BURDUR MEHMET AKİF ERSOY ÜNİVERSİTESİ

Gölhisar Uygulamalı Bilimler Yüksekokulu

NESNE TABANLI PROGRAMLAMA II DERSİ

PROJE KONUSU: Kanser Teşhisi ve Sınıflandırması

Öğrenci Ad-Soyad

1- Ömer Avcı
2- Buğra Karaahmetoğlu
3- Yunus Emre Çağan
4- Semai Miraç Arıcı
5- Sümeyye Üstünsoy

MAYIS 2024

BURDUR

.

1. GİRİŞ

Bu proje, meme kanseri tanısı için Naïve Bayes sınıflandırıcılarının uygulanmasını ve karşılaştırılmasını içermektedir. Kullanılan veri kümesi, meme kanseri teşhisi konulan hastalara ait tıbbi bilgileri içermektedir. Bu veri kümesi, meme kanserinin benign (iyi huylu) ve malignant (kötü huylu) olmak üzere iki sınıfa ayrılmış durumdadır. Veri seti, her hastaya ait çeşitli tıbbi parametreleri ve bu parametrelerin kanser teşhisindeki önemini içermektedir.

Naive Bayes sınıflandırıcıları, olasılık teorisine dayanan basit ama etkili yöntemlerdir. Bu projede, Bernoulli Naive Bayes, Gaussian Naive Bayes ve Multinomial Naive Bayes sınıflandırıcıları kullanılmıştır. Her bir sınıflandırıcı, veri setinin farklı özelliklerini ele alarak model oluşturmakta ve bu modellerin performansları karşılaştırılmaktadır.

Veri kümesi, makine öğrenimi modellerinin eğitim ve test süreçlerinde kullanılmak üzere train-test split yöntemi ile ayrılmıştır. Eğitim ve test verilerinin ayrılması, modelin genelleme yeteneğini değerlendirmek için kritik öneme sahiptir. Bu projede, veri setinin %80'i eğitim için, %20'si ise test için kullanılmıştır.

Bu raporda, her bir sınıflandırıcının performansı, doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru gibi metriklerle değerlendirilmiştir. Ayrıca, her sınıflandırıcının sonuçları, karışıklık matrisleri (confusion matrices) ve bu matrislerin görselleştirilmesi ile detaylandırılmıştır.

2. GEREÇ VE YÖNTEM

Naive Bayes sınıflandırıcıları, Bayes teoremi üzerine kurulu olup, özellikle metin sınıflandırma ve tıbbi teşhis gibi alanlarda yaygın olarak kullanılmaktadır. Naive Bayes algoritmaları, verilerin bağımsız özellikler varsayımı altında basit ama etkili sınıflandırma modelleri oluşturur.

Naive Bayes, Thomas Bayes'in 18. yüzyılda geliştirdiği olasılık teorisine dayanır. 1763 yılında ölümünden sonra yayımlanan Bayes'in Teoremi, bir olayın gerçekleşme olasılığını önceki bilgileri kullanarak hesaplamayı mümkün kıldı. 20. yüzyılın ortalarında, özellikle 1950'lerde ve 1960'larda, Bayes'in Teoremi'ni temel alan istatistiksel yöntemler, özellikle makine öğrenimi ve bilgi teorisi alanlarında önem kazandı.

Bernoulli Naive Bayes

Bernoulli Naive Bayes, ikili (binary) özelliklerle çalışmak üzere tasarlanmıştır. Bu yöntem, her özelliğin var olup olmamasını değerlendirir. Bernoulli Naive Bayes, özellikle metin sınıflandırma ve belge tanıma gibi alanlarda yaygın olarak kullanılmaktadır. Bir veri noktasının y sınıfına ait olma olasılığı şu şekilde hesaplanır:

$$P(y|x) \propto P(y) \prod_{i=1}^{n} P(x_i|y)^{x_i}(1 - P(x_i|y))^{(1-x_i)}$$

P(y): Sınıf y'nin önsel olasılığıdır. Her sınıfın veri setindeki oranı hesaplanır.

xi: Özellik vektörünün i'inci bileşenidir. Bu bileşen 0 veya 1 olabilir.

$P(xi|y)$: Özellik xi'nin sınıf y 'ye koşullu olasılığıdır. Her özellik için sınıf koşullu olasılıkları hesaplanır. Örneğin, sınıf y verildiğinde xi özelliğinin 1 olma olasılığı.

Gaussian Naive Bayes

Gaussian Naive Bayes, özelliklerin normal dağılıma sahip olduğu varsayımı ile çalışır. Bu yöntem, sürekli verilerle ve çalışmak için uygundur tıbbi teşhis, biyolojik veri analizi gibi alanlarda kullanılmaktadır. Gaussian Naive Bayes, verilerin ortalama ve standart sapmalarını kullanarak sınıflandırma yapar. Bir veri noktasının y sınıfına ait olma olasılığı şu şekilde hesaplanır:

$$P(y|x) \propto P(y) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{y,i}^2}} \exp\left(-\frac{(x_i-\mu_{y,i})^2}{2\sigma_{y,i}^2}\right)$$

P(*y*): Sınıf y'nin önsel olasılığıdır.
xi: Özellik vektörünün i'inci bileşenidir. Bu bileşen pozitif bir tamsayı olabilir.
$P(xi|y)$: Özellik xi'nin sınıf y 'ye koşullu olasılığıdır.

Multinomial Naive Bayes

Multinomial Naive Bayes, özellikle sayısal veri setleri ve sayım verileri ile çalışmak için uygundur. Bu yöntem, özellikle belge sınıflandırma ve doğal dil işleme gibi alanlarda yaygın olarak kullanılmaktadır. Multinomial Naive Bayes, özelliklerin frekanslarına dayanarak sınıflandırma yapar.

$$P(y|x) \propto P(y) \prod_{i=1}^{n} P(x_i|y)^{x_i}$$

P(*y*): Sınıf y 'nin önsel olasılığıdır.
xi: Özellik vektörünün i 'inci bileşenidir. Bu bileşen pozitif bir tamsayı olabilir.
$P(xi|y)$: Özellik xi 'nin sınıf y 'ye koşullu olasılığıdır.

2.1 Veri Seti Tanımı

Tablo 1. Örnek tablo Göğüs kanseri verisine ait olan tablodan ilk 7 sütunundan bir kesittir.

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	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2

Bu tablo, göğüs kanseri verisine ait özellikleri ve değerlerini içermektedir. Tabloda, farklı hastalara ait örnek numaraları ve bu numaralara karşılık gelen hücre özelliklerinin değerleri bulunmaktadır. Hücre özellikleri arasında hücre büyüklüğünün düzensizliği, hücre şeklinin düzensizliği, marjinal adezyon, tek epitel hücre boyutu, çıplak çekirdekler, sönük kromatin, normal nukleoller ve mitozlar yer almaktadır. Her bir özellik için numerik değerler verilmiştir.

Veri setinin sütunları şu şekildedir:
Sample code number: Hastaya ait örnek kod numarası
Clump Thickness: Hücre yığılma kalınlığı
Uniformity of Cell Size: Hücre boyutunun düzenliliği
Uniformity of Cell Shape: Hücre şeklinin düzenliliği
Marginal Adhesion: Kenar yapışkanlığı
Single Epithelial Cell Size: Tek epitel hücre boyutu
Bare Nuclei: Çıplak çekirdek
Bland Chromatin: Düz kromatin
Normal Nucleoli: Normal çekirdekçikler
Mitoses: Mitoz sayısı
Class: Kanser sınıfı (benign: iyi huylu, malignant: kötü huylu)

2.2 Karşılaştırma

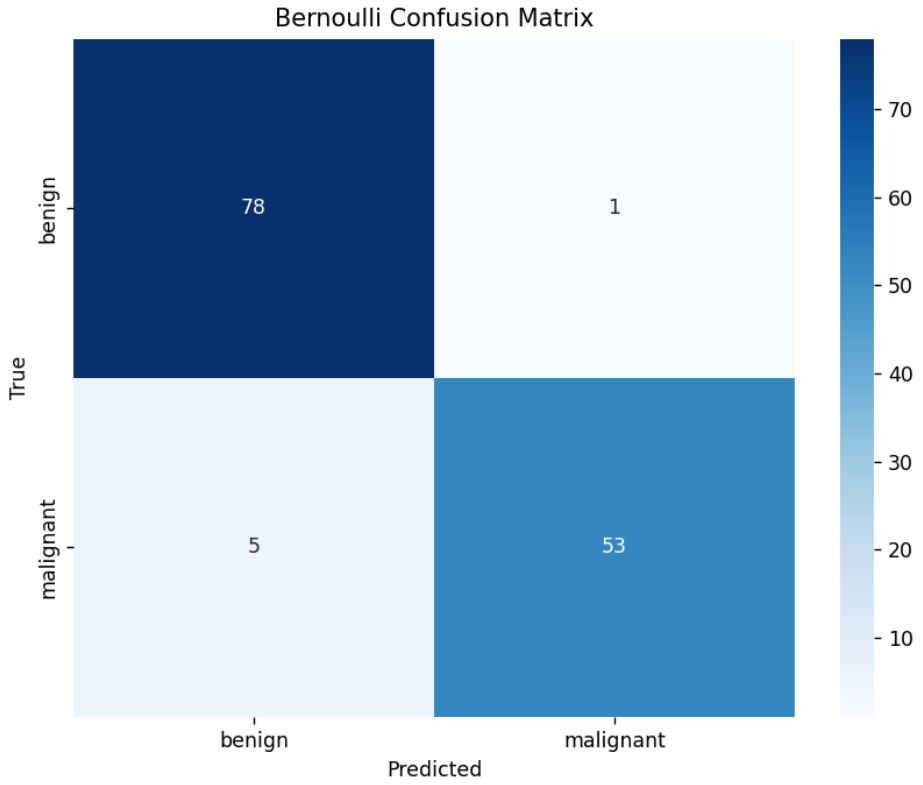
Referanslar	**Sınıflandırıcılar (Classifiers)**	**Doğruluk (Accuracy)**
Lavanya ve Rani [1]	CART + Exhaustive (FST)	95.13%
Chaurasia ve Pal [2]	Simple Logistic	74.5%
Asri ve sınıf arkadaşları [3]	SVM	97.13%
Amrane ve sınıf arkadaşları [4]	kNN	97.51%
Sarkar ve Nag [5]	C4.5	96.71%
Houfani ve sınıf arkadaşları [6]	MLP & LR	97.9%
Sumbaly ve sınıf arkadaşları [7]	J48	94.36%
Sulyman Age Abdulkareema [8]	XGBoost + RFE	99.02%
Bizim Araştırmamız	Bernoulli NB	95.62%
Bizim Araştırmamız	Multinomial NB	95.62%
Bizim Araştırmamız	Gaussian NB	88.32%

Tablo 2: Bu veri setiyle daha önce yapılmış çalışmalar.

3. BULGULAR

Bu bölümde, Naive Bayes sınıflandırıcıları ile elde edilen sonuçlar, performans metrikleri ve görselleştirmeler sunulmuştur. Her sınıflandırıcı için doğruluk, kesinlik, duyarlılık ve F1 skoru gibi metrikler hesaplanmış ve karşılaştırılmıştır. Ayrıca, karışıklık matrisleri kullanılarak her modelin sınıflandırma performansı detaylandırılmıştır.

3.1 Bernoulli Sınıflandırıcısı

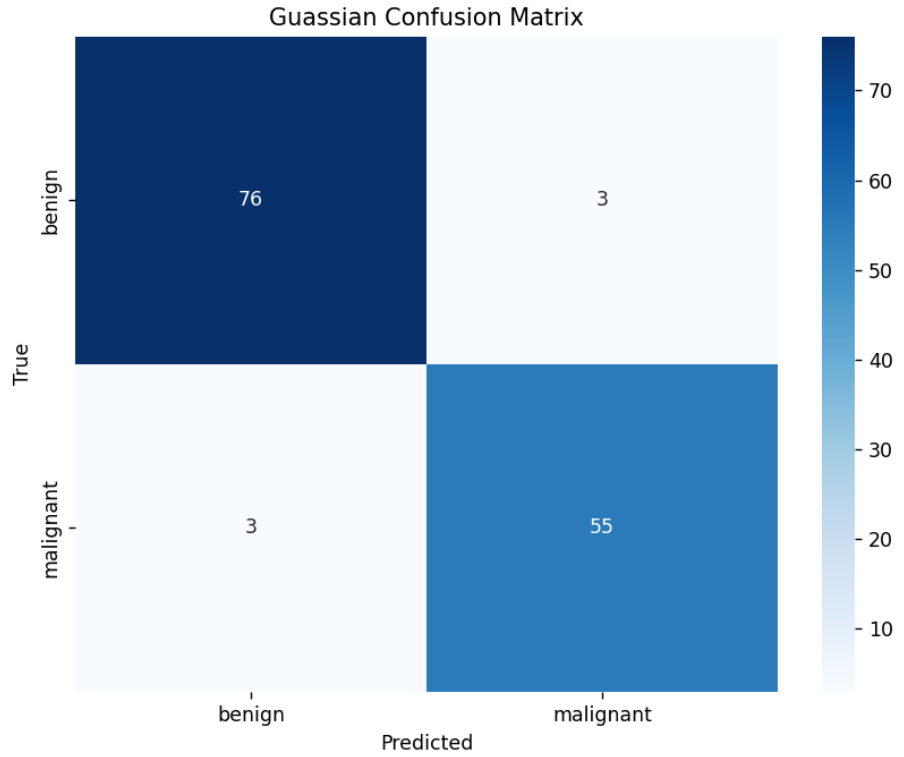


(1) Şekil 1 Bernoulli

Şekil (1): Bu görselde Bernoulli matrisi iyi huylu ve kötü huylu durumları tahmin etmek için kullanılmıştır. Sol üst kare, koyu mavi renkte ve içinde “78” yazıyor. Bu, benign durumlar için 78 doğru pozitif tahminini temsil ediyor. Sağ üst kare, açık mavi renkte ve içinde “1” yazıyor. Bu, malign durumlar için yanlış pozitif tahminini temsil ediyor. Sol alt kare, açık mavi renkte ve içinde "5" yazıyor. Bu, malign durumlar için yanlış negatif tahminleri gösteriyor. Sağ alt kare, mavi renkte ve içinde "53" yazıyor. Bu, malign durumlar için doğru pozitif tahminini temsil ediyor.

Accuracy : 0.9562043795620438
Precision : 0.9814814814814815
Recall : 0.9137931034482759
F1 Score : 0.9464285714285714

3.2 Gaussian Sınıflandırıcısı

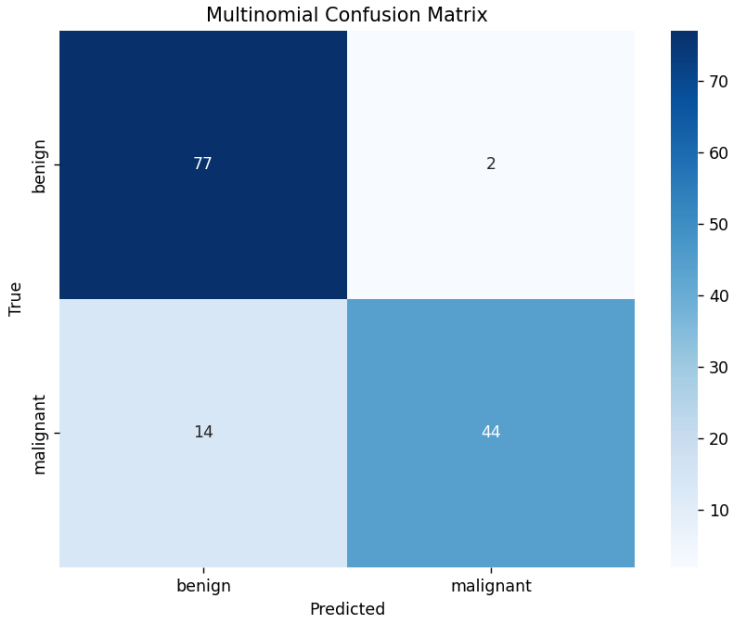


(2) Şekil 2: Gaussian

Şekil 2: Bu görselde Gaussian matrisi iyi huylu ve kötü huylu durumları tahmin etmek için kullanılmıştır. Sol üst kare, koyu mavi renkte ve içinde “76” yazıyor. Bu, benign durumlar için 76 doğru pozitif tahminini temsil ediyor. Sağ üst kare, açık mavi renkte ve içinde “3” yazıyor. Bu, malign durumlar için yanlış pozitif tahminini temsil ediyor. Sol alt kare, açık mavi renkte ve içinde "3" yazıyor. Bu, malign durumlar için yanlış negatif tahminleri gösteriyor. Sağ alt kare, mavi renkte ve içinde "55" yazıyor. Bu, malign durumlar için doğru pozitif tahminini temsil ediyor.

Accuracy : 0.9562043795620438
Precision : 0.9482758620689655
Recall : 0.9482758620689655
F1 Score : 0.9482758620689655

3.3 Multinomial Sınıflandırıcısı



(3) Şekil 3: Multinomial

Bu görselde Multinomial matrisi iyi huylu ve kötü huylu durumları tahmin etmek için kullanılmıştır. Sol üst kare, koyu mavi renkte ve içinde “77” yazıyor. Bu, benign durumlar için 77 doğru pozitif tahminini temsil ediyor. Sağ üst kare, açık mavi renkte ve içinde “2” yazıyor. Bu, malign durumlar için yanlış pozitif tahminini temsil ediyor. Sol alt kare, açık mavi renkte ve içinde "14" yazıyor. Bu, malign durumlar için yanlış negatif tahminleri gösteriyor. Sağ alt kare, mavi renkte ve içinde "44" yazıyor. Bu, malign durumlar için doğru pozitif tahminini temsil ediyor.

Accuracy : 0.8832116788321168
Precision : 0.9565217391304348
Recall : 0.7586206896551724
F1 Score : 0.8461538461538461

.

REFERANSLAR

[1] D. Lavanya ve K. U. Rani, “Analysis of feature selection with classification: Breast cancer datasets,” Indian Journal of Computer Science and Engineering (IJCSE), vol. 2, no. 5, pp. 756–763, 2011.

[2] V. Chaurasia ve S. Pal, “Data mining techniques: to predict and resolve breast cancer survivability,” International Journal of Computer Science and Mobile Computing IJCSMC, vol. 3, no. 1, pp. 10–22, 2014.

[3] H. Asri, H. Mousannif, H. Al Moatassime, ve T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” Procedia Computer Science, vol. 83, pp. 1064–1069, 2016.

[4] M. Amrane, S. Oukid, I. Gagaoua ve T. Ensarİ, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineering's' Meeting (EBBT), Istanbul, 2018, pp. 1-4, doi: 10.1109/EBBT.2018.8391453.

[5] S. K. Sarkar ve A. Nag, “Identifying patients at risk of breast cancer through decision trees,” International Journal of Advanced Research in Computer Science, vol. 8, no. 8, pp. 88–91, 2017.

[6] D. Houfani, S. Slatnia, O. Kazar, N. Zerhouni, H. Saouli, ve I. Remadna, “Breast cancer classification using machine learning techniques: a comparative study”, Medical Technologies Journal, vol. 4, no. 2, pp. 535–544.

[7] R. Sumbaly, N. Vishnusri, ve S. Jeyalatha, “Diagnosis of breast cancer using decision tree data mining technique,” International Journal of Computer Applications, vol. 98, no. 10, 2014.

[8] S. Age Abdulkareema, Z. Olorunbukademi Abdulkareemb, “An Evaluation of the Wisconsin Breast Cancer Dataset using Ensemble Classifiers and RFE Feature Selection Technique” International Journal of Sciences: Basic and Applied Research (IJSBAR) (2021) Volume 55, No 2, pp 67-80.

KAYNAKLAR

https://scikit-learn.org/stable/modules/naive_bayes.html

https://www.kaggle.com/code/dskagglemt/breast-cancer-using-naive-bayes

https://iq.opengenus.org/gaussian-naive-bayes/

https://iq.opengenus.org/bernoulli-naive-bayes/

https://www.geeksforgeeks.org/multinomial-naive-bayes/

https://core.ac.uk/download/pdf/387567227.pdf

KAYNAK KOD ADRESİ

https://github.com/oomeravcii/CancerPrediction-with-NaiveBayes

https://github.com/Semai-Mirac/Naive_Bayes_Grub2

https://github.com/sumeyyeustunsoy/naivebayes.py

https://github.com/Bugrakaraahmetoglu/CancerPrediction-with-NaiveBayes

https://github.com/ynsmr-cgn/CancerPrediction-with-NaiveBayes