357 kayıt iyi huylu, 212 kayıt kötü huyludur.)
- 32 adet öznitelik bulunmaktadır
- Kayıtlarda eksik veriler bulunmamaktadır

Verilerin sınıflandırılması için kullanılan algoritma Şekil 2.1'de bulunan akış diyagramı ile sunulmuştur.



Şekil 2.1 Akış diyagramı

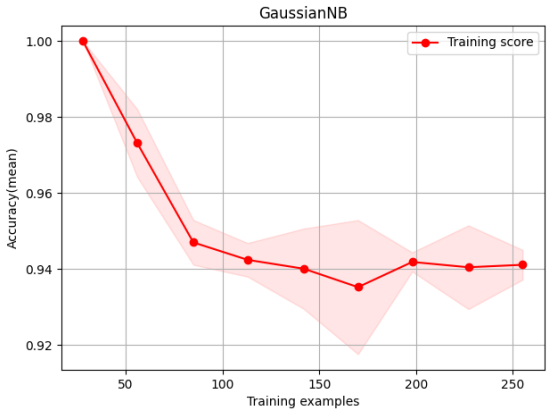
Veri seti üzerindeki tüm işlemler 4GB RAM ve AMD Phenom II X2 570 işlemci ve UBUNTU 20.04 LTS sürümlü bir işletim sistemine sahip makinede gerçekleştirilmiştir.

3. SONUÇ

Geliştirilen model Bayes teoreminden elde edilen sonuçlara göre iyi huylu (B) veya kötü huylu(M) tümör teşhisinde bulunur. Bu veri setinde öznitelikler arasında bağıntı vardır ve bu bağıntılar kullanılan sınıflandırıcıda yok sayılarak işlemler gerçekleştirilmiştir. Tüm verilerin %70'lik kısmı eğitim seti olarak, %30'luk kısmı ise test seti olarak ayrılmıştır.

Modeli eğitim seti ile eğittikten sonra elde edilen doğruluk değeri **% 95.321** olarak bulunmuştur.

Model sırasıyla 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 oranlarında ayrılan eğitim verileriyle eğitilerek elde edilen doğruluk değerleriyle bir öğrenme eğrisi oluşturulmuştur. Oluşturulan eğri Şekil 2.2 ile sunulmuştur.



Şekil 2.2 Modelin Öğrenme Eğrisi

Veri seti üzerinde 10-fold çapraz doğrulama tekniği uygulanmış ve elde edilen değerler:

- Doğruluk-Accuracy (mean): % 93.717
- Standart Sapma: % 3.403

olarak hesaplanmıştır.

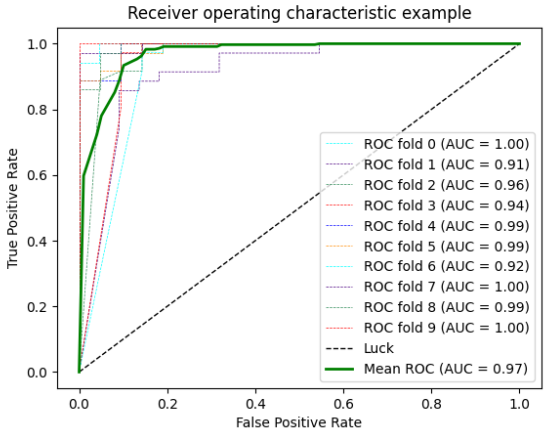
Model için 10-fold tekniği ile ROC eğrileri çizdirilmiş ve eğrilerin altında kalan alanlar sırasıyla: [0.997, 0.913, 0.965, 0.942, 0.989, 0.988, 0.921, 0.996, 0.921, 1] olarak bulunmuştur.

Eğri altında kalan alanların ortalama değeri ile standart sapması:

- AUC (mean): % 96.761
- Standart Sapma: % 3.156

olarak hesaplanmıştır.

Grafiğe dökülen AUC değerleri Şekil 2.3 ile sunulmuştur.



Şekil 2.3 ROC Eğrileri

Modelin test verileri üzerinde hesaplanan karışıklık (confusion) matrisi:

- True Positive(TP): 60
- False Negative(FN): 5
- False Positive(FP): 3
- True Negative(TN): 103

şeklinde hesaplanmıştır.

Modelin aynı veri seti ve aynı modelleme yöntemi kullanılarak 2019 yılında yapılan bir çalışma[7] ile sonuçlarının kıyaslanması Çizelge 3.1 ile sunulmuştur.

Özellik	Geçerli Makale	Diğer Akademik Makale
Accuracy	**% 95.321**	% 94.751
10-Fold Acc.	**% 93.677**	% 93.435
10-Fold Std	**% 2.844**	% 5.031
AUC Mean	% 96.761	% **98.7**
AUC Std	% 3.156	% **1.4**
Precision	**% 95.238**	% 93
Recall	% 92.307	**% 98**
F1	% 93.75	**% 95**
Sensitivity	% 92.307	**% 98**
Specificity	**% 97.169**	% 87

Çizelge 3.1 Verileri Karşılaştırma
(Specificity ve sensivity değerleri makalede özel olarak belirtilmemiş olup karışıklık matrisinden hesaplanmıştır. Kalın olarak belirtilen veriler daha kabul edilebilir olduğu anlamına gelir.)

4. KAYNAKÇA

[1] https://www.ntv.com.tr/saglik/meme-kanseri-dunyada-her-yil-2-1-milyon-kadini-etkiliyor,hpmaGtGjJk6o0VEcL6YSEg

[2] https://www.researchgate.net/publication/272863357_Diagnosis_of_Breast_Cancer_using_Decision_Tree_Data_Mining_Technique

[3] https://hsgm.saglik.gov.tr/tr/kanser-nedir-belirtileri.html

[4] https://tr.wikipedia.org/wiki/Meme_kanseri

[5] https://tr.wikipedia.org/wiki/Bayes_teoremi

[6] https://www.researchgate.net/publication/335444489_Meme_Kanseri_Teshisi_Icin_Yeni_Bir_Skor_Fuzyon_Yaklasimi

[7] https://www.researchgate.net/publication/335079111_Breast_Cancer_diagnosis_using_machine_learning_classification_methods_using_Hadoop