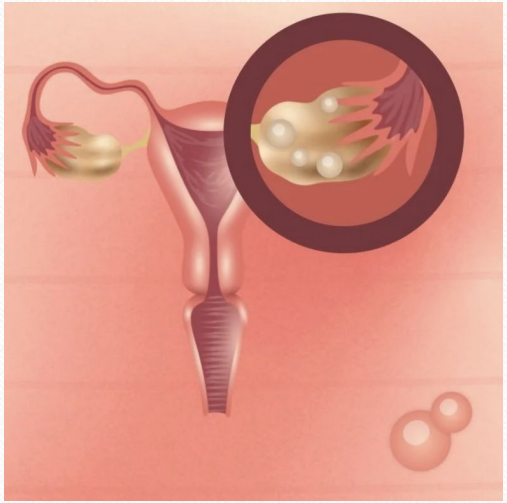
POLYCYSTIC OVARY SYNDROME PREDICTION USING CLASSIFICATION MODELS



BY:
SANDHYA NAGARAJAN IYER
SAKSHI RATHI
JAYESH PRASAD ANANDAN

CONTENTS

- Problem Statement
- Dataset Description
- Hormonal and Phenotypical factor description
- Data Model
  - Data Preprocessing
  - EDA - visualisation analysis
  - PCA
  - Prediction Models
- Inference

PROBLEM STATEMENT

- In our project, we are trying to detect if a patient has PCOS using given phenotypical and hormonal features
- PCOS is a hormonal disorder for women causing enlarged ovaries and small cysts on the outer layer.
- PCOS results in high levels of androgen in women, plus a lot of abnormal hormonal levels resulting in irregular menstrual periods, excess hair growth, acne, infertility, and weight gain.

DATASET

Polycystic ovary syndrome (PCOS)

Data was collected from 10 different hospitals across Kerala, India

The dataset contains all physical and clinical parameters to determine PCOS and infertility related issues.

The dataset contains 541 rows x 45 columns

For every Yes/No, Yes = 1 ; No= 0
Blood Group indications:
A+ = 11
A- = 12
B+ = 13
B- = 14
O+ =15
O- = 16
AB+ =17
AB- = 18
Blood pressure entered as systolic and diastolic separately
RBS means Random glucose test
Beta-HCG cases are mentioned as Case I and II.

Hormonal and Phenotypical factor description

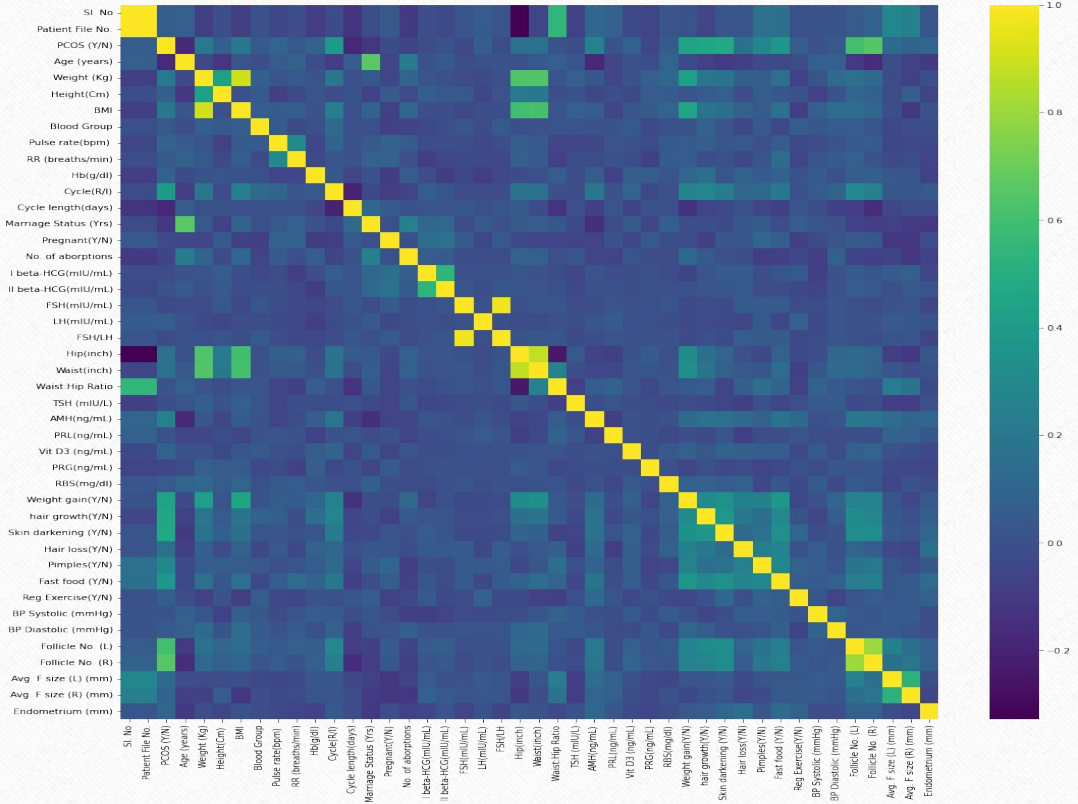
Hormone	Name	Normal Range in an Adult Female
I beta-HCG(mIU/mL)	human chorionic gonadotropin	Nonpregnant: less than 5 mIU/mL Pregnant: above 25 mIU/mL
II beta-HCG(mIU/mL)	human chorionic gonadotropin	Nonpregnant: less than 5 mIU/mL Pregnant: above 25 mIU/mL
FSH(mIU/mL)	Follicle-stimulating hormone	1.4-9.9 mIU/mL during the first half of the menstrual cycle Rise up to 17.2 mIU/mL during ovulation During pregnancy, FSH levels drop to ≤ 0.1 mIU/mL
LH(mIU/mL)	Luteinizing hormone	Weeks one and two of the menstrual cycle: 1.37 to 9 IU/L Week two, before ovulation: 6.17 to 17.2 IU/L. Weeks three and four of the menstrual cycle: 1.09 to 9.2 IU/L.
TSH (mIU/L)	Thyroid-stimulating hormone	0.5 to 5.0 mIU/L
AMH(ng/mL)	Anti-Müllerian hormone	1.5 – 4.0 ng/ml
PRL(ng/mL)	Prolactin	Nonpregnant women: less than 25 ng/mL Pregnant women: 80 to 400 ng/mL
Vit D3 (ng/mL)	Vitamin D3 Test	between 20 and 40 ng/mL
PRG(ng/mL)	Progesterone Test	Female (mid-cycle): 5 to 20 ng/mL Pregnancy 1st trimester: 11.2 to 90.0 ng/mL
RBS(mg/dl)	Random Blood Sugar Test	less than 140 mg/dL

IMPORTANT FACTORS AFFECTED BY PCOS

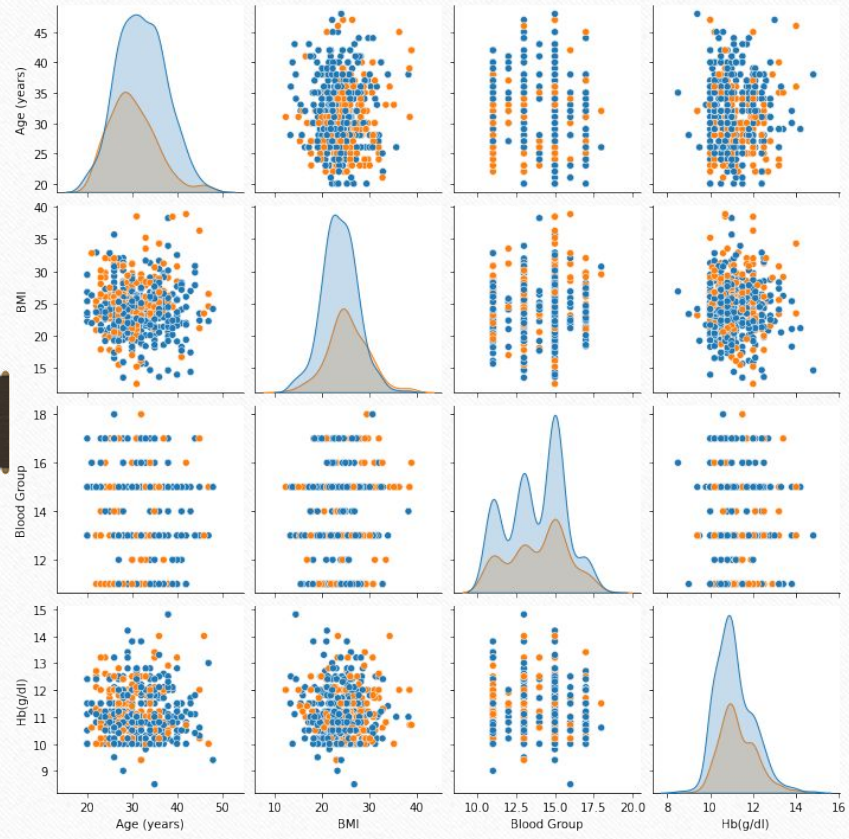
- Follicle stimulating hormone(FSH) and Luteinizing hormone(LH)
  - The ratio between LH and FSH increases to around 2 or 3 resulting in ovulation issues
- Thyroid-stimulating hormone(TSH)
  - To avoid problems like underactive or overactive thyroid resulting in irregular periods
- Anti-Müllerian hormone(AMH)
  - PCOS results in increased AMH levels which may stop

Exploratory Data Analysis

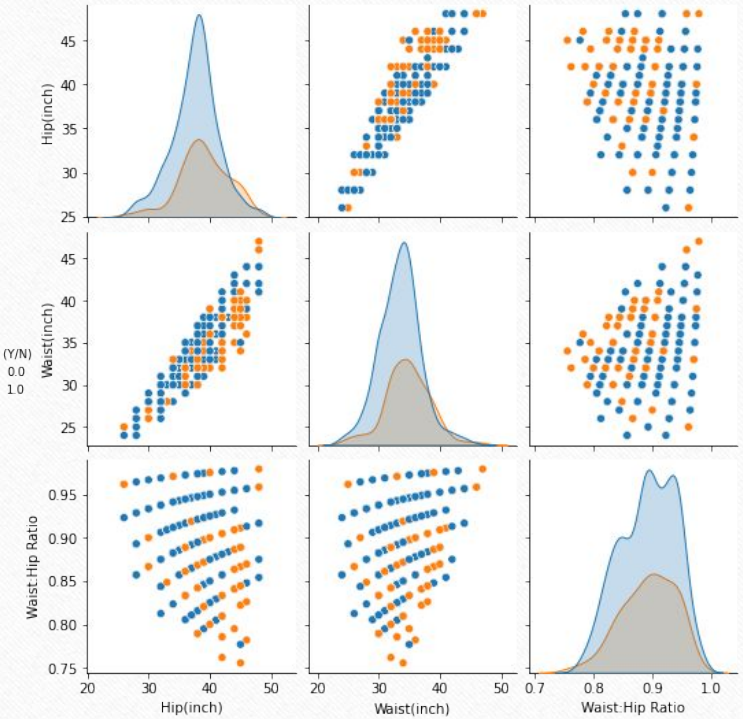
HEAT MAP FOR THE WHOLE DATASET



PAIRPLOTS FOR GROUP OF FEATURES

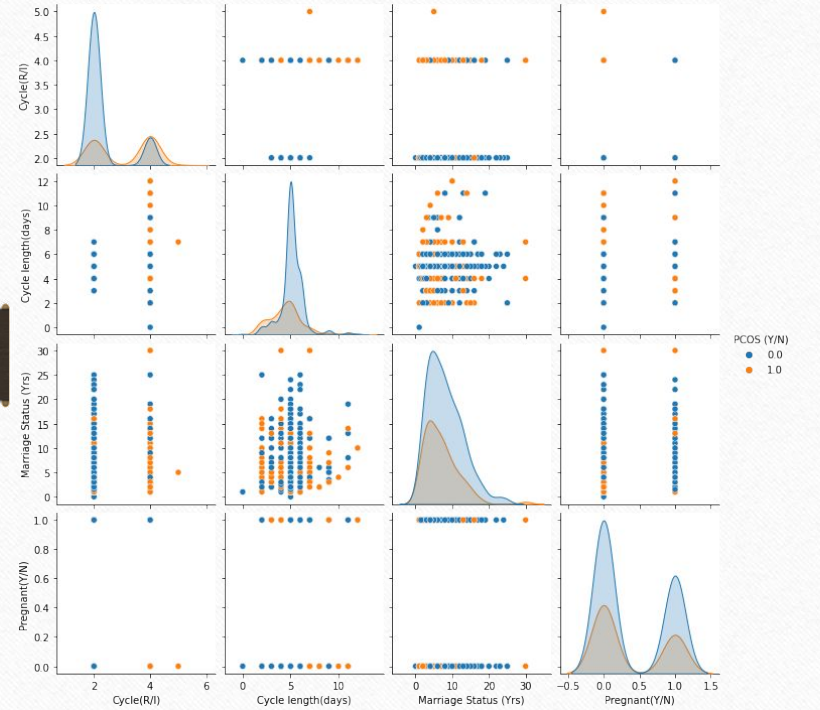


Trends of Phenotypes

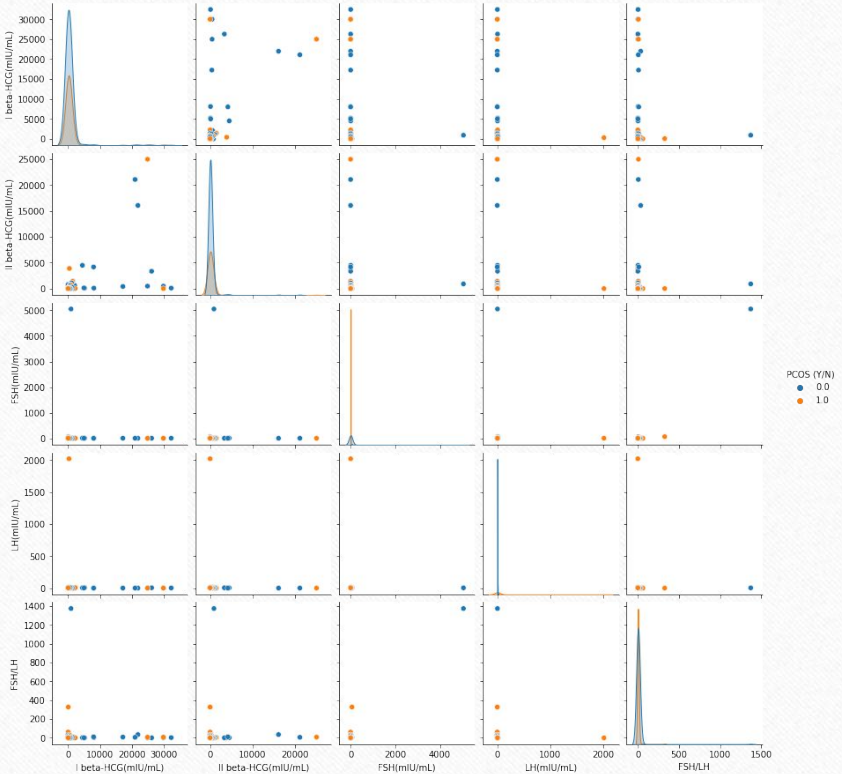


Trends of Hip and Waist Measurement

PAIRPLOTS FOR GROUP OF FEATURES

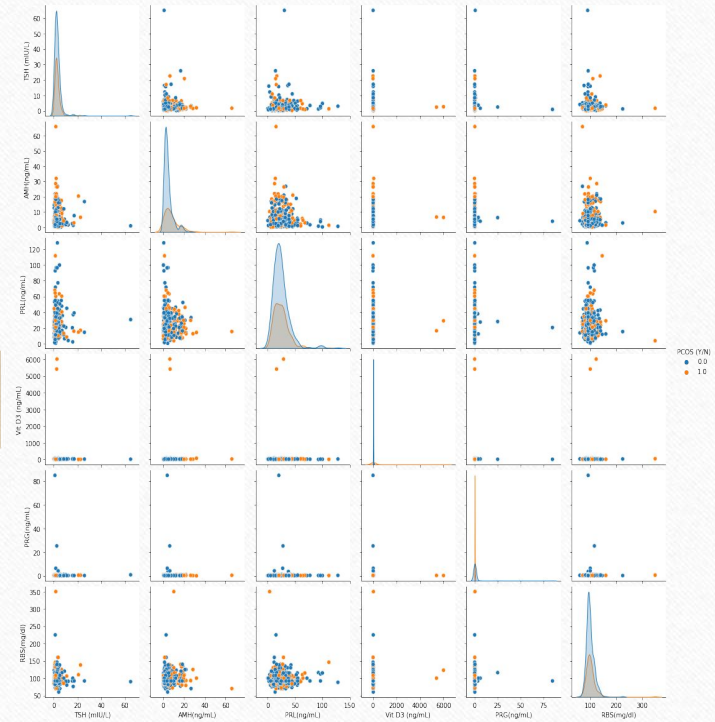


Trends of Menstrual Cycle and Pregnancy

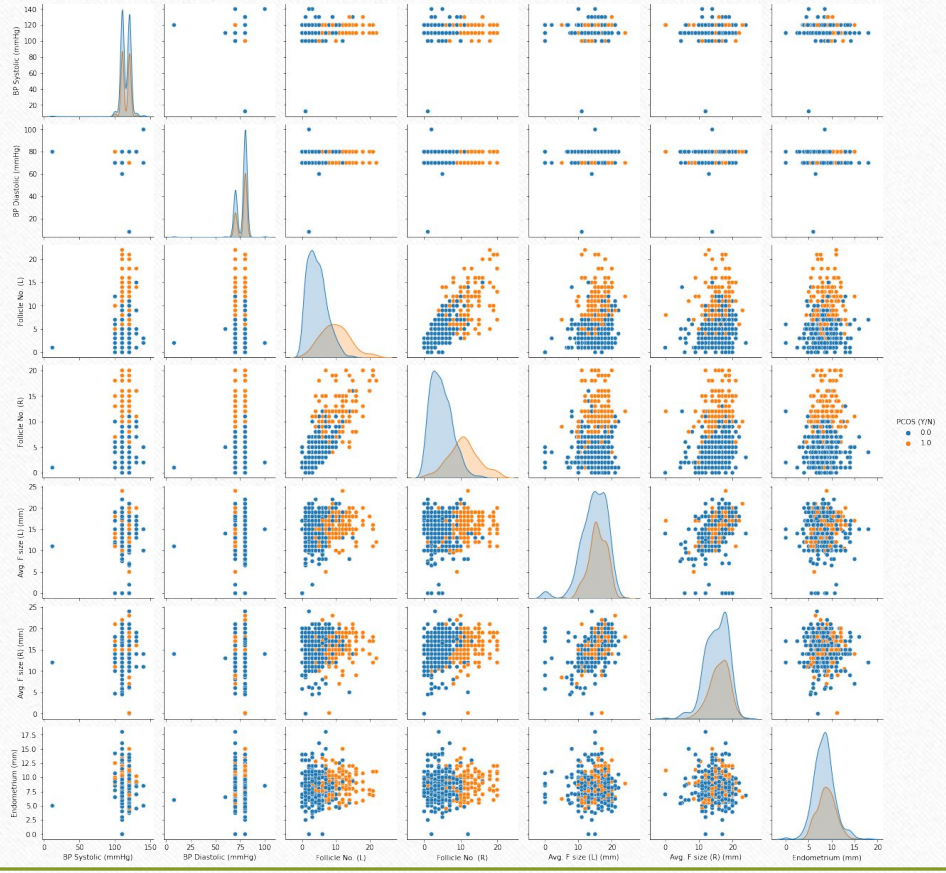


Trends of Hormone Levels

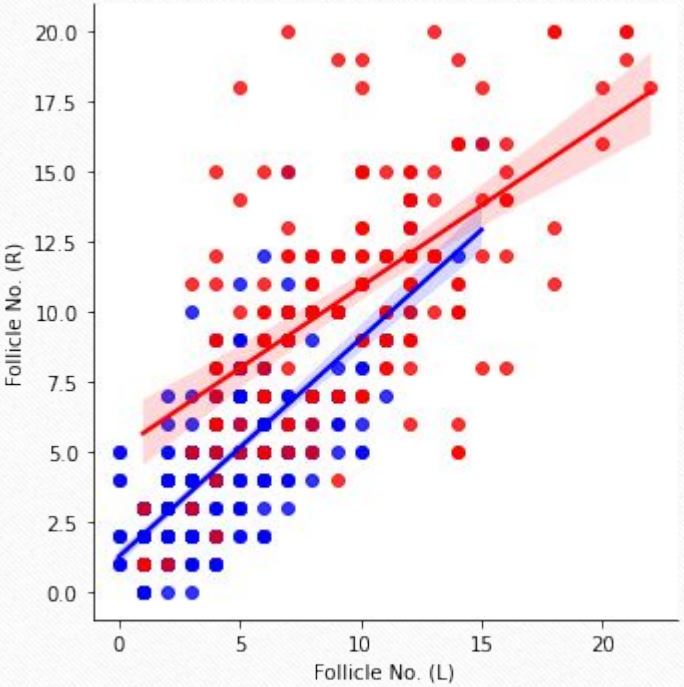
PAIRPLOTS FOR GROUP OF FEATURES



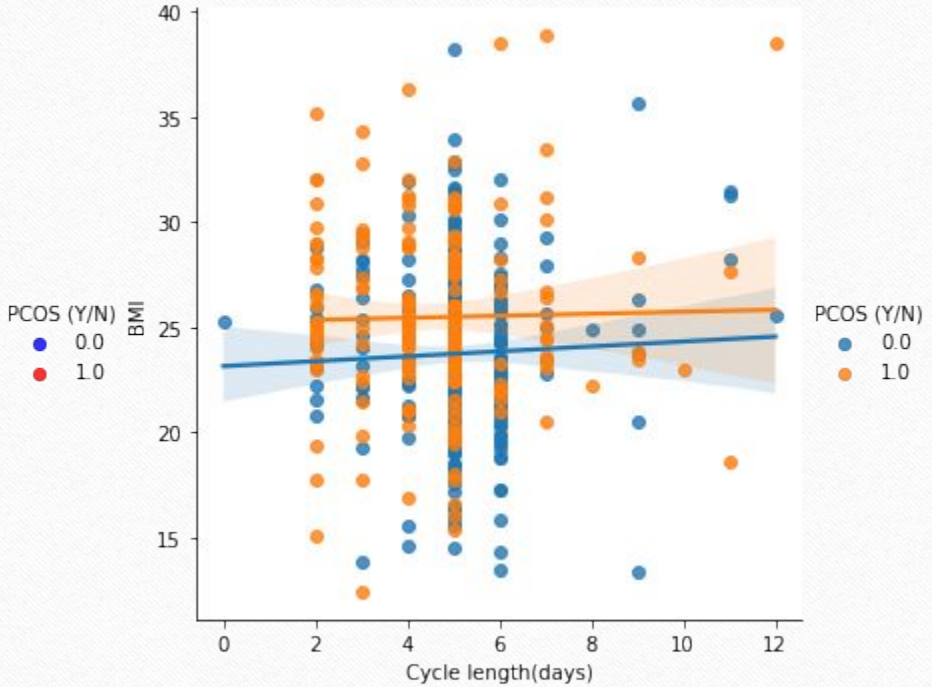
Trends of Hormonal and Blood Sugar Levels



Trends of Blood Pressure and Follicle Measurements

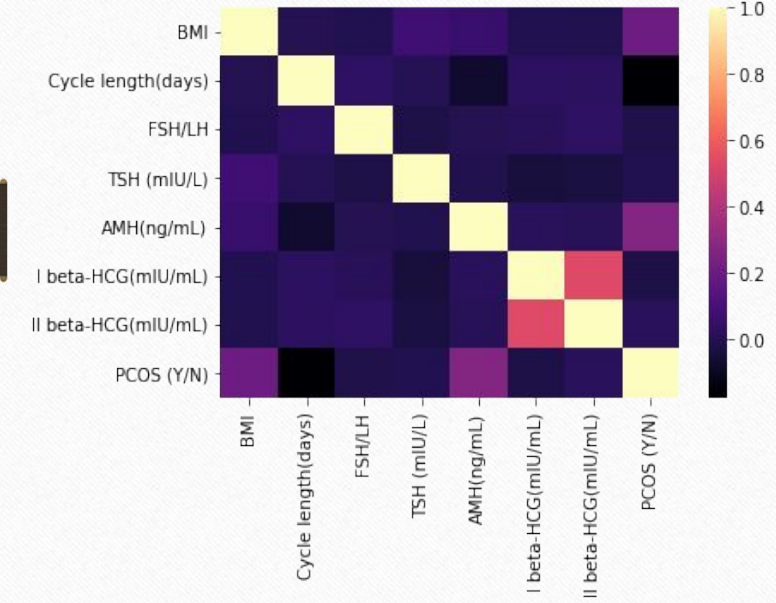


Analysis of No. of Follicles present

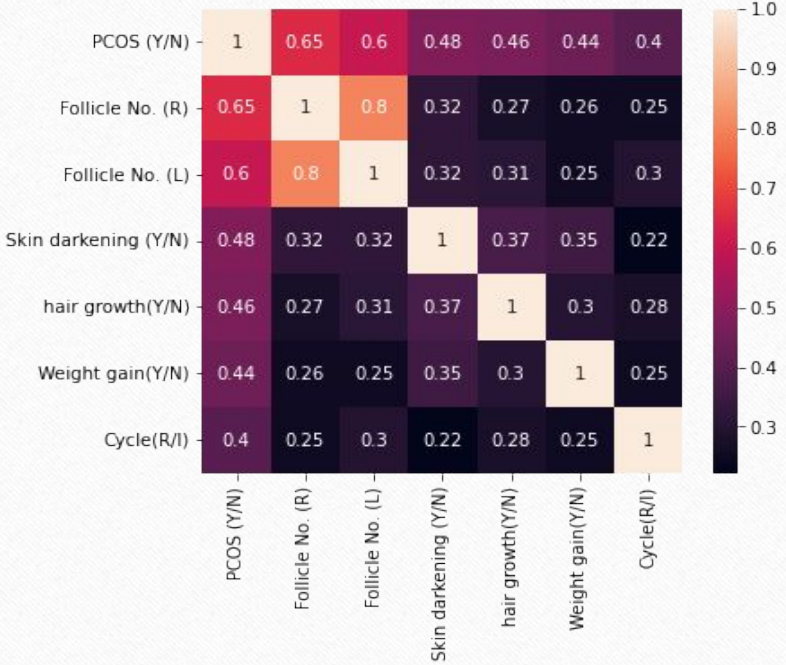


Analysis of Cycle Length vs BMI

HEAT MAP FOR IMPORTANT FACTORS RESULTING IN PCOS:



HEAT MAP FOR HIGHLY CORRELATED FACTORS RESULTING IN PCOS:



Principal Component Analysis (PCA)

Implemented PCA on the dataset and observed the following:

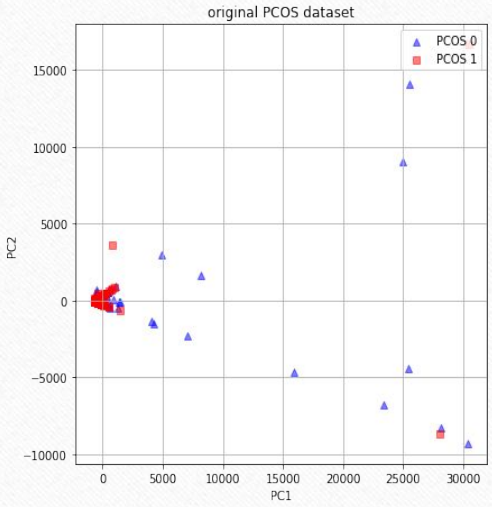
Using 2 Components:

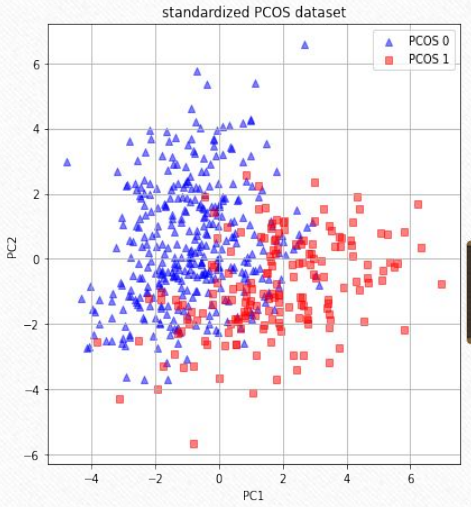
Explainable variance = 0.86186924 0.12186443

Total Variance Explained = **98.3733 %**

singular_values = 80765.10016782
30369.77196088

Prediction accuracy for the standardized test dataset with PCA and Random Forest : **86.84%**





PREDICTION MODELS

Ensemble Learning Methods used for classifying and predicting PCOS:

- RandomForest
- Decision Tree
- Bagging Classifier
- ADABoost Classifier

BAGGING CLASSIFIER

Estimators = 1500

Training Results:

CONFUSION MATRIX:

```
[[233   0]
 [  0 118]]
```

ACCURACY SCORE:

1.0000

CLASSIFICATION REPORT:

	0.0	1.0	accuracy	macro avg	weighted avg
precision	1.0	1.0	1.0	1.0	1.0
recall	1.0	1.0	1.0	1.0	1.0
f1-score	1.0	1.0	1.0	1.0	1.0
support	233.0	118.0	1.0	351.0	351.0

Testing Results:

CONFUSION MATRIX:

```
[[119  12]
 [ 11  48]]
```

ACCURACY SCORE:

0.8789

CLASSIFICATION REPORT:

	0.0	1.0	accuracy	macro avg	weighted avg
precision	0.915385	0.800000	0.878947	0.857692	0.879555
recall	0.908397	0.813559	0.878947	0.860978	0.878947
f1-score	0.911877	0.806723	0.878947	0.859300	0.879224
support	131.000000	59.000000	0.878947	190.000000	190.000000

BOOST CLASSIFIER

Estimators = 30

Training Results:

CONFUSION MATRIX:

```
[[230   3]
 [  6 112]]
```

ACCURACY SCORE:

0.9744

CLASSIFICATION REPORT:

	0.0	1.0	accuracy	macro avg	weighted avg
precision	0.974576	0.973913	0.974359	0.974245	0.974353
recall	0.987124	0.949153	0.974359	0.968139	0.974359
f1-score	0.980810	0.961373	0.974359	0.971092	0.974276
support	233.000000	118.000000	0.974359	351.000000	351.000000

Testing Results:

CONFUSION MATRIX:

```
[[120  11]
 [ 12  47]]
```

ACCURACY SCORE:

0.8789

CLASSIFICATION REPORT:

	0.0	1.0	accuracy	macro avg	weighted avg
precision	0.909091	0.810345	0.878947	0.859718	0.878428
recall	0.916031	0.796610	0.878947	0.856320	0.878947
f1-score	0.912548	0.803419	0.878947	0.857983	0.878660
support	131.000000	59.000000	0.878947	190.000000	190.000000

RANDOM FOREST CLASSIFIER

Estimators = 1000

Training Results:

CONFUSION MATRIX:

```
[[233   0]
 [  0 118]]
```

ACCURACY SCORE:

1.0000

CLASSIFICATION REPORT:

	0.0	1.0	accuracy	macro avg	weighted avg
precision	1.0	1.0	1.0	1.0	1.0
recall	1.0	1.0	1.0	1.0	1.0
f1-score	1.0	1.0	1.0	1.0	1.0
support	233.0	118.0	1.0	351.0	351.0

Testing Results:

CONFUSION MATRIX:

```
[[125   6]
 [ 11  48]]
```

ACCURACY SCORE:

0.9105

CLASSIFICATION REPORT:

	0.0	1.0	accuracy	macro avg	weighted avg
precision	0.919118	0.888889	0.910526	0.904003	0.909731
recall	0.954198	0.813559	0.910526	0.883879	0.910526
f1-score	0.936330	0.849558	0.910526	0.892944	0.909385
support	131.000000	59.000000	0.910526	190.000000	190.000000

INFERENCE

- The Random Forest based model gives a us a decent accuracy of 91%, we can further choose specific features which contribute more towards PCOS prediction, to build a better model and more data will also help us in improving the accuracy of our model.
- From the feature selection and EDA, we can infer that features like Follicle number, weight gain, cycle (R/I), skin darkening and hair growth have a relatively higher correlation coefficient and contribute the most towards prediction of PCOS.