

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

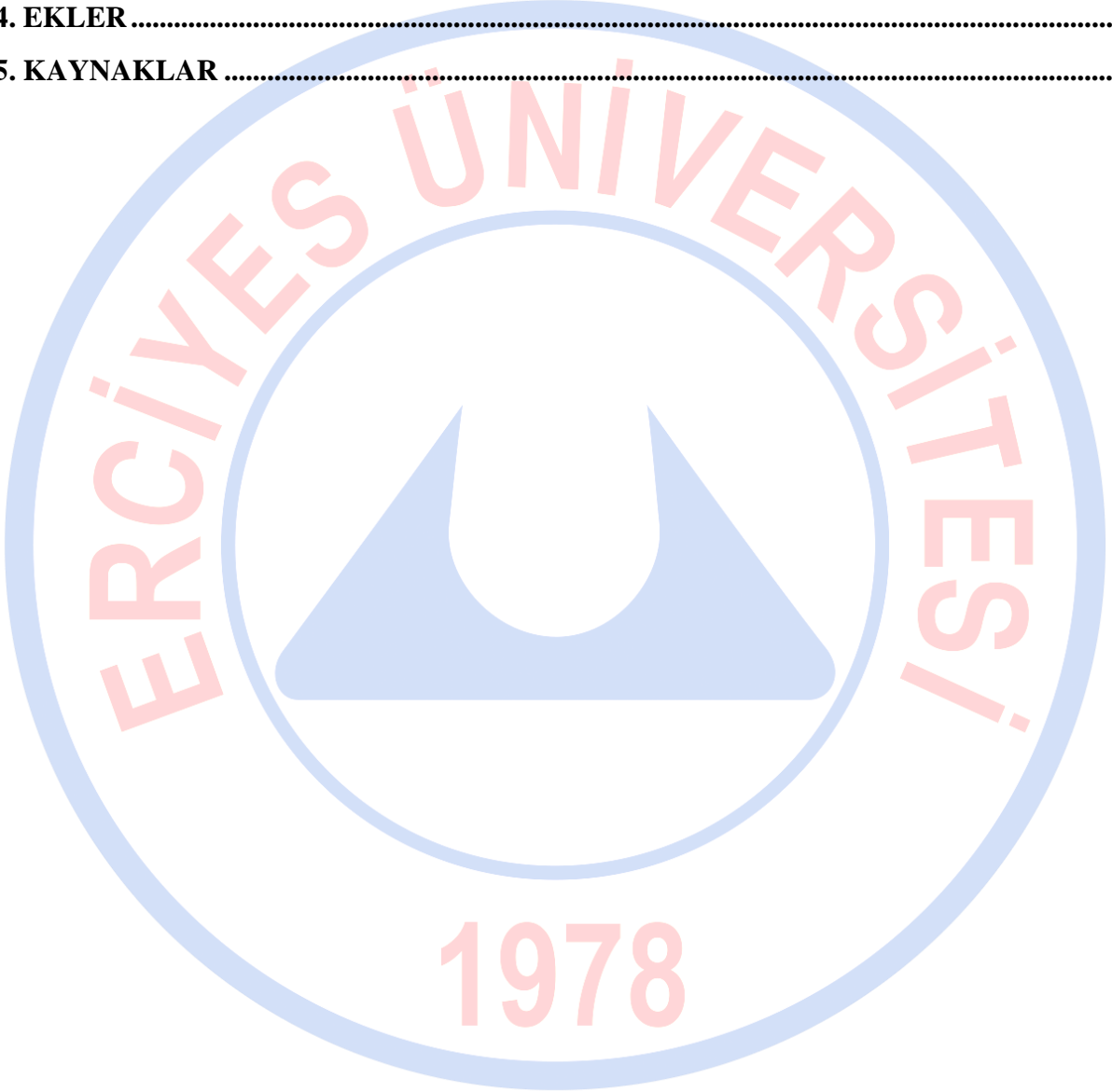
Dr. Öğr. Üyesi Özkan Ufuk Nalbantoğlu

ARALIK 2022

KAYSERİ

İÇİNDEKİLER

1. ÖZET.....	3
2. VERİ SETİ YÜKLEME VE DÜZENLEME	3
3. ALGORİTMA DEĞERLENDİRMELERİ.....	6
3.1. KARMAŞIKLIK MATRİSLERİ.....	6
3.2. SINIFLANDIRMA RAPORLARI.....	7
3.3. SONUÇ TABLOSU	9
4. EKLER	9
5. KAYNAKLAR	9



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 (kansersiz olmayan) ve kötü huylu (kansersiz) 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.

[illegible]

Şekil 1. Veri Seti

Verideki 1. Öznitelik olan ‘id’ kolonunu kullanmayacağım için sildim. Daha sonra 2. Öznitelik olan ‘diagnosis’ kolonunu yani hücrenin malign (kanserli) veya benign (kancersiz) olduğunu tahmin etmek için M ve B değerlerini sayısallaştırdım. Kalan 30 özneliği kullanarak ‘diagnosis’ özneliğini 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.

```

11 """VERİLER YÜKLEDİM, İNCELEDİM, GÖRSELLEŞTİRDİM..."""
12 veriler = pd.read_csv('veri.csv')
13 veriler.isna().sum()
14 veriler.describe()
15 veriler.info()
16
17 #ID KOLONUNU SİLİYORUM
18 veriler.drop('id',inplace =True,axis = 1)
19
20 #GÖRSELLEŞTİRİYORUM
21 plt.figure(figsize=(4,4))
22 sns.countplot(data = veriler,x = 'diagnosis')
23
24 sns.pairplot(veriler, hue="diagnosis", vars=["radius_mean", "texture_mean", "perimeter_mean", "radius
25 plt.show()
26
27 #DIAGNOSIS KOLONUMU SAYISALLAŞTIRIYORUM M/B --> 1/0
28 diagnosis = veriler[["diagnosis"]]
29 diagnosis = preprocessing.LabelEncoder().fit_transform(diagnosis)
30
31 #DIAGNOSIS KOLONUNU TAHMİN ETTİRECEĞİM İÇİN KALAN VERİLERİ AYIRIYORUM
32 kalan = veriler.iloc[:,2:].values
33
34 #DIAGNOSIS KOLONUNU SAYILAŞTIRDIKTAN SONRA TEKRAR KALAN VERİLERLE BİRLEŞTİRİYORUM
35 bir= pd.DataFrame(data= diagnosis, index= range(569), columns= ["diagnosis"])
36 iki= pd.DataFrame(data= kalan, index= range(569))
37 veri=pd.concat([bir,iki], axis=1)
38

```

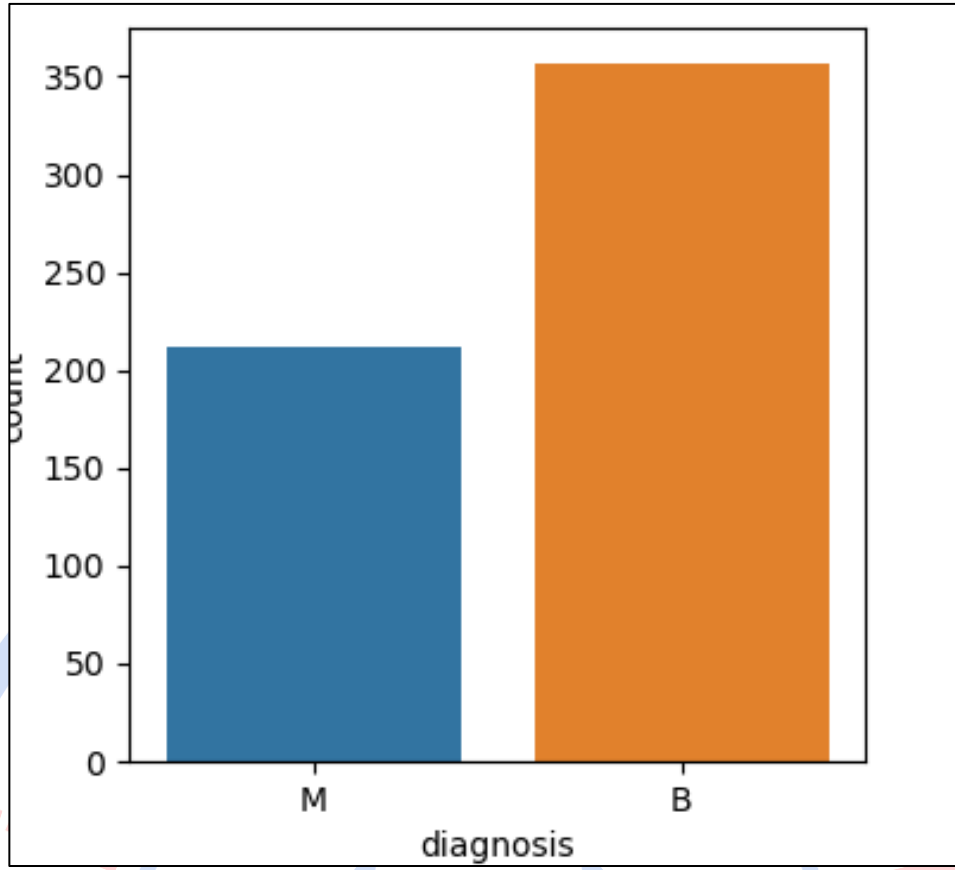
Şekil 2. Veri Yükleme ve Düzenleme

veri - DataFrame

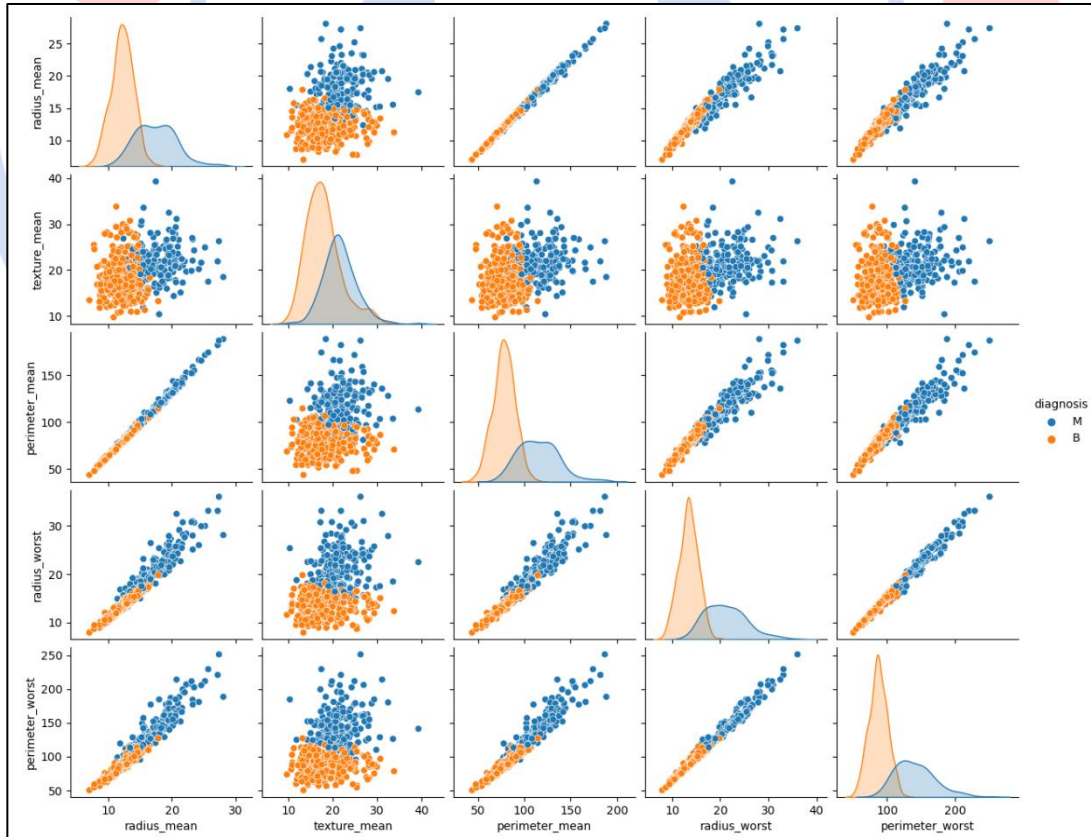
index	diagnosis	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0	1	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03083	0.006193	25.38	17.33	184.6	2019	0
1	1	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01389	0.003532	24.99	23.41	158.8	1956	0
2	1	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.004571	23.57	25.53	152.5	1709	0
3	1	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009288	14.91	26.5	98.87	567.7	0
4	1	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.005115	22.54	16.67	152.2	1575	0
5	1	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08889	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672	0.01137	0.02165	0.005882	15.47	23.75	103.4	741.6	0
6	1	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254	0.01039	0.01369	0.002179	22.88	27.66	153.2	1686	0
7	1	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488	0.01448	0.01486	0.005412	17.06	28.14	110.6	897	0
8	1	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03582	0.03553	0.01226	0.02143	0.003749	15.49	30.73	106.2	739.3	0
9	1	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217	0.07743	0.01432	0.01789	0.01008	15.09	40.68	97.65	711.4	0
10	1	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269	0.01101	0.007591	0.0146	0.003042	19.19	33.88	123.8	1150	0
11	1	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06882	0.5058	0.9849	3.564	54.16	0.005771	0.04061	0.02791	0.01282	0.02008	0.004144	20.42	27.28	136.5	1299	0
12	1	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297	0.0809	0.0409	0.04484	0.01284	20.96	29.94	151.7	1332	0
13	1	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126	0.05051	0.01992	0.02981	0.003002	16.84	27.66	112	876.5	0
14	1	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501	0.01628	0.01961	0.008093	15.03	32.01	108.8	697.7	0
15	1	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424	0.04741	0.0109	0.01857	0.005466	17.46	37.13	124.1	943.2	0
16	1	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162	0.01998	0.01109	0.0141	0.002085	19.07	30.88	123.4	1138	0
17	1	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026	0.02501	0.03188	0.01297	0.01689	0.004142	20.96	31.48	136.8	1315	0
18	1	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893	0.03391	0.01521	0.01356	0.001997	27.32	30.88	186.8	2398	0
19	0	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.008462	0.0146	0.02387	0.01315	0.0198	0.0023	15.11	19.26	99.7	711.2	0
20	0	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.004097	0.01898	0.01698	0.00649	0.01678	0.002425	14.5	20.49	96.09	630.5	0
21	0	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0.009606	0.01432	0.01985	0.01421	0.02027	0.002968	10.23	15.66	65.13	314.9	0
22	1	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096	3.384	44.91	0.006789	0.05328	0.06446	0.02252	0.03672	0.004394	18.07	19.80	125.1	980.9	0
23	1	23.84	137.2	1484	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127	4.303	93.99	0.004728	0.01259	0.01715	0.01038	0.01883	0.001987	29.17	35.59	188	2615	0

Format | Resize | Background color | Column min/max | Save and Close | Close

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)

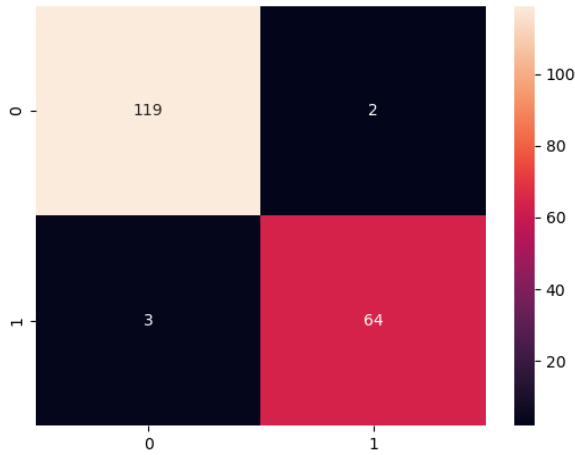
```
38  
39 """EĞİTİM/TEST VERİLERİ AYRILIYOR..."""  
40 x_train, x_test, y_train, y_test = train_test_split(kalan, diagnosis, test_size=0.33, random_state=0)  
41  
42 """ÖLÇEKLEME YAPIYORUM..."""  
43 sc=StandardScaler()  
44 x_train = sc.fit_transform(x_train)  
45 x_test = sc.transform(x_test)  
46
```

Şekil 6. Ölçekleme İşlemi

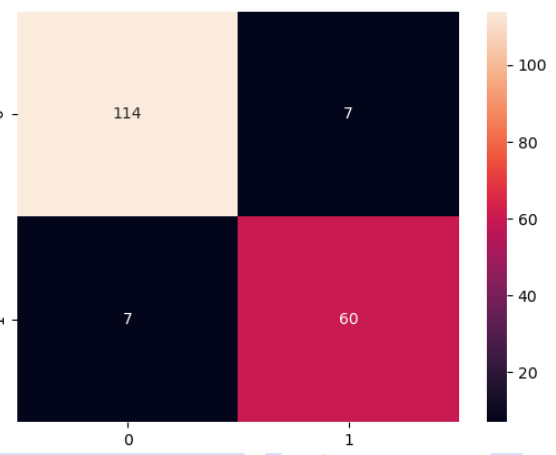
3. ALGORİTMA DEĞERLENDİRMELERİ

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

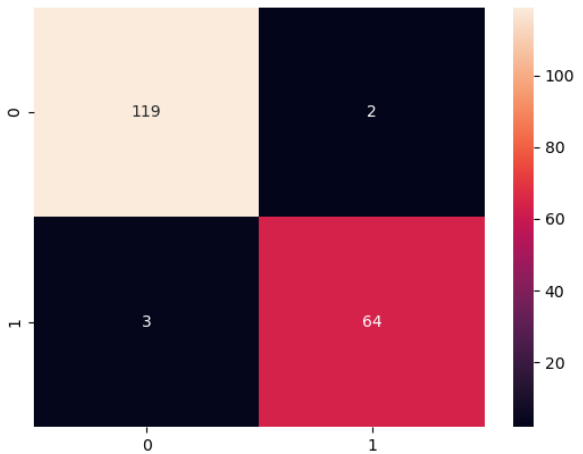
3.1. KARMAŞIKLIK MATRİSLERİ



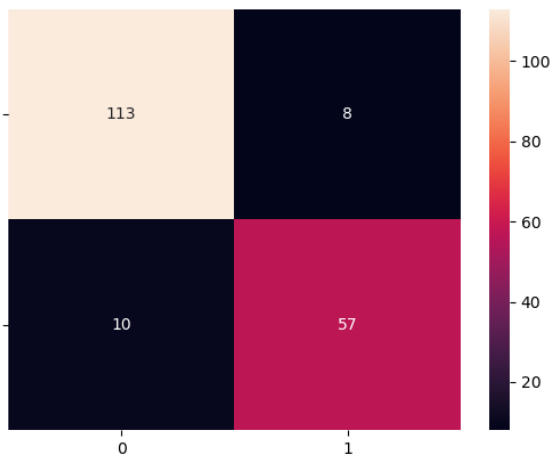
Şekil 7. LGR Karmaşıklık Matrisi



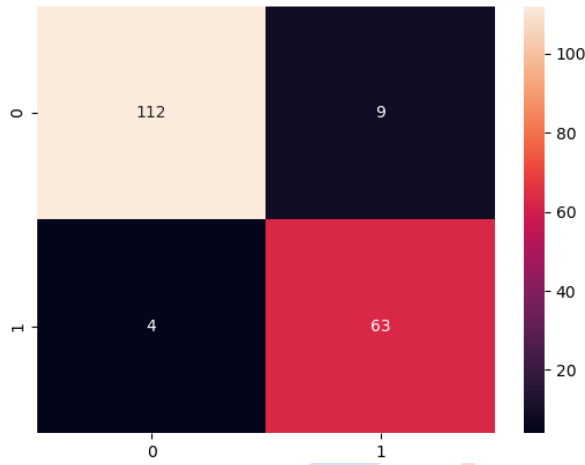
Şekil 7. KNN Karmaşıklık Matrisi



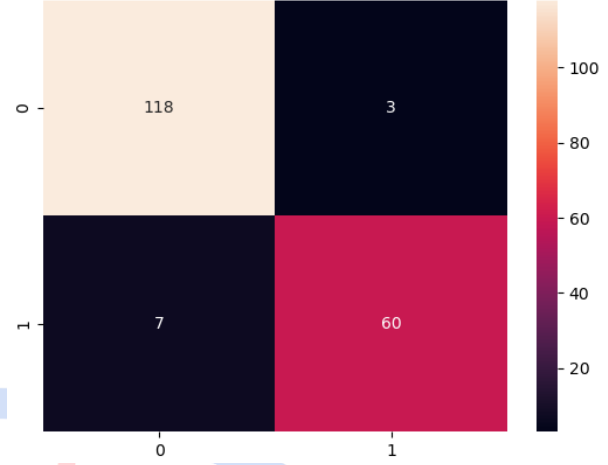
Şekil 8. SVC Karmaşıklık Matrisi



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>