T.C.
ERCİYES ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

2022-2023 YILI GÜZ DÖNEMİ
INTRODUCTION THE PATTERN RECOGNITION DERSİ

FİNAL ÖDEVİ RAPORU

SINIFLANDIRMA ALGORİTMALARI İLE MEME KANSERİ TESPİTİ

Hazırlayan
1030510211
Öznur Hasoğlu
Öğretim Üyesi
Dr. Öğr. Üyesi Özkan Ufuk Nalbantoğlu

ARALIK 2022
KAYSERİ

İÇİNDEKİLER

1. ÖZET

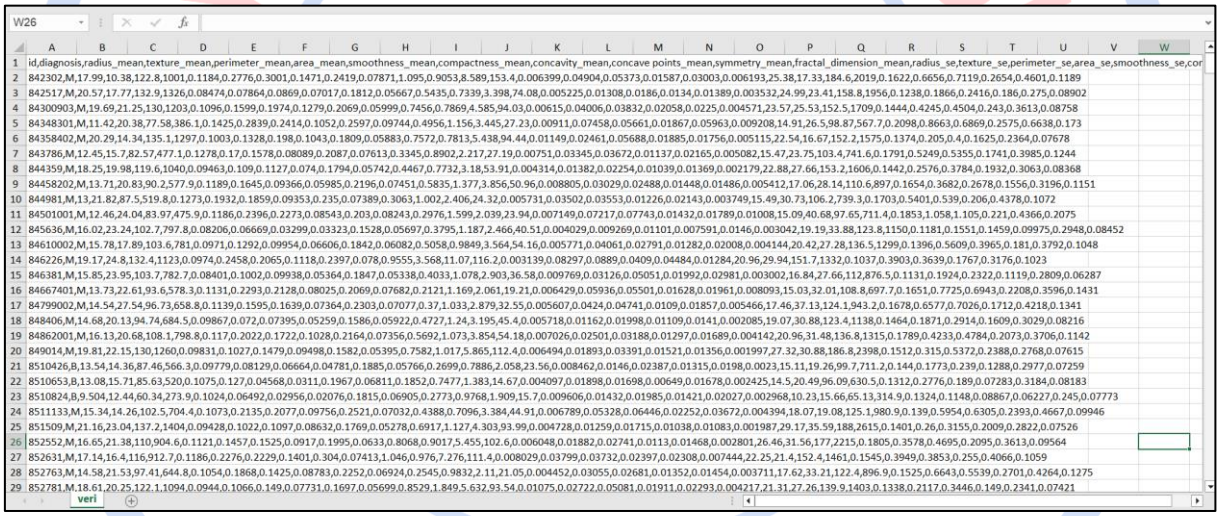
Bu proje Introduction the Pattern Recognition dersi final ödevi kapsamında hazırlanmıştır.

Proje, makine öğrenimi algoritmalarını kullanarak meme bölgesindeki tümör hücrelerinden alınan 32 öznitelikli 570 örnekten oluşan bir veri setiyle, hücreleri iyi huylu (kanserli olmayan) ve kötü huylu (kanserli) olarak sınıflandırma üzerine bir program yazmaktır. Program, Python programlama dili aracılığıyla Spyder geliştirme ortamında yazılmıştır.

Projeyi yaparken öncelikle veri setini düzenledim. Daha sonra verileri Logistic Regression(LGR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), Desicion Tree Classifier (DTC), Random Forest Classifier (RFC) algoritmaları kullanarak eğittim ve sonuçları belirli metrikler üzerinden karşılaştırdım. Özetle sunduğum bu aşamaların detayları raporda ilgili başlıklarda verilmiştir. Çalışır program dosyasını GitHub hesabıma yükleyerek eklerde Github adresimi paylaşmış bulunmaktayım.

2. VERİ SETİ YÜKLEME VE DÜZENLEME

Proje için verilen 570 örnek ve 32 öznitelikten oluşan Şekil 1'de gördüğünüz 'csv' formatındaki veri setini kullandım.

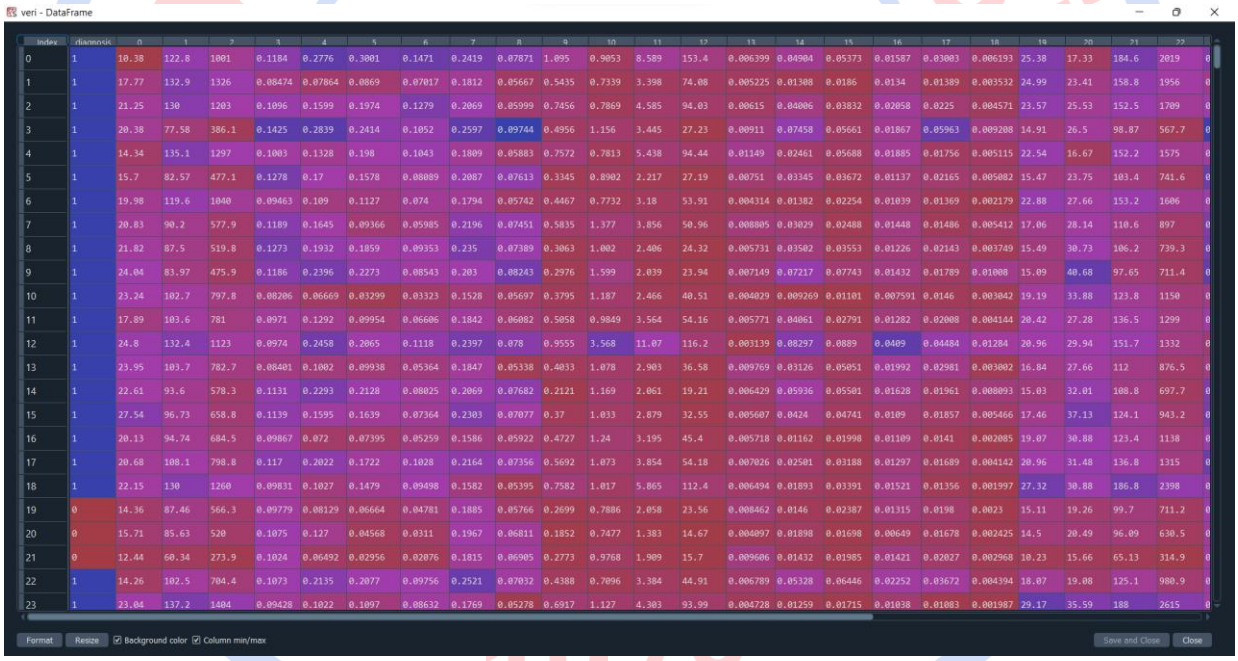


Şekil 1. Veri Seti

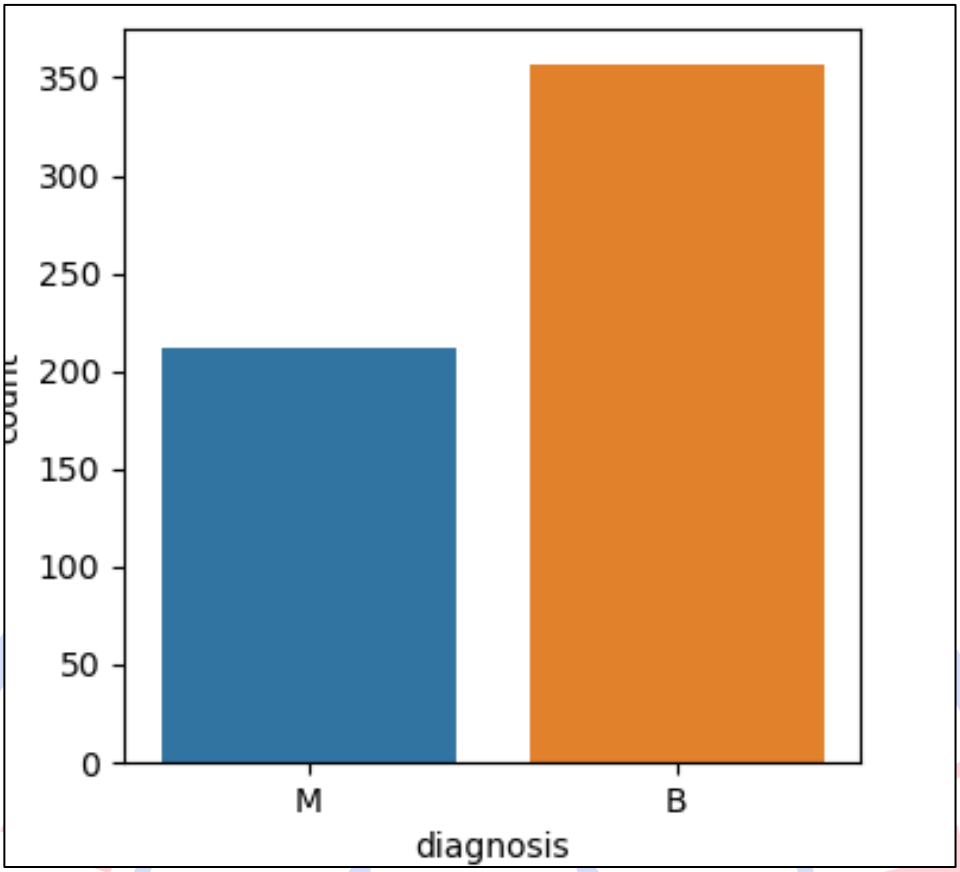
Verideki 1. Öznitelik olan 'id' kolonunu kullanmayacağım için sildim. Daha sonra 2. Öznitelik olan 'diagnosis' kolonunu yani hücrenin maligant (kanserli) veya benign (kansersiz) olduğunu tahmin etmek için M ve B değerlerini sayısallaştırdım. Kalan 30 özniteliği kullanarak 'diagnosis' özniteliğini tahmin etmek üzere programlama yaptım. Bahsettiğim veri düzenleme işlemlerini gerçekleştirdiğim kodlar Şekil 2'de, verinin düzenlenmiş hali Şekil 3'te ve veriler üzerinde yapılmış bazı görselleştirmeler Şekil 4 ve Şekil 5'te verilmiştir.

```python
"""VERİLER YÜKLEDİM, İNCELEDİM, GÖRSELLEŞTİRDİM..."""
veriler = pd.read_csv('veri.csv')
veriler.isna().sum()
veriler.describe()
veriler.info()

#ID KOLONUNU SİLİYORUM
veriler.drop('id',inplace  =True,axis = 1)

#GÖRSELLEŞTİRİYORUM
plt.figure(figsize=(4,4))
sns.countplot(data = veriler,x = 'diagnosis')

sns.pairplot(veriler, hue="diagnosis", vars=["radius_mean", "texture_mean", "perimeter_mean", "radius
plt.show()

#DIAGNOSIS KOLONUMU SAYISALLAŞTIRIYORUM M/B --> 1/0
diagnosis = veriler[["diagnosis"]]
diagnosis = preprocessing.LabelEncoder().fit_transform(diagnosis)

#DIAGNOSIS KOLONUNU TAHMİN ETTİRECEĞİM İÇİN KALAN VERİLERİ AYIRIYORUM
kalan = veriler.iloc[:,2:].values

#DIAGNOSIS KOLONUNU SAYILAŞTIRDIKTAN SONRA TEKRAR KALAN VERİLERLE BİRLEŞTİRİYORUM
bir= pd.DataFrame(data= diagnosis, index= range(569), columns= ["diagnosis"])
iki= pd.DataFrame(data= kalan, index= range(569))
veri=pd.concat([bir,iki], axis=1)
```

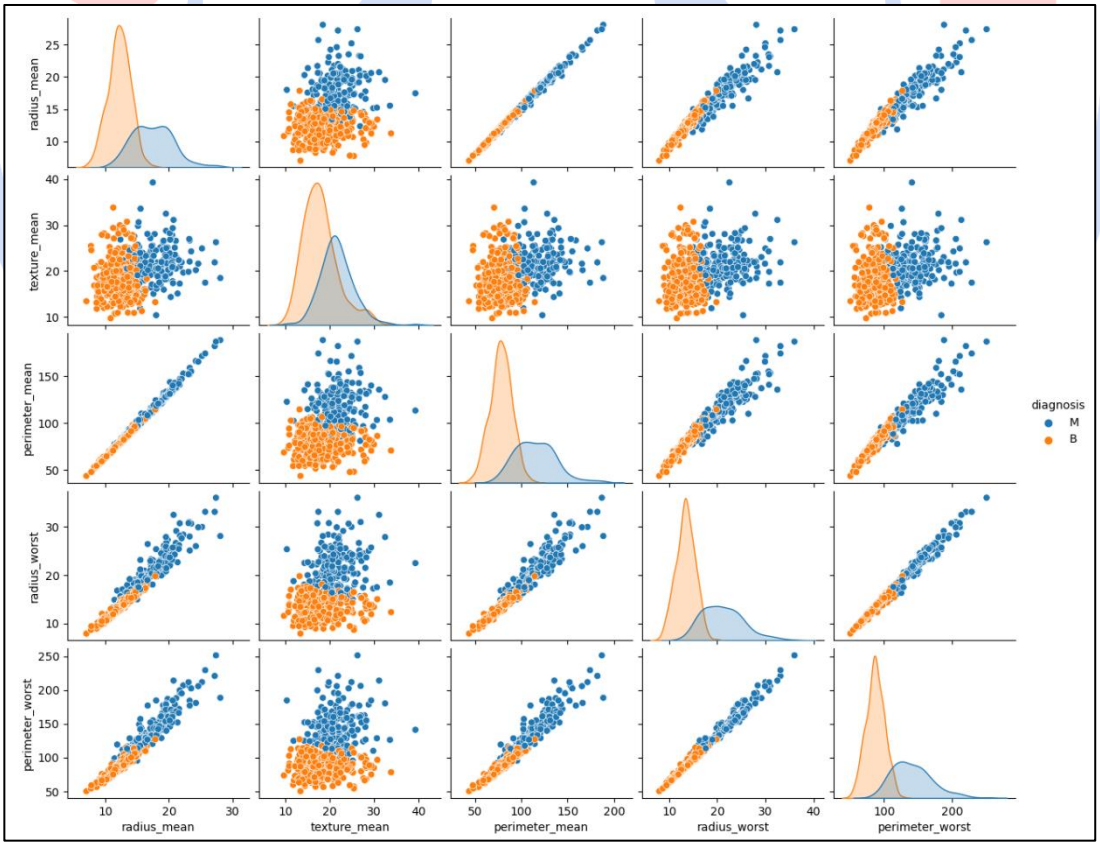
Şekil 2. Veri Yükleme ve Düzenleme



Şekil 3. Düzenleniş Veri



Şekil 4. Veri Sınıf Dağılımı



Şekil 5. Bazı Öznitelikler Arası Pairplot Grafiği

Eğitim ve test verilerini ayırdıktan sonra ölçekleme işlemi uyguladım. (Şekil 6)

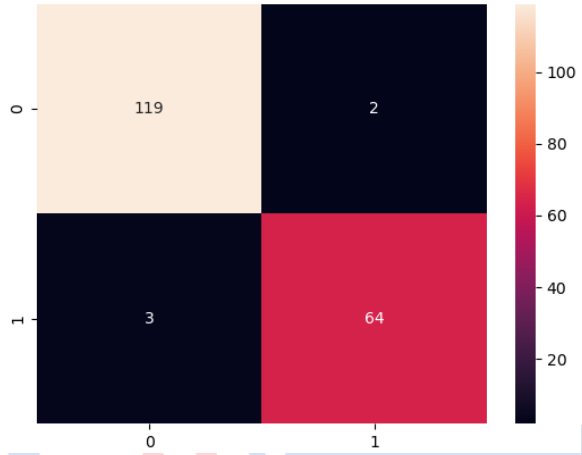
```
38
39  """EĞİTİM/TEST VERİLERİ AYRILIYOR..."""
40  x_train, x_test,y_train,y_test = train_test_split(kalan,diagnosis,test_size=0.33, random_state=0)
41
42  """ÖLÇEKLEME YAPIYORUM..."""
43  sc=StandardScaler()
44  x_train = sc.fit_transform(x_train)
45  x_test = sc.transform(x_test)
```

Şekil 6. Ölçekleme İşlemi

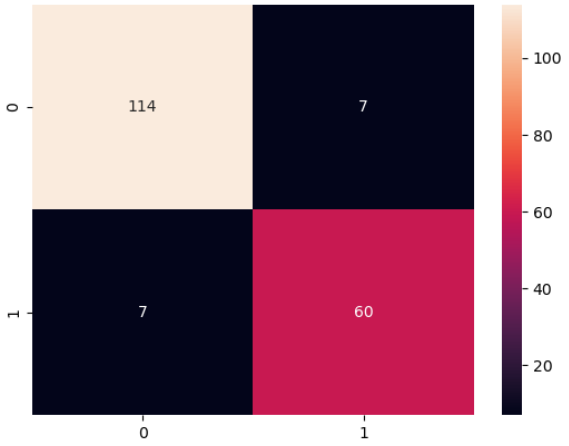
3. ALGORİTMA DEĞERLENDİRMELERİ

Seçtiğim algoritmaları ayrı ayrı denedim. Karmaşıklık matrisi, sınıflandırma raporu ve çapraz doğrulama kullanarak performanslarını karşılaştırdım. Kodları dosyaya ek olarak vereceğim için rapora eklemedim.

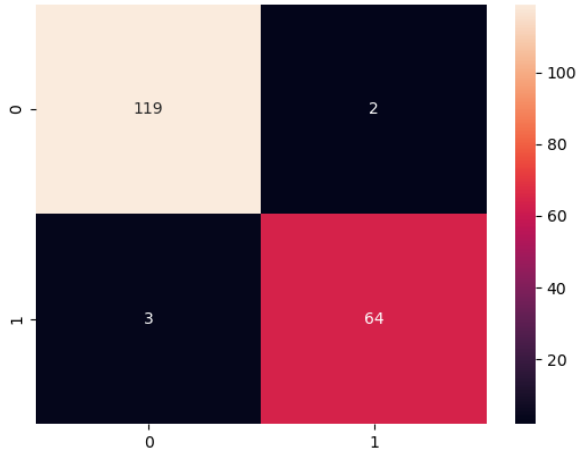
3.1. KARMAŞIKLIK MATRİSLERİ



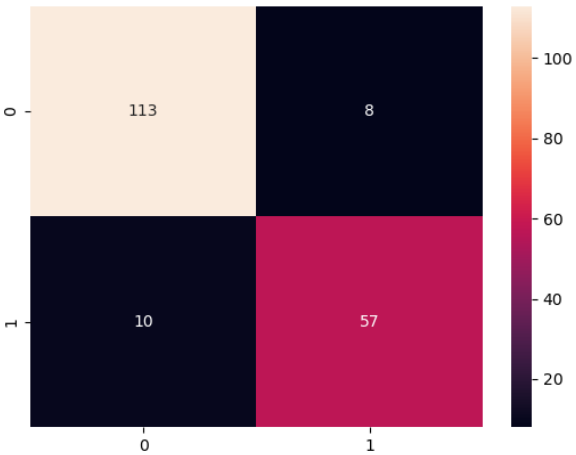
Şekil 7. LGR Karmaşıklık Matrisi



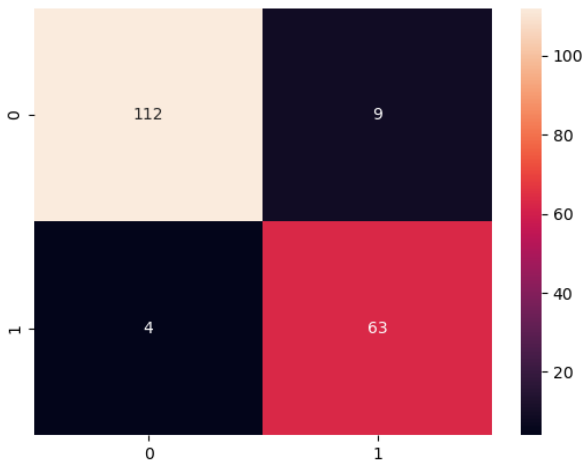
Şekil 7. KNN Karmaşıklık Matrisi



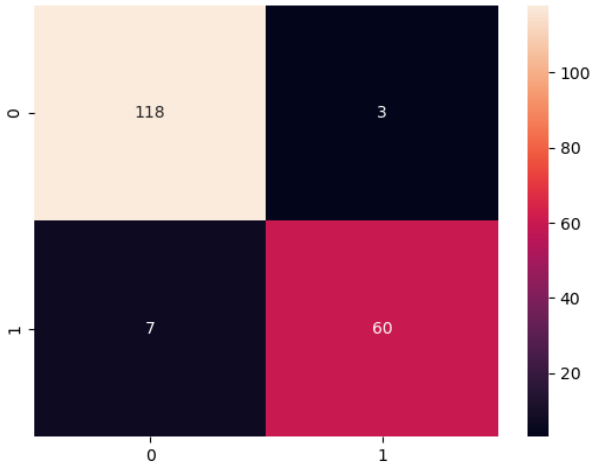
Şekil 8. SVC Karmaşıklık Matrisi



Şekil 9. GNB Karmaşıklık Matrisi



Şekil 10. DTC Karmaşıklık Matrisi



Şekil 11. RFC Karmaşıklık Matrisi

3.2. SINIFLANDIRMA RAPORLARI

LGR	precision	recall	f1-score	support
0	0.98	0.98	0.98	121
1	0.97	0.96	0.96	67
accuracy			0.97	188
macro avg	0.97	0.97	0.97	188
weighted avg	0.97	0.97	0.97	188

Roc eğrisi altındaki alan (AUC): 0.9693474774885902

Şekil 12. LGR Sınıflandırma Raporu

KNN	precision	recall	f1-score	support
0	0.94	0.94	0.94	121
1	0.90	0.90	0.90	67
accuracy			0.93	188
macro avg	0.92	0.92	0.92	188
weighted avg	0.93	0.93	0.93	188

Roc eğrisi altındaki alan (AUC): 0.91883557419514

Şekil 13. KNN Sınıflandırma Raporu

SVC	precision	recall	f1-score	support
0	0.98	0.98	0.98	121
1	0.97	0.96	0.96	67
accuracy			0.97	188
macro avg	0.97	0.97	0.97	188
weighted avg	0.97	0.97	0.97	188

Roc eğrisi altındaki alan (AUC): 0.9693474774885902

Şekil 14. SVC Sınıflandırma Raporu

GNB	precision	recall	f1-score	support
0	0.92	0.93	0.93	121
1	0.88	0.85	0.86	67
accuracy			0.90	188
macro avg	0.90	0.89	0.89	188
weighted avg	0.90	0.90	0.90	188

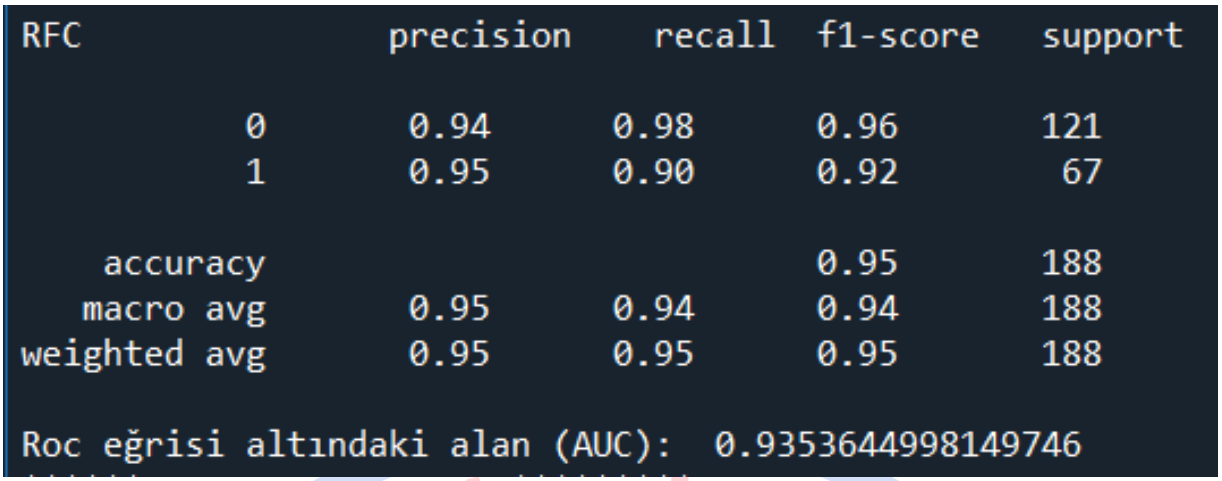
Roc eğrisi altındaki alan (AUC): 0.8923152830886888

Şekil 15. GNB Sınıflandırma Raporu

DTC	precision	recall	f1-score	support
0	0.96	0.92	0.94	121
1	0.86	0.93	0.89	67
accuracy			0.92	188
macro avg	0.91	0.92	0.91	188
weighted avg	0.92	0.92	0.92	188

Roc eğrisi altındaki alan (AUC): 0.9213642531145924

Şekil 16. DTC Sınıflandırma Raporu



RFC	precision	recall	f1-score	support
0	0.94	0.98	0.96	121
1	0.95	0.90	0.92	67
accuracy			0.95	188
macro avg	0.95	0.94	0.94	188
weighted avg	0.95	0.95	0.95	188

Roc eğrisi altındaki alan (AUC): 0.9353644998149746

Şekil 17. RFC Sınıflandırma Raporu

3.3. SONUÇ TABLOSU

	LGR	KNN	SVC	GNB	DTC	RFC
Accuracy	**0.97**	**0.93**	**0.97**	**0.90**	**0.92**	**0.95**
Cross Val.	**0.98**	**0.96**	**0.97**	**0.94**	**0.91**	**0.96**
Cross Val. S.S.	**0.02**	**0.02**	**0.02**	**0.05**	**0.05**	**0.03**
Sensitivity(ReCall)	**0.97**	**0.92**	**0.97**	**0.89**	**0.92**	**0.94**
Specifiticy	**0.96**	**0.90**	**0.96**	**0.85**	**0.93**	**0.90**
AUC	**0.96**	**0.91**	**0.96**	**0.89**	**0.92**	**0.93**

Tablo 1. Değerlendirme Sonuçları

Tabloyu incelediğimde en tutarlı ve uygun modelin küçük bir farkla Logistic Regression algoritması ile eğittiğim model olduğu görülmektedir. Sonuç olarak **Logistic Regression** algoritmasını seçerek **%97 doğruluk** oranında bir model eğitmiş oldum. Detaylı açıklamaları programımda yorum satırlarında görebilirsiniz.

4. EKLER

Hasoğlu, Öznur. "Introduction-to-Pattern-Recognition-Final-Project". *GitHub*. Yayın Tarihi: 20 Ocak 2022

https://github.com/oznurhasoglu/Introduction-to-Pattern-Recognition-Final-Project-

5. KAYNAKLAR

Yasser, M. "Breast Cancer Dataset". *KAGGLE*. Yayın Tarihi: 2022

https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset