POLİKİSTİK OVER SENDROMU (PCOS) TAHMİNİ

İÇERİK

1. VERİ SETİ

2. KEŞİFSEL VERİ ANALİZİ

3. VERİ ÖN İŞLEME

4. MODELLEME

5. SONUÇLAR

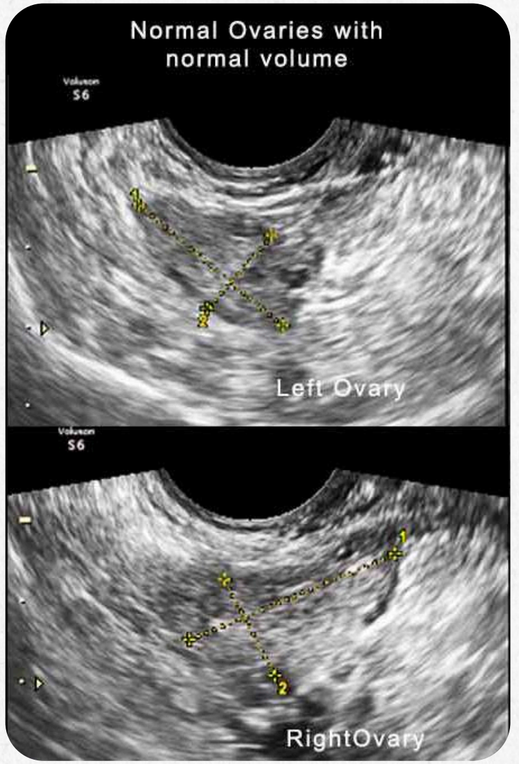
PCOS ?

Polikistik over sendromu, kadınlarda en sık görülen hormonal bozuklukların başında yer almaktadır. PCOS, üreme çağındaki kadınların tahmini olarak %8-13'ünü etkilemektedir. Dünya çapında etkilenen kadınların %70'e kadarı teşhis edilemiyor. PCOS, anovulasyonun en yaygın nedeni ve kısırlığın önde gelen nedenidir.

Polikistik over sendromunun belirtileri kişiden kişiye farklılık gösterebilir. Semptomlar zamanla değişebilir.

Olası semptomlar:

- Düzensiz menstrüal döngü dönemler
- Kısırlık
- Sivilce veya yağlı cilt
- Yüzde veya vücutta aşırı kıllanma
- Saç dökülmesi veya saç incelmesi
- Özellikle göbek çevresinde kilo alımı



PCOS'lu kişilerin aşağıdakiler de dahil olmak üzere başka sağlık sorunlarına sahip olma olasılığı daha yüksektir:

- Tip 2 diyabet
- Hipertansiyon
- Yüksek kolesterol
- Kalp hastalığı
- Endometrial kanser

Veri Seti

#	Column	Non-Null Count	Dtype
0	Sl. No	541 non-null	int64
1	Patient File No.	541 non-null	int64
2	PCOS (Y/N)	541 non-null	int64
3	Age (yrs)	541 non-null	int64
4	Weight (Kg)	541 non-null	float64
5	Height(Cm)	541 non-null	float64
6	BMI	541 non-null	float64
7	Blood Group	541 non-null	int64
8	Pulse rate(bpm)	541 non-null	int64
9	RR (breaths/min)	541 non-null	int64
10	Hb(g/dl)	541 non-null	float64
11	Cycle(R/I)	541 non-null	int64
12	Cycle length(days)	541 non-null	int64
13	Marraige Status (Yrs)	540 non-null	float64
14	Pregnant(Y/N)	541 non-null	int64
15	No. of abortions	541 non-null	int64
16	I beta-HCG(mIU/mL)	541 non-null	float64
17	II beta-HCG(mIU/mL)	541 non-null	object
18	FSH(mIU/mL)	541 non-null	float64
19	LH(mIU/mL)	541 non-null	float64
20	FSH/LH	541 non-null	float64
21	Hip(inch)	541 non-null	int64
22	Waist(inch)	541 non-null	int64
23	Waist:Hip Ratio	541 non-null	float64
24	TSH (mIU/L)	541 non-null	float64
25	AMH(ng/mL)	541 non-null	object
26	PRL(ng/mL)	541 non-null	float64
27	Vit D3 (ng/mL)	541 non-null	float64
28	PRG(ng/mL)	541 non-null	float64
29	RBS(mg/dl)	541 non-null	float64
30	Weight gain(Y/N)	541 non-null	int64
31	hair growth(Y/N)	541 non-null	int64
32	Skin darkening (Y/N)	541 non-null	int64
33	Hair loss(Y/N)	541 non-null	int64
34	Pimples(Y/N)	541 non-null	int64
35	Fast food (Y/N)	540 non-null	float64
36	Reg.Exercise(Y/N)	541 non-null	int64
37	BP _Systolic (mmHg)	541 non-null	int64
38	BP _Diastolic (mmHg)	541 non-null	int64
39	Follicle No. (L)	541 non-null	int64
40	Follicle No. (R)	541 non-null	int64
41	Avg. F size (L) (mm)	541 non-null	float64
42	Avg. F size (R) (mm)	541 non-null	float64
43	Endometrium (mm)	541 non-null	float64
44	Unnamed: 44	2 non-null	object

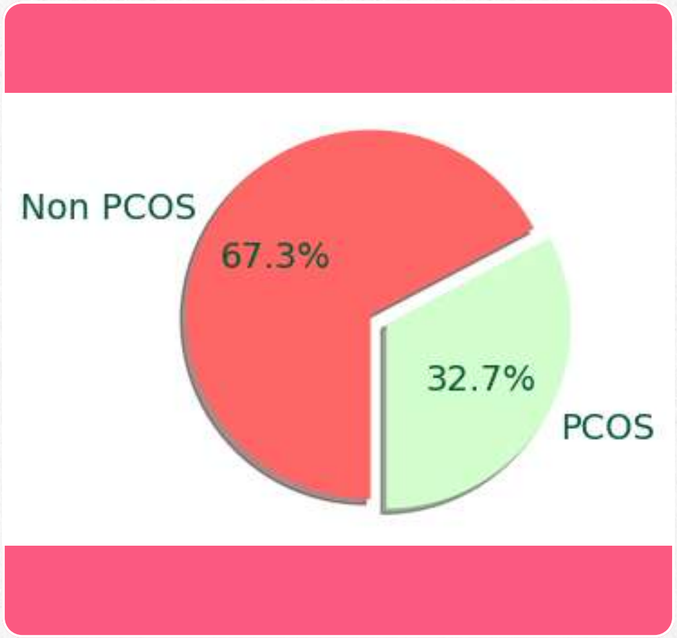
Bu çalışma için kullanılan açık kaynaklı Polikistik Over Sendromu verileri Kaggle'dan alındı.

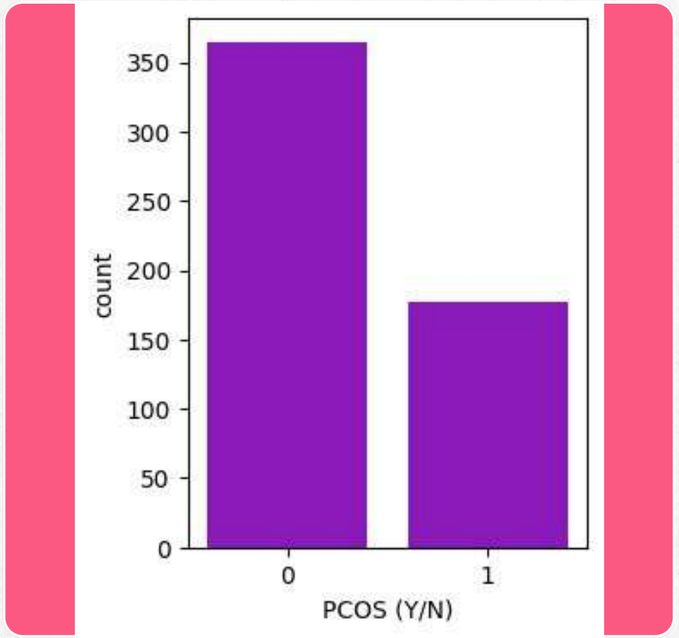
Veri Seti : PCOS Dataset

- 541 satır
- 45 sütun

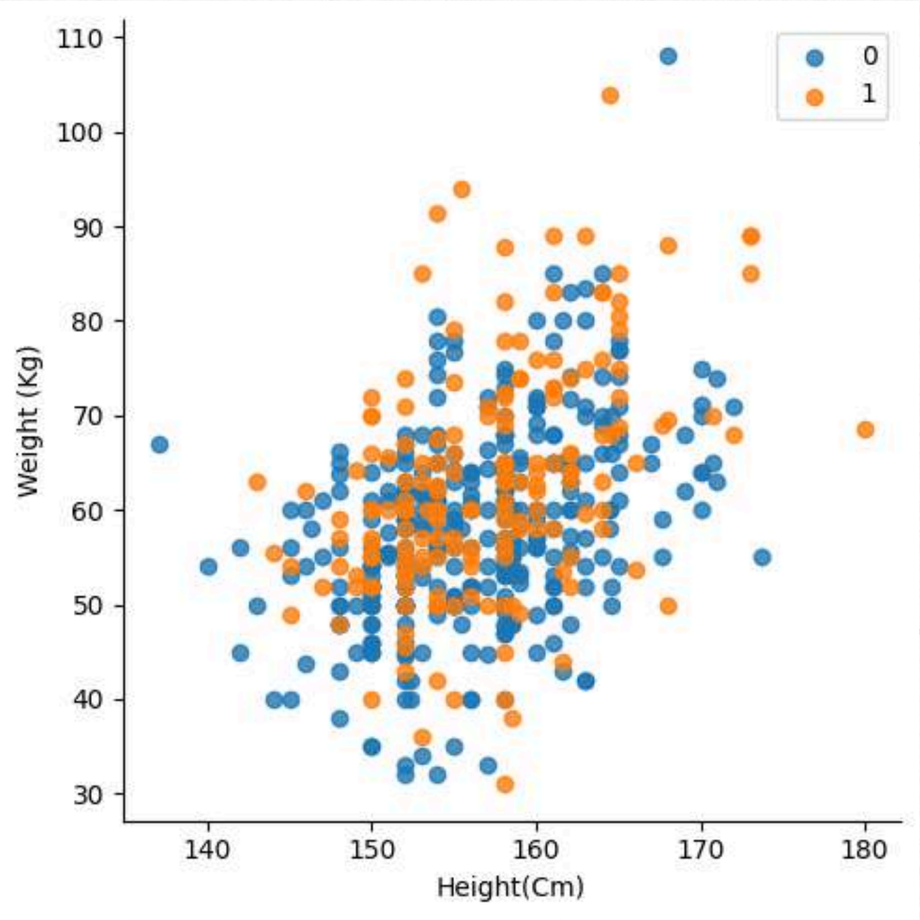
Veri setindeki her satır bir hastayı temsil eder, her sütun ise hastalara ait özelliklerini içerir.

Hedef Değişken Dağılımı

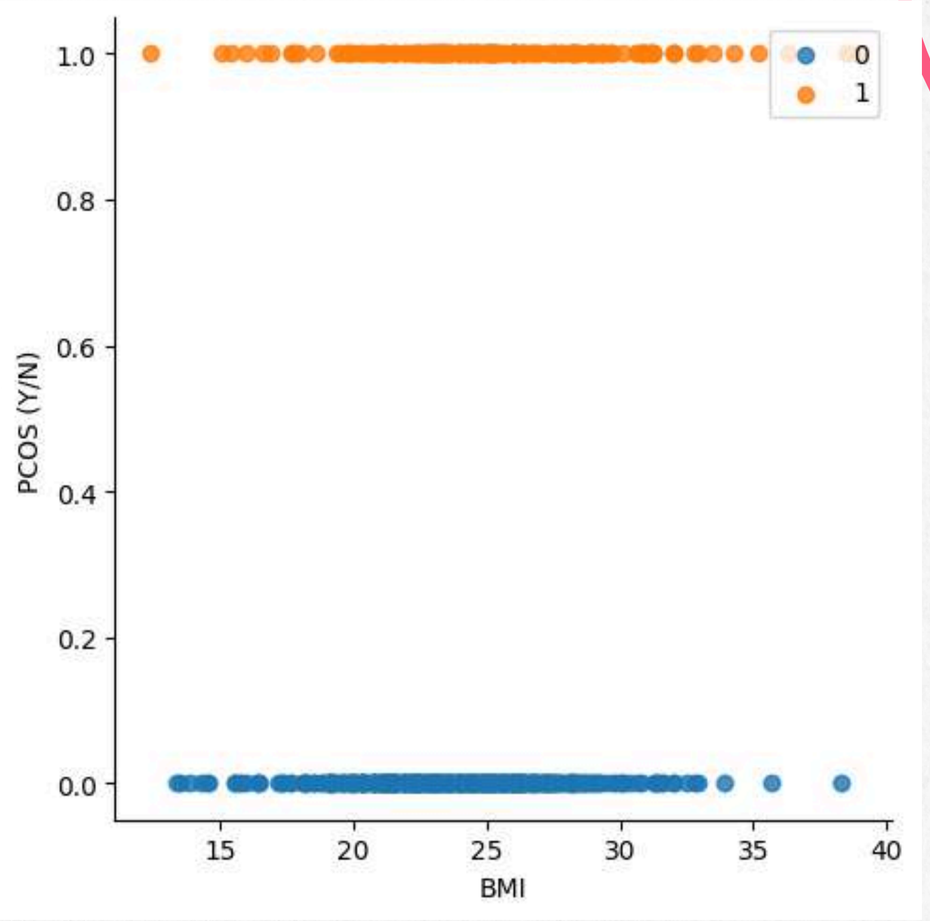




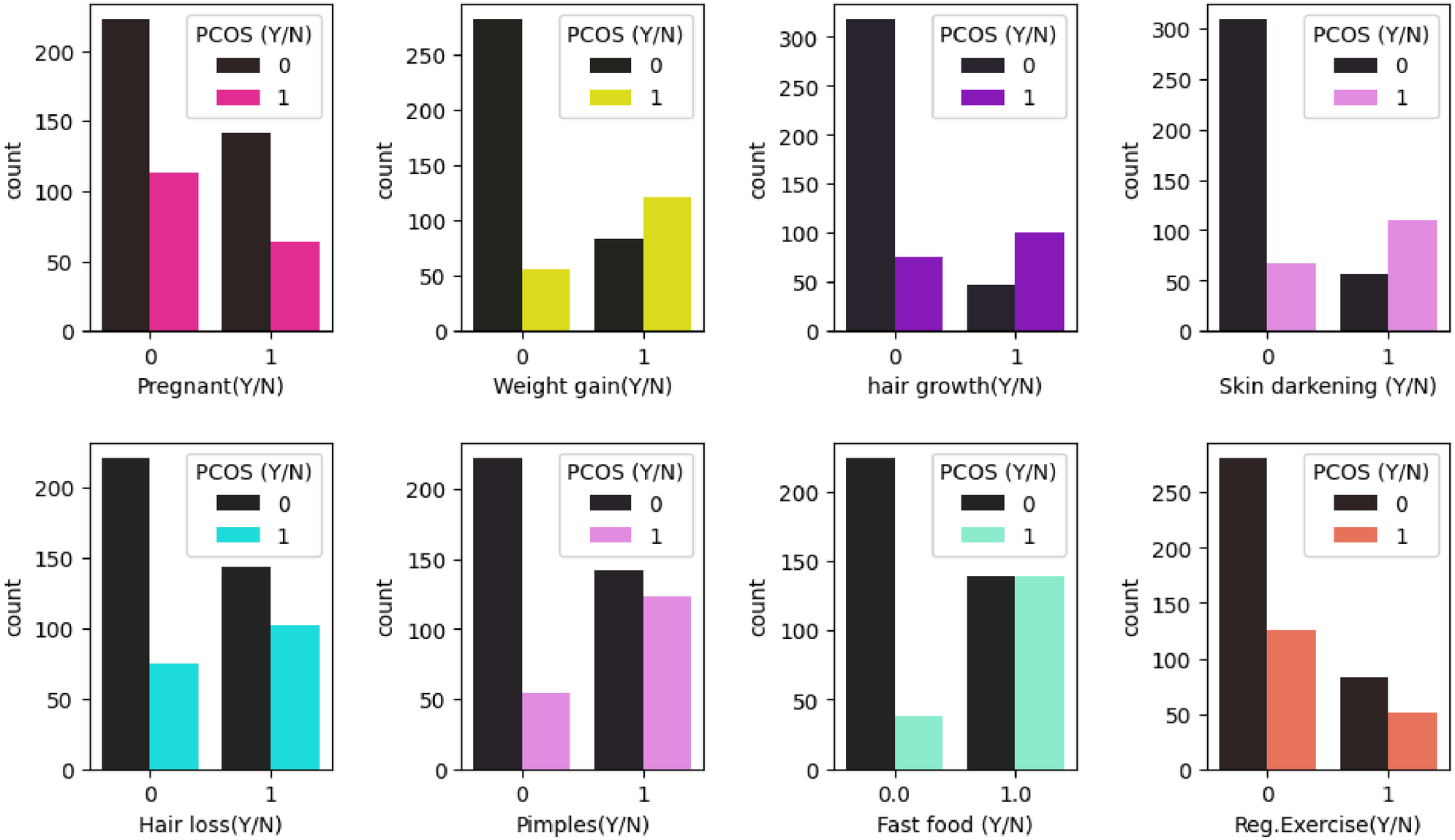
(0) PCOS olmayan : 364 kişi
(1) PCOS: 174 kişi

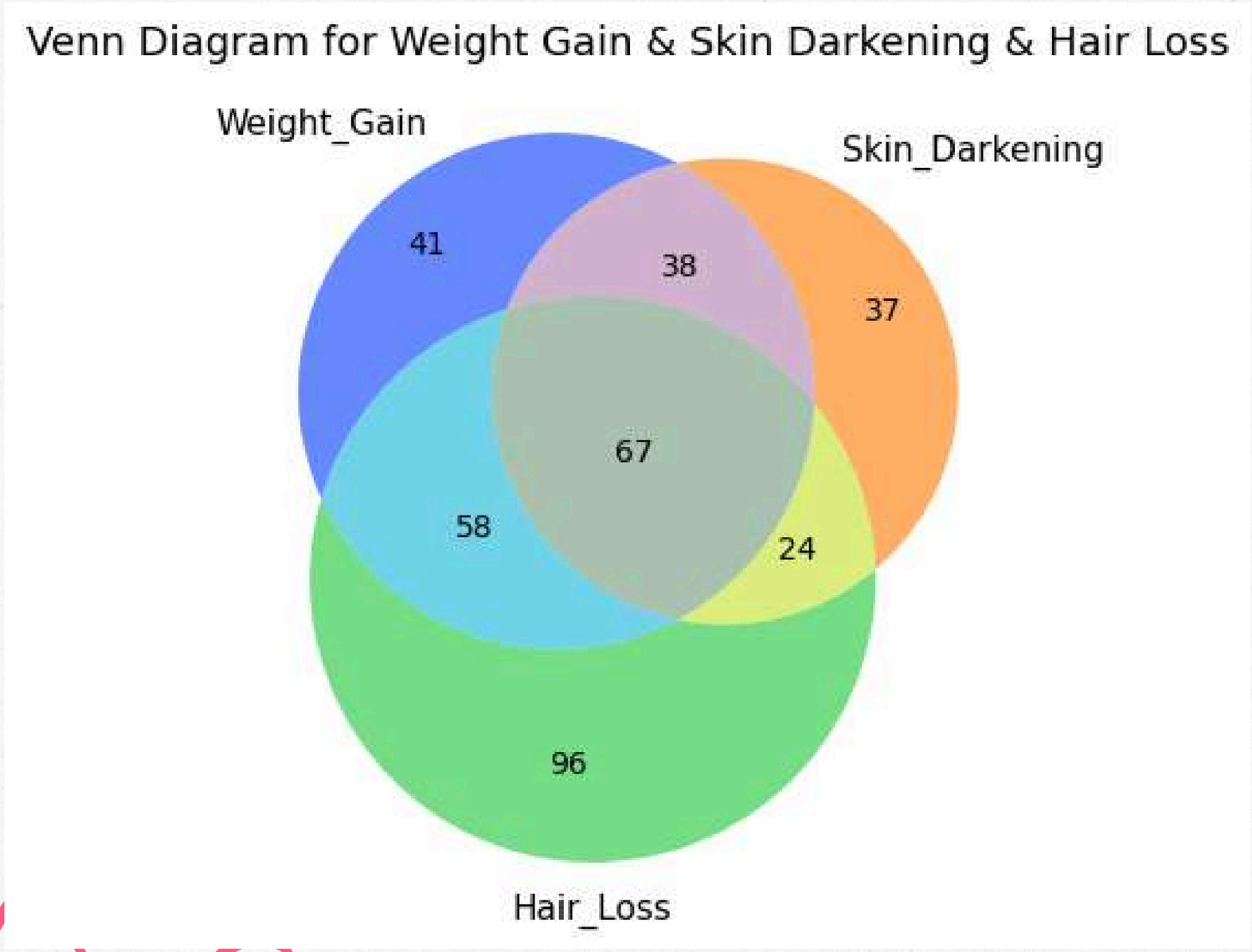


Zayıf veya şişman olmak direkt olarak PCOS'u etkilememektedir.

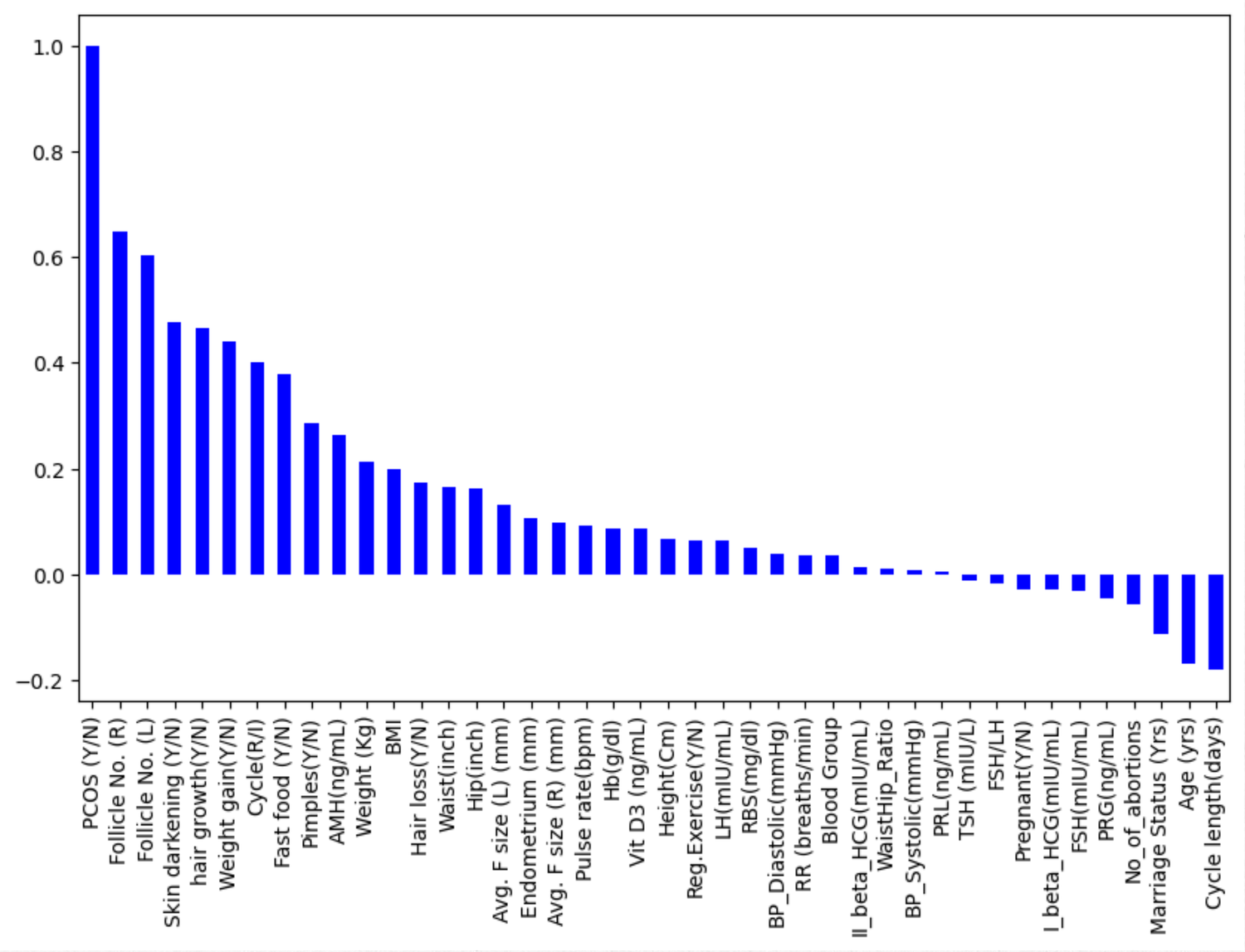


Bu sebeple herhangi bir BMI değerine sahip olan bir kişi PCOS olabilir veya olmayabilir.


PCOS (Y/N)
0
1
count
Pregnant(Y/N)
Weight gain(Y/N)
hair growth(Y/N)
Skin darkening (Y/N)
Hair loss(Y/N)
Pimples(Y/N)
Fast food (Y/N)
Reg.Exercise(Y/N)



- 187 kişide bu belirtilerden en az ikisi bulunmaktadır.
- Üç belirtiye sahip olan kişi sayısı ise 67'dir.



Folikül sayısı, cilt kararması PCOS hastalığı ile pozitif ilişkili iken, evlilik durumu, yaş, regl döngüsü uzunluğu ise PCOS hastalığı ile negatif ilişkilidir.

Veri Ön İşleme

- Tekrarlayan veri yoktu.
- Gereksiz sütunlar atıldı.

```
df = data.drop(["Sl. No", "Patient File No.",  "Unnamed: 44"], axis=1)
```

- Feature isimleri düzenlendi.

```
df.rename(columns={'Height(Cm) ': 'Height(Cm)'}, inplace=True)
df.rename(columns={'Marraige Status (Yrs)': 'Marriage Status (Yrs)'},
df.rename(columns={'Pulse rate(bpm) ': 'Pulse rate(bpm)'}, inplace=True)
df.rename(columns={'II    beta-HCG(mIU/mL)': 'II_beta_HCG(mIU/mL)'}, inplace=True)
df.rename(columns={' Age (yrs)': 'Age (yrs)'}, inplace=True)
df.rename(columns={'  I   beta-HCG(mIU/mL)': 'I_beta_HCG(mIU/mL)'}, inplace=True)
df.rename(columns={'No. of abortions': 'No_of_abortions'}, inplace=True)
df.rename(columns={'BP _Systolic (mmHg)': 'BP_Systolic(mmHg)'}, inplace=True)
df.rename(columns={'BP _Diastolic (mmHg)': 'BP_Diastolic(mmHg)'}, inplace=True)
df.rename(columns={'Waist:Hip Ratio': 'WaistHip_Ratio'}, inplace=True)
```

- Veri tipleri değiştirildi.

```
df.loc[df['II_beta_HCG(mIU/mL)'] == '1.99.', 'II_beta_HCG(mIU/mL)'] = 1.99
df.loc[df['AMH(ng/mL)'] == 'a', 'AMH(ng/mL)'] = np.nan # eksik değer

df['II_beta_HCG(mIU/mL)'] = pd.to_numeric(df['II_beta_HCG(mIU/mL)'], errors='coerce').astype('float64')
df['AMH(ng/mL)'] = pd.to_numeric(df['AMH(ng/mL)'], errors='coerce').astype('float64')
```

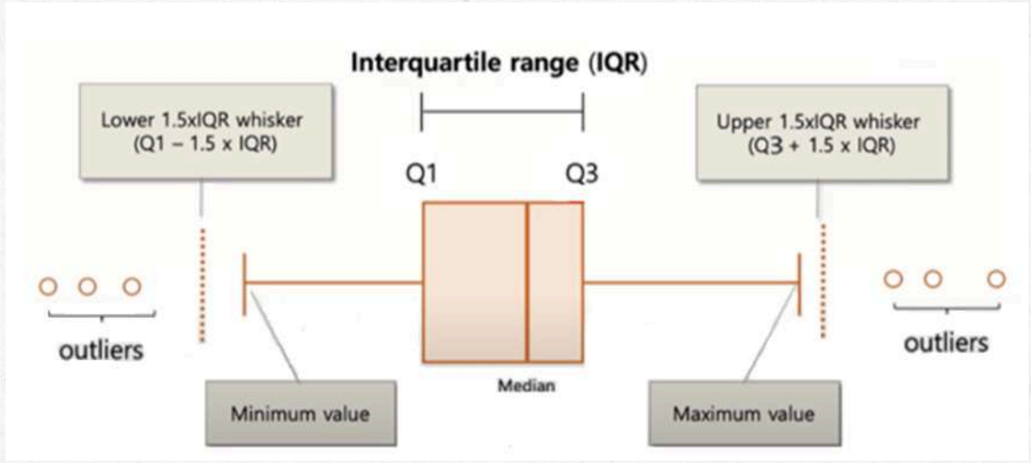
- Bazı featurelar veri setinden çıkarıldı.

```
df.drop(["I_beta_HCG(mIU/mL)", "II_beta_HCG(mIU/mL)"], axis=1, inplace=True)
```

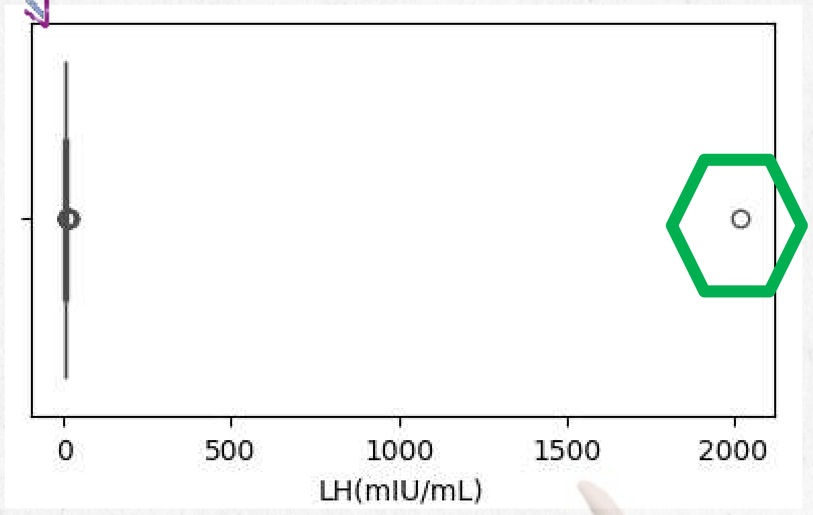
- Eksik değerler dolduruldu.

```
df["Marriage Status (Yrs)"].fillna(df["Marriage Status (Yrs)"].median(), inplace=True)
df["AMH(ng/mL)"].fillna(df["AMH(ng/mL)"].median(), inplace=True)
df["Fast food (Y/N)"].fillna(df["Fast food (Y/N)"].mode()[0], inplace=True)
```

Temel Modelleme...



- Aykırı değerler değerlendirildi.
- Ölçeklendirme işlemi uygulandı.



Modelleme

Model	Cross Validated Score (Mean)	Cross Validated Score (Std)
Random Forest	0.900535	0.022300
XGBoost	0.884389	0.029680
Logistic Regression	0.856482	0.015656
Decision Tree	0.798770	0.038840
KNN	0.782438	0.019610
SVM	0.673617	0.003765

Seçilen modeller:
Random Forest, XGBoost, SVM

Standard Scaler uygulandıktan ve Outlier analizi yapıldıktan sonra elde edilen sonuçlar

Model	Cross Validated Score (Mean)	Cross Validated Score (Std)
Random Forest	0.900508	0.018545
XGBoost	0.877412	0.020879
SVM	0.868110	0.015233
Logistic Regression	0.863539	0.020417
KNN	0.840283	0.029654
Decision Tree	0.798770	0.038840

En iyi sonuçlar **XGBoost** modelinden elde edilmiştir.

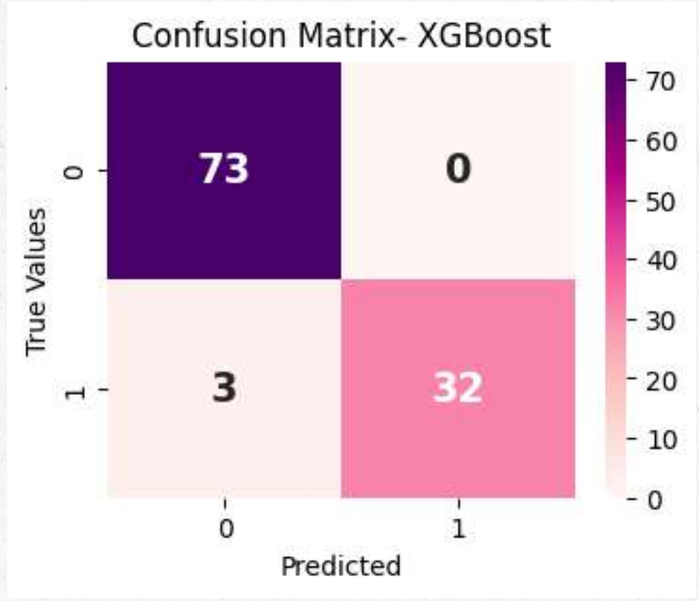
	Model	Train Accuracy	Test Accuracy	F1-Score	AUC Score	Precision	Recall
0	SVM	0.960648	0.925926	0.923071	0.972994	0.933242	0.925926
1	Random Forest	1.000000	0.953704	0.953526	0.986301	0.953534	0.953704
2	XGBoost	1.000000	0.962963	0.962669	0.986693	0.963143	0.962963

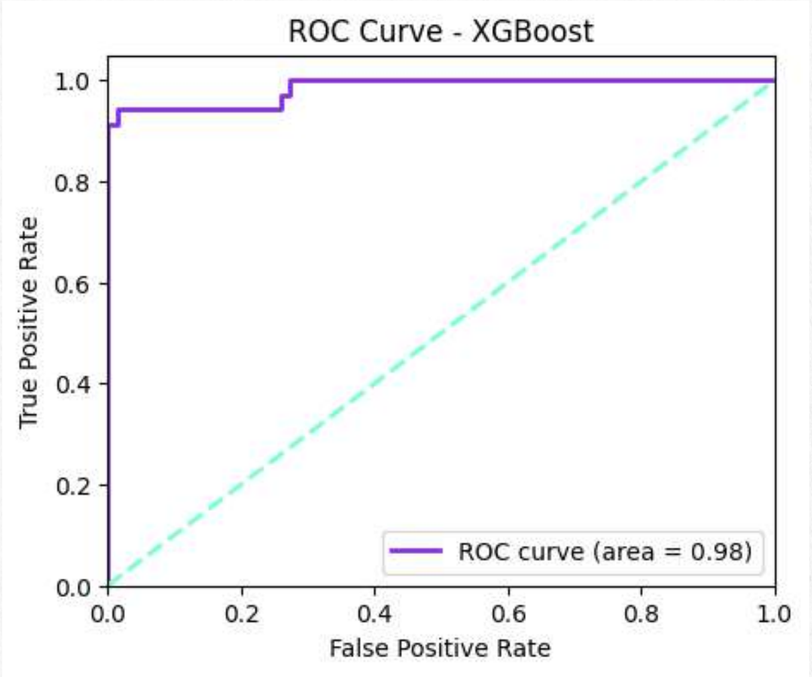
Hiper Parametre Ayarlaması Sonrası...

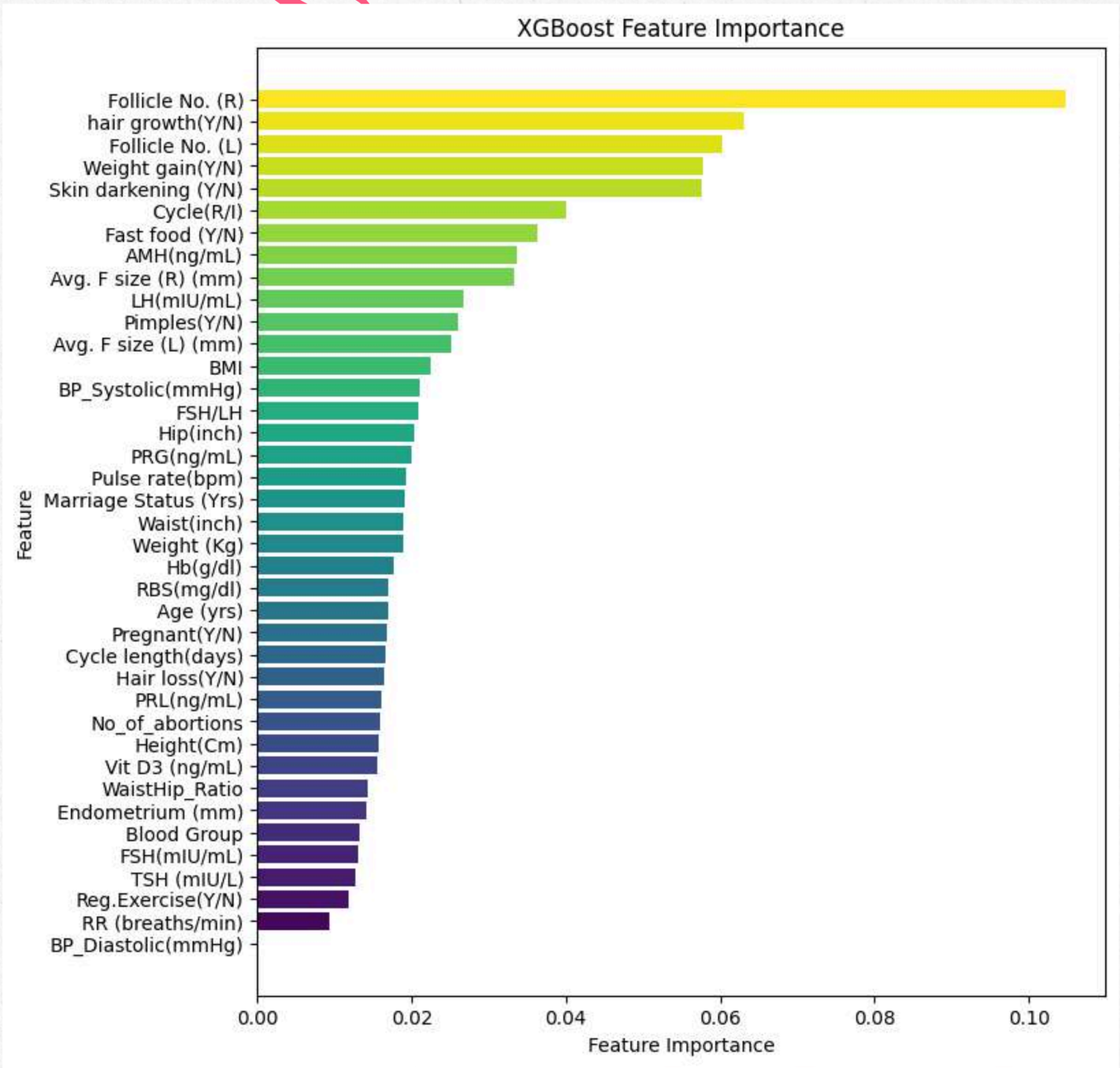
- Parametre araması için **GridSearchCV** kullanıldı.
- Çapraz doğrulama için **StratifiedKFold** kullanıldı. (n_splits=5)

En İyi Parametreler
'eta': 0.05
'max_depth': 3
'n_estimators': 150
'subsample': 0.8

	Model	Train Accuracy	Test Accuracy	F1-Score	AUC Score	Precision	Recall
0	SVM	1.000000	0.916667	0.913943	0.955382	0.920833	0.916667
1	Random Forest	0.997685	0.953704	0.953526	0.983953	0.953534	0.953704
2	XGBoost	0.990741	0.972222	0.971880	0.984344	0.973319	0.972222


Confusion Matrix- XGBoost
73
0
3
32
True Values
Predicted
0
1
70
60
50
40
30
20
10
0


ROC Curve - XGBoost
True Positive Rate
False Positive Rate
ROC curve (area = 0.98)
0.0
0.2
0.4
0.6
0.8
1.0


XGBoost Feature Importance
Feature
Feature Importance
Follicle No. (R)
hair growth(Y/N)
Follicle No. (L)
Weight gain(Y/N)
Skin darkening (Y/N)
Cycle(R/I)
Fast food (Y/N)
AMH(ng/mL)
Avg. F size (R) (mm)
LH(mIU/mL)
Pimples(Y/N)
Avg. F size (L) (mm)
BMI
BP_Systolic(mmHg)
FSH/LH
Hip(inch)
PRG(ng/mL)
Pulse rate(bpm)
Marriage Status (Yrs)
Waist(inch)
Weight (Kg)
Hb(g/dl)
RBS(mg/dl)
Age (yrs)
Pregnant(Y/N)
Cycle length(days)
Hair loss(Y/N)
PRL(ng/mL)
No_of_abortions
Height(Cm)
Vit D3 (ng/mL)
WaistHip_Ratio
Endometrium (mm)
Blood Group
FSH(mIU/mL)
TSH (mIU/L)
Reg.Exercise(Y/N)
RR (breaths/min)
BP_Diastolic(mmHg)
0.00
0.02
0.04
0.06
0.08
0.10

TEŞEKKÜRLER

LEYLA TÜLÜ