BURSA TEKNIK ÜNİVERSİTESİ

MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ

VERİ MADENCILIĞI (BLM0463)

Proje Adı: DECİSİON BASED TREE SINIFLANDIRMA METODU KULLANARAK KRONİK BÖBREK HASTALIĞI (CDK) VERİ SETİN VERİ ANALİZİ

FARHAN AHMAD

20360859096

İÇİNDEKİLER

GİRİŞ....3
VERİ SETİ GENEL BİLGİLER....3
METODOLOJİ:....3
SAYISAL ÖZELLİKLER ANALİZİ....5
KATEGORİK ÖZELLİKLER ANALİZİ....5
SG (Specific Gravity) Analizi....5
AL(albumin) Analizi....6
SU (Sugar) Analizi....6
RBC (Red Blood Cells) Analizi....7
PC (Pus Cell) Analizi....7
PCC (Pus Cell clumps) Analizi....8
BA (Bacteria) Analizi....8
HTN (Hypertension) Analizi....9
DM (Diabetes Mellitus) Analizi....9
CAD (Coronary Artery Disease) Analizi....10
APPET (Appetite) Analizi....10
PE (Pedal Edema) Analizi....11
ANE (Anemia) Analizi....11
DECİSİON TREE GÖRSELLEŞTİRME....12
KORELASYON ANALİZİ....13
MODEL PERFORMANSI....13
MODEL KARŞILAŞTIRMASI....14
ÖZELLİK ÖNEMLİLİK ANALİZİ....15
SONUÇ....15
KAYNAKÇA....16

KRONİK BÖBREK HASTALIĞI VERİ ANALİZ RAPORU

GİRİŞ

Bu rapor, kronik böbrek hastalığı (CKD) veri seti üzerinde yapılan kapsamlı bir analizi içermektedir. Analizde, hasta ve sağlıklı bireylerin çeşitli tıbbi ölçümlerini kullanarak hastalık tespiti için bir makine öğrenmesi modeli geliştirilmiştir. Rapor, veri analizi, görselleştirmeler ve model performans değerlendirmelerini içermektedir.

Temel Bilgileri:

VERİ SETİ GENEL BİLGİLER

- Toplam Örnek Sayısı: 399
- CKD (Kronik Böbrek Hastası) Sayısı: 149
- Normal (Hasta Olmayan) Sayısı: 250
- Toplam Özellik Sayısı: 24

METODOLOJİ:

Veri Seti Bilgileri:

Başlangıçta veri seti şu sütunlardan oluşmaktadır:

Sütun Adı	**Açıklama**	**Tür**
age	yaş	nümerik
bp	kan basıncı	nümerik
sg	özgül ağırlık (1.005–1.025 arası)	kategorik
al	albümin seviyesi (0–5)	kategorik
su	şeker seviyesi (0–5)	kategorik
rbc	kırmızı kan hücresi durumu (normal/abnormal)	kategorik
pc	irin hücresi durumu (normal/abnormal)	kategorik

pcc	irin hücresi topaklanması (present/notpresent)	kategorik
ba	bakteri varlığı	kategorik
bgr	rastgele kan şekeri	nümerik
bu	kan üre seviyesi	nümerik
sc	serum kreatinin	nümerik
sod, pot, hemo, pcv, wc, rc	çeşitli kan değerleri	nümerik
htn, dm, cad, appet, pe, ane	sağlık durumları (yes/no)	kategorik
class	hedef değişken: hasta mı değil mi? (ckd/notckd)	kategorik

VERİNİN İLK 5 SATIRI:

```
First 5 rows:
    age    bp     sg   al   su     rbc        pc         pcc  ... rbcc  htn   dm  cad  appet   pe  ane class
0  48.0  80.0  1.020  1.0  0.0     NaN    normal  notpresent  ...  5.2  yes  yes   no   good   no   no   ckd
1   7.0  50.0  1.020  4.0  0.0     NaN    normal  notpresent  ...  NaN   no   no   no   good   no   no   ckd
2  62.0  80.0  1.010  2.0  3.0  normal    normal  notpresent  ...  NaN   no  yes   no   poor   no  yes   ckd
3  48.0  70.0  1.005  4.0  0.0  normal  abnormal     present  ...  3.9  yes   no   no   poor  yes  yes   ckd
4  51.0  80.0  1.010  2.0  0.0  normal    normal  notpresent  ...  4.6   no   no   no   good   no   no   ckd

[5 rows x 25 columns]
```

EKSİK DEĞER:

age	9
bp	12
sg	47
al	46
su	49
rbc	152
pc	65
pcc	4
ba	4
bgr	44
bu	19
sc	17
sod	87
pot	88
hemo	52
pcv	71
wbcc	106
rbcc	131
htn	2
dm	2
cad	2

appet	1
pe	1
ane	1
class	0

SAYISAL ÖZELLİKLER ANALİZİ

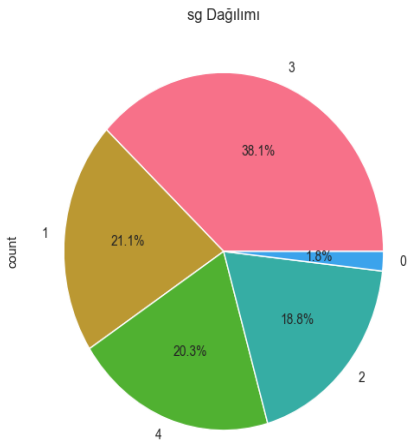
Sayısal özelliklerin istatistiksel analizi aşağıdaki gibidir:

	age	bp	bgr	bu	sc	sod	pot	hemo	pcv	wbcc	rbcc
count	399	399	399	399	399	399	399	399	399	399	399
mean	51.49	76.59	145.16	56.71	3	137.62	4.58	12.53	39.07	8293.48	4.74
std	16.96	13.5	75.33	49.46	5.63	9.21	2.82	2.72	8.17	2530.78	0.84
min	2	50	22	1.5	0.4	4.5	2.5	3.1	9	2200	2.1
25%	42	70	101	27	0.9	135	4	10.85	34	6950	4.5
50%	54.5	80	121	42	1.3	138	4.4	12.6	40	8000	4.8
75%	64	80	150	62.5	2.75	141	4.8	14.65	44	9350	5.1
max	90	180	490	391	76	163	47	17.8	54	26400	8

KATEGORİK ÖZELLİKLER ANALİZİ

SG (Specific Gravity) Analizi

- 1.005(0) = 7
- 1.010(1) = 84
- 1.015(2) = 75
- 1.020(3) = 152
- 1.025(4) = 81

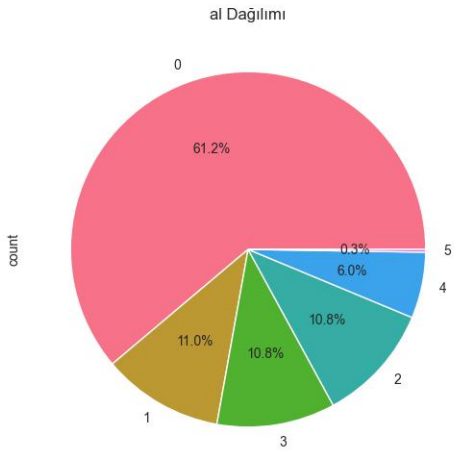


Şekil: sg özelliğinin pasta grafiği dağılımı

AL(albumin) Analizi

al özelliğinin dağılımı:

- “0” = 244
- “1” = 44
- “2” = 43
- “3” = 43
- “4” = 24
- “5” = 1

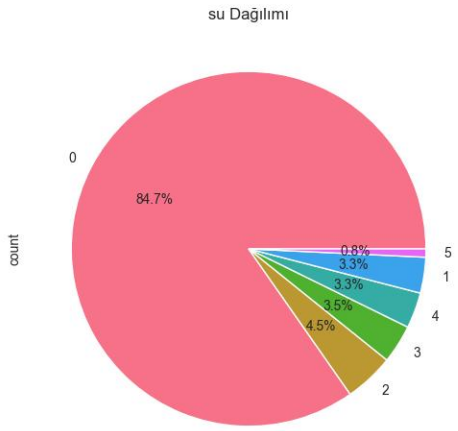


Şekil: al özelliğinin pasta grafiği dağılımı

SU (Sugar) Analizi

su özelliğinin dağılımı:

- “0” = 338
- “1” = 13
- “2” = 18
- “3” = 14
- “4” = 13
- “5” = 3

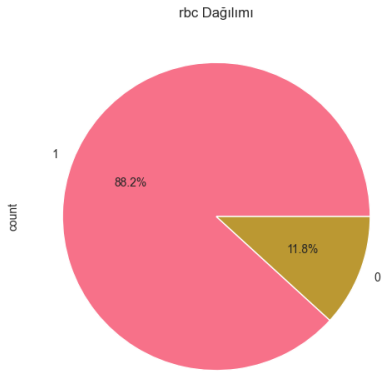


Şekil: su özelliğinin pasta grafiği dağılımı

RBC (Red Blood Cells) Analizi

rbc özelliğinin dağılımı:

- abnormal (1) = 352
- normal (0) = 47

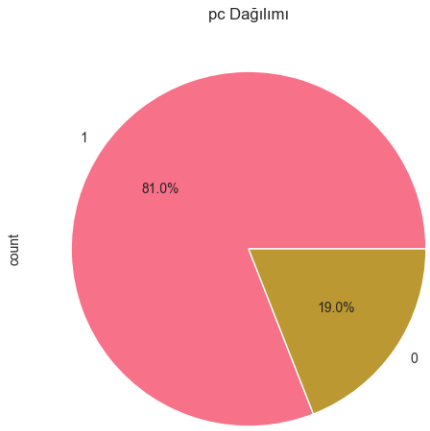


Şekil: rbc özelliğinin pasta grafiği dağılımı

PC (Pus Cell) Analizi

pc özelliğinin dağılımı:

- Abnormal (1) = 323
- normal (0) = 76

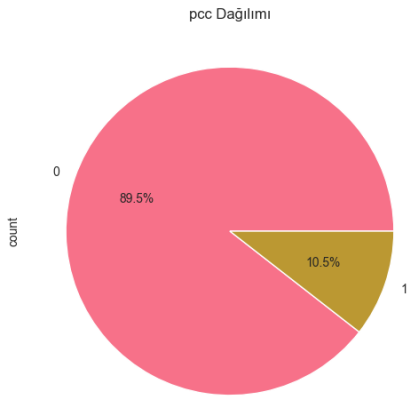


Şekil: pc özelliğinin pasta grafiği dağılımı

PCC (Pus Cell clumps) Analizi

pcc özelliğinin dağılımı:

- present (0) = 357
- notpresent (1) = 42

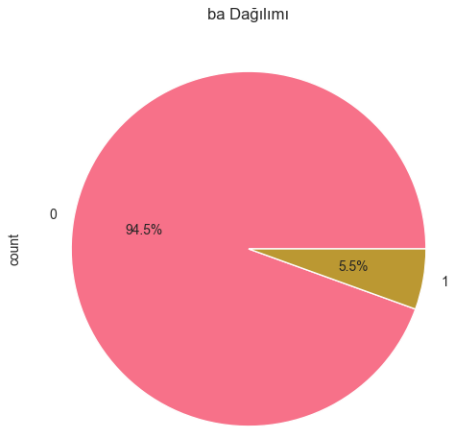


Şekil: pcc özelliğinin pasta grafiği dağılımı

BA (Bacteria) Analizi

ba özelliğinin dağılımı:

- present (0) = 377
- notpresent (1) = 22

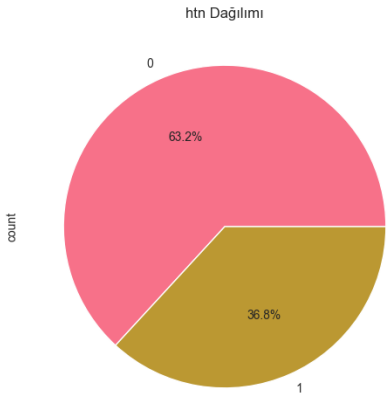


Şekil: ba özelliğinin pasta grafiği dağılımı

HTN (Hypertension) Analizi

htn özelliğinin dağılımı:

- yes (0) = 252
- no (1) = 147

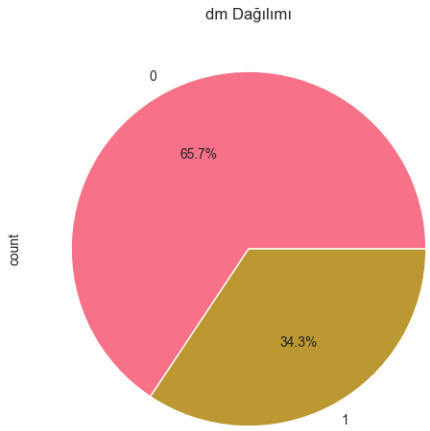


Şekil: htn özelliğinin pasta grafiği dağılımı

DM (Diabetes Mellitus) Analizi

dm özelliğinin dağılımı:

- yes (0) = 262
- no (1) = 137

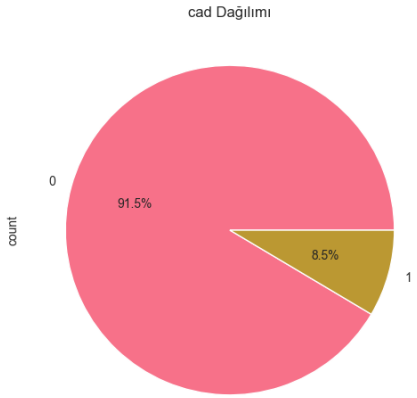


Şekil: dm özelliğinin pasta grafiği dağılımı

CAD (Coronary Artery Disease) Analizi

cad özelliğinin dağılımı:

- yes (0) = 365
- no (1) = 34

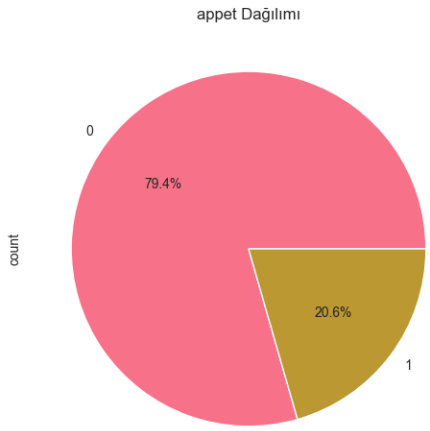


Şekil: cad özelliğinin pasta grafiği dağılımı

APPET (Appetite) Analizi

appet özelliğinin dağılımı:

- good (0) = 317
- poor (1) = 82

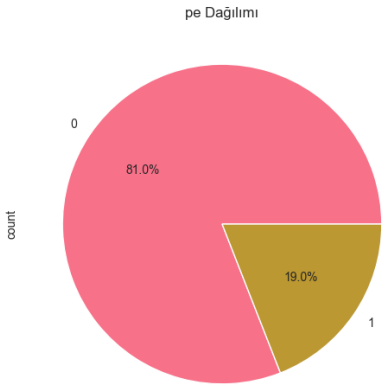


Şekil: appet özelliğinin pasta grafiği dağılımı

PE (Pedal Edema) Analizi

pe özelliğinin dağılımı:

- yes (0) = 323
- no (1) = 76

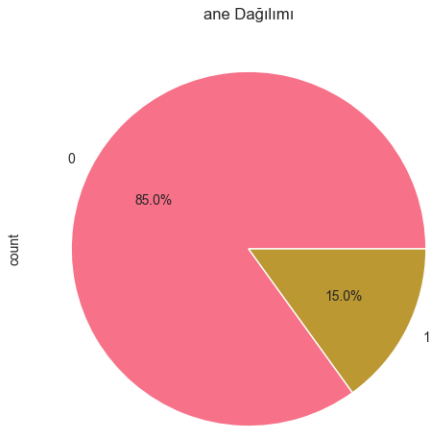


Şekil: pe özelliğinin pasta grafiği dağılımı

ANE (Anemia) Analizi

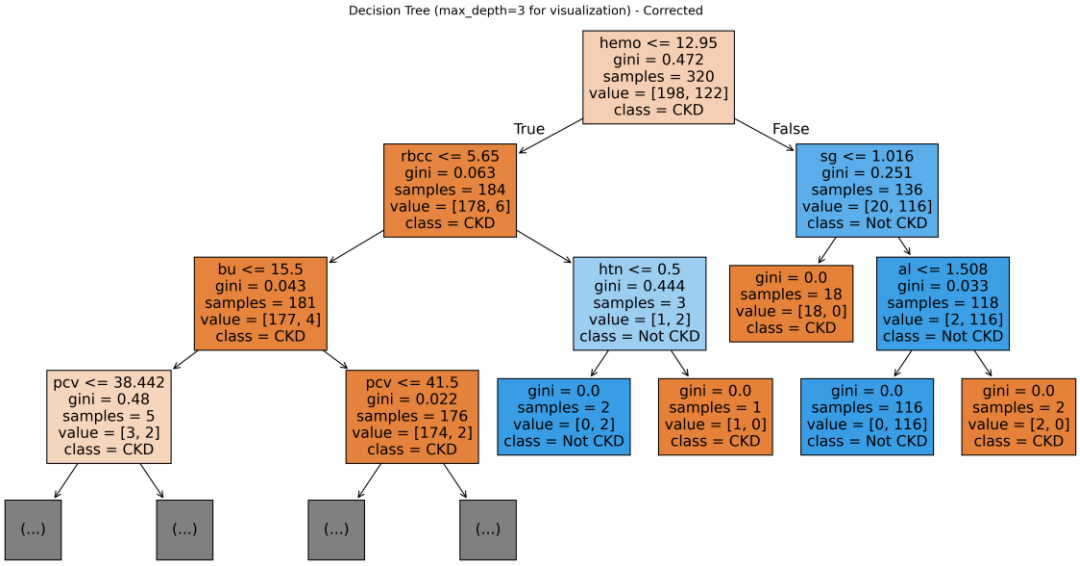
ane özelliğinin dağılımı:

- yes (0) = 339
- no (1) = 60



Şekil: ane özelliğinin pasta grafiği dağılımı

DECİSİON TREE GÖRSELLEŞTİRME

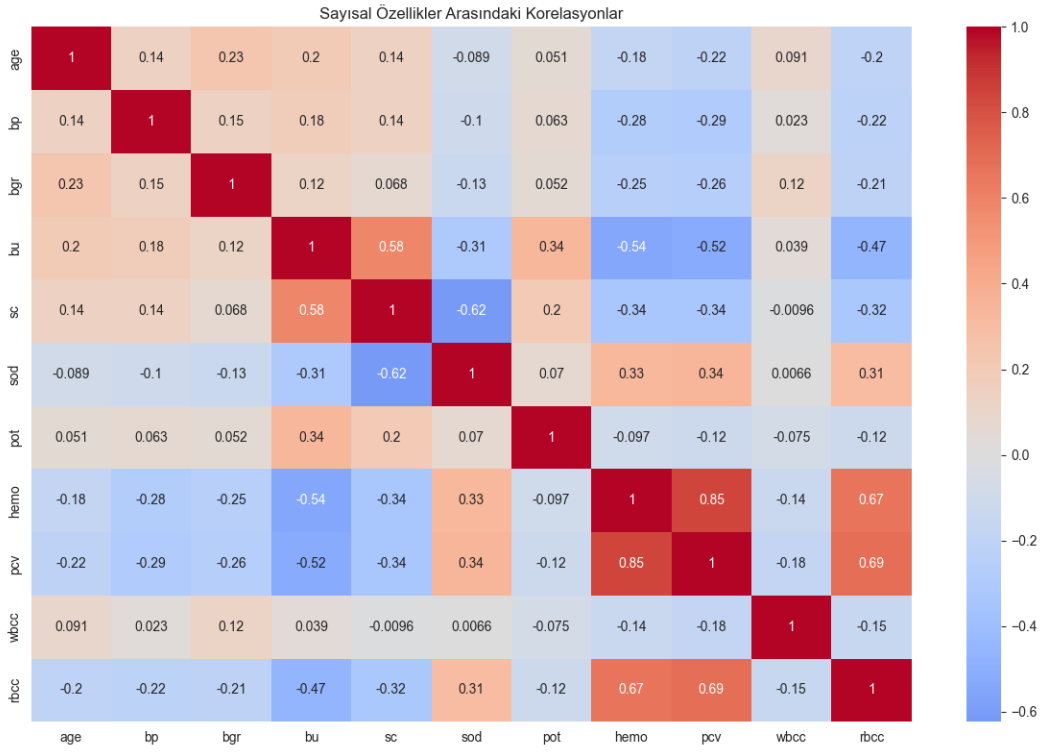


Şekil: Desicion Tree görseleştirme

Decision Tree Doğrululuk: 0.9875

KORELASYON ANALİZİ

Sayısal özellikler arasındaki korelasyonlar, özelliklerin birbirleriyle olan ilişkilerini göstermektedir. Korelasyon değerleri -1 ile 1 arasında değişmekte olup, 1'e yakın değerler güçlü pozitif ilişkiyi, -1'e yakın değerler güçlü negatif ilişkiyi göstermektedir.



Şekil: Sayısal özellikler arasındaki korelasyon matrisi

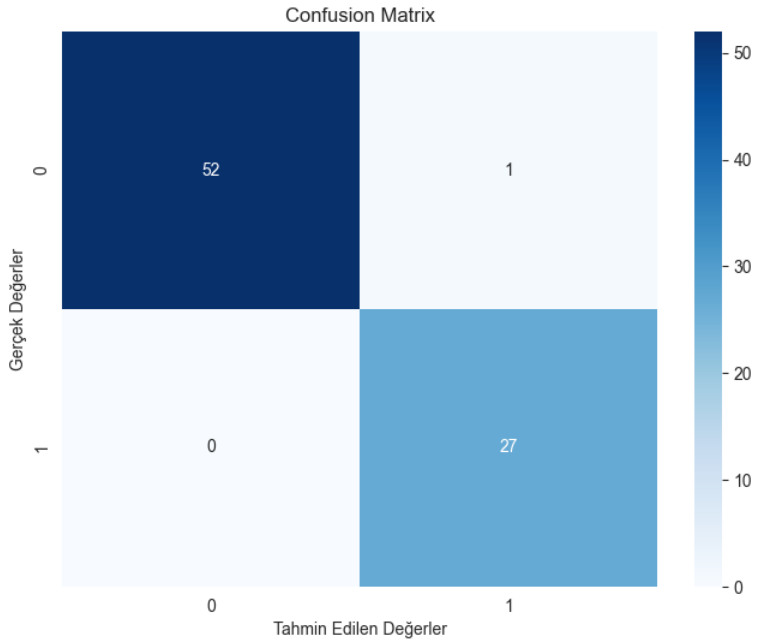
MODEL PERFORMANSI

Cross-validatör Sonuçları:

- Ortalama CV Skoru: 0.9875 (+/- 0.0332)
- Bu sonuç, modelin farklı veri parçaları üzerindeki tutarlılığını göstermektedir.

Sınıflandırma Performans Metrikleri:

Class	Precision	Recall	F1-Score	Support
NotCDK	1	0.98	0.99	53
CKD	0.96	1	0.98	27
Accuracy			0.99	80
Macro Avg	0.98	0.99	0.99	80
Weighted Avg	0.99	0.99	0.99	80



Şekil: Karmaşıklık Matrisi (Confusion Matrix)

- Sol üst: Doğru Negatif (TN)- Sağlıklı bireylerin doğru tahmin edilmesi
- Sağ alt: Doğru Pozitif (TP)- Hasta bireylerin doğru tahmin edilmesi
- Sol alt: Yanlış Negatif (FN)- Hasta bireylerin sağlıklı olarak yanlış tahmin edilmesi
- Sağ üst: Yanlış Pozitif (FP)- Sağlıklı bireylerin hasta olarak yanlış tahmin edilmesi

MODEL KARŞILAŞTIRMASI

- Random Forest Accuracy: 1.000
- Naive Bayes Accuracy: 0.988
- Support Vector Machine Accuracy: 0.950

ÖZELLİK ÖNEMLİLİK ANALİZİ

Özellik önemliliği analizi, hangi özelliklerin hastalık tespitinde daha etkili olduğunu göstermektedir. Yüksek önem derecesine sahip özellikler, modelin tahminlerinde daha büyük rol oynamaktadır.

Özellik Önem Derecesi

Niteklik Numarası	Nitelik Adı	Özellik Önem Derecesi
14	hemo	0.686983
2	sg	0.209909
3	al	0.026096
10	bu	0.021989
15	pcv	0.018355
17	rbcc	0.016258
13	pot	0.010618
18	htn	0.008849
11	sc	0.000942
6	pc	0
7	pcc	0
4	su	0
0	age	0
1	bp	0
5	rbc	0
9	bgr	0
12	sod	0
8	ba	0
16	wbcc	0
19	dm	0
20	cad	0
21	appet	0
22	pe	0
23	ane	0

SONUÇ

Yapılan analiz sonucunda:

1. Model 0.9875 doğruluk oranı ile başarılı bir performans göstermiştir.
2. En önemli özellikler: hemo, sg, al
3. Kategorik ve sayısal özelliklerin dağılımları dengeli bir veri seti olduğunu göstermektedir.

KAYNAKÇA

1- Yaptığım analizin Kodları
https://github.com/ahmadfarhan203/CDKDataAnalysis/tree/main

2- Sunum Videosu
https://youtu.be/cK_xjS9jwgo

3- Veri seti:
Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic Kidney Disease Data Set. UCI Machine Learning Repository. https:/doi.org/10.24432/C5G020

4- Akademik destekleyici kaynaklar:
Polat, K., & Günes, S. (2007). An expert system approach based on PCA and ANFIS to diagnosis of chronic kidney disease. Digital Signal Processing, 17(4), 702–710. https://doi.org/10.1016/j.dsp.2006.09.005

Kora, A. D., & Kalva, K. (2015). Hybrid Bacterial Foraging and PSO for detecting kidney disease. Procedia Computer Science, 57, 722–729. https://doi.org/10.1016/j.procs.2015.07.507