2021

DETECTION & CLASSIFICATION OF PCOD/PCOS USING MACHINE LEARNING ALGORITHMS

ATHULYA SHANTY
2048030
2MDS

CHRIST UNIVERSITY | Bangalore

TABLE OF CONTENTS

Detection and Classification of PCOD/ PCOS

INTRODUCTION

The chosen domain is Healthcare in women. Ovaries are an important part in the reproductive system of females. The reproductive system of women is controlled by the complex interplay of primarily five reproductive hormones namely estrogen, gonadotropin-releasing hormone, follicle stimulating hormone, progesterone and luteinizing hormone. An imbalance within these hormones leads to a hormonal disorder called the polycystic ovary syndrome (PCOS) or polycystic ovarian disease (PCOD) among women of reproductive age. The signs and symptoms of this disease include anovulation, menstrual dysfunction and signs of hyperandrogenism. Other signs and symptoms include hirsutism, infertility, obesity, metabolic syndrome, and diabetes .Women with PCOS/PCOD have high chances of hypertension. PCOS/PCOD is clinically diagnosed with an ultrasound abdomen scanning.

Polycystic ovary syndrome (PCOS) is very common in women these days mainly due to poor lifestyle choices. It is discovered that PCOS, an endocrine disorder found among the females of childbearing age has become a critical reason for infertility. PCOS can induce abnormalities in the ovaries, with high danger of abortion, infertility, heart problems, diabetes, uterus cancer etc. The signs of PCOS include cysts in ovaries, overweight, menstrual disorder, high levels of male hormones, pimples, hair fall and hirsutism. It is not easy to determine PCOS because of its different combinations of symptoms in different women and various criteria involved in the diagnosis. The time needed for various biochemical tests and ovary scanning; also the financial expenses have become a hardship to the patients. An early diagnosis opens the door to the future care and treatment. Therefore getting a warning message for the ones who are likely to be under the disease is a helpful strategy in getting avoided with the syndrome by maintaining a good and healthy lifestyle.

My aim is to build a model using machine learning algorithms which provides at most accuracy to predict the presence of PCOD/PCOS in women. This project aims to detect the polycystic ovaries Disease/Syndrome in the ovaries of a woman in order to recommend an immediate attention and need of diagnosis through a set of questionnaires. The diagnosis of PCOS/PCOD is important as this disease can cause problems with menstrual periods and make it difficult for females to conceive. If not treated it can cause insulin resistant diabetes, obesity and high cholesterol leading to heart disease.

This project aims to give an alert or warning to the females with PCOS/PCOD at its earliest stage possible. Even though there are no certain medicines for this syndrome/disease, lifestyle modifications such as diet, exercise and weight loss are considered to be the first-line treatment for women with PCOS/PCOD.

ABOUT THE DATA

The data was collected from a population of females living across the world. The source of data collection was primary and was carried out using Google forms which were distributed with the help of social media and messaging services. The google form contained 6 sections such as the main section, Personal details, Symptoms, Personal details & history, lifestyle and Medical history which contain sub questions that is curated to identify certain key criteria of PCOS/ PCOD which in turn help in developing a model to predict if the female has PCOD / PCOS or not. The collected data contains 882 rows and 27 variables.

TABLE 1: DATASET DETAILS

DATASET DETAILS		
Sl NO	**Criteria**	**Details**
1	Name	Uncleaned_data.xlsx
2	Type	Multivariate
3	No of rows	882
4	No of columns	27
5	Missing Values	Yes
6	Target Type	Available (Binary Categorical) for 212 observations and Not Available for 511 observations.
7	Applicable Technique	Classification

TABLE 2 : FEATURES DESCRIPTION

FEATURES DESCRIPTION		
Sl NO	**Feature Name**	**Description**
1	Timestamp	Indicates the date and time of the response
2	What is your name?	Takes the name of the responder
3	Select your Gender	Takes 3 unique values such as Female, Male and other
4	Pick your age limit	Takes 4 unique values such as Less than 18, 19-34, 35-50 and 51 years & above
5	How would you describe your Body Physic ?	Takes 4 unique values such as I am underweight, I am at a healthy weight, I am a bit overweight, and I am obese.
6	Have you done an ultrasound abdomen scanning and what does your report say?	Takes 4 unique values such as Normal results, Cysts in ovary, I don't remember, and No scanning was done
7	Do you notice any of these right before your period begins?	Takes 8 unique values such as Bloating, Breast pain, Constipation, Diarrhoea, Headache, Sleeplessness, Mood swings and None
8	Are you experiencing irregular or late periods?	Takes 2 unique values such as Yes and No
9	Are you experiencing painful periods?	Takes 2 unique values such as Yes and No
10	Are you experiencing Excessive bleeding?	Takes 2 unique values such as Yes and No
11	How often do you get your periods?	Takes 3 unique values such as Less than 21 days, 21-40 days, and More than 40 days

12	How long does your period last?	Takes 4 unique values such as Less than 3 days, 3-5 days, 5-7 days, and More than 7 days
13	How would you like to rate your period pains	Takes 4 unique values such as Mild, Moderate, Severe and No period pain
14	Do you notice any clots during your periods?	Takes 3 unique values such as Yes-Small Clots, Yes-Large clots, and No clots
15	Do you have the habit of consuming alcohol?	Takes 2 unique values such as Yes and No
16	Do you have the habit of smoking?	Takes 2 unique values such as Yes and No
17	Are you under any stress?	Takes 2 unique values such as Yes and No
18	Do you exercise regularly?	Takes 2 unique values such as Yes and No
19	Is your mother diagnosed with PCOS/PCOD?	Takes 2 unique values such as Yes and No
20	Do you suffer from diabetes?	Takes 2 unique values such as Yes and No
21	Do you suffer from Hypothyroidism?	Takes 2 unique values such as Yes and No
22	Do you experience excessive growth of facial and body hair?	Takes 2 unique values such as Yes and No
23	Do you have Acne/ Hyper-pigmentation?	Takes 2 unique values such as Yes and No
24	What is your marital status	Takes 2 unique values such as Married and Unmarried
25	How many kids do you have?	Takes 4 unique values such as 0, 1, 2 , and More than 2
26	Which work profile matches yours?	Takes 3 unique values such as Student, Employed and Unemployed
27	Are you diagnosed with PCOD/PCOS?	The target variable takes 3 unique values such as Yes, No , and Didn’t check.

IMPLEMENTATION DETAILS

The raw dataset contained 882 observations and 27 variables. The lengthy column names were renamed to shorter ones. The duplicate values in the dataset were removed. Even though the desired population is women, the survey was also attended by some of the males whom must be removed for further analysis. Therefore there comes a need to filter the dataset such that the required dataset contains only the data from females. The same was done by removing the observations with 'other' and 'male' as gender. Also, there were some unwanted variables for our study and they were removed. There were missing data in the dataset. Dropping the missing values was not a solution as it dropped the entire dataset and those were handled by replacing the null values with the string 'unknown'. The column named 'kids' contained both numerical and string data which will result in errors in further processes. Therefore the numbers 0,1,2 were replaced to strings 'zero', 'one' and 'two'. After cleaning the data, the observations got reduced to 723 and the variables got reduced to 24.

The dataset had 511 observations which contained only independent variables whereas 212 observations contained both independent variables and dependent variable. These were separated into two different data frames for further analysis. There are no numerical variables found in the dataset whereas there were 24 categorical variables. These 24 categorical variables were converted to numerical variables by label encoding. The dataframe which contained both dependent and independent variables were divided into dependent variable and independent variables.

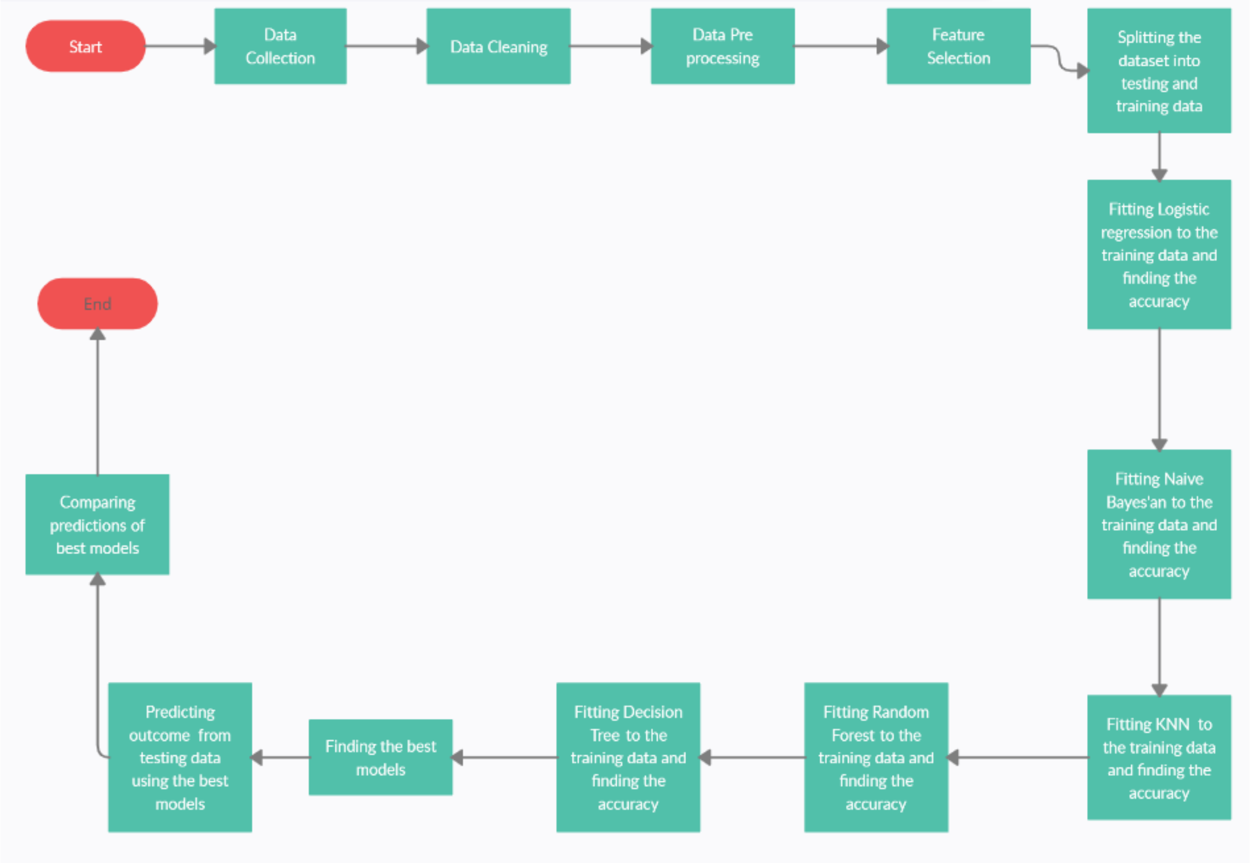
The dependent variable is 'diagnosis' and independent variables are 'age', 'body_type', 'scanning', 'before_period', 'irregular_period', 'painful_period', 'bleeding', 'period_cycle', 'period_duration', 'period_pain', 'clots', 'alcohol', 'smoking', 'stress', 'exercise', 'Hereditary', 'diabetes', 'hypothyroidism', 'hair_growth', 'acne', 'marital_status', 'kids' and 'work'. Bar plots were plotted with their frequency for unique category of the categorical variables for a better understanding of the modelling data.

The correlation matrix gave the correlation between the variables and the heatmap of the same gave a better way of visualization. The relevant features for the study were chosen based on the correlation between variables. The variables with correlation more than 0.2 were considered to be the important variables.

The variable to be predicted was categorical and hence different algorithms such as Naive Bayes classification, Logistic regression, KNN classification, Random Forest classifier and Decision tree classifier were used for modelling the data. The data frame which have both the independent and dependent variables were splitted to X_train, y_train, X_test and y_test. Each model was trained by inputting X_train, y_train values of the training dataset. Model evaluation and the confusion matrix were drawn for each algorithm.

The data frame which had only dependent variables were inputted to the two models with the highest accuracy for predicting the diagnosis. The count of predicted positives and negatives for PCOS/PCOD were calculated from predictions of both models and they were plotted in bar graph. The predicted diagnosis by the best models were also compared for a better understanding.

METHODOLOGY

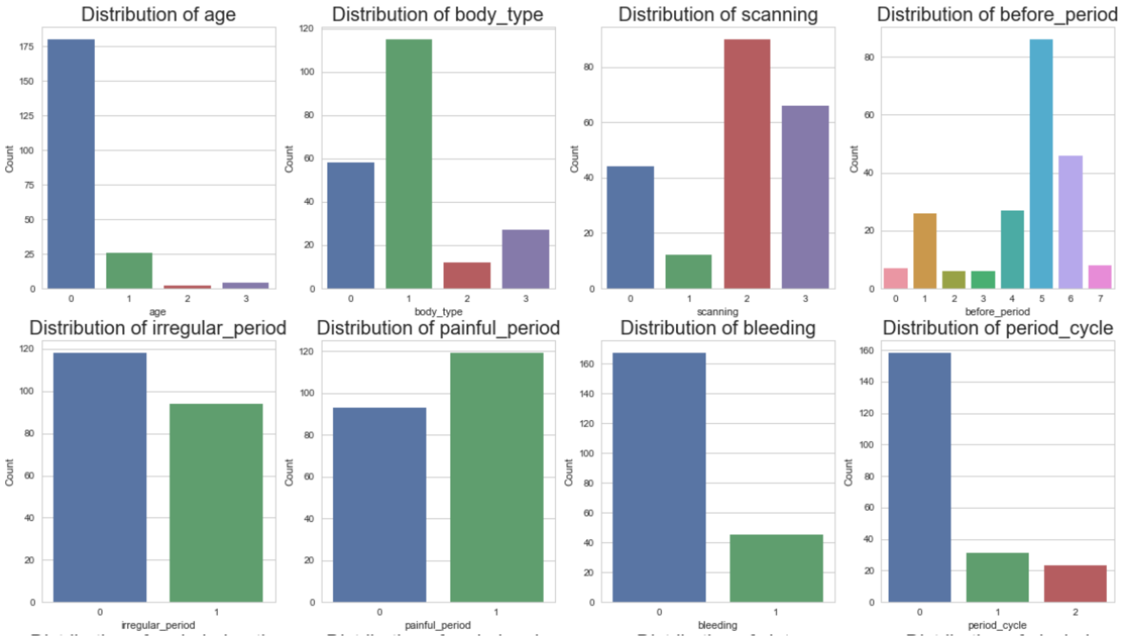


Start
Data Collection
Data Cleaning
Data Pre processing
Feature Selection
Splitting the dataset into testing and training data
Fitting Logistic regression to the training data and finding the accuracy
Fitting Naive Bayes'an to the training data and finding the accuracy
Fitting KNN to the training data and finding the accuracy
Fitting Random Forest to the training data and finding the accuracy
Fitting Decision Tree to the training data and finding the accuracy
Finding the best models
Predicting outcome from testing data using the best models
Comparing predictions of best models
End

FIGURE 1 : FLOWCHART SHOWING THE PROCESSES UNDERGONE BY THE DATASET

RESULTS AND DISCUSSION

- The lengthy column names was renamed to shorter ones.
- Removing duplicates removed 113 duplicate records and resulted in 763 observations with 27 variables.
- Dropping ‘male’ and ‘other’ from gender resulted in 723 observations with 27 variables.
- Removal of unwanted columns such as ‘Timestamp’, ’name’, etc resulted in 723 observations with 24 variables.
- Handling missing data was done by filling the null values with the variable ‘unknown’.
- The dataset contained 24 categorical variables and no numerical variables. The categorical variables are 'age', 'body_type', 'scanning', 'before_period', 'irregular_period', 'painful_period', 'bleeding', 'period_cycle', 'period_duration', 'period_pain', 'clots', 'alcohol', 'smoking', 'stress', 'exercise', 'Hereditary', 'diabetes', 'hypothyroidism', 'hair_growth', 'acne', 'marital_status', 'kids', 'work', and 'diagnosis'.
- The categorical variables have been converted to numerical by Label Encoding.
- The count of unique value of every variables were plotted in a bar graph.



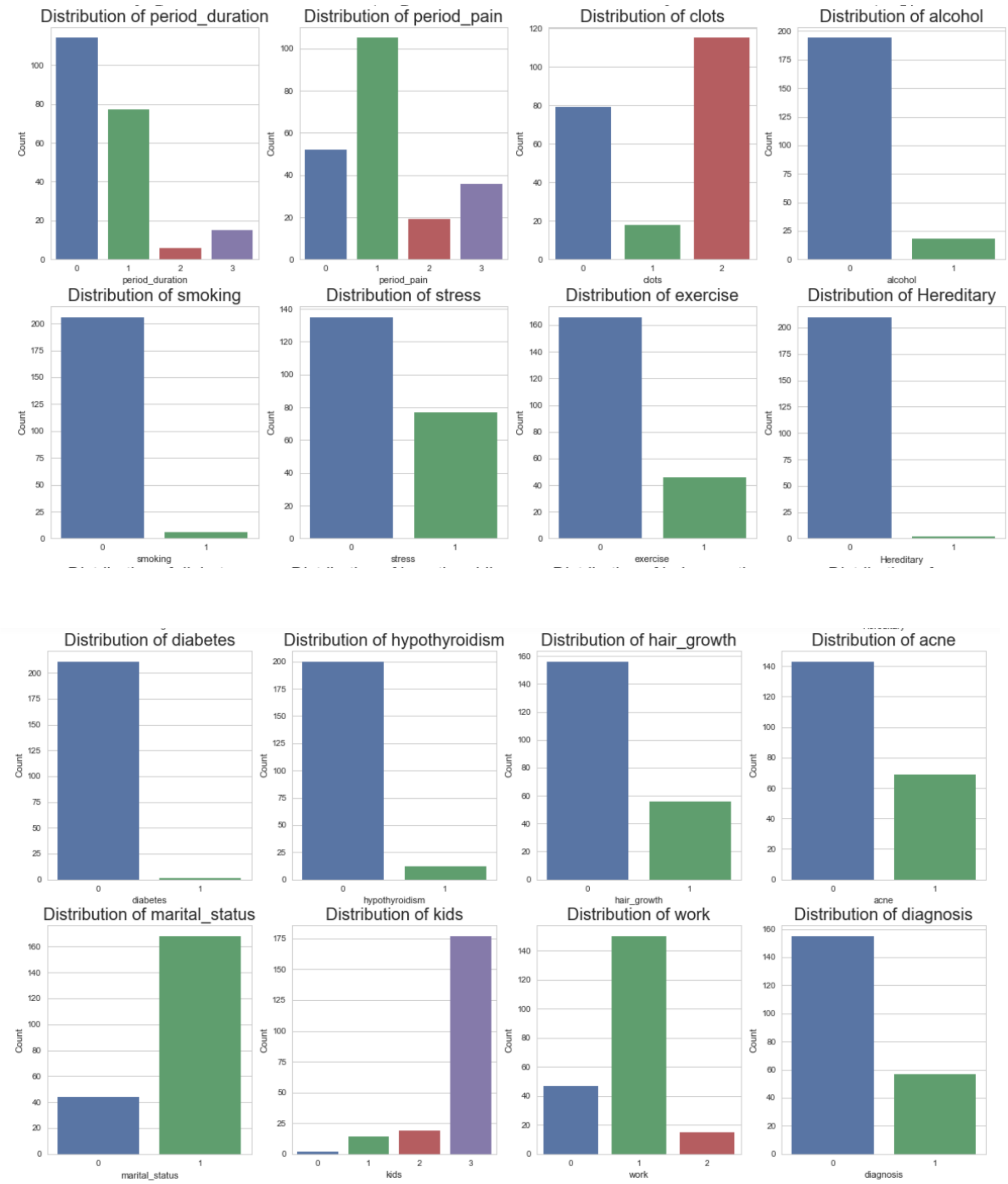


FIGURE 2 : BAR PLOT INDICATING COUNT OF UNIQUE VALUES OF EACH VARIABLE

- The dependent variable is diagnosis and the independent variables are 'age', 'body_type', 'scanning', 'before_period', 'irregular_period', 'painful_period', 'bleeding', 'period_cycle', 'period_duration', 'period_pain', 'clots', 'alcohol', 'smoking', 'stress', 'exercise', 'Hereditary', 'diabetes', 'hypothyroidism', 'hair_growth', 'acne', 'marital_status', 'kids', 'work'.

- Correlation heat map has been plotted.

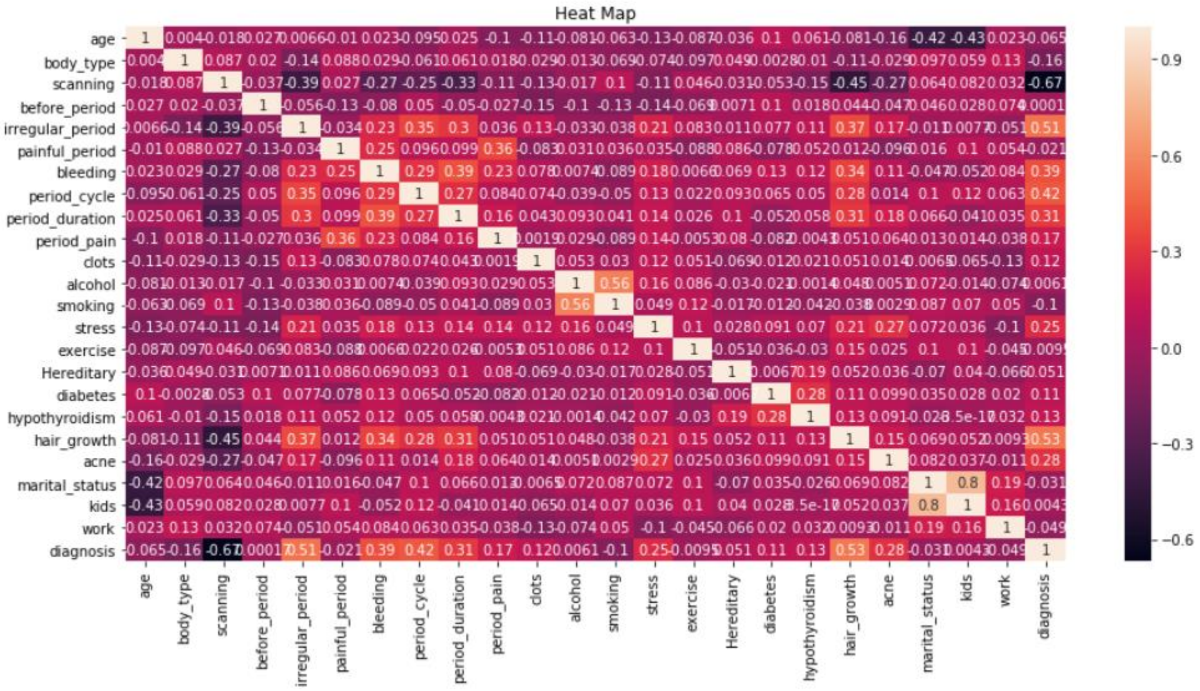
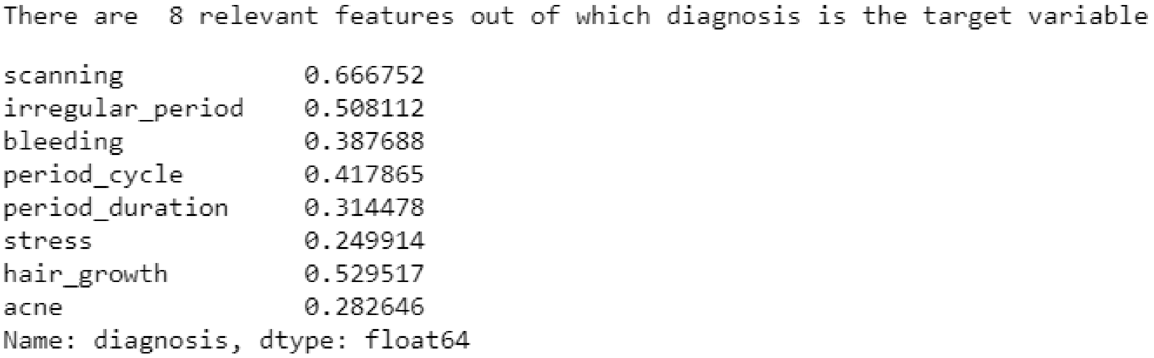


FIGURE 3 : CORRELATION HEAT MAP

- The relevant features was found to be 'scanning', 'irregular_period', 'bleeding', 'period_cycle', 'stress', 'hair_growth' and 'acne' through the method of correlation with cut-off as 0.2.

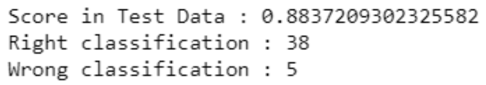


```
There are  8 relevant features out of which diagnosis is the target variable

scanning             0.666752
irregular_period     0.508112
bleeding             0.387688
period_cycle         0.417865
period_duration      0.314478
stress               0.249914
hair_growth          0.529517
acne                 0.282646
Name: diagnosis, dtype: float64
```

FIGURE 4 : CORRELATION OF RELEVANT FEATURES

- The dataset which have both dependent and independent variable was splitted into Xtrain, ytrain, Xtest and ytest for modelling.
- The accuracy for Naïve Bayesian Model is 88.37%.
- The confusion matrix for Naïve Bayesian Model is



```
Score in Test Data : 0.8837209302325582
Right classification : 38
Wrong classification : 5
```

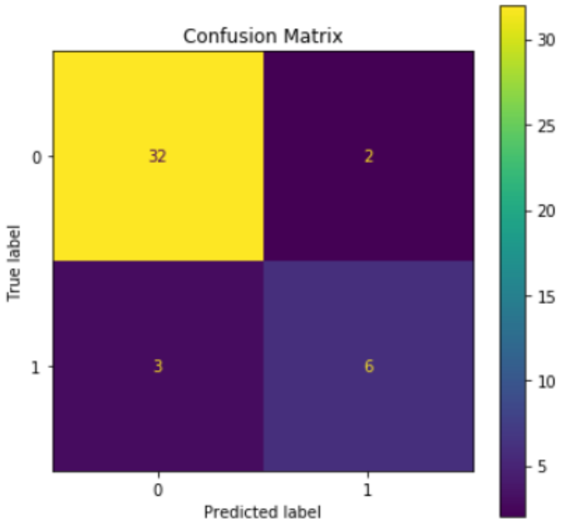
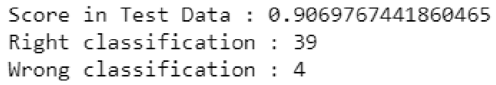


FIGURE 5 :CONFUSION MATRIX FOR NAÏVE BAYESIAN MODEL

- The accuracy for Logistic Regression model is 90.70%
- The confusion matrix for Logistic Regression Model is



```
Score in Test Data : 0.9069767441860465
Right classification : 39
Wrong classification : 4
```

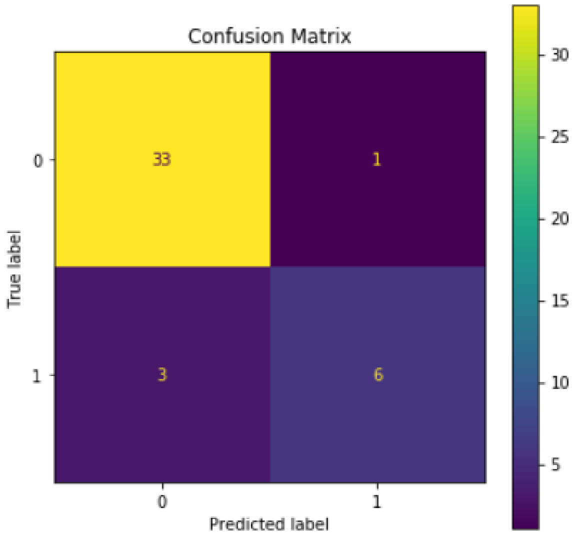
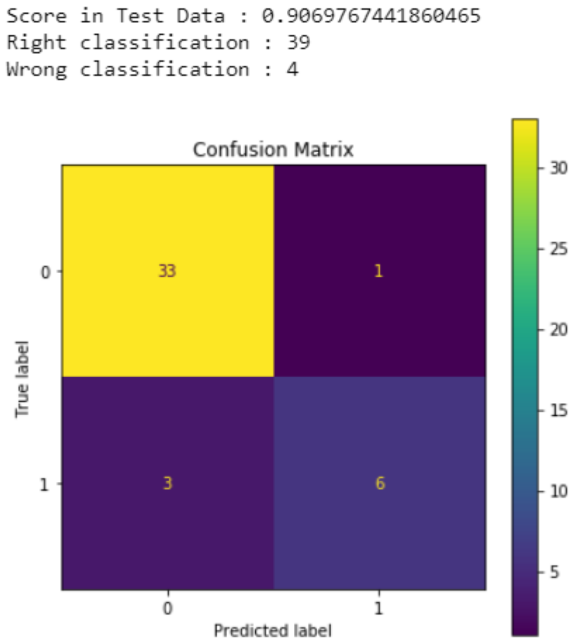


FIGURE 6 :CONFUSION MATRIX FOR LOGISTIC REGRESSION MODEL

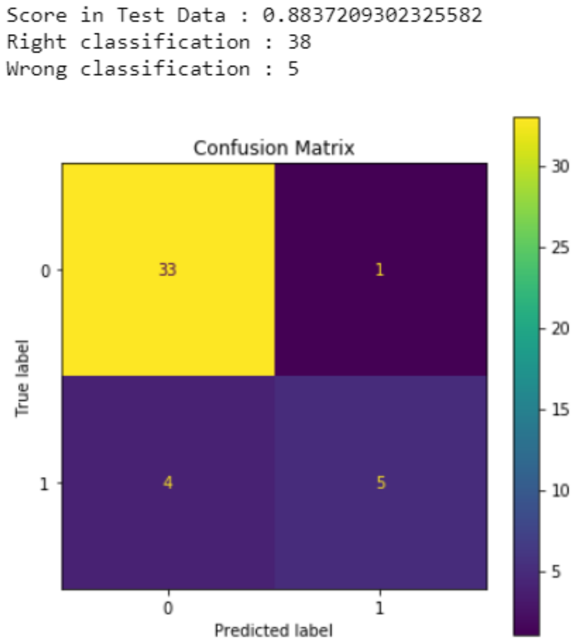
- The accuracy for KNN model is 90.70%.
- The confusion matrix for KNN Model is



```
Score in Test Data : 0.9069767441860465
Right classification : 39
Wrong classification : 4
```

FIGURE 7 :CONFUSION MATRIX FOR KNN MODEL

- The accuracy for Random Forest Classifier is 88.37%.
- The confusion matrix for Random Forest Model is



```
Score in Test Data : 0.8837209302325582
Right classification : 38
Wrong classification : 5
```

FIGURE 8 :CONFUSION MATRIX FOR RANDOM FOREST MODEL

- The accuracy for Decision Tree model is 88.37%.
- The decision tree is given below.

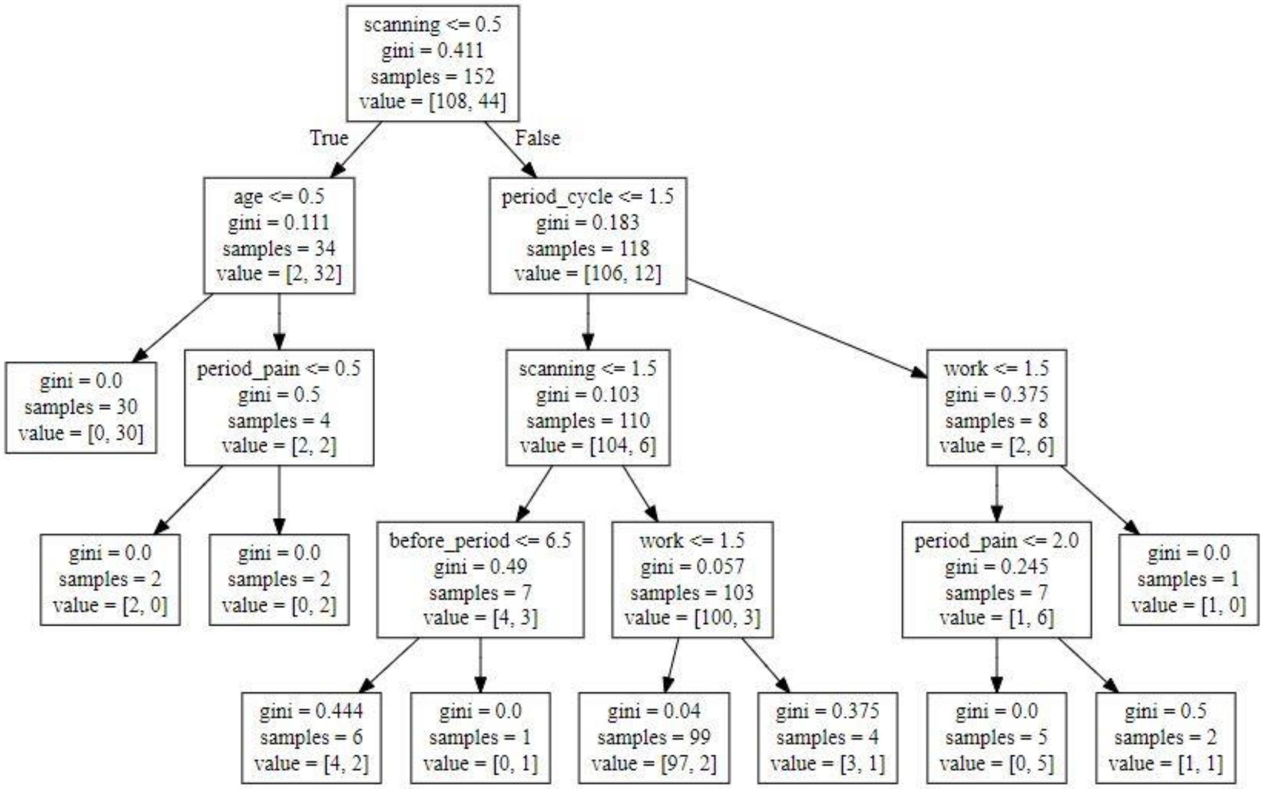
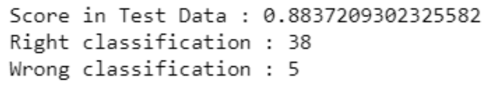


FIGURE 9: DECISION TREE CLASSIFICATION

- The confusion matrix for Decision Tree Model is



```
Score in Test Data : 0.8837209302325582
Right classification : 38
Wrong classification : 5
```

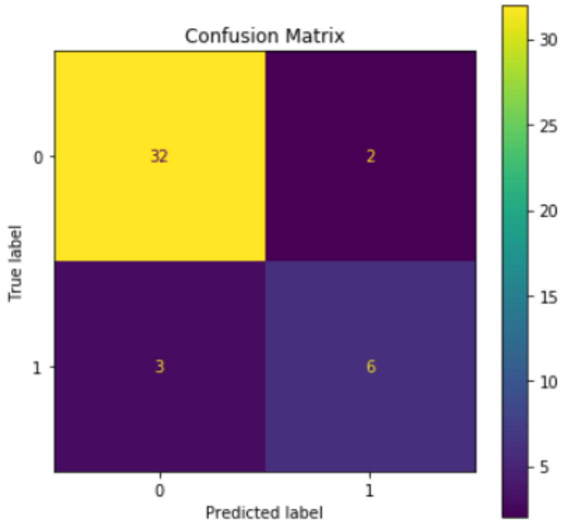
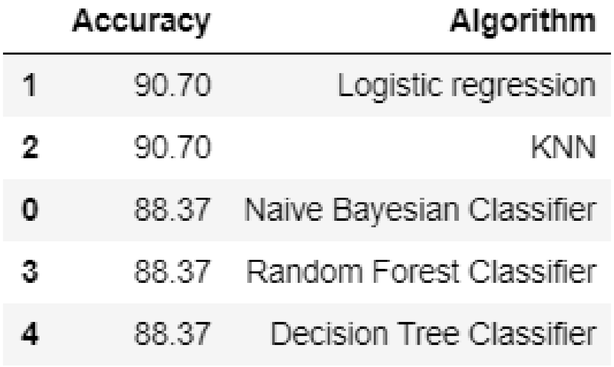


FIGURE 10 :CONFUSION MATRIX FOR DECISION TREE MODEL

- There were a total of 5 models and their accuracy were compared with each other to find the best model(s).

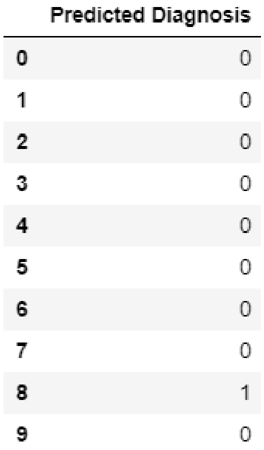
TABLE 3: COMPARISON OF MODEL ACCURACY



	Accuracy	Algorithm
1	90.70	Logistic regression
2	90.70	KNN
0	88.37	Naive Bayesian Classifier
3	88.37	Random Forest Classifier
4	88.37	Decision Tree Classifier

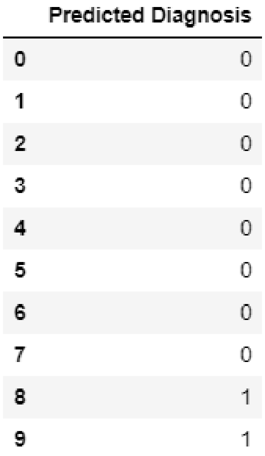
- The best models were found to be Logistic Regression and KNN with 90.70% of accuracy.
- The diagnosis were predicted for the test data using the models with high accuracy. In the prediction of diagnosis, 1 denotes a female with PCOS/PCOD and 0 denotes a female with no PCOD/PCOS.

TABLE 4 : PREDICTION OF 1st 10 ROWS USING LOGISTIC REGRESSION MODEL



	Predicted Diagnosis
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	1
9	0

TABLE 5: PREDICTION OF 1st 10 ROWS USING KNN MODEL



	Predicted Diagnosis
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	1
9	1

- Prediction of diagnosis using Logistic Regression predicted 75 have PCOS/PCOD whereas 436 have no PCOD/PCOS.

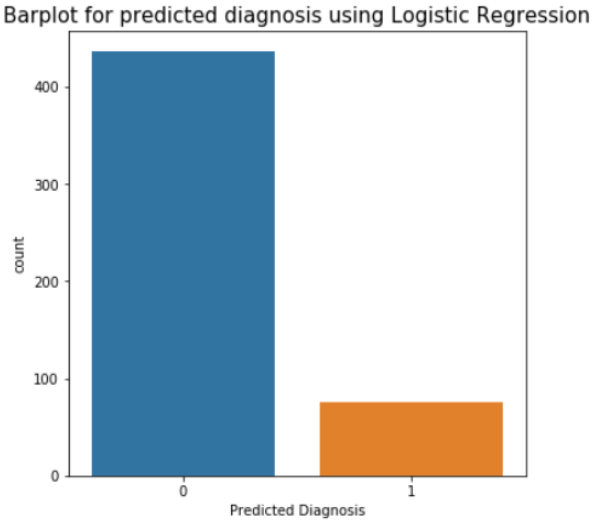


FIGURE 11 : BARPLOT DEPICTING COUNT OF UNIQUE VALUES OF PREDICTED DIAGNOSIS USING LOGISTIC REGRESSION MODEL

- Prediction of diagnosis using KNN predicted 61 have PCOS/PCOD whereas 450 have no PCOD/PCOS.

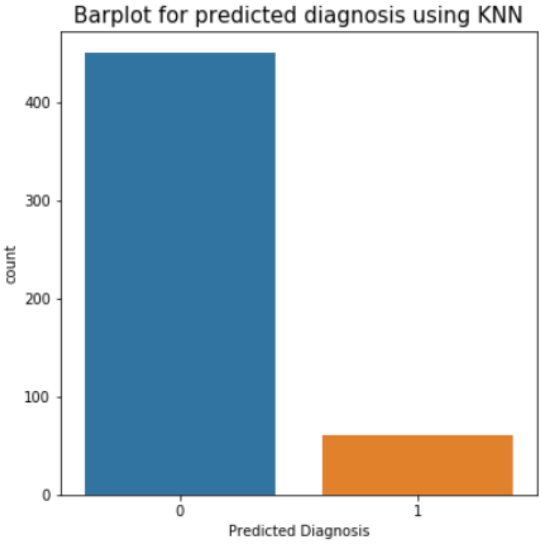


FIGURE 12 : BARPLOT DEPICTING COUNT OF UNIQUE VALUES OF PREDICTED DIAGNOSIS USING KNN MODEL

- The diagnosis predicted by both the Logistic Regression model and KNN model were compared. 485 predictions were predicted the same by both the models and 26 predictions were predicted differently by the models.

LIMITATIONS

- A minimum of 20 cysts in the ovaries were considered to be PCOD/PCOS. But we were not able to collect the number of cysts in each ovary which would have made the model much more accurate.
- The training data set was comparatively very less to the testing dataset. Reaching out to the ones with the disease should have been increased for a better accuracy and prediction.

CONCLUSION

The present world women population is widely affected by preterm abortions, infertility, anovulation etc. It is observed that polycystic ovary syndrome (PCOS), a condition seen among the women of reproductive age is having a major influence in the cause of infertility. PCOS/PCOD can increase the risk of infertility, metabolic syndrome, sleep apnea, endometrial cancer, and depression. It is an endocrine disorder characterized by changes in the female hormone levels and the abnormal production of male hormones. This condition leads to ovarian dysfunction with increased risk of miscarriage and infertility. The symptoms of PCOS include obesity, irregular menstrual cycle, and excessive production of male hormone, acne, and hirsutism. A healthy lifestyle, eating a healthy diet and exercising regularly will help the women with PCOS/PCOD.

In this project, I have developed models with high accuracy for predicting or detecting the presence of PCOS/PCOD in women. Machine Learning can be used in healthcare systems for diagnostic purposes with much accuracy and precision. Correct diagnosis is the baseline of any proper treatment and in this project, I have used machine learning approaches like KNN, Naive Bayes Classification, Random Forest, Decision Tree and Logistic Regression to diagnose PCOS. A model with 90.70% of accuracy has been built for the prediction of PCOD/PCOS using the Logistic Regression model and KNN model. The models Logistic Regression and KNN performed well than Naïve Bayesian, Random Forest, and Decision tree. These models with high accuracy helps in the diagnosis of the syndrome/disease at it earliest stage possible and would be able to warn them beforehand. Out of 511 predictions, 85 predictions were predicted as same by both the models whereas 26 predictions were not predicted as the same.The time needed for various biochemical tests and ovary scanning; also the financial expenses have become a hardship to the patients. An early diagnosis opens the door to the future care and treatment. The proposed models predict the diagnosis of PCOS/PCOD in women at the earlier stage as possible.

FUTURE WORK

- I am planning to further continue this project by collecting more data which would increase the accuracy rate and improve the prediction.
- I am planning to perform unsupervised learning algorithm on the data where dependent variable is absent. I also wanted to compare the predictions of both supervised and unsupervised learning algorithms.
- I am planning to develop a mobile application or web page which will take all the variables as inputs and give the predictions as output with a warning message and steps to recover from PCOS/PCOD.

ACKNOWLEDGEMENT

I would like to express my special thanks to Dr. Tijo Thomas (MBBS) for his able guidance and support in completing my project. I would also like to extend my gratitude to my teacher Mrs. Ummesalma M for giving me an opportunity to do this project on Machine Learning and guiding me until the completion of my project. I would like to extend my gratitude to the Coordinator, Head of the department, Dean of Sciences and the Christ University Management.

BIBLIOGRAPGHY & REFERENCES

[1]https://www.healthline.com/health/polycystic-ovary-disease#health-effects

[2]https://www.yashodahospitals.com/diseases-treatments/pcod-pcos-symptoms-causes-treatment/