FERGUSSON COLLEGE (AUTONOMOUS), PUNE

2020 – 2021

A Project Report On

"Analysis of PCOS (Polycystic Ovarian Syndrome) in women and Comparative study between women with PCOS and women without PCOS"

Submitted to

Fergusson College (Autonomous), Pune 411004

In the partial fulfilment of MSc IMCA III (Semester V)

Submitted by

Aditi Pujari

Janhavi Gunjal

Shriya Bhonde

Under the Guidance of

Prof. Hrishikesh Khaladkar

Assistant Professor, Fergusson College (Autonomous), Pune-04

C E R T I F I C A T E

This is to certify that the students of MSc IMCA III carrying out the project work entitled "Analysis of PCOS in women and Comparative study between women with PCOS and women without PCOS" have completed their work satisfactorily as a partial fulfilment of the Degree of Post Graduation from "Fergusson College (Autonomous)" during the academic year 2020-2021.

Mr. Hrishikesh Khaladkar

(Project Guide)

Dr. V. V. Acharya

(Head of the Department)

ACCEPTANCE

This is to certify that undersigned have accessed and evaluated the project "**Analysis of PCOS in women and Comparative study between women with PCOS and women without PCOS**" submitted by Ms. Aditi Pujari, Ms. Janhavi Gunjal and Ms. Shriya Bhonde. The project report has been accepted for the partial fulfilment of the degree of **MSc IMCA** from **Fergusson College (Autonomous), Pune**.

External Examiner:

Name: ----------

Date: ----------

Industrial Examiner:

Name: ----------

Date: ----------

Internal Examiner:

Name: ----------

Date: ----------

PLACE: PUNE

DATE: ----------

ACKNOWLEDGEMENT

This report is presented at the end of the Semester V.

We herewith acknowledge the encouragement, guidance and supervision by **Mr. Hrishikesh Khaladkar** Sir, our project guide and express him our deepest gratitude for his constant support and insights.

A special note of thanks to **Dr. Nagare Sir** for his help, guidance in the medical aspect of the project and for his valuable time. We take this opportunity to express him our sincere gratitude.

We are thankful to **Ms. Nazreen Khan** who have been our guiding light throughout the project, for her constant help, support and valuable inputs, which made our project better and better at each step.

We would like to extend our deep esteems to all the ladies for their valuable time and their responses which were indeed essential for collecting the sample data for the project analysis.

We would also like to thank Head of the Department, Dr. V. V. Acharya Sir.

Thanking You,

Aditi Pujari

Janhavi Gunjal

Shriya Bhonde

MSc IMCA

Fergusson College (Autonomous), Pune

DECLARATION

We declare that the project titled "Analysis of PCOS (Polycystic Ovarian Syndrome) in women and Comparative study between women with PCOS and women without PCOS", submitted by us, for the partial fulfilment of our Master's degree during (2020-2021) is our original work.

We further declare that the analysis has been carried out based on the primary data collected by us. We have given our best and hope that our project work may be helpful.

- Aditi Pujari
- Janhavi Gunjal
- Shriya Bhonde

Place: Pune

Date:

INDEX

INTRODUCTION

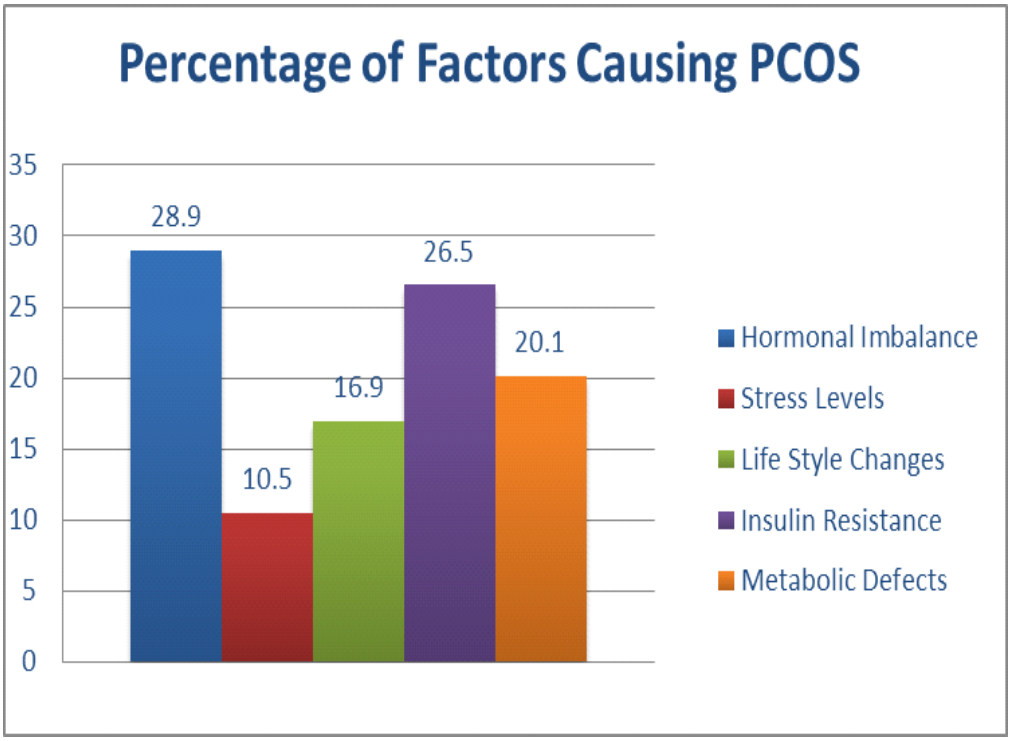
Nowadays, in this fast moving world we all have become prey to many factors like pollution, lifestyle changes, stress, bad dietary habits, excessive use of technology etc. All these factors affect us in some or the other way. The majority group of people is the young generation who is being seriously affected by all these.

The major effects seen in women/girls is the obesity, fatigue, mood swings, depression, hair loss, acne, irregular periods, skin darkening, unwanted hair growth, etc. of whom the root cause can be any deficiency, hormonal imbalance, etc. The study shows high possibility of occurrence of PCOS (Polycystic Ovarian Syndrome) to women mainly in the reproductive age.

PCOS (Polycystic Ovarian Syndrome) is a 'syndrome' or group of symptoms that affects ovaries and ovulation. The major effects of PCOS seen in women is the irregular periods, mood swings, hair loss, weight gain, central obesity, acne, low confidence levels, etc.

The other effects seen are the infertility, insulin resistance leading to Type II Diabetes, increasing male hormones in women, hirsutism, risk of miscarriage, high abortion rate, trouble conceiving, etc.

The below image clearly depicts how different factors have major effect on PCOS.



Many researchers claim that PCOS is only because of bad dietary habits, sedentary lifestyle, lack of exercise, stress, etc.

The effect can be reversed by following proper diet, making some good lifestyle changes and regular exercise like any aerobic activity like yoga, aerobics, Zumba etc, can be very fruitful and having a positive and peaceful approach.

Seeing this disorder faced by many girls or women our age typically, made us think upon and hence thought of choosing this topic as our project subject and hence create awareness.

The analysis is carried out on two different datasets, and hence the conclusion is different in both the cases. One dataset is from www.kaggle.com regarding PCOS which has majorly the clinical factors along with some physical parameters too.

The second dataset is from the data collected through google forms keeping in mind the lifestyle, dietary habits, exercising habits of women, etc. It has more of physical parameters rather than the clinical factors.

- The dataset PCOS_data_without_infertility.xlxs on www.kaggle.com

  This data has more of the clinical parameters and physical parameters as well.

- The data collected through google forms, keeping in mind the physical and clinical parameters also the lifestyle, work culture, dietary habits of women

Analysis of PCOS (Polycystic Ovarian Syndrome) in women and Comparative Study between women with PCOS and Healthy Women

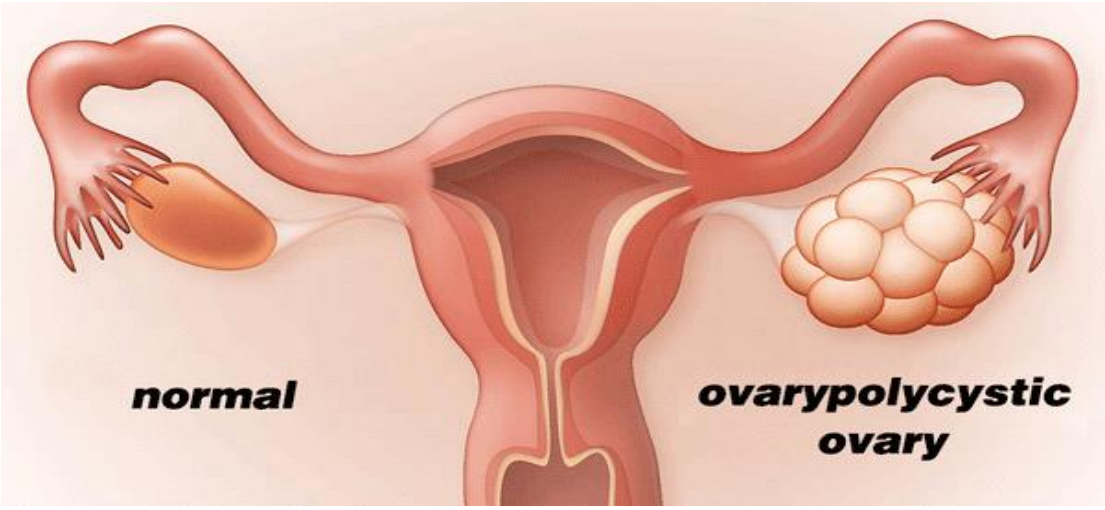
Aim :

1. Factors majorly affecting PCOS
2. Comparative study between women with PCOS and women without PCOS

What is PCOS?

- PCOS (Polycystic Ovarian Syndrome) is a 'syndrome' or group of symptoms that affects ovaries and ovulation.
- The ovaries release eggs to be fertilized by a man's sperm. The release of an egg each month is called ovulation.
- PCOS affects a woman's ovaries, the reproductive organs that produce estrogens and progesterone — hormones that regulate the menstrual cycle. The ovaries also produce a small amount of male hormones called androgens, causing hormonal imbalance in a woman.

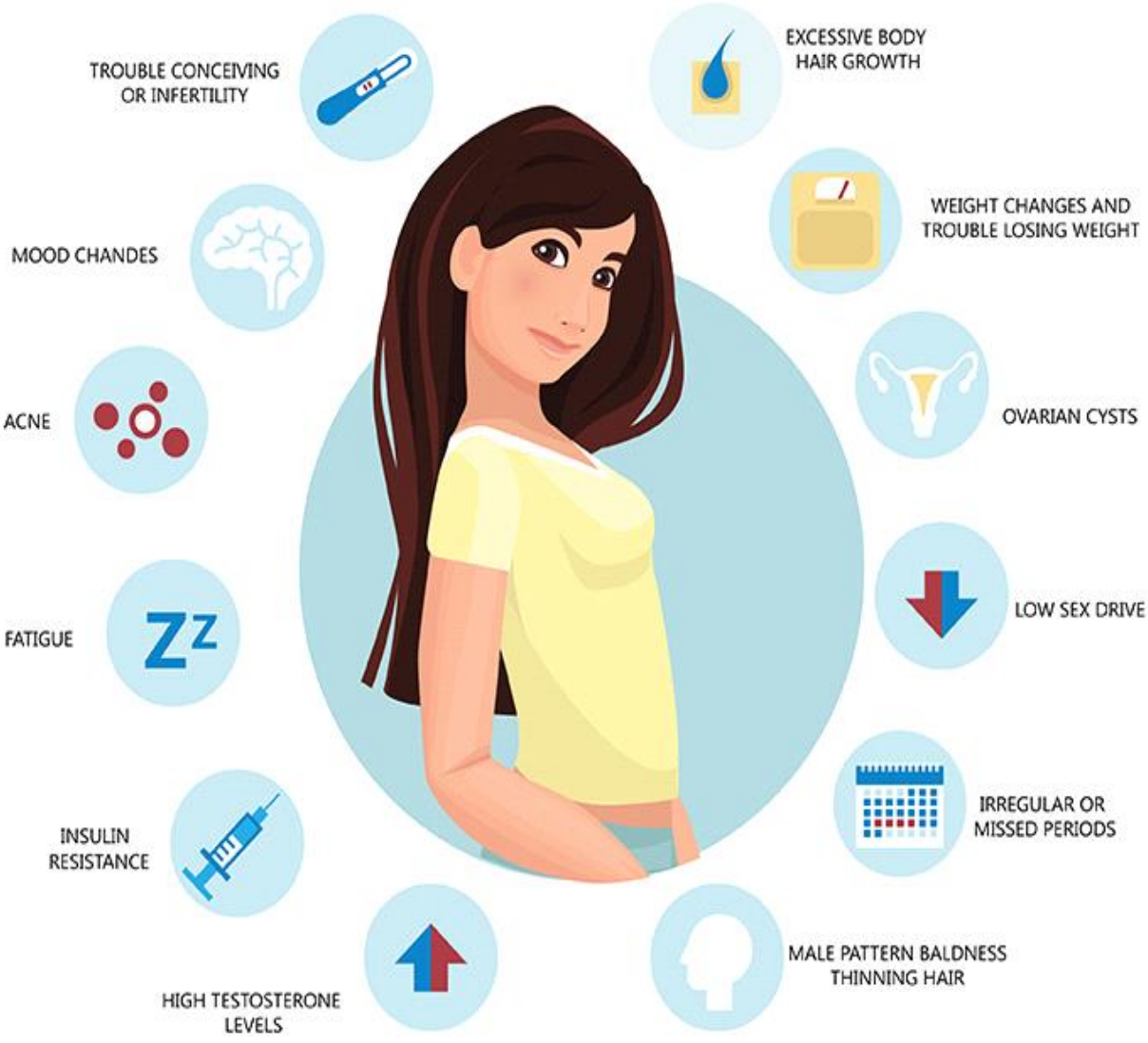
Normal Ovary v/s Polycystic Ovary



- The three main features are:
    - Cysts in the ovaries
    - High levels of male hormones
    - Irregular or skipped period

- The common symptoms are:
    -
    - Irregular periods
    - Heavy/Scanty bleeding
    - Acne
    - Unwanted hair growth
    - Male pattern baldness
    - Weight gain or weight loss
    - Skin darkening
    - Hair thinning
    - High level of male hormones

PCOS SYMPTOMS



Kerala Dataset Analysis

Dataset Used

PCOS_data_without_infertility.xlsx

Source

https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos

Data Information

The dataset contains all physical as well as clinical parameters to determine PCOS and infertility related issues. The data is collected from 10 different hospitals across Kerala, India.

Data attributes

- PCOS (Y/N)
- Age (years)
- Weight (kg)
- Height (cm)
- BMI (Body Mass Index)
- Blood Group
- Pulse rate (bpm)
- RR (Respiratory rate) (breaths/min)
- Hb (Haemoglobin) (g/dl)
- Cycle (R/I)
- Cycle length (days)
- Marriage status (years)
- Pregnant (Y/N)
- No. of abortions
- FSH (Follicle Stimulating Hormone)
- LH (Luteinizing Hormone)
- FSH/LH
- Hip (inch)
- Waist (inch)

- Waist Hip Ratio
- TSH (Thyroid Stimulating Hormone)
- AMH (Anti-Mullerian Hormone)
- PRL (Prolactin) (ng/dl)
- Vit D3 (ng/dl)
- PRG (Progesterone) (ng/dl)
- RBS (Random Blood Sugar) (ng/dl)
- Weight gain (Y/N)
- Hair growth (Y/N)
- Skin darkening (Y/N)
- Hair loss (Y/N)
- Pimples (Y/N)
- Fast food (Y/N)
- Regular exercises (Y/N)
- BP_Systolic (mmHg)
- BP_Diastloic (mmHg)
- Follicle No. (L)
- Follicle No. (R)
- Avg. F size (L) (mm)
- Avg F size (R) (mm)
- Endometrium

Attributes explanation

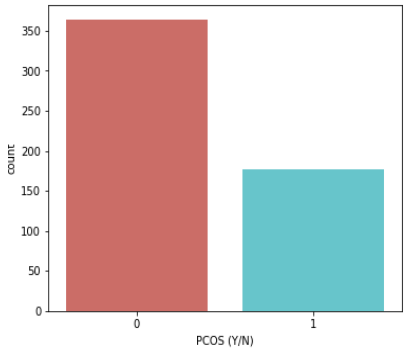
The medical parameters are as follows:

- Pulse rate: The number of times heart beats in a minute.
- RR (Respiratory rate): The respiration rate is the number of breaths a person takes per minute.
- Hb (Haemoglobin): Haemoglobin is a protein found in red blood cells. It gives blood its red colour and its job is to carry oxygen throughout the body.

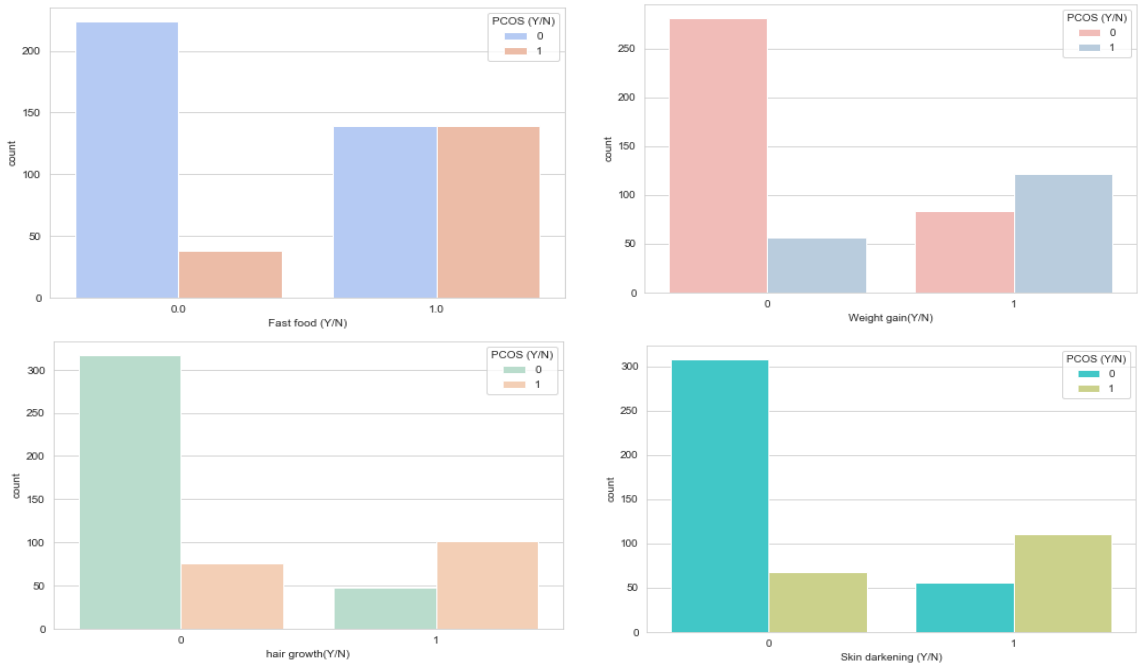
- FSH (Follicle Stimulating Hormone): FSH plays an important role in sexual development and functioning. In women, FSH helps control the menstrual cycle and stimulates the growth of eggs in the ovaries.
- LH (Luteinizing Hormone): Luteinizing hormone (LH) is produced and released in the anterior pituitary gland. This hormone is considered a gonadotrophic hormone because of its role in controlling the function of ovaries in females and testes in males, which are known as the gonads.
- TSH (Thyroid Stimulating Hormone): The thyroid is a butterfly-shaped gland in the throat. It produces hormones that help regulate many bodily functions, such as metabolism, heart rate, and body temperature.
- AMH (Anti-Mullerian Hormone): Anti-Mullerian Hormone (AMH) is a hormone secreted by cells in developing egg sacs (follicles). The level of AMH in a woman's blood is generally a good indicator of her ovarian reserve.
- PRL (Prolactin): Prolactin is a hormone produced by your pituitary gland which sits at the bottom of the brain whose two primary responsibilities are milk production and development of mammary glands within breast tissues.
- PRG (Progesterone): Progesterone is a hormone released by the corpus luteum in the ovary. It plays important roles in the menstrual cycle and in maintaining the early stages of pregnancy.
- Follicle No.: Ovarian follicles are small sacs filled with fluid that are found inside a woman’s ovaries. They secrete hormones which influence stages of the menstrual cycle. Follicle No. is the count of follicle count of egg containing follicles in the ovaries.
- Endometrium: The endometrium is the inner lining of the uterus. Each month, the endometrium thickens and renews itself, preparing for pregnancy. If pregnancy doesn't occur, the endometrium sheds in a process known as menstruation.

Visualizations

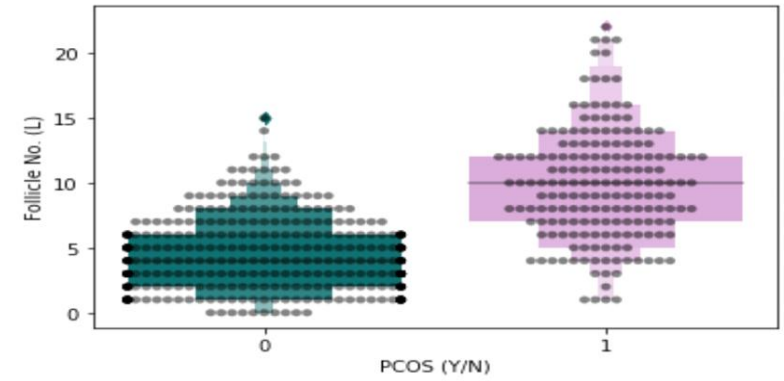
For visualizations, the comparative study of women with and without PCOS was done. Based on the attributes that majorly affected PCOS, count plots were plotted with hue as PCOS(Y/N). Given below is the count plot for PCOS(Y/N).

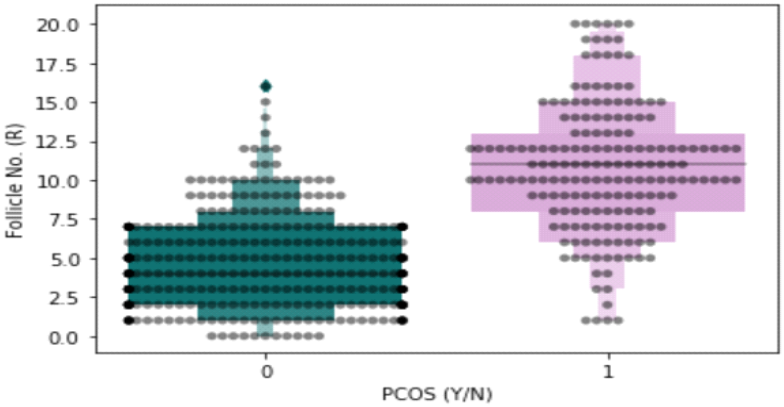


Following are the comparative plots for PCOS(Y/N) based on attributes like Fast Food, Hair growth, Skin darkening, Weight gain.

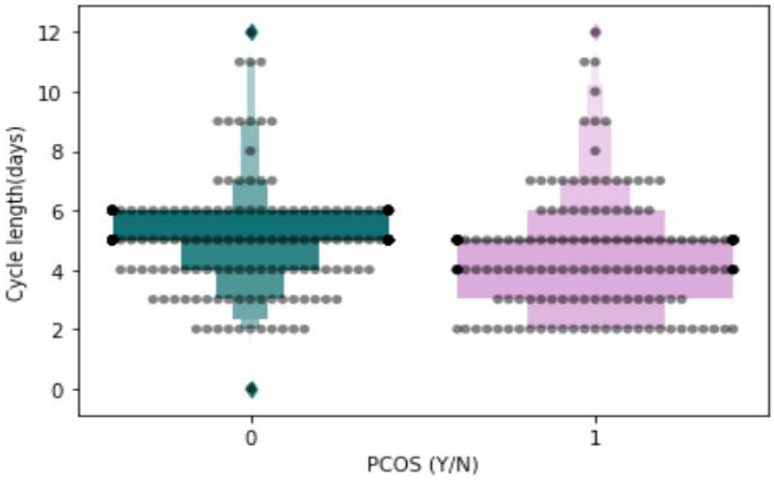


We can see that the amount of unwanted hair growth in women with PCOS is more as compared to women without PCOS. Also, the amount of consumption of fast food, weight gain and skin darkening is quite high. This proves that these are the basic factors causing PCOS in women.

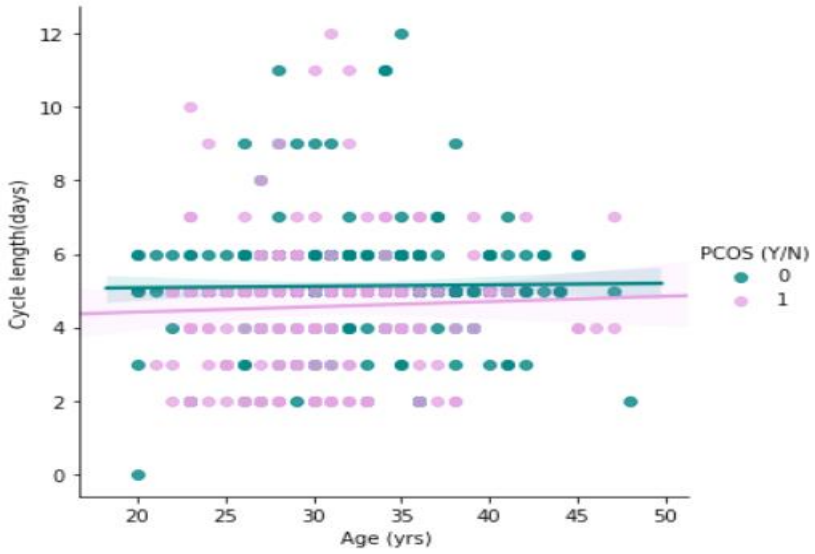




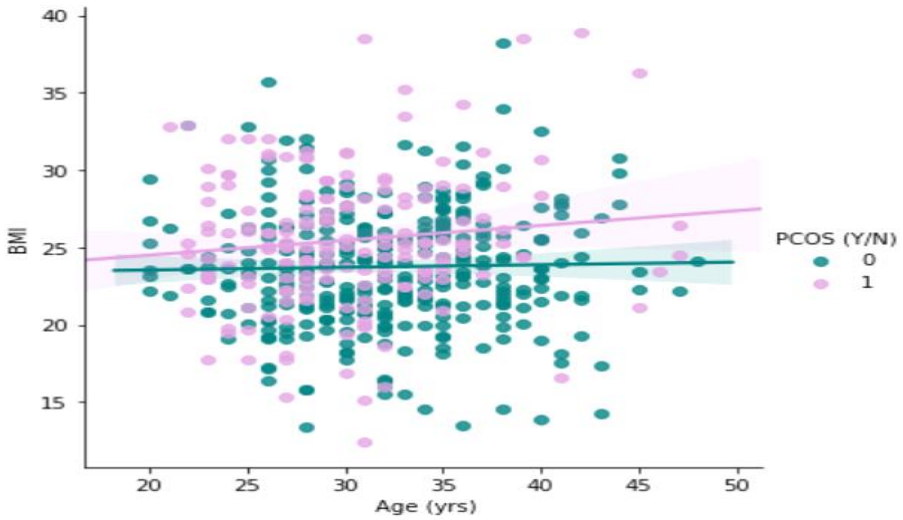
The number of follicles in women with PCOS is higher, as expected and are unequal as well.



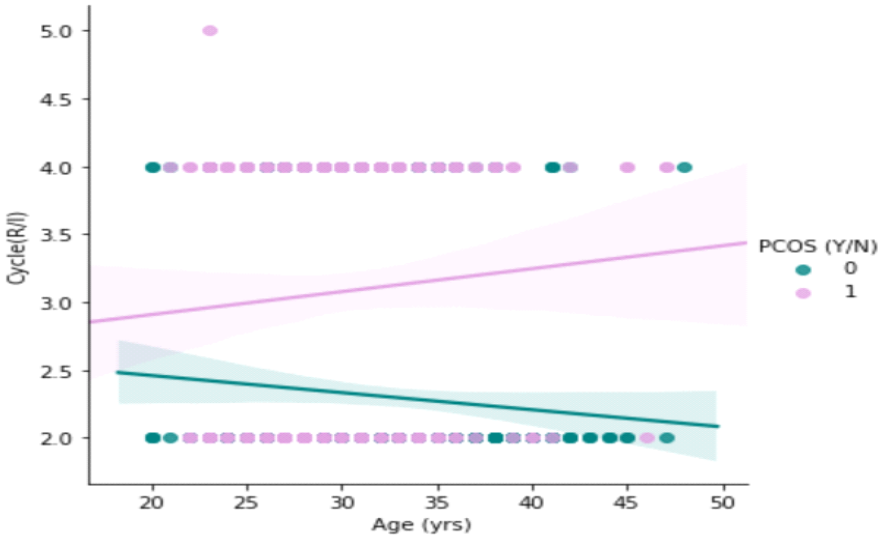
Women with PCOS having cycle lengths 2,4 and 5 (days) have maximum frequency and Women without PCOS having cycle lengths 5 and 6 (days) have maximum frequency.



The length of menstrual cycle is overall consistent over different ages for normal cases. Whereas in the case of PCOD the length increased with age.



Body mass index (BMI) is showing consistency for normal cases. Whereas for PCOS the BMI increases with age.



4 indicates irregular menstrual cycle

2 indicates a regular menstrual cycle

The mensural cycle becomes more regular for normal cases with age. Whereas, for PCOS the irregularity increases with age.

Modelling

For analyzing the data, **Logistic Regression** algorithm was used.

Logistic Regression is a statistical modal that uses a logistic function to model a binary dependent variable. The algorithm uses logit function to model the probability of a certain class. Each object is assigned a probability between 0 and 1. Mathematically, a binary logistic model has a dependent variable with two possible values (pass/fail, success/failure) labelled as 0 and 1. The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier.

Why Logistic Regression?

\Logistic Regression algorithm is used for classification problems where the dependent variable is binary (0 or 1). As, the data set used has the target variable as binary and the problem is of classification, this algorithm was the best choice and would give better results.

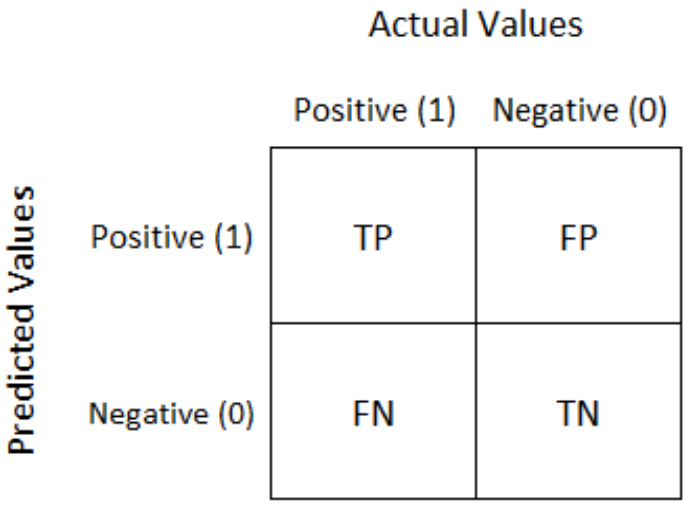
Results

Results of this modelling technique contains following two things -

- Confusion Matrix
- Classification Report

1) **Confusion Matrix** - A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

It basically looks like -



where,

TP (True Positive) - You predicted positive and it's true.

FP (False Positive) - You predicted positive and it's false.

FN (False Negative) - You predicted negative and it's false.

TN (True Negative) - You predicted negative and it's true.

2) **Classification Report** - The classification report visualizer displays the precision, recall, F1, and support scores for the model.

- **Recall** - Out of all the positive classes, how much we predicted correctly. It should be high as possible.

  Formula - Recall = TP/(TP+FN)

- **Precision -** Out of all the positive classes we have predicted correctly, how many are actually positive.

  Formula - Precision = TP/(TP+FP)

- **F-measure -** It is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean.

  Formula - F-measure=2*Precision*Recall/(Precision+Recall)

- **Support** - Support is the number of actual occurrences of the class in the specified dataset.

Following are the images of Confusion matrix and Classification report after fitting the regression model and ROC (Receiver operating characteristic) curve.

Confusion Matrix

```
from sklearn.metrics import confusion_matrix

confusion_matrix = confusion_matrix(y_test, y_pred)
print(confusion_matrix)
```

```
[[104   5]
 [ 12  42]]
```

Classification Report

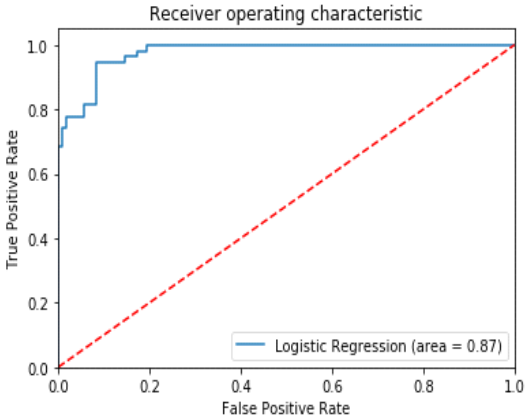
```
from sklearn.metrics import classification_report

print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.90      0.95      0.92       109
           1       0.89      0.78      0.83        54

    accuracy                           0.90       163
   macro avg       0.90      0.87      0.88       163
weighted avg       0.90      0.90      0.89       163
```



The accuracy of the logistic regression on test data set was 90%.

Interpretation of the ROC curve

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters True Positive Rate (TPR) and False Positive Rate (FPR) where TPR = TP / (TP + FN) and FPR = FP / (FP + TN). The area under the curve (AUC) ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0 and one whose predictions are 100% correct has an AUC of 1.0. As, the area is 0.87 implies that 87% of the predictions were correct.

PCOS data collection using google forms and Analysis

Primary Data Collection

Collecting data manually is always time consuming and that too in such pandemic situation, it was not feasible to do so.

We collected data online by generating a Google form and shared it across the population of age group in between 12-50 years which included information of both – women with PCOS and women without PCOS. Along with that, we designed a manual which included all the attributes along with the explanation of each attribute i.e., how the form needs to be filled and how the questions need to be answered. The form was open for about a month and we got exact 300 observations. Also, we observed that the data came out to be unbalanced i.e., we got 79.7% data of women without PCOS and 20.3% data of women with PCOS.

Link of Google Form:

https://forms.gle/pmtxMFCkupVxEvXLA

Link of Manual:

https://drive.google.com/file/d/1-1_e-a18zUFOypPQs231EoguIJA5cjIk/view?usp=drivesdk

Data

There were different sections i.e.., different set of questions based on PCOS and non PCOS and based on marriage i.e.., different set of questions for married and unmarried women. Sections included information of health and lifestyle, menstrual cycle, physical changes, personal and about PCOS and awareness regarding PCOS.

Attributes:

- Age
- Height (in cm)
- Weight (in kg)
- Sleeping hours
- Working hours
- Smoking

- Alcohol Consumption
- Diabetes
- Hypertension
- Junk food
- Mood swings
- Thyroid
- Stress
- Exercise
- First period age
- Period flow
- Period cycle
- Length of a period
- Stress during periods
- Stress affecting periods
- Hair loss
- Hair thinning
- Unwanted hair growth
- Weight gain
- Weight loss
- Acne
- Skin darkening
- Obesity
- Marital status
- Years of marriage
- Abortion/miscarriage
- Awareness about PCOS
- PCOS status
- PCOS detected - age

- Heredity
- Type of treatment
- Interval between periods

Data Cleaning

After data collection, the main part that comes is data cleaning. After observation, we found that there are some outliers in the data (in some attributes) that need to be cleaned.

For example, in height column, the expected input was in cm but there were some entries which were in inches, so height attribute was a part of data cleaning. Similarly, some attributes had outliers so those entries were replaced by a measure of central tendency i.e.., median.

Description and Coding of attributes

There are 37 attributes in all out of which 27 are categorical attributes and rest are numeric attributes. We found that all attributes are not required for data modelling techniques since they don't contribute much in modelling and the results may highly vary if those attributes are included but those attributes are used for visualization purpose.

Attributes which are removed for data modelling are:

- Diabetes
- Marital status
- Years of marriage
- Awareness about PCOS
- PCOS detected-age
- Abortion/Miscarriage
- Heredity
- Type of treatment
- Interval between periods

Following coding is done during analysis:

- YES – 1, NO - 0

- Periods: Regular – 1, Irregular – 0
- Period flow: Light - 0, Normal – 1, Heavy – 2

Data Modelling Techniques

Since, our target variable is categorical i.e.., the outcome for PCOS is either yes or no, there are some best suitable modelling techniques which must be used for analysis when target is categorical.

So, some techniques which we included in our project are:

- Logistic Regression
- Ensemble Learning techniques

We included 3 Ensemble Learning techniques in our project -

- Random forest algorithm
- Naïve Bayes classifier
- Ada Boost Classifier

I) Logistic Regression

Classification is among the most important areas of machine learning, and logistic regression is one of its basic methods. As discussed above about Logistic regression, the assumptions of Logistic regression are proved in this dataset too.

Result of Logistic Regression

Since, the data is unbalanced; the accuracy score is not of that importance. The classification report plays an important role when the data is unbalanced.

Accuracy score: 0.84

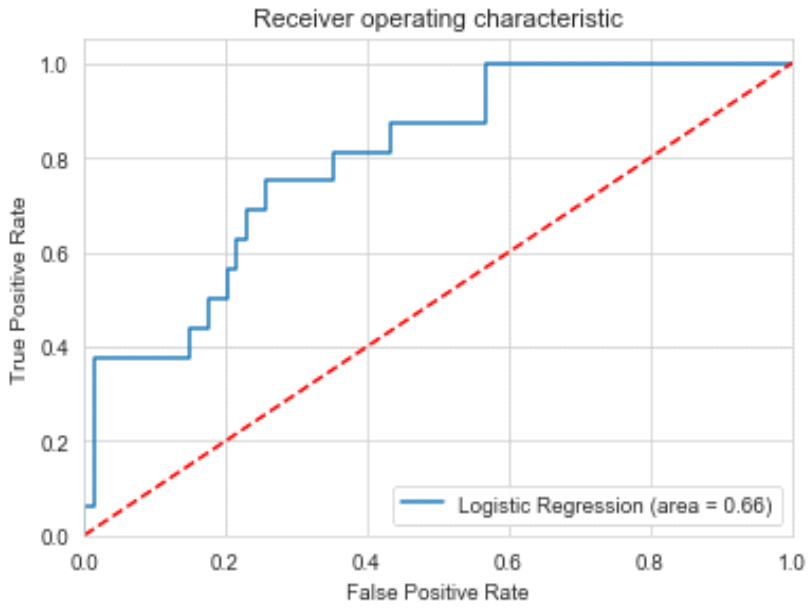
Confusion matrix: [70 4]

[10 6]

Classification report:

	Precision	Recall	F1-score	Support
0	0.88	0.95	0.91	74
1	0.60	0.38	0.46	16
Accuracy			0.84	90
Macro avg	0.74	0.66	0.69	90
Weighted avg	0.83	0.84	0.83	90

ROC Curve:



II) ENSEMBLE LEARNING TECNIQUES

Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve particular computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one. Other applications of ensemble learning include assigning a confidence to the decision made by the model, selecting optimal (or near optimal) features, data fusion, incremental learning, non stationary learning and error-correcting.

There are various ensemble learning techniques, but for our project we chose 3 out of all.

1) Random Forest Algorithm

- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **Ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- As the name suggests, **"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
- The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Why Random Forest?

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Results of Random Forest Algorithm

Confusion Matrix: `[67 1]`

`[16 6]`

Classification Report:

	Precision	**Recall**	**F1-Score**	**Support**
0	0.81	0.99	0.89	68
1	0.86	0.27	0.41	22
Accuracy			0.81	90
Macro avg	0.83	0.63	0.65	90
Weighted avg	0.82	0.81	0.77	90

2) Naive Bayes Classifier

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, which can be described as:

Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

We have used Gaussian model which is a type of Naïve Bayes Classifier, another model can also be used.

Gaussian: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

Results of Naive Bayes Classifier

Confusion Matrix:
```
[65 10]
[9 6]
```

Classification Report:

	Precision	Recall	F1-Score	Support
0	0.88	0.87	0.87	75
1	0.38	0.40	0.39	15
Accuracy			0.79	90
Macro avg	0.63	0.63	0.63	90
Weighted avg	0.79	0.79	0.79	90

3) Ada Boost Classifier

- **AdaBoost**, which stays for 'Adaptive Boosting', is a machine learning meta-algorithm which can be used in conjunction with many other types of learning algorithms to improve performance. AdaBoost helps you **combine multiple "weak classifiers" into a single "strong classifier"**.
- AdaBoost algorithms can be used for both classification and regression problem**.**
- Ada Boosting is best used to boost the performance of decision trees and this is based on binary classification problems.
- AdaBoost was originally called AdaBoost.M1 by the author. More recently it may be referred to as discrete Ada Boost. As because it is used for classification rather than regression.

Result of AdaBoost Classifier

Confusion Matrix: `[62 11]`

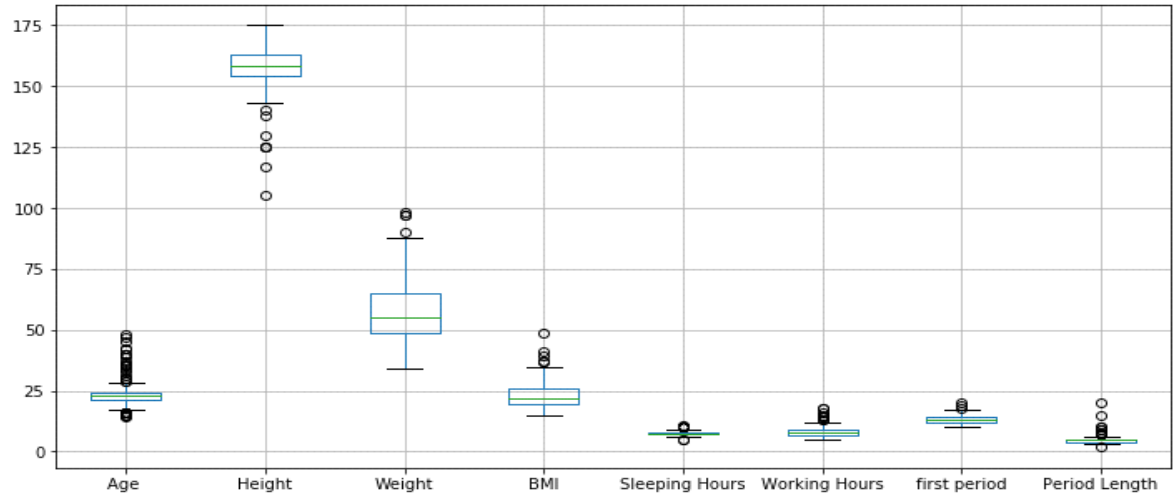
`[12 5]`

Classification Report:

	Precision	**Recall**	**F1-Score**	**Support**
0	0.84	0.85	0.84	73
1	0.31	0.29	0.30	17
Accuracy			0.74	90
Macro avg	0.58	0.57	0.57	90
Weighted avg	0.74	0.74	0.74	90

Interpretation of visualizations on PCOS data

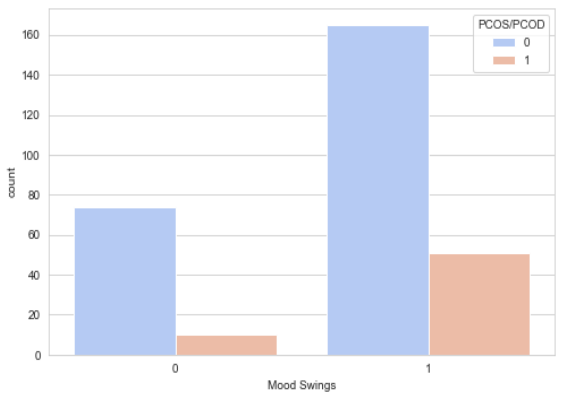
Data Visualization is an important skill. It provides an important suite of tools for gaining a qualitative understanding of data. Visuals accurately communicate and portray data and help us understand the patterns. Different ways of showcasing includes trends, pie charts, histograms, count plots and many more. Following are some graphs that help us understand our data.

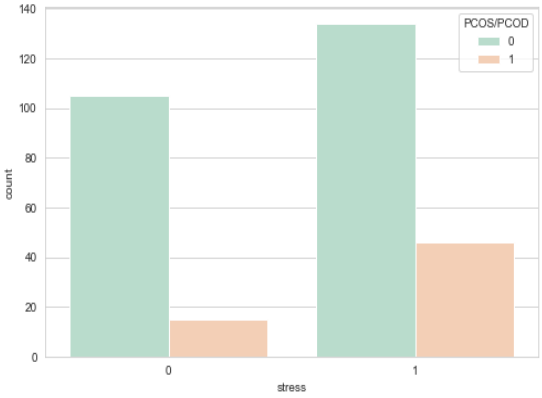


The above image shows the box plot for numeric attributes in the data. ***Box plots*** are a measure of how well distributed the data in a data set is. It divides the data set into three quartiles. This *graph* represents the minimum, maximum, median, first quartile and third quartile in the data set. The outliers can also be easily located using this plot.

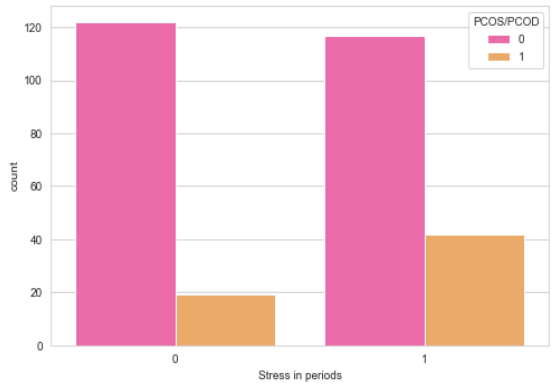
Comparative study between women with and without PCOS

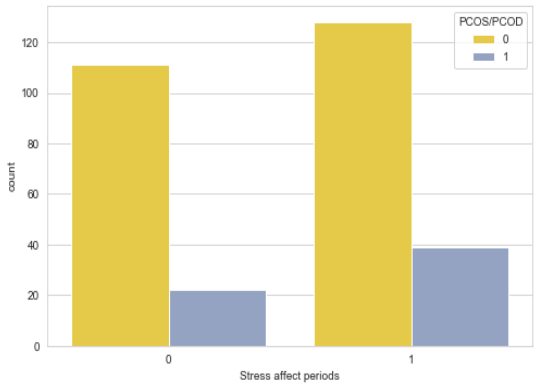
Following figures shows the count plot of attributes based on PCOS.



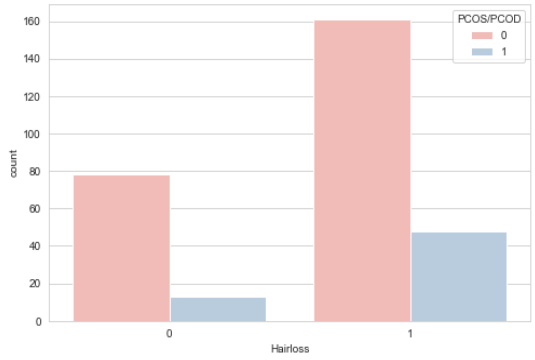


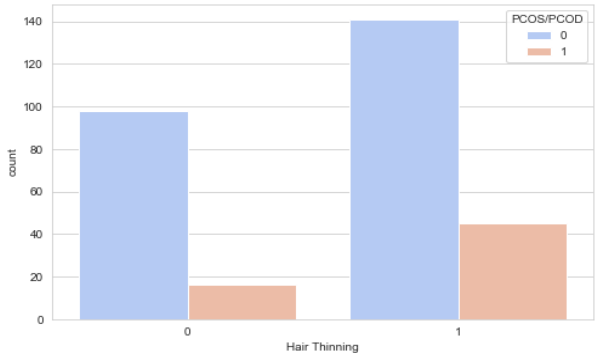
We can see that the count of mood swings and stress in women with PCOS is considerably high as compared to women without PCOS. Similarly, count of stress in periods and stress affects periods in women with PCOS is more than women without PCOS.



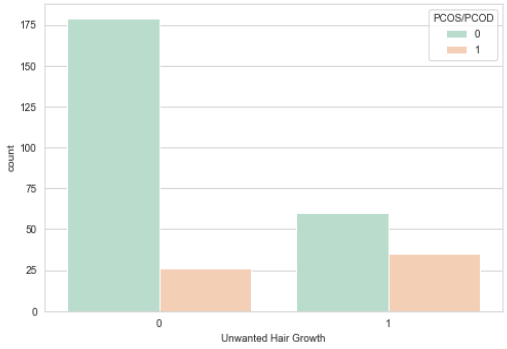


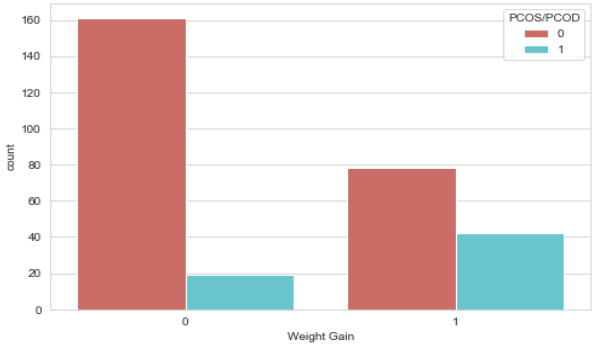
The following graphs show that the amount of hair loss and hair thinning in women with PCOS is more than women without PCOS. This results that hair loss and hair thinning are important indications of having PCOS.

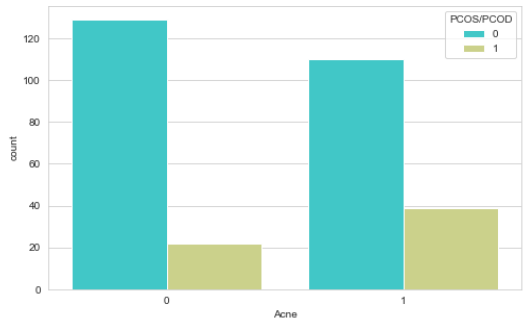


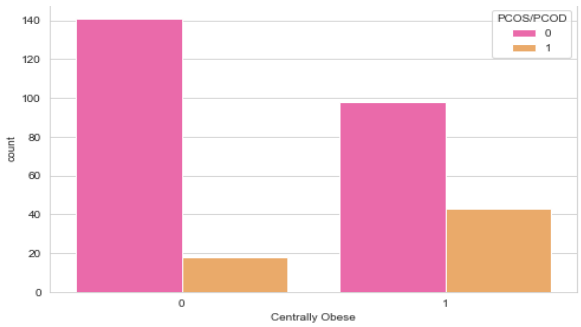


Similarly, unwanted facial hair growth, sudden weight gain, acne and obesity are important factors that may cause PCOS in women. Through the graphs provided below it is clear that the count of these attributes in women having PCOS is high than women without PCOS.







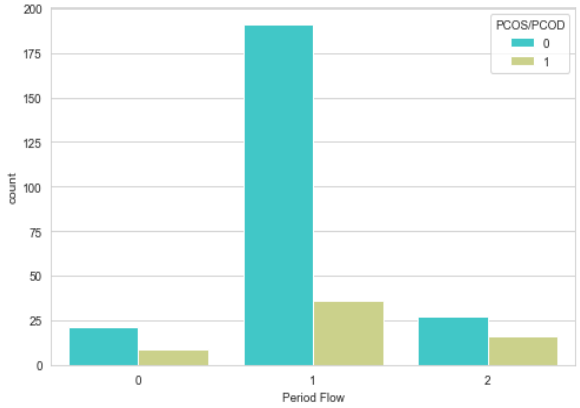


The following count plot shows the period flow of women with and without PCOS.

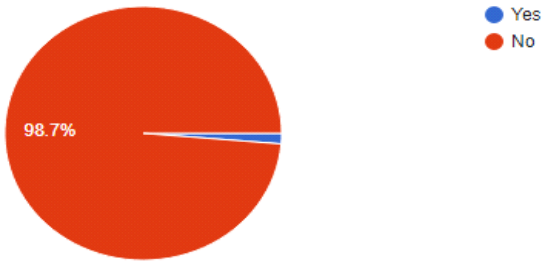
0 represent Light flow

1 represents Normal flow

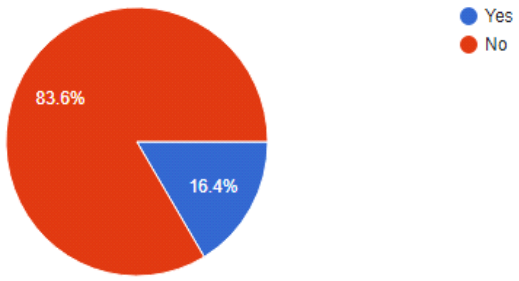
2 represent Heavy flow



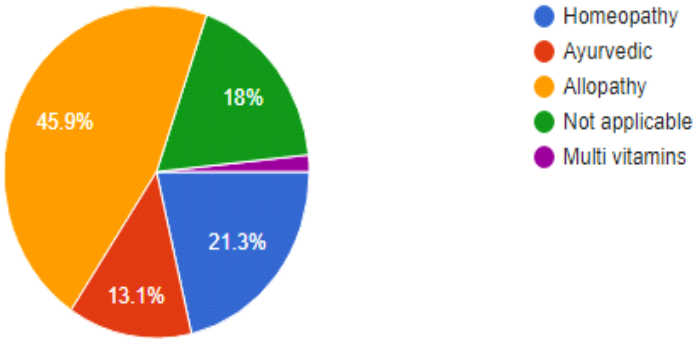
The pie chart below shows the percentage of diabetes in women. 98.7% women do not have diabetes while the count of women having diabetes is only 2.3%. This brings us to a conclusion that there is no correlation between PCOS and diabetes.



The pie chart below shows the count of women whose family had a member, diagnosed with PCOS i.e family history if any or hereditary.



The chart below shows the treatment which women with PCOS underwent. It shows that out of 20.3% women 45.9% underwent for allopathic medication, 21.3% homeopathy, 13.1% ayurvedic, 1.7% multivitamins and 18% women do not take any medications. Despite of having PCOS, 18% women do not take any kind of medications, this ignorance might lead to serious illness in women.



Conclusion

- After comparing the results of different methods applied on the dataset, we may conclude that Logistic Regression gives better **Accuracy** score (84%) than other methods.
- Considering **F1-score**, we see that Logistic Regression gives us a better result (46%).
- Random Forest Classifier gives us better **Precision** as compared with other methods (86%).
- Naive Bayes Classifier gives us better **Recall** (40%).

	Precision	**Recall**	**F1-Score**	**Accuracy**
1	0.86	0.40	0.46	0.84
Method	Random Forest Classifier	Naive Bayes Classifier	Logistic Regression	Logistic Regression

Limitations

Since, the data was unbalanced we used the same dataset for our analysis without any manipulations. No manipulation or extension in the dataset was done to make it balanced.

References

- https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8
- https://builtin.com/data-science/random-forest-algorithm
- https://youtu.be/0FLY-hfzHvk
- https://youtu.be/Az9lWdqebaU
- https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
- https://pubmed.ncbi.nlm.nih.gov/16790100/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3742326/
- https://www.rbmojournal.com/article/S1472-6483(10)61182-0/pdf
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4796416/
- https://www.yourhormones.info/hormones/anti-muellerian-hormone/
- https://www.hormone.org/your-health-and-hormones/glands-and-hormones-a-to-z/hormones/anti-mullerian-hormone-amh
- https://www.yourhormones.info/hormones/thyroid-stimulating-hormone/
- https://www.uofmhealth.org/health-library/ug1836
- https://www.sciencedirect.com/science/article/pii/S1110569016301510#:~:text=1.,to%20as%20high%20as%2026%25.
- https://www.womenshealth.gov/a-z-topics/polycystic-ovary-syndrome
- https://towardsdatascience.com/machine-learning-part-17-boosting-algorithms-adaboost-in-python-d00faac6c464
- https://blog.paperspace.com/adaboost-optimizer/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3742326/
- https://clinicaltrials.gov/ct2/show/NCT01859663
- https://www.ijpsonline.com/articles/a-casecontrolled-comparative-hospitalbased-study-on-the-clinical-biochemical-hormonal-and-gynecological-parameters-in-polycystic-o-3367.html
- https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/