Polycystic Ovary Syndrome and Menstrual Cycle Duration Prediction
Machine Learning Hackathon

Ananya Appan(IMT2017004), Seelam Lalitha(IMT2017027)
Swasti Shreya Mishra(IMT2017043)

November 24, 2019

1 Problem Statement

1.1 PCOS Prediction

Predicting the occurrence of PCOS (Polycystic ovary syndrome) in women.

1.2 Menstrual Cycle Duration Prediction

Predicting the length of menstrual cycle in women between menarche and menopause.

2 Novelty of Problem Statement

Polycystic ovary syndrome is a disorder involving infrequent, irregular or prolonged menstrual periods, and often excess male hormone (androgen) levels.

Complications of PCOS can include infertility ,miscarriage and premature birth. It can also lead to Non-alcoholic steatohepatitis - a severe liver inflammation caused by fat accumulation in the liver. In worse cases, abnormal uterine bleeding can also occur,leading to endometrial cancer (Cancer of the uterine lining). Apart from all this, it can manifest in our everyday lives in the form of sleeplessness, depression, anxiety and loss of appetite.

Not many are aware of this disorder, Even if they are, the symptoms are not very evident. No single test exists to diagnose PCOS.

The process of trying to diagnose PCOS can be both time-consuming and confusing for several reasons. Many disagree about how best to confirm if a person actually has PCOS. The sticking point: Does a woman have to have high levels of androgens – "male" hormones such as testosterone that is also produced

naturally by the female body – to have PCOS? Some groups say yes, others say no.

The cause for PCOS is unknown till date. There are controversies around what physical and biological factors are responsible for its manifestation. Thus, we decided that it would be a good idea to figure this out using ML.

Another unpredictable factor which a lot of women face is the estimation of the duration of their menstrual cycle. This depends on a lot of things including sleep cycle, food intake and living environment. A sudden change in any of these could lead to an unforeseen change in one's menstrual cycle duration.

Being girls, we decided to try to predict this as well. We felt that this is an issue which , due to it's "sensitiveness", is not talked about much. We felt that some awareness should be created about this, and that some action should be taken. Being mentally prepared always helps, and knowing for how long a cycle will last would definitely be useful.

3 Approach to the Problem Statement

There has not been much work on this until now in the field of data science. We worked on a dataset from 10 different hospitals across Kerala, India. The dataset contains all physical and clinical parameters to determine PCOS and infertility related issues. The dataset size is however small, only 542 rows and 42 columns. The link to the dataset can be found(kaggle.com) here.

Link to the project repository : GitHub

4 Analysis of the dataset

4.1 Data Pre-Processing

4.1.1 Inference of Columns

- PCOS (Y/N) - Whether the candidate has been diagnosed with PCOS or not
- Age (years) - Age of the candidate in years
- Weight (kg) - Weight of the candidate in kilograms
- Height (cm) - Height of the candidate in centimeters
- BMI - Body Mass Index of the candidate
- Blood Group - Blood Group already encoded as integers
- Pulse rate(bpm) - Pulse rate of the candidate in beats per minute.

- RR (breaths/min) - Respiratory Rate of the candidate.
- Hb(g/dl) - Haemoglobin levels of candidate in grams pre deciliter.
- Cycle length(days) - Duration of menstruation cycle.
- Marriage Status(Yrs) - The married life of the candidate in years.
- No. of abortions - The number of abortions of candidate.
- FSH(mlU/mL) - Follicle Stimulating Hormone of candidate.
- LH(mlU/mL) - Luteinizing Hormone of candidate.
- FSH/LH - Ratio of FSH and LH of the candidate.
- Hip(inch) - measure of candidate's hip in inches.
- Waist(inch) - measure of candidate's waist in inches.
- Waist/Hip ratio - ratio of measure of Waist and Hip
- TSH(mlU/mL) - Thyroid Stimulating Hormone.
- AMH(ng/mL) - Anti-Mullerian Hormone levels of candidate.
- PRL(ng/mL) - Prolactin levels of candidate.
- Vit D3 - Vitamin D3 levels of candidate.
- RBS(mg/dl) - Random Blood Sugar levels.
- BP_Systolic(mm/Hg) - Systolic Blood Pressure levels of candidate.
- BP_Diastolic(mm/Hg) - Diastolic Blood Pressure levels of candidate.
- Follicle no. (R) - Follicles generated by right ovary.
- Follicle no. (L) - Follicles generated by left ovary.
- Avg F size (L) - average size of follicles produced by left ovary.
- Avg F size (R) - average size of follicles produced by right ovary.
- Endometrium (mm) - endometrium thickness of candidate in mm.

4.1.2 Data Encoding

All the values in the columns have been encoded appropriately. So, we did not have to encode any data explicitly.

4.1.3 Missing Value Detection

The number of missing values in our dataset were only 2. These were found in the columns Marraige Status (Yrs) and Fast food (Y/N). Since there were only two, we decided to remove the rows entirely instead of imputing the missing data.

4.1.4 Removal of Extra Columns

By analysing the data, we realized that a lot of columns provided were redundant. An instance of this would be BMI, when columns for height and weight were already provided. Thus, the following columns were removed.

1)"Unnamed: 42" : This column was filled with only NaNs. Also, being "unnamed", we couldn't find it significant.

2)"Sl. No" , "Patient File No.": Index which in no logical way would contribute to the occurrence of PCOS.

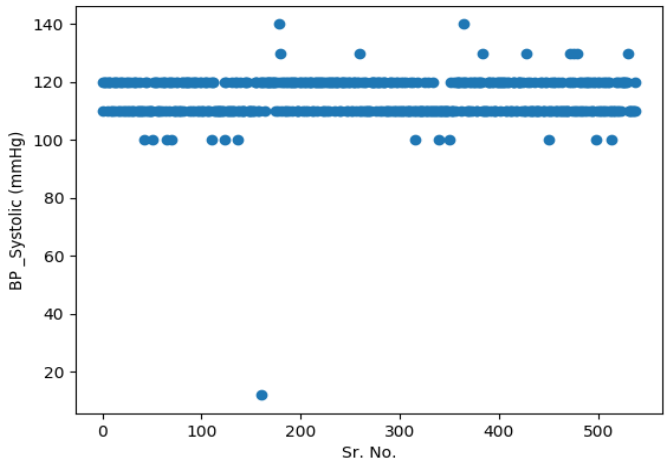
3)"FSH/LH" : Columns FSH(mIU/mL) and LH(mIU/mL) already exist. Hence, this column was redundant.

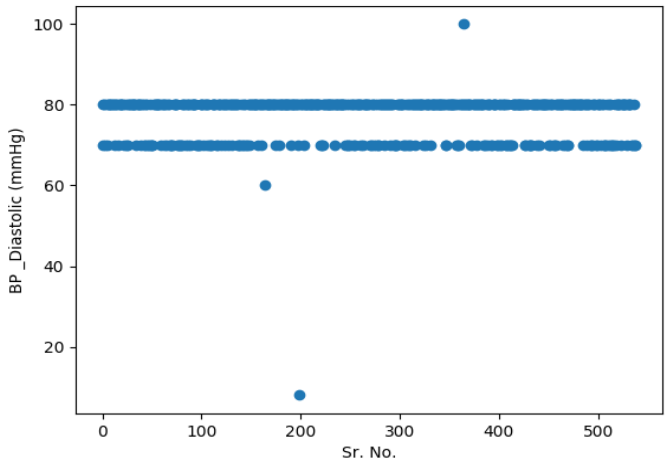
4)"Waist:Hip Ratio" : Columns Waist(inch) and Hip(inch) already exist. Hence, this column was redundant.

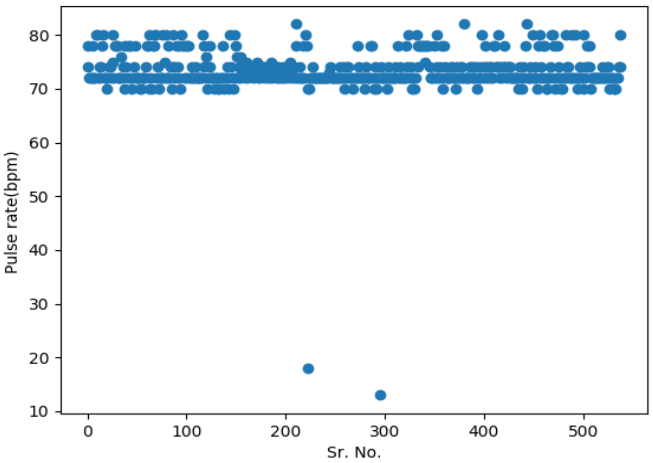
4.2 Exploratory Data Analysis

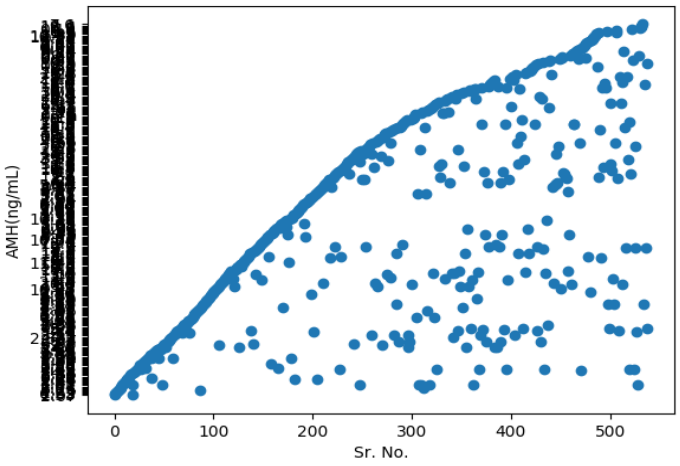
4.2.1 Outlier Detection

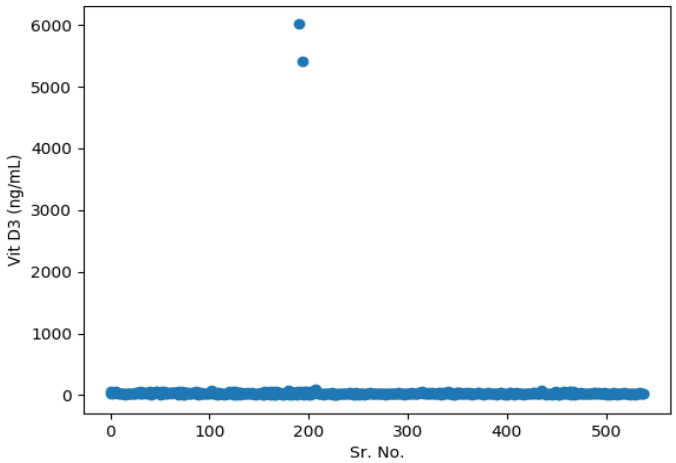
Histograms for each feature were plotted to detect any visible outliers.

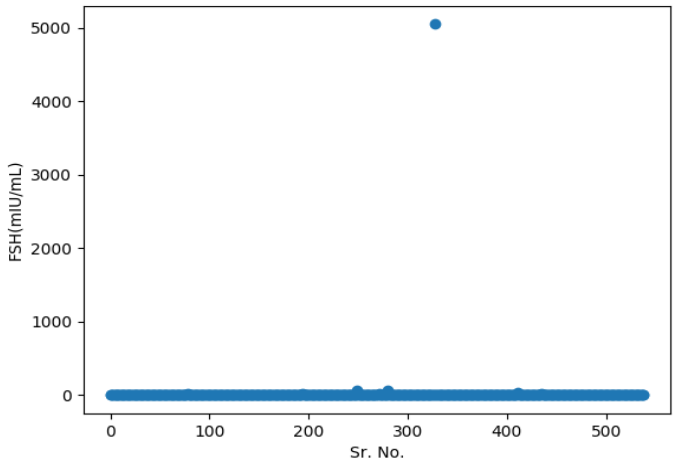

BP _Systolic (mmHg)
140
120
100
80
60
40
20
0
100
200
300
400
500
Sr. No.

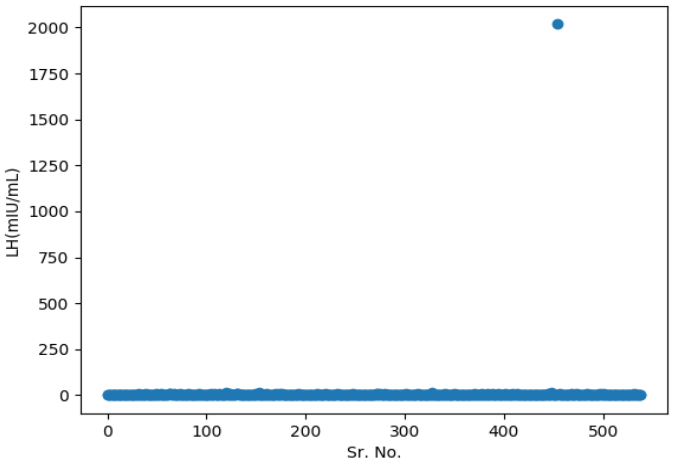

BP _Diastolic (mmHg)
100
80
60
40
20
0
100
200
300
400
500
Sr. No.

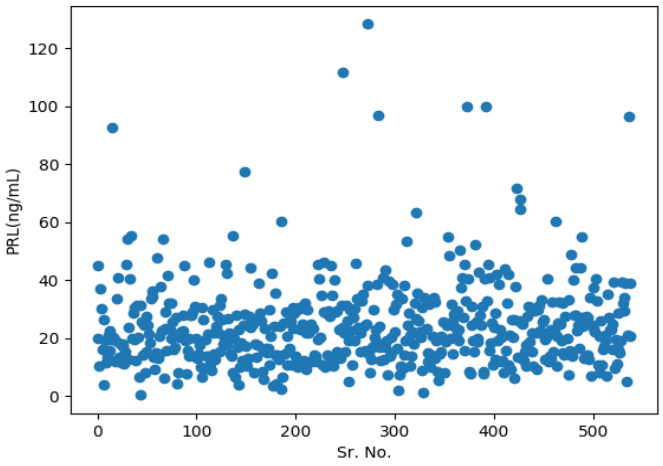

Pulse rate(bpm)
80
70
60
50
40
30
20
10
0
100
200
300
400
500
Sr. No.

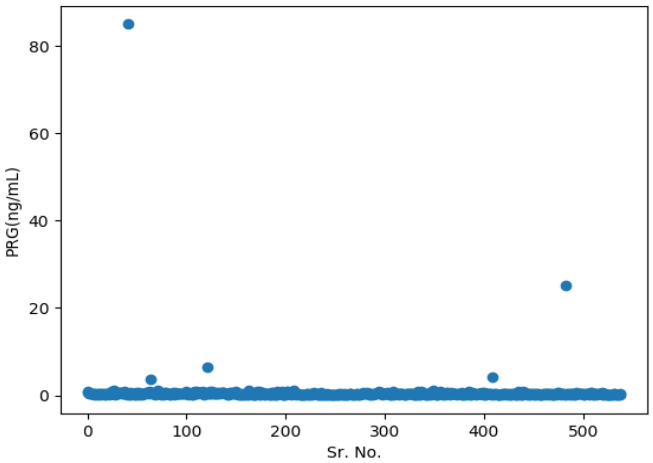

AMH(ng/mL)
0
100
200
300
400
500
Sr. No.


Vit D3 (ng/mL)
6000
5000
4000
3000
2000
1000
0
0
100
200
300
400
500
Sr. No.


FSH(mIU/mL)
5000
4000
3000
2000
1000
0
0
100
200
300
400
500
Sr. No.


LH(mIU/mL)
2000
1750
1500
1250
1000
750
500
250
0
0
100
200
300
400
500
Sr. No.

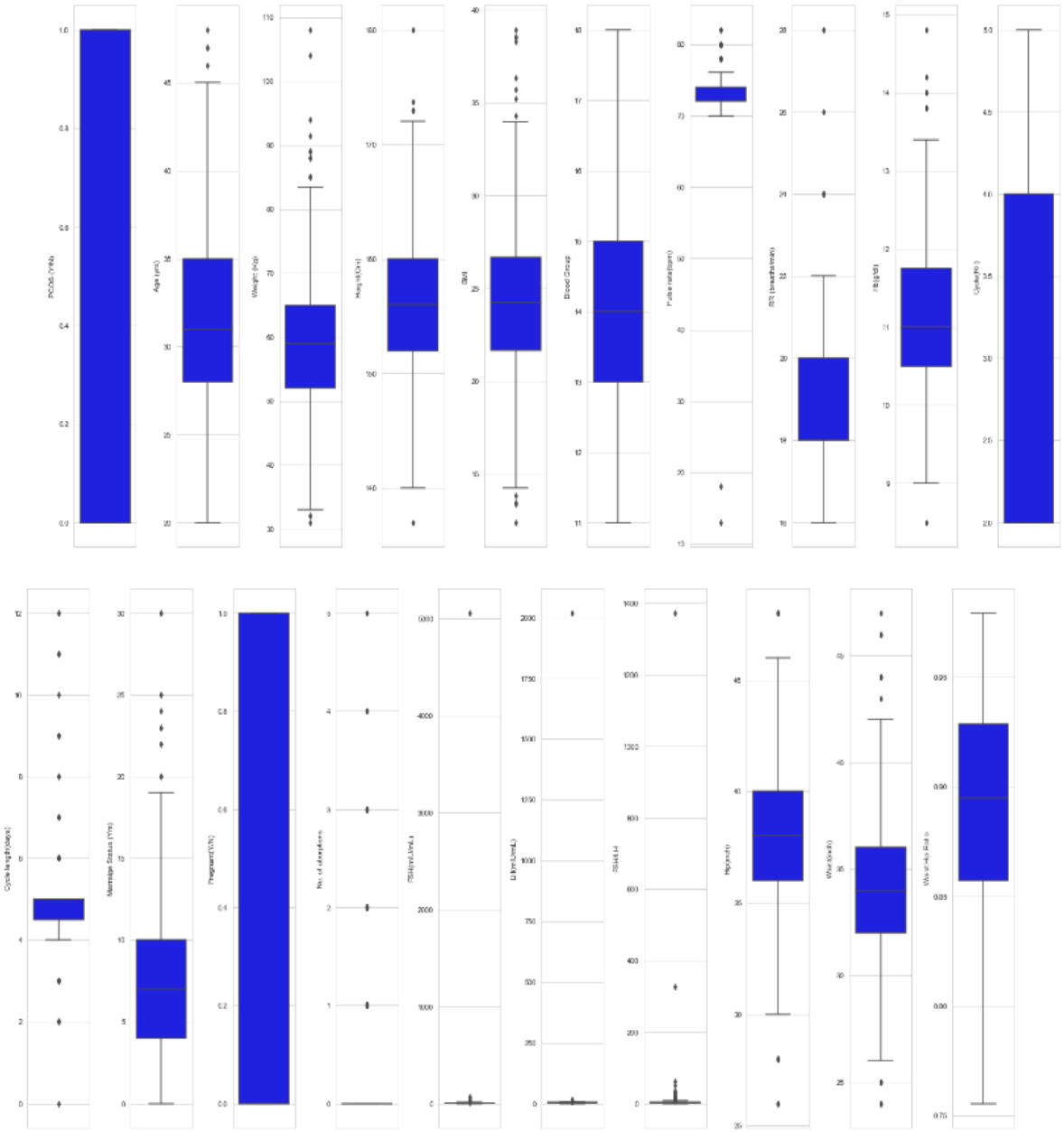

PRL(ng/mL)
120
100
80
60
40
20
0
0
100
200
300
400
500
Sr. No.

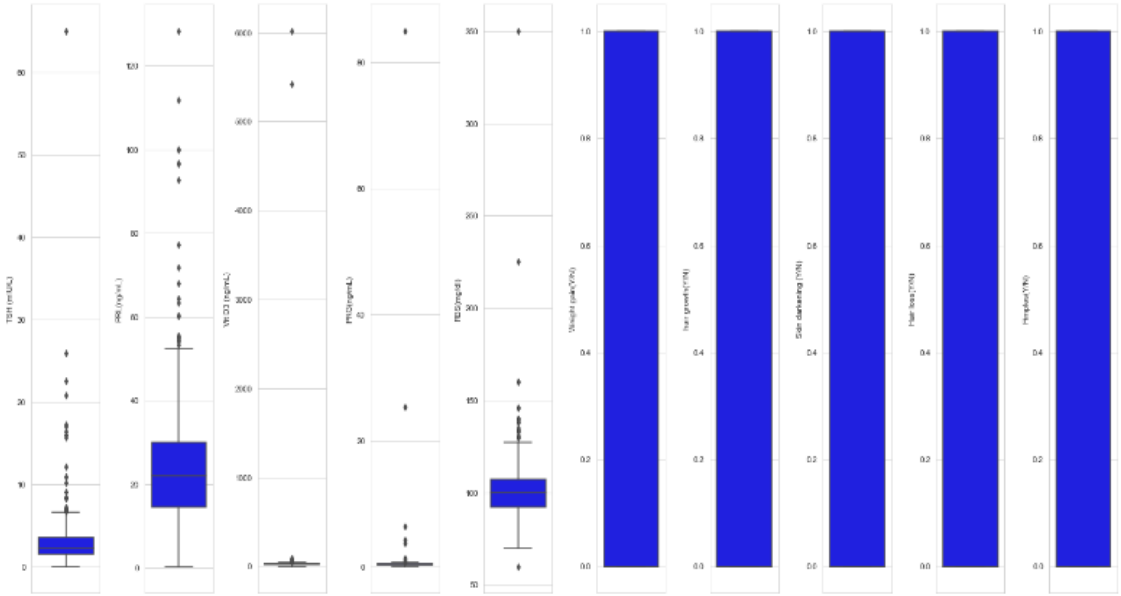


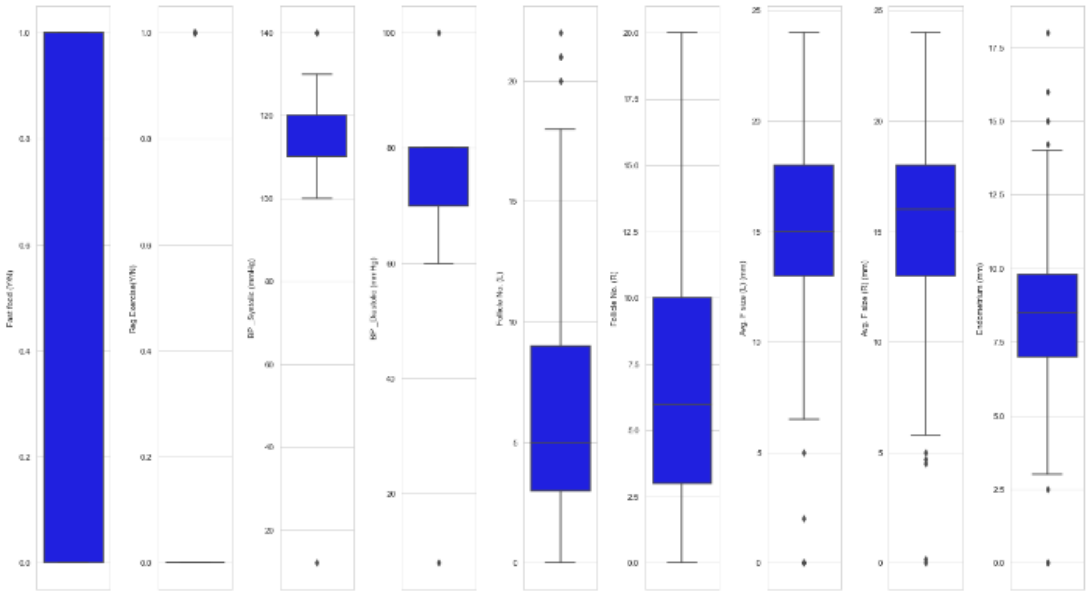
In addition to this, we printed the minimum and maximum for each column. Based on this, by printing the columns in ascending order , we figured out obvious outliers in the given data. The following is the minimum and maximum value for each of the corresponding columns:

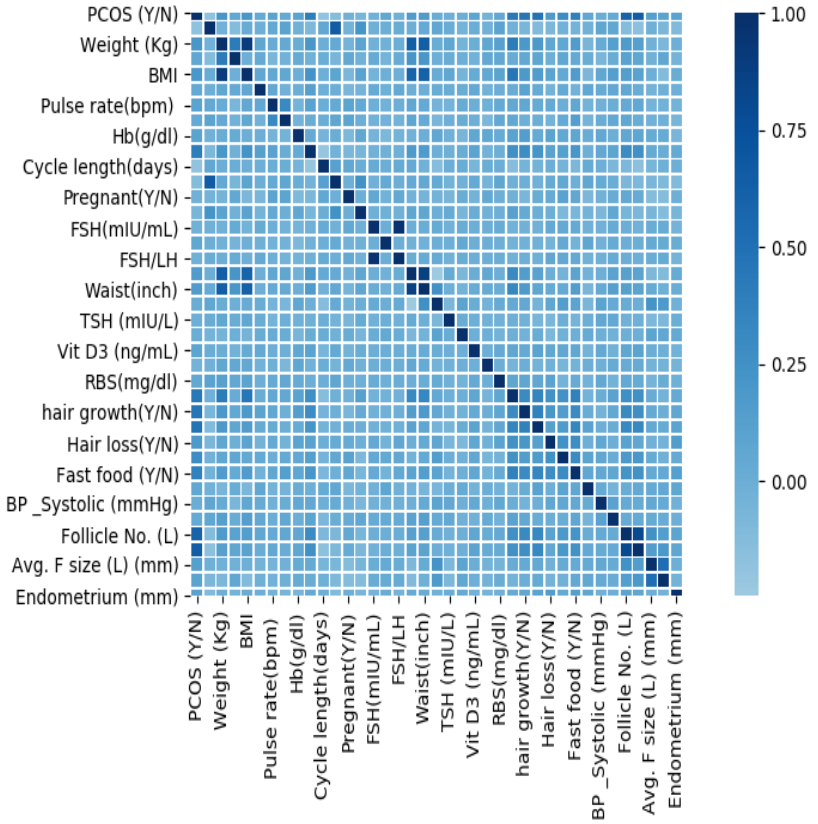
Column	**Minimum**	**Maximum**
BP Systolic(mmHg)	12	140
BP Diastolic(mmHg)	8	100
Pulse rate(bpm)	13	82
AMH(ng/mL)	0.1	a
Vit D3(ng/mL)	0.0	6014.66
FSH(mIU/mL)	0.21	5052.0
LH(mIU/mL)	0.02	2018.0
PRL(ng/mL)	0.4	128.24
PRG(ng/mL)	0.047	85.0

Boxplots for each column were plotted. The following figure shows the boxplots for all the considered columns.









By plotting the correlation matrix, we see that largely, there is not much correlation between the columns. A higher correlation is seen between Marriage Status (yrs) and Age (yrs) which is only to be expected. Other features with high correlation include BMI and Weight, and Weight and Waist / Hip Ratio. These redundant columns have been removed.

5 Model Building

The problem we are tackling is a classification problem. Furthermore, since, we know the expected output, it is a supervised Machine Learning problem. We had to be very careful about the given dataset. Firstly, the dataset was very small, having around 540 data points. Thus we had to be careful about retaining rows while removing outliers. Secondly, the dataset is very clean. Thus, not much data pre-processing was required.

We used the following models to classify the given data

5.1 Logistic Regression

PCOS Prediction:
Average Accuracy : 88.30996309963089
Maximum Accuracy : 94.85714285714286

Logistic regression gives a probability of occurrence of PCOS based on a threshold value, which is it's output. The baseline model in case of Logistic Regression is to predict the most frequent outcome as the outcome for all data points. In our case logistic regression was apt for mainly two reasons -

- From the co-relation matrix, it was evident that most of the features given were independent of each other. Logistic Regression tries to model the probability of an event occurring depending on the values of the independent variables.
- We require binary classification of output and logistic regression does it based on a threshold value

5.2 Random Forest

PCOS Prediction:
Average Accuracy : 87.03004744333153
Maximum Accuracy : 94.28571428571428

Menstrual Cycle Prediction:
Average Mean Absolute Error : 1.070294148655771
Average Mean Squared Error : 2.526651203654895
Average Root Mean Squared Error : 1.5850801604340476

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
The reason behind implementing the model is that, it is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier.

But after implementing the model and training it we found that the accuracy of the Logistic Regression model was better as compared to that of the Random Forest model. We can infer the following from this:

- Since random forests have been observed to overfit for some datasets with noisy classification/regression tasks, our dataset might be noisy.

- The no free lunch theorem tells us that if one algorithm outperforms another in one metric, it will lose in another metric. Therefore, our data might be simple(requires a linear model) and trying to fit a non-linear model worsens the accuracy over the train set.

For **Menstrual Cycle Prediction**, we have used Random Forest Regressor instead of Random Forest Classifier. The predicted days are further converted into integers by taking the ceil() of the floating point values of the days.

5.3 Naive Bayes

PCOS Prediction:
Average Accuracy : 83.33895624670537
Maximum Accuracy : 91.42857142857143

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

The reason behind implementing the Gaussian Naive Bayes model is as follows:

- They require a small amount of training data to estimate the necessary parameters. Considering the amount of available data, Naive Bayes model seemed like the apt model for us.
- Naive Bayes Classifier assumes strong independence condition among features. From the correlation matrix, it is evident that the features aren't highly correlated.
- The decoupling of the class conditional feature distributions in a Naive Bayes Classifier means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality. Since, we have a significant number of columns, we need to find the columns that contributes the most.

5.4 K-Nearest Neighbours

PCOS Prediction:
Average Accuracy : 73.00052714812864
Maximum Accuracy : 81.71428571428572

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

KNN hinges on this assumption being true enough for the algorithm to be useful. It captures the idea of similarity (sometimes called distance, proximity, or closeness) by calculating the distance between points on a graph. This

distance can be calculated in many ways, the most popular being the simple Euclidian distance. Other methods include Minowski Distance and Manhattan Distance.

It's quite tricky to predict the number of classes to use while working with KNN. In our case, we know that the output in binary, and hence, will use only two classes (K = 2).

However, the accuracy of the classifier was only around 73%. This may be because of the following reasons.

- KNN is generally used when we don't have the actual output, i.e, in unsupervised learning problem. However, in our case, we did have the actual output. We have not made use of this here.
- The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase. The fact that we have a very small data-set works in our favour here. But at the same time, the variables we are working with are independent, and more difficult to work with.

5.5 Linear Regression

Menstrual Cycle Prediction:
Average Mean Absolute Error : 1.0281075382182392
Average Mean Squared Error : 2.5819926199262015
Average Root Mean Squared Error : 1.5995378693133153

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y).

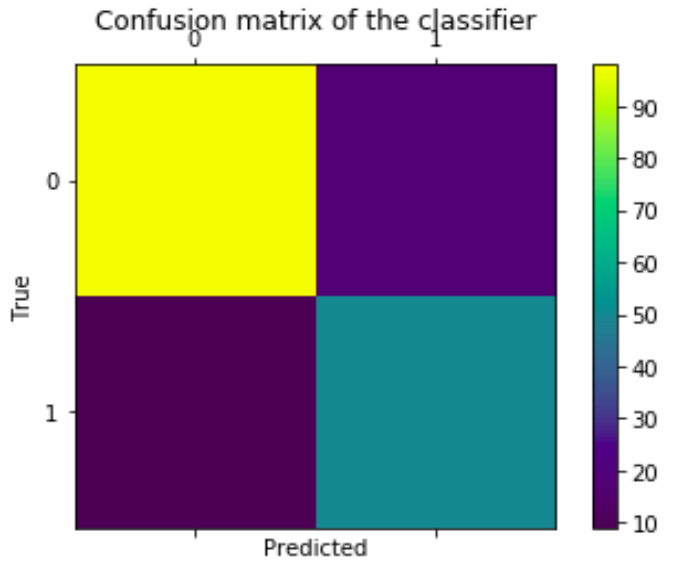
The reason behind our implementation of the Linear Regression model is that we had already implemented the Logistic Regression for PCOS Prediction and observed that a linear model fit the dataset well.
As expected, Linear Regression produced a better accuracy than that of Random Forest Regressor.

6 Result

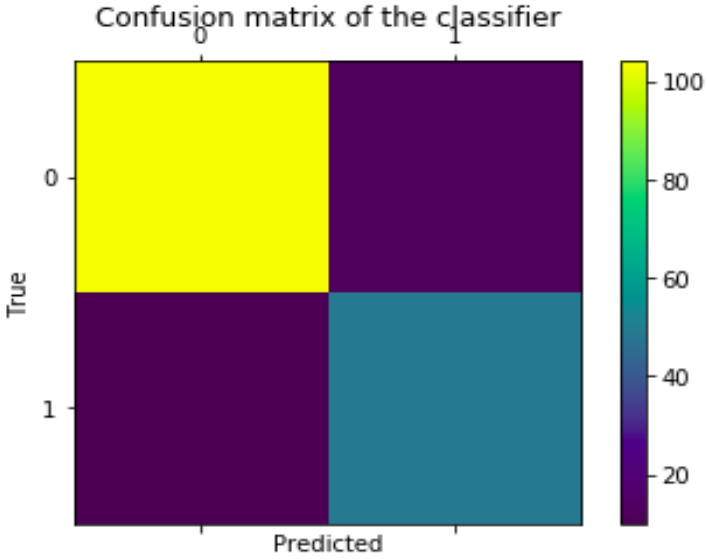
6.1 PCOS Detection

The below are the confusion matrix plots of the 4 models we have implemented:

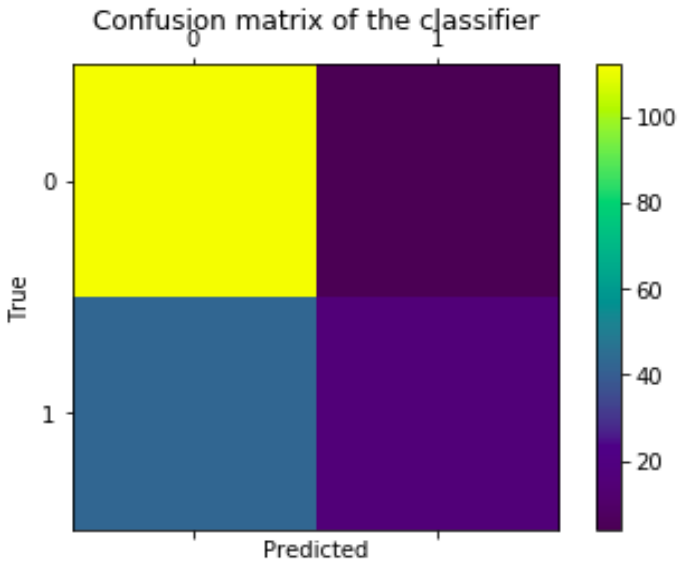
Logistic Regression



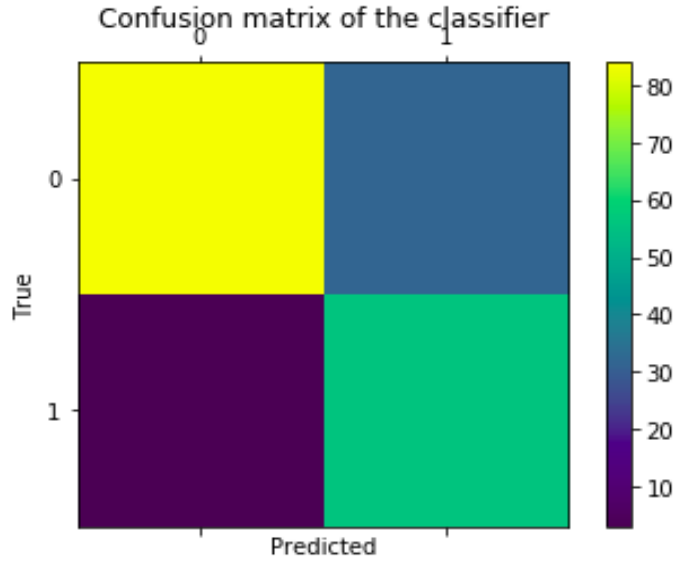
Random Forest



K-Nearest Neighbours



Gaussian Naive Bayes



The best results were obtained when we used Logistic Regression as our Machine Learning model. With an average accuracy of 88.3% and a maximum accuracy of 94% , we got good results even for the small data set we had to work with.

6.2 Menstrual Cycle Duration Prediction

The best results were obtained when Linear Regression was used. The average RMSE was found to be 1.59.