POLYCYSTIC OVARY SYNDROME PREDICTION USING CLASSIFICATION MODELS

Authors:

SANDHYA NAGARAJAN IYER
SAKSHI RATHI
JAYESH PRASAD ANANDAN

CSCI-B 565 DATA MINING

Prof. Yuzhen Ye

12th December 2022

INDIANA UNIVERSITY
BLOOMINGTON

CONTENTS

1. ABSTRACT 3
2. INTRODUCTION 4
3. METHODOLOGY 6
    - I. Data Preprocessing 6
    - II. Exploratory Data Analysis 6
    - III. Feature Extraction 15
    - IV. Classification 17
4. RESULTS 18
5. DISCUSSION 21
6. REFERENCES 21

ABSTRACT

To diagnose whether an individual is affected by Polycystic Ovary Syndrome, commonly known as PCOS, based on the data collected from 10 different hospitals in Kerala, India. The dataset is available in Kaggle and is used for our project.

Polycystic ovary syndrome (PCOS) is a condition in which the ovaries produce an abnormal amount of androgens, male sex hormones that are usually present in women in small amounts. The name polycystic ovary syndrome describes the numerous small cysts (fluid-filled sacs) that form in the ovaries. However, some women with this disorder do not have cysts, while some women without the disorder do develop cysts. The exact cause of PCOS is not clear. Many women with PCOS have insulin resistance. This means the body can't use insulin well. Insulin levels build up in the body and may cause higher androgen levels. Obesity can also increase insulin levels and make PCOS symptoms worse. PCOS may also run in families. It's a common syndrome seen in women to have PCOS.

Keywords:

Prediction, Classification, Exploratory Data Analysis, Pairplots, Scatter Plots,Data Preprocessing, Feature Extraction, PCOS, Random Forest, Bagging Classifier, Adaboost classifier, estimators, Follicle number, accuracy, diagnose, dataset, phenotypes, hormones, features, disorder, duplicate values , null values , Principal Component Analysis, cysts, androgen, symptoms, heatmap, physical, attributes

INTRODUCTION

Polycystic ovary syndrome is a common hormonal problem in women. Women with this condition may not be able to ovulate and may have high levels of androgen causing a lot of physical and hormonal imbalance in the body. With PCOS, the ovaries are surrounded by small sacs of fluid called cysts. These cysts may not develop in every case. These cysts contain immature eggs called follicles which in turn do not produce eggs regularly causing ovulation issues.

The main aim of the project is to determine if a given patient has PCOS or not using phenotypical and hormonal data collected from 10 hospitals in Kerala.

PCOS can be detected using certain irregularities seen in this phenotypical and hormonal data. In this dataset we are using major hormones like Follicle stimulating hormone and phenotypes like weight gain and skin darkening which are majorly seen physical changes in women with PCOS.

As mentioned above, PCOS results in high levels of androgen in the body. Androgens are a group of sex hormones which triggers puberty and ensures the role of reproductive health. So, testosterone is a popularly known androgens. This hormone is the reason for facial and body hair, voice deepening, Adam's Apple etc in men. So, when a woman has more androgen in her body than to produce estrogen it results in male like phenotypical changes as well as issues with her menstrual cycle. It may also be a cause for weight gain and insulin resistance.

The next two most important hormones that show irregularity in PCOS are Follicle Stimulating Hormone (FSH) and Luteinizing Hormone (LH). These two hormones are produced by the pituitary gland found in the brain.These hormones are produced and sent to the ovaries resulting in a surge of LH which is an ovulation

signal in women. Whereas, in women with PCOS since there is high LH in the body already, there is no LH surge happening leading to no egg being released.
This creates a major hormonal disruption and interferes with follicle development. Therefore, the women may have an irregular menstruation cycle and no ovulation.

It is seen normally that the ratio between LH and FSH is seen to be between 1 to 2 whereas in women with PCOS it is seen to be in the range of 2 to 3 or more which is not good.

Then we see the Anti-Müllerian hormone (AMH). This hormone develops antral follicles on the ovary helping it to finally mature and potentially release an egg. Since it controls the early egg too much of this hormone may result in no ovulation. The normal range is 1.5 – 4.0 ng/ml.

So, using these major hormonal attributes we find that the physical attributes like BMI, Age, Weight, Height, Cycle length etc are affected. Using both these characteristics we use different prediction models to determine if the patient has PCOS or not.

METHODOLOGY

Dataset: The PCOS dataset 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 2 dataset samples without infertility and with fertility.m Data samples without infertility contains 540 samples with 45 columns and data with infertility contains 540 samples and 6 columns. So, we merge these two datasets to build our model.

I. Data Preprocessing:

Renaming the columns: Renaming the column names that weren't formatted correctly. Example: 'II beta-HCG(mIU/mL) to :'II beta-HCG(mIU/mL)' , 'Marraige Status (Yrs)': 'Marriage Status (Yrs)'.

Encoding categorical variables: Real data can be categorical string values, but since the machine learning model performs mathematical operations, we need to convert categorical data to integer format to perform predictions. We use the pd.to_numeric function for this.

Check null values - replace them with median value: Handling missing values appropriately is very important as the machine learning algorithms don't don't support data with missing values and it may lead to drawing incorrect inferences. Here, we replace the null values by the median of the feature column.

Check duplicate values - Duplicate values can result from combining two or more datasets. So, we tend to drop the duplicate values.

II. Exploratory Data Analysis:

To understand the dataset, we have implemented a heatmap to see the correlations between the different features of the dataset. We can observe from the plot that some of the attributes are highly correlated and most of them are not correlated that much.

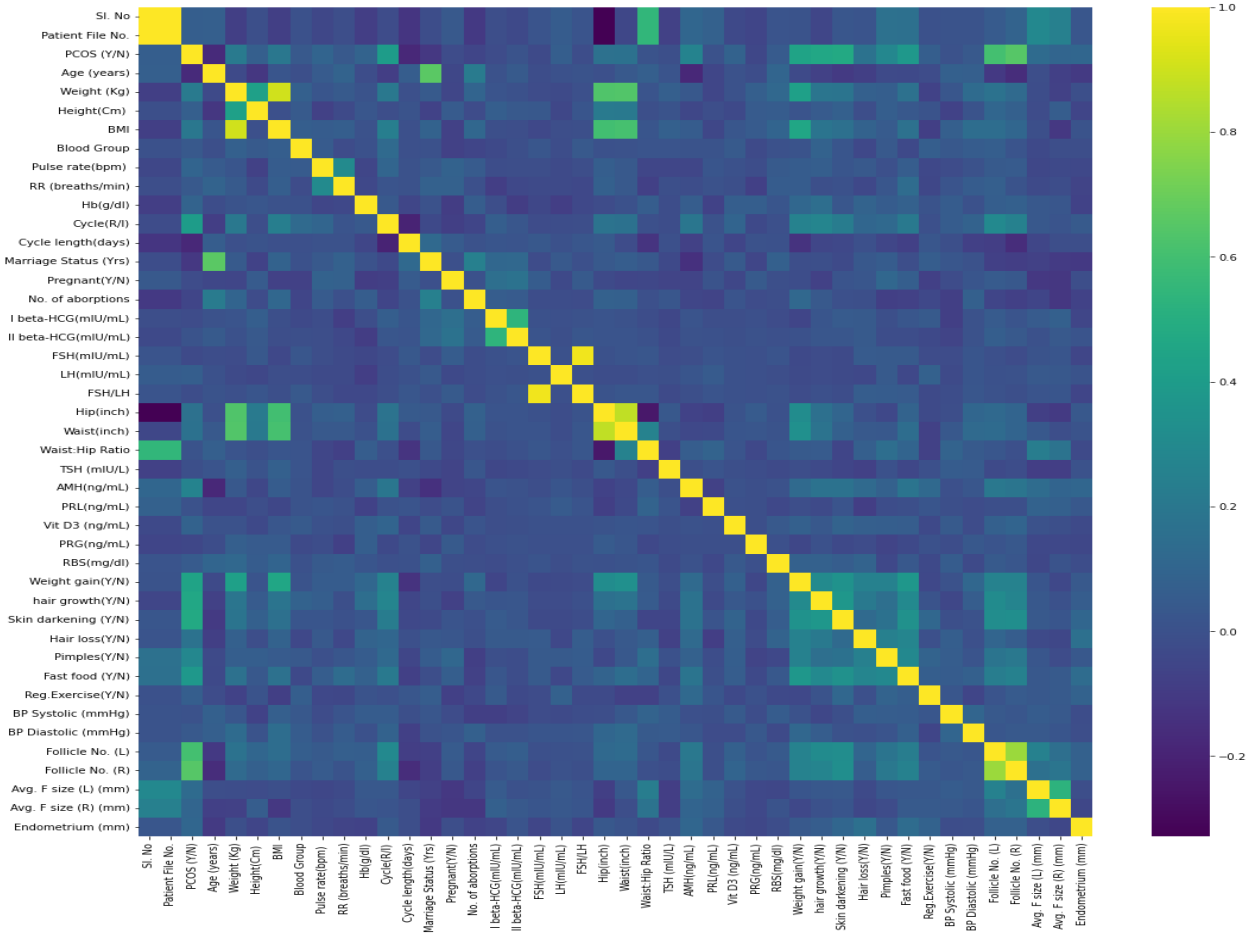

Fig 1. Heatmap for the whole dataset

In the plot, considering the target attribute PCOS(Y/N), we can observe that this column is highly correlated with

1. Cycle(R/I)
2. Weight Gain (Y/N)
3. Hair Growth (Y/N)
4. Skin Darkening (Y/N)
5. Fast Food (Y/N)

Scatterplots between feature columns with hue as PCOS Y/N with x and y as Blood group and marriage status resp - plotting the box plot and swarm plot.

Pairplots :

To analyze the trends between the columns we use pairplots from the seaborn package. We developed pairplots to observe how every attribute is distributed and how they compare with other attributes.

Pairplots of combinations of related features.

1. ‘Age (years)’, ‘BMI’, ‘Blood Group’, ‘Hb(g/dl)’

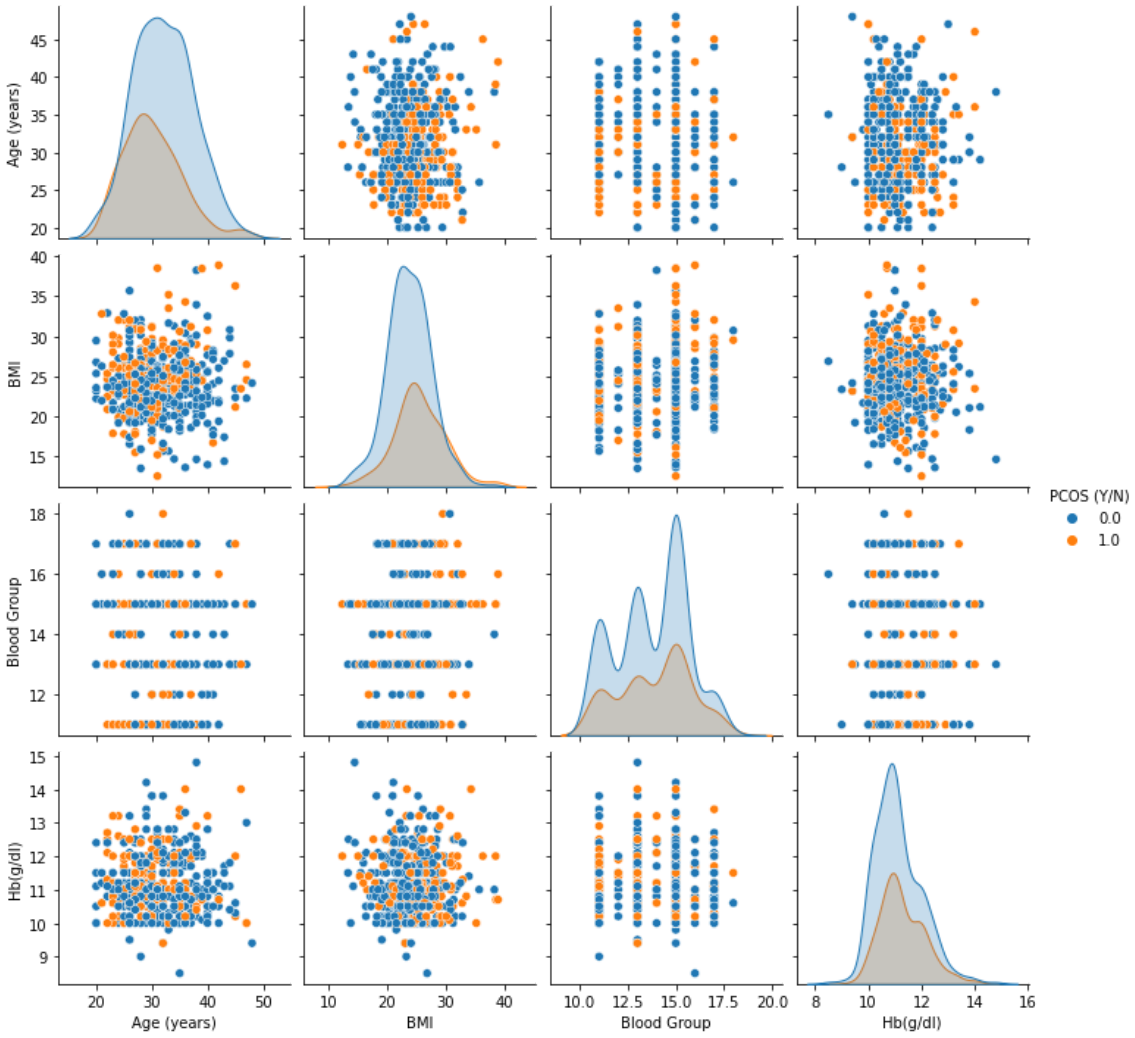

Fig 2.1 Distribution of Age, BMI, Blood Group, Hb(g/dl)

We can observe that the data contains people affected by PCOS is having mean of 35 years age, and those who are not affected ranges above 40 years old, which is typically when they reach menopause stage. The blue points indicate that the individual is not affected by PCOS, and the orange indicates that the individual is affected by PCOS.

2. Hip(inch)','Waist(inch)','Waist:Hip Ratio'

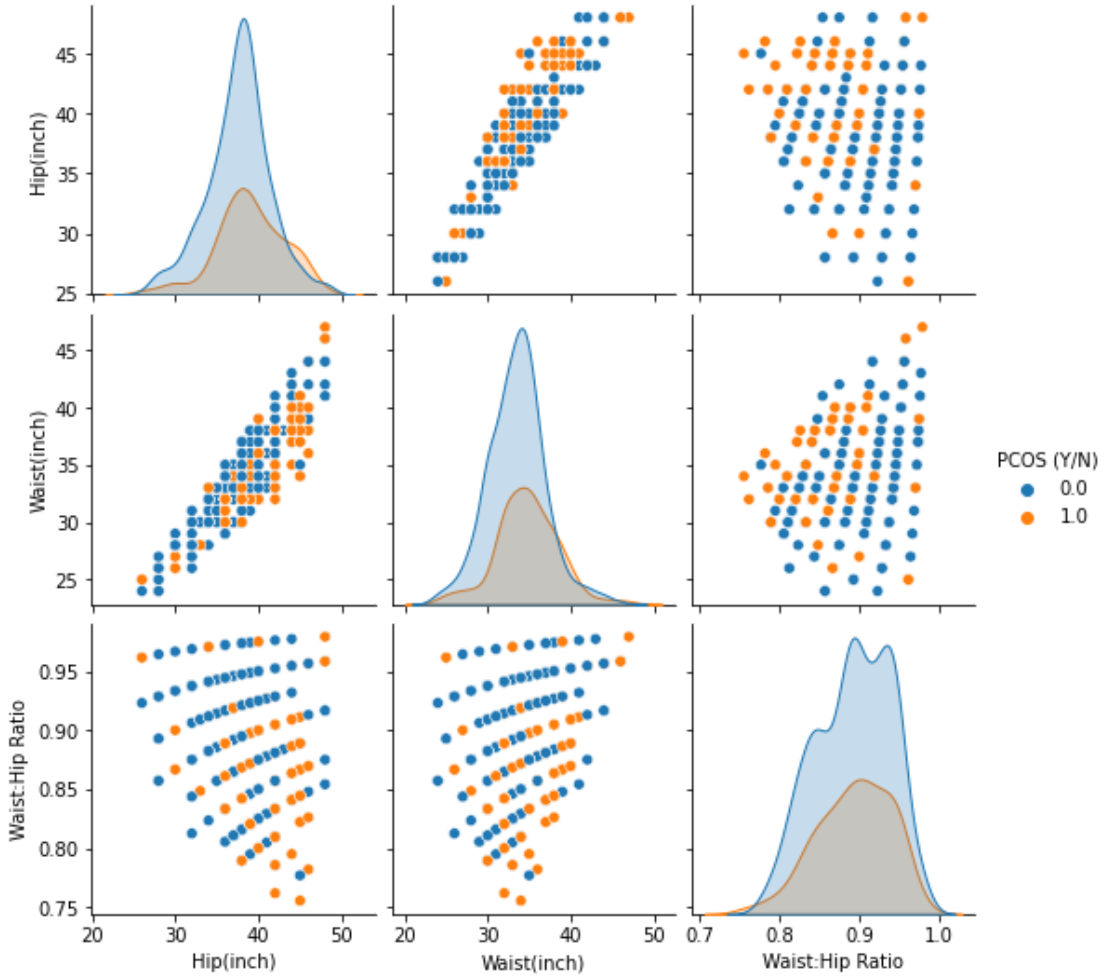

Fig. 2.2 Hip(inch), Waist(inch), Waist:Hip Ratio

We can observe that all the three attributes are normally distributed.

	Hip (inch)	Waist (inch)	Waist Hip Ratio
Mean	37.9926	33.841035	0.89189
Variance	15.7153	12.9137	0.002142

The mean and variance speaks about the distribution of the attributes. We can observe that the Variance of Waist and Hip Ratio is negligible.

3. 'Cycle(R/I)','Cycle length(days)','Marriage Status (Yrs)','Pregnant(Y/N)'

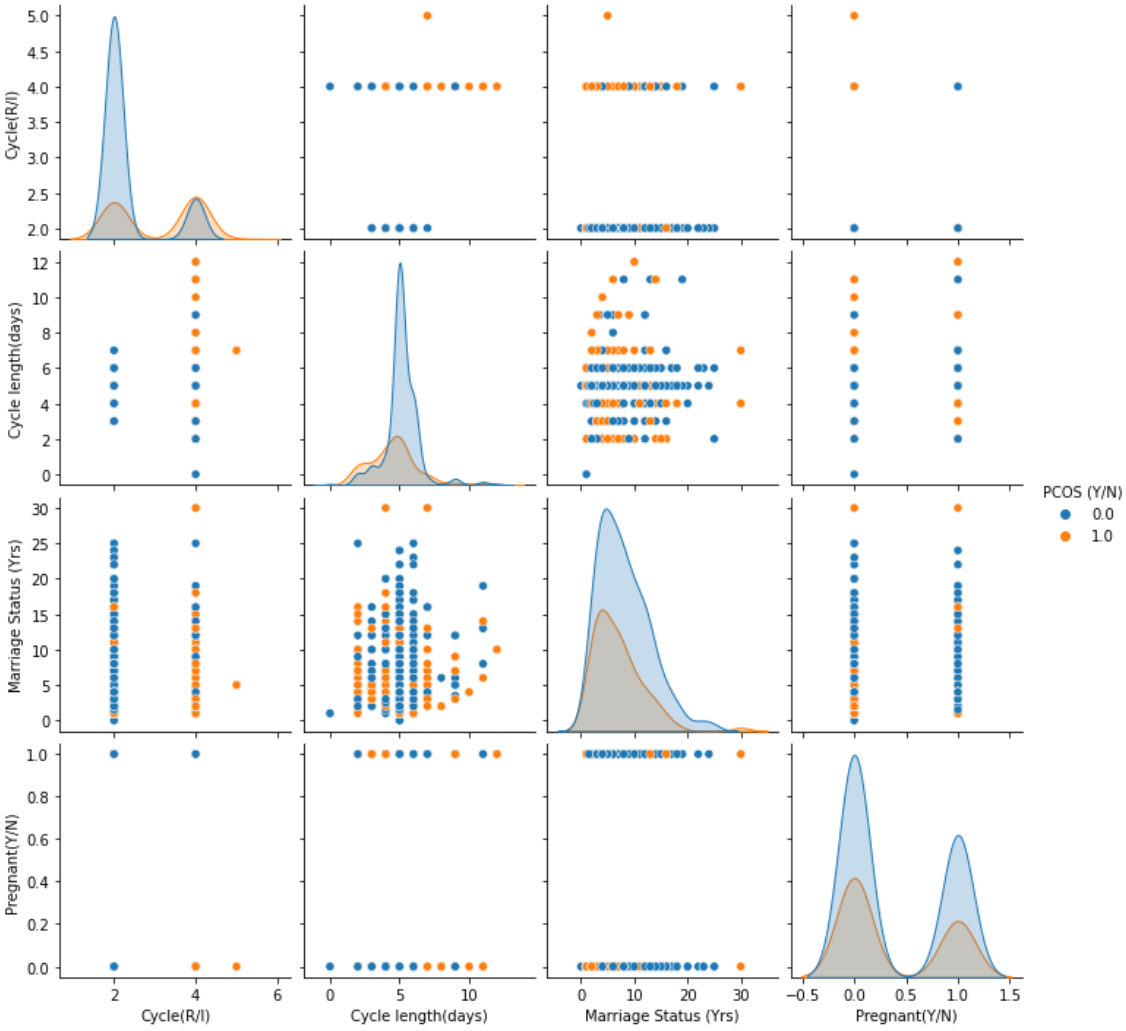

Fig 2.3 Cycle(R/I), Cycle length(days), Marriage Status (Yrs), Pregnant(Y/N)

We can observe from the plot that the Cycle length of individuals with PCOS is low compared to healthy individuals.

Also we can understand that the number of people having PCOS and getting pregnant is lower than women getting pregnant without PCOS. It validates our understanding that PCOS affects women to get pregnant.

4. I beta-HCG(mIU/mL)', 'II beta-HCG(mIU/mL)' , 'FSH(mIU/mL)', 'LH(mIU/mL)', 'FSH/LH'

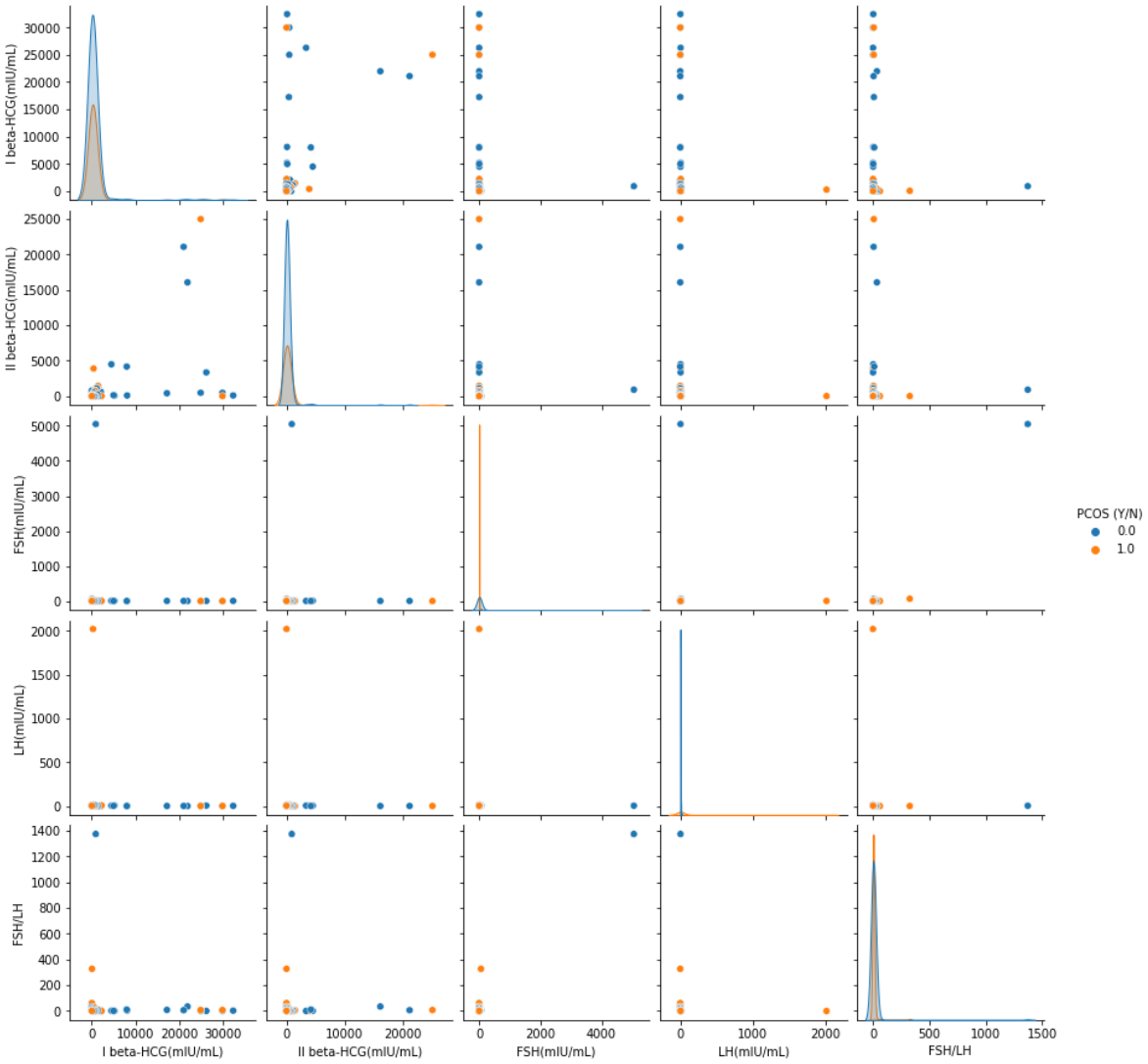

Fig 2.4 Distribution of HCG(mIU/mL), FSH(mIU/mL), LH(mIU/mL) and 'FSH/LH'

We can observe that the Luteinizing hormone (LH) level for PCOS affected individuals is very high. Similarly the Follicle Stimulating hormone (FSH) levels are also very high for PCOS having women.

5. TSH (mIU/L)', 'AMH(ng/mL)', 'PRL(ng/mL)', 'Vit D3 (ng/mL)', 'PRG(ng/mL)', 'RBS(mg/dl)'

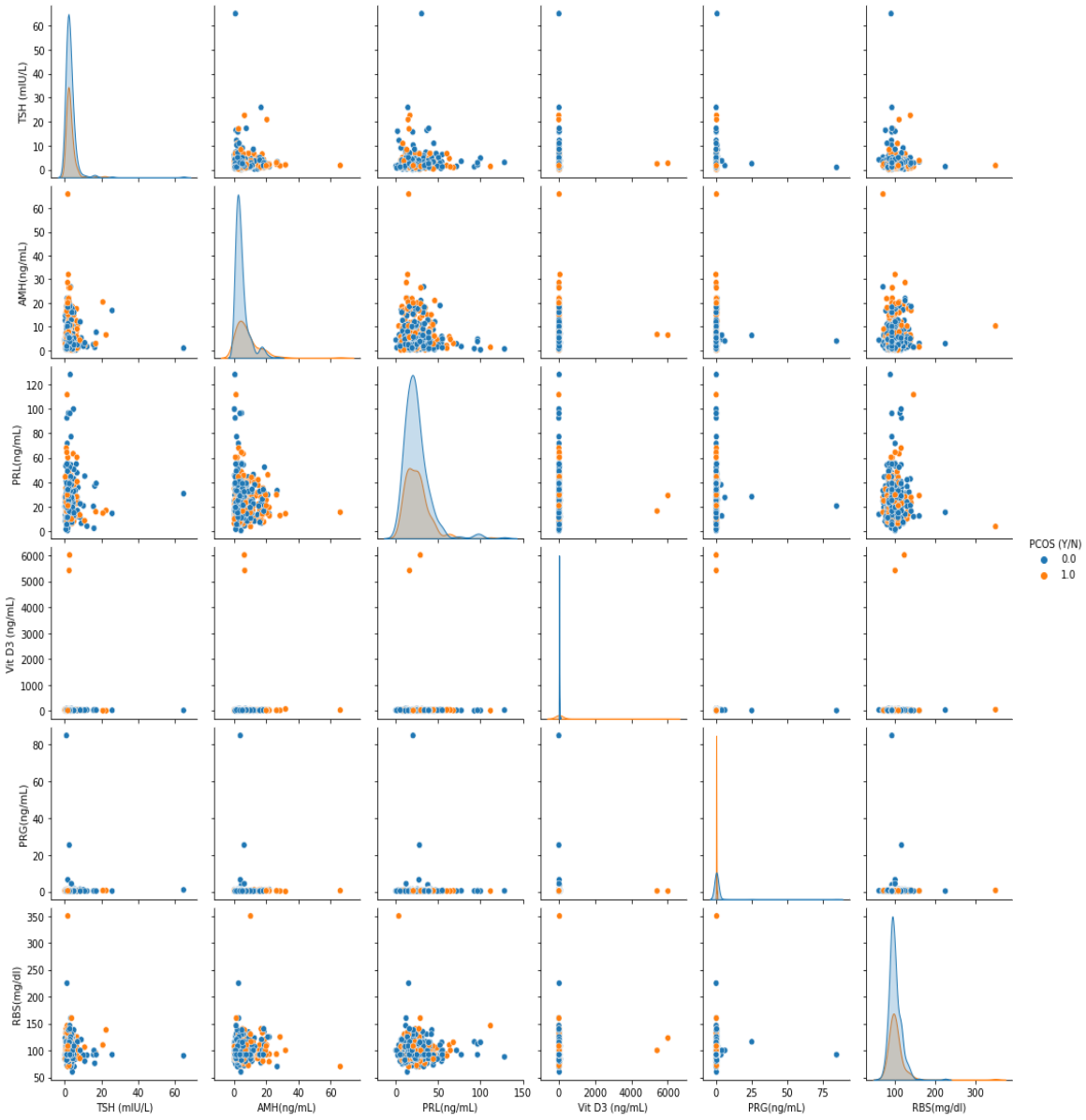

Fig. 2.5 Distribution of TSH (mIU/L)', 'AMH(ng/mL)', 'PRL(ng/mL)', 'Vit D3 (ng/mL)', 'PRG(ng/mL)' and 'RBS(mg/dl)'

It is evident from the plot that the Progesterone (PRG) levels of PCOS having individuals is very high. Also the levels of Vitamin D3 levels in PCOS affected women are higher due to intake of supplements to tackle increase in AMH levels.

6. 'BP Systolic (mmHg)', 'BP Diastolic (mmHg)', 'Follicle No. (L)', 'Follicle No. (R)', 'Avg. F size (L) (mm)', 'Avg. F size (R) (mm)', 'Endometrium (mm)'

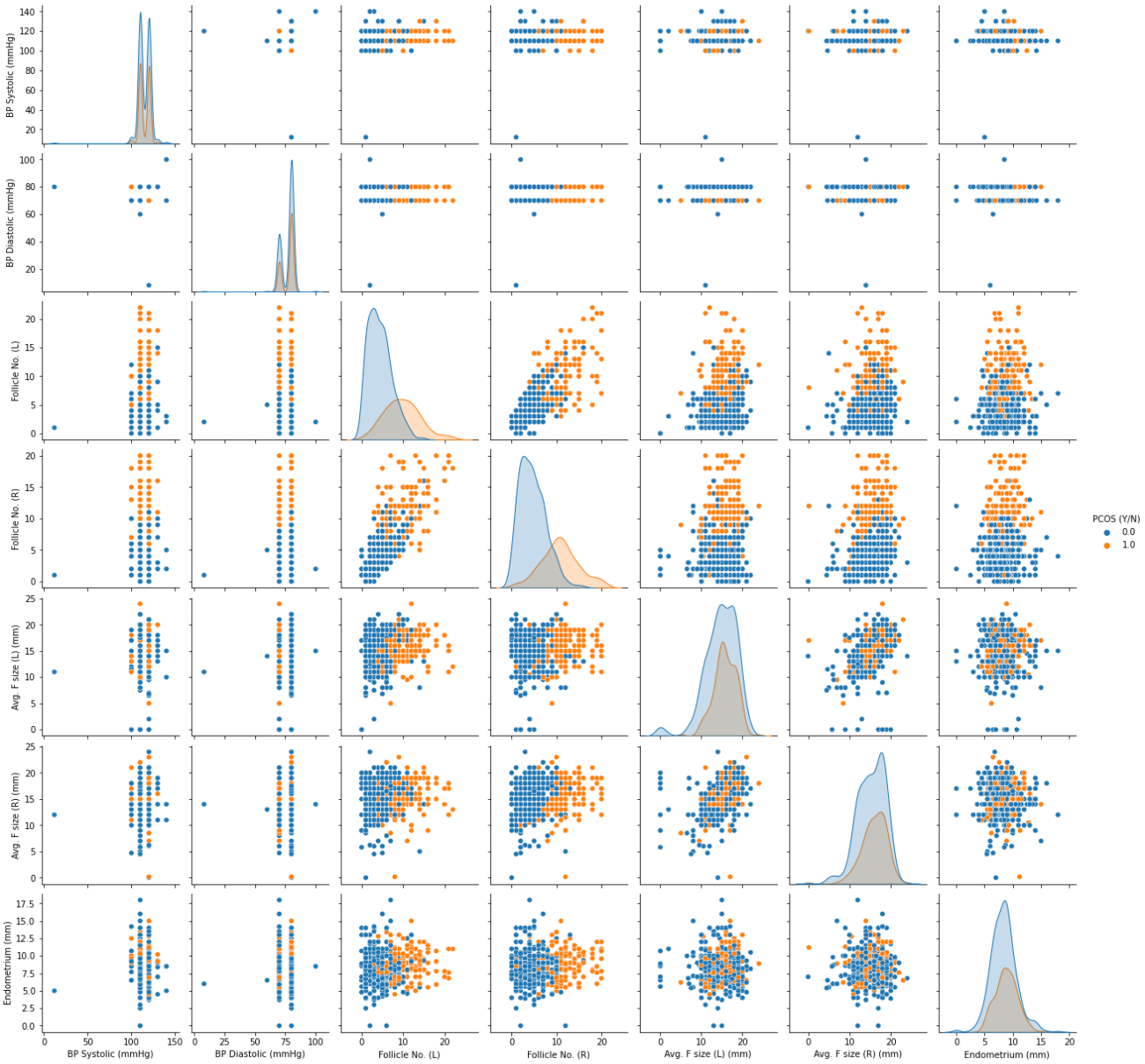

Fig. 2.6 Distribution of BP Systolic (mmHg)', 'BP Diastolic (mmHg)', 'Follicle No. (L)', 'Follicle No. (R)', 'Avg. F size (L) (mm)', 'Avg. F size (R) (mm)', 'Endometrium (mm)'

We can observe that the average size of the follicle is larger in case of PCOS affected individuals. Also the number of follicles present in both Left and Right ovaries are higher due to accumulation of immature follicles in the ovaries.

Principal Component Analysis (PCA):

By analyzing the dataset, We observed that many of the columns involved just binary values which represent True or False. This resulted in highly sparse dataset with high dimensionality. Hence, we used PCA to reduce the dimensionality of the dataset to 2 components.

Comparison of scatter plots of standardized and original dataset after dimension reduction.

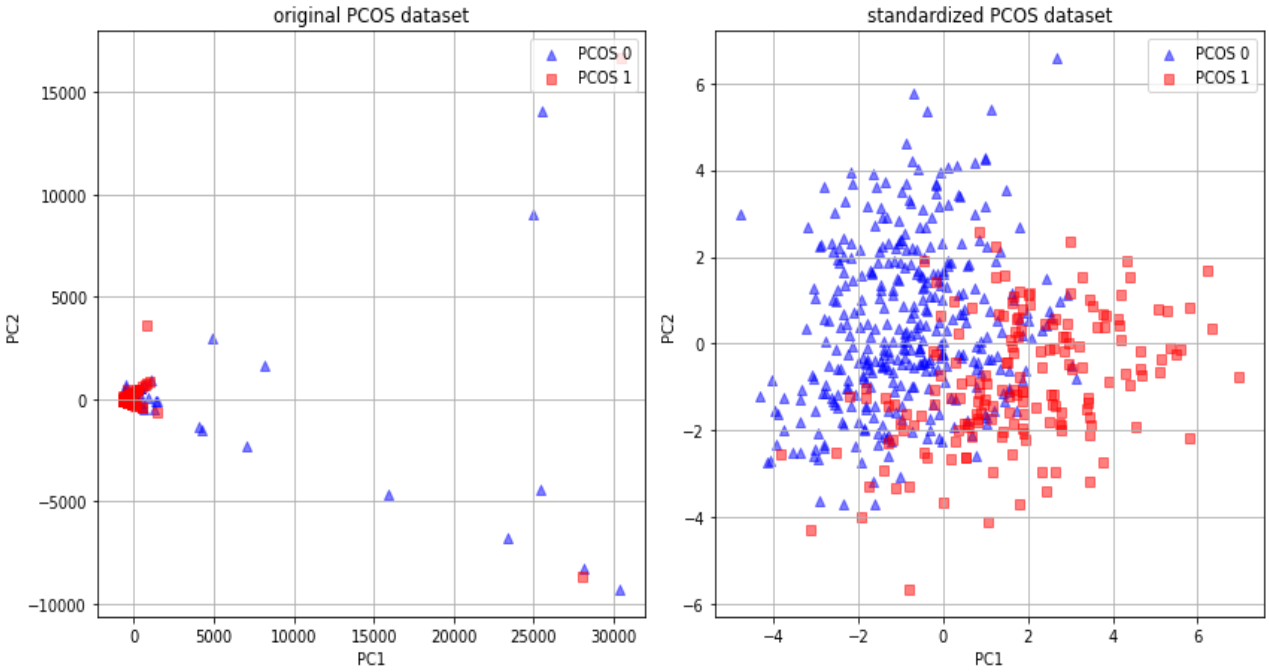

Fig 3 PCA on the dataset with and without standardized dataset with 2 components

The PCOS Classes are distinct and are fairly separable. We can see that our first two principal components explain most of the variance in this dataset (98.37%)! This is an indication of the total information represented compared to the original data. We can further use this standardized data for our prediction models, to predict if an individual has PCOS or not.

III. Feature Extraction

We have around 45 features that can help in determining the detection of PCOS in a woman, for example, age, body mass index, weight, blood pressure, etc. So, we plot a heatmap to visualize data and to find the correlation matrix for the features.

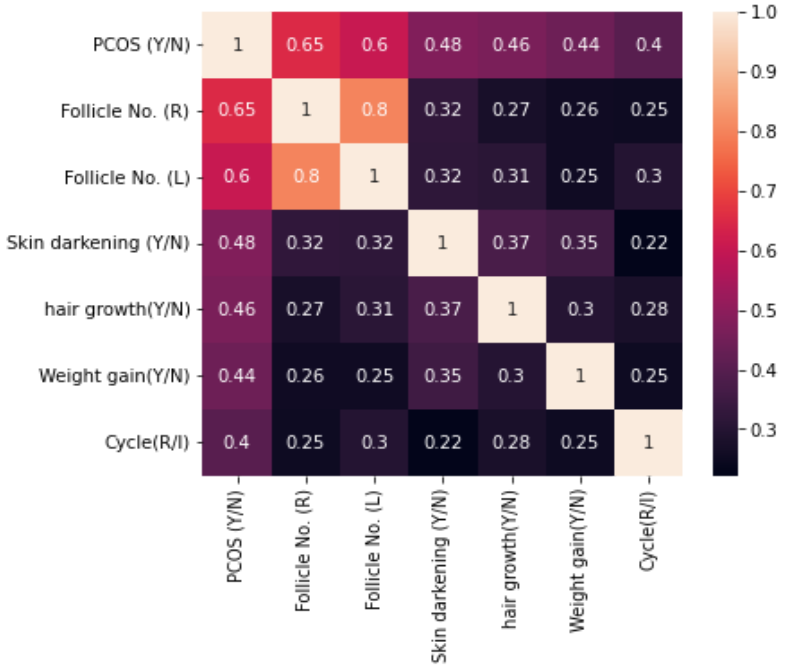

Fig. 4.1 Heatmap for highly correlated attributes from the dataset

We can observe that the Follicle No. in Left and Right and Skin Darkening plays a vital role in predicting the PCOS Status.

Features where correlation coefficient above 0.4 : 'PCOS (Y/N)', 'Follicle No. (R)', 'Follicle No. (L)','Skin darkening (Y/N)', 'hair growth(Y/N)', 'Weight gain(Y/N)', 'Cycle(R/I)'.

After analyzing the dataset, we now understand the major attributes which can help us identify the impact of PCOS. In order to observe the relationship between these attributes we now perform EDA for those specific attributes.

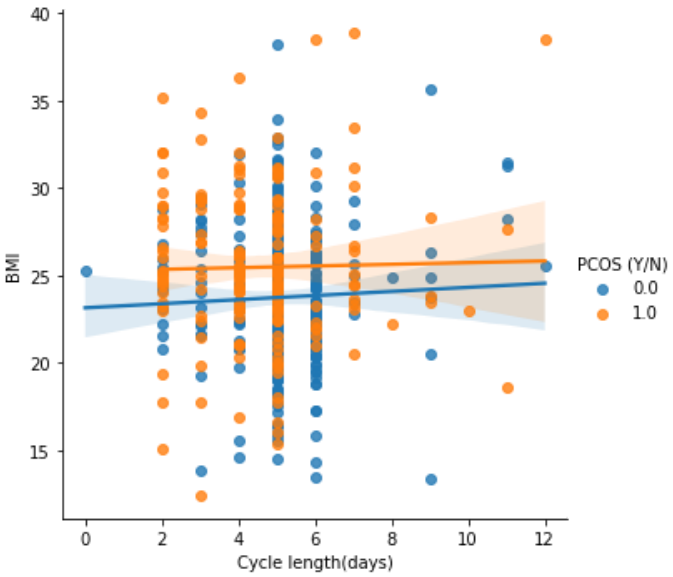

Fig 5.1 Scatter plot of Cycle length(days) and BMI.

We can observe that the BMI of PCOS affected women are higher and the cycle length of PCOS affected women are also higher than normal women.

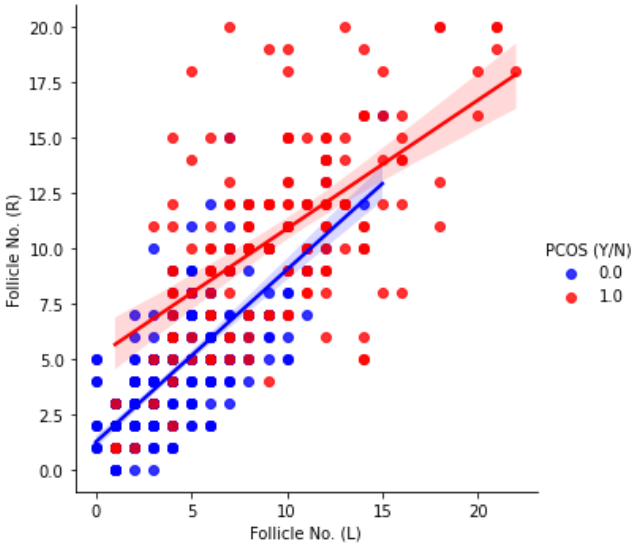

Fig 5.2 Scatter plot of Follicle No. (L) and Follicle No. (R)

We can observe from the plot that the number of follicles in PCOS affected women are higher than normal women.

IV. Classification

Split the dataset: Since we have a limited amount of data, in order to avoid overfitting, We split the dataset into a train and test set with 35% data in the test. The next step is to build a learning model. We try different classifiers and for every classifier we find accuracy, confusion matrix, cross validation score and a classification report which contains precision, f1 score, recall and support for training set and testing set. We build models with the following classifiers:

Random Forest: It consists of many decision trees. It uses bagging and features randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Bagging Classifier: It is a parallel ensemble method, the base learners are generated in parallel simultaneously. It is a way to decrease the variance of the prediction model by generating additional data in the training stage.

Adaboost classifier:It is a sequential ensemble method. Different learners learn sequentially with early learners fitting simple models to the data and then the data is analyzed for errors.It decreases the bias error and builds strong predictive models.
In the next section we'll discuss the performance of each classifier.

RESULTS

BAGGING CLASSIFIER

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

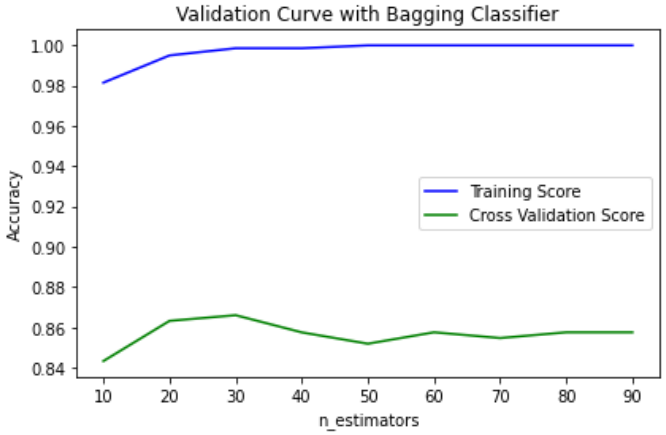

Fig 6 Validation Curve for Bagging Classifier

Bagging Classifier with 50 estimators gave a test accuracy of 87%. As the number of estimators increase, the accuracy seems to increase too. We have performed cross validation for the dataset to observe the trend and to confirm that the model is not overfitting.

ADABOOST CLASSIFIER

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

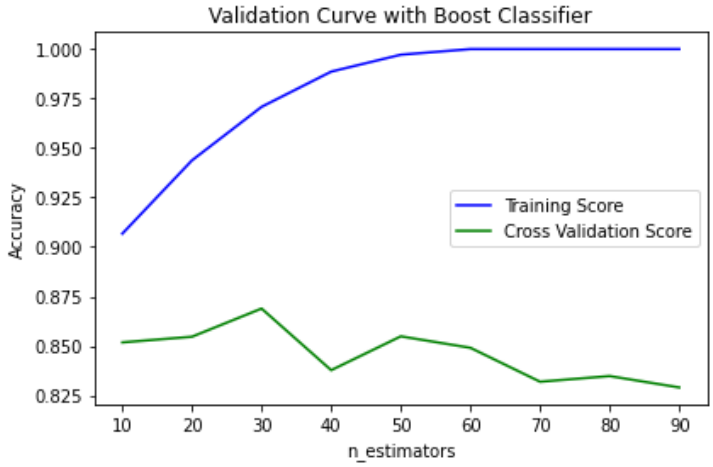

Fig 7 Validation Curve for Boost Classifier

Adaboost Classifier with 30 base parameters yields an accuracy of 87% on test data. Varying the number of estimators: Compared to bagging, the accuracy of the boosting ensemble improves rapidly with the number of base estimators. We also performed cross validation for the dataset and found that with increase in estimators, the model tends to overfit. Hence we used just 30 estimators for our model.

RANDOM FOREST CLASSIFIER

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

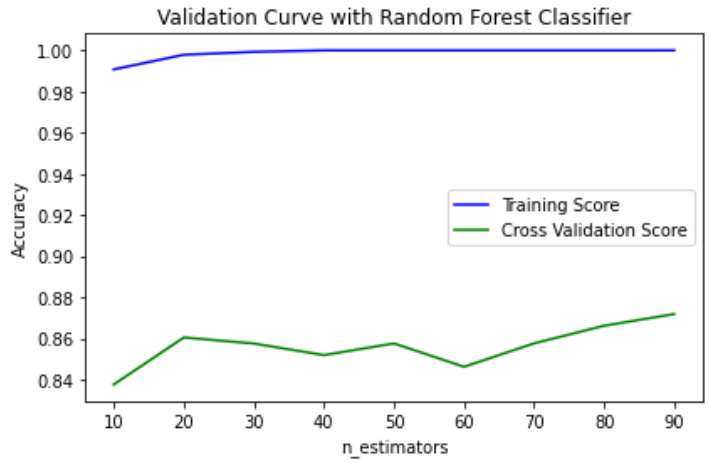

Fig 8 Validation Curve for Random Forest Classifier

The Random Forest Classifier with 100 estimators gives test accuracy of 91%. The Random Forest model turns out to be a better classification model, due to its random feature selection to train the individual trees, which makes the trees more independent and results in a better prediction. As we can see the model is not overfitted and the accuracy tends to increase with increase in estimators.

DISCUSSION

We can infer from the accuracy reports of these classifiers that the Random Forest based model gives 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. This increases our model performance. 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.

REFERENCES

1. Dataset: Polycystic ovary syndrome (PCOS) https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos
2. Sirmans SM, Pate KA. Epidemiology, diagnosis, and management of polycystic ovary syndrome. Clin Epidemiol. 2013 Dec 18;6:1-13. doi: 10.2147/CLEP.S37559. PMID: 24379699; PMCID: PMC3872139.
3. Halvard Gjønnaess (1994) Ovarian electrocautery in the treatment of women with polycystic ovary syndrome (PCOS): Factors affecting the results, Acta Obstetricia et Gynecologica Scandinavica, 73:5, 407-412, DOI: 10.3109/00016349409006253
4. Diamanti-Kandarakis Evanthia, Christakou Charikleia and Marinakis Evangelos, Phenotypes and Environmental Factors: Their Influence in PCOS, Current Pharmaceutical Design 2012; 18(3) . https://dx.doi.org/10.2174/138161212799040457
5. Rocha AL, Oliveira FR, Azevedo RC, Silva VA, Peres TM, Candido AL, Gomes KB, Reis FM. Recent advances in the understanding and management of polycystic ovary syndrome. F1000Res. 2019 Apr 26;8:F1000 Faculty Rev-565. doi: 10.12688/f1000research.15318.1. PMID: 31069057; PMCID: PMC6489978.
6. Kujanpää L, Arffman RK, Pesonen P, et al. Women with polycystic ovary syndrome are burdened with multimorbidity and medication use independent of body mass index at late fertile age: A population-based cohort study. Acta Obstet Gynecol Scand. 2022;101:728-736. doi: https://dx.doi.org/10.1111/aogs.14382
7. Ma. Karen Celine Ilagan-Vega, Ourlad Alzeus G. Tantengco, Elizabeth Paz-Pacheco, A bibliometric analysis of polycystic ovary syndrome research in Southeast Asia: Insights and implications, Diabetes & Metabolic Syndrome: Clinical Research & Reviews,Volume 16, Issue 2, 2022, 102419, ISSN 1871-4021 https://doi.org/10.1016/j.dsx.2022.102419