Project Report

On

# A PREDICTION MODEL FOR DETECTING THE RISK OF PCOS

*Submitted*

*in partial fulfilment of the requirements for the degree of*

MASTER OF SCIENCE

*in*

APPLIED STATISTICS AND DATA ANALYTICS

*by*


**DEVI D**

**(Register No. SM20AS010)**


**(2020 - 2022)**

*Under the Supervision of*

**Ms. VRINDA MURALEEDHARAN**



DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI - 682011


APRIL 2022

**ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM**


# CERTIFICATE

This is to certify that the dissertation entitled, **A PREDICTION MODEL FOR DETECTING THE RISK OF PCOS is a bonafide record of the work done by Ms. DEVI D under my guidance as partial fulfillment of the award of the degree of Master of Science in Applied Statistics And Data Analytics at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.**

Date:
Place: Ernakulam

**Ms. VRINDA MURALEEDHARAN**
Assistant Professor,
Department of Mathematics and Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

**Ms. Shanty. B.P**
Assistant professor & HOD,
Department of Mathematics,
St. Teresa's College(Autonomous),
Ernakulam.

**External Examiners**

1:..... 2: .....

# DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **Ms. VRINDA MURALEEDHARAN**, Assistant Professor, Department of Mathematics, St. Teresa's College(Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam. **DEVI D**

Date: **SM20AS010**

# ACKNOWLEDGEMENTS

Primarily I would like to thank the supreme power the almighty god for being able to complete this project with success. It is my privilege to express my profound gratitude to my project guide Ms. VRINDA MURALEEDHARAN, Assistant professor, Department of Mathematics and Statistics. St.Teresa's College Ernakulam, for the valuable support and guidance throughout my project.

I am extremely thankful for the teaching and non-teaching staff of the Department of Mathematics and Statistics.

I am also very thankful to HOD for the valuable suggestions, and critical examination of work during the progress.

I would like to express my gratitude to each and everyone who helped and encouraged me throughout this project and made this a successful one

Ernakulam. DEVI D

Date: **SM20AS010**

# ABSTRACT

Polycystic ovary syndrome (PCOS) has been determined as one of the serious health problems among women which affects women's fertility and leads to crucial health conditions. The condition once detected cannot be cured but treatment can help relieve its affects. Identifying PCOS is very challenging, as it must consider many gynecological & metabolic parameters to diagnose it. The clinical tests and scanning are time consuming and expensive. These are the reasons why most of the women neglect the early symptoms of PCOS. To rectify these problems this study put forward a symptom recognition and evaluation method which helps the patients to predict or identify PCOS at initial stages. Logistic regression and Random Forest techniques were used here in this study. The reason behind to build multiple models is to find out the best one for the given dataset.

# Contents

# Chapter 1

# INTRODUCTION

## 1.1 INTRODUCTION

PCOD (Polycystic Ovarian Disease) or PCOS (Polycystic Ovarian Syndrome) has taken over the female world. It is currently a big problem in India, affecting a large number of young females. PCOD strikes mostly at an early age, hence it affects a large number of young adults. This is a common endocrine condition that affects women of reproductive age and has an unknown cause. The average age of people with pcod is between 18 and 45 years old. It is critical for young people to understand this disease at its onset, along with the causes and implications in the future.

PCOD is characterized by multiple tiny cysts in the ovaries. It enlarges the ovaries and cause an excess of androgen and oestrogen hormones to be produced, resulting in bodily issues.

Polycystic Ovarian Disease (PCOD), also known as Polycystic Ovary Syndrome (PCOS) is a very common condition that affects 5% to 10% of women between the ages of 12 and 45. It's an issue caused by an imbalance in a woman's hormones. It can make it difficult for her to conceive and cause complications with her monthly periods. No ovulation, irregular periods, acne and hirsutism are the main characteristics. It can cause insulin resistant diabetes, obesity and high cholesterol leading to heart disease if not treated.

The ovaries normally produce female sex hormones and a small quantity of male sex hormones. During each menstrual cycle, these help regulate the correct growth of eggs in the ovaries. An imbalance in these sex hormones is linked to polycystic ovary syndrome. They begin to produce significantly more androgens during PCOS. As a result of this patients stop ovulating, develop acne, and develop excess facial and body hair. Follicles are egg-containing sacs within the ovaries. During each menstrual cycle, one or more eggs are usually released. This is referred to as ovulation. The eggs in these follicles do not develop and are not discharged from the ovaries in people who have pcos. Instead, they can cause a tiny cysts in the ovary, hence the name polycystic ovaries. PCOS appears to run in families, so if other women in our family have PCOS , irregular periods ,or diabetes , our chances of getting it are increased.

Acne, Weight gain and losing weight , Extra hair on the face and body are common symptoms of PCOD/PCOS. Women frequently develop thicker and darker facial hair, patches of thick, darker, velvety skin, thinning hair on the scalp, irregular periods. Women with PCOS typically have fewer than nine periods per year. Some women do not have periods, while others have very heavy bleeding, fertility problems.

The prediction of PCOS is not a trivial one, because there are so many symptoms connected with PCOS. However , it is not necessary for all symptoms to be appeared in PCOS patient. Only a few women get the same set of symptoms, which can vary according to their lifestyle. There is currently no single test to diagnose PCOS.

### 1.1.1 Motivation of the Study

PCOS is difficult to diagnose since numerous gynecological and metabo lic characteristics must be considered. Clinical examinations and scanning are time consuming and costly. These are the reasons why the majority of women ignore the early signs of PCOS. To address the issues, this study proposed a method for symptom recognition and evaluation that can assist patients in predicting or detecting PCOS in its early phases. It only suggests a clinical test if the patient is in a high-risk situation. Various data science and machine learning algorithms were

used to create this system. Medical parameters required to run the program in set to an optimum which further reduces the cost.

### 1.1.2 Objective of the Study

- The prime objective of the study is to get familiarize some machine learning technique like logistic regression and random forest.
- To compare the performance of this model and determine the best model among them.

### 1.1.3 Source of the Data

The secondary data is collected from the site Kaggle. It contains samples from ten different hospitals across Kerala, India. The identities of the patients have not been disclosed. The data contains 541 samples and 43 variables. Only a few variables are selected for the study.

# Chapter 2

# LITERATURE REVIEW

## 2.1 LITERATURE REVIEW

Polycystic Ovarian Syndrome (PCOS) is a complex health disorder that affects women of reproductive age. It can be diagnosed based on clinical symptoms such as increased BMI, elevated hormone levels, hair loss, acne, skin darkening, cycle length, endometrial thickness, high blood pressure levels, and so on. Several studies have been conducted in order to discover the causes and symptoms of this disease, as well as to create diagnostic techniques that can detect PCOS early. Here is a summary of some of the papers that were relevant to our research.

The work by Subrata et al (2020) is about using data to diagnose polycystic ovarian syndrome (PCOS) in women. The dataset is subjected to a variety of classifiers, including gradient boosting, random forest, logistic regression, and hybrid random forest and logistic regression (RFLR). The top ten factors are well enough to indicate PCOS disease, according to the findings. PCOS patients can be reliably classified with RFLR.

Hassan et al. (2020) employed machine learning techniques such as Support Vector Machine, CART, Naive Bayes Classification, Random Forest, and Logistic Regression to diagnose PCOS from patient clinical data. As explanatory factors, clinical symptoms such as increased BMI, higher hormone levels, hair loss, acne, skin darkening, hirsutism, cycle duration, endometrial thickness, high blood pressure levels, and so on are used. In this study, the Random Forest algorithm was found to have the highest accuracy in the diagnosis of PCOS.

Namrata (2020) compared PCOS and associated factors using two distinct classifiers: linear and nonlinear. KNN is a linear classifier, while the model of Logistic Regression is a nonlinear classifier. F1 score was found to be more effective for logistic regression than KNN classifier in her study, and the superior model was chosen as logistic regression.

# Chapter 3

# METHODOLOGY

Building an efficient statistical model, comparing the performance of various exiting algorithms on the dataset must be carried out. Here, logistic regression and random forest are two statistical techniques used for the detection of PCOS. The description of the techniques used in the project are mainly taken from Chatterjee S, Ali S H (2006), Montgomery D.C, et al (2012), Draper N.R and Harry S (2003)

## 3.1 LOGISTIC REGRESSION

Logistic regression is an appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analysis, the logistic regression is a predictive analysis. Logistic regression is used to describe the data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

### 3.1.1 Models with a binary response variable

Consider the situation where the response variable in a regression problem takes on only two possible values, 0 and 1. These could be arbitrary assignments resulting from observing a qualitative response. Suppose that the model has the form,

$$y_i = x_i'\beta + \varepsilon_i$$

where $x_i' = [1, x_{i1}, x_{i2}, x_{ik}], \beta' = [\beta_0, \beta_1, \beta_2, \beta_k]$ and the response variable $y_i$ takes on the value either 0 or 1.We will assume that the response variable $y_i$ is a Bernoulli random variable with probability distribution as follows:

| $y_i$ | Probability |
|---|---|
| 1 | $p(y_i = 1) = \pi_i$ |
| 0 | $p(y_i = 0) = 1 - \pi_i$ |

Now since $E(\varepsilon_i) = 0$, the expected value of the response variable is

$$E(y_i) = 1(\pi) + 0(1 - \pi_i) = \pi_i$$

This implies that

$$E(y_i) = x_i'\beta = \pi_i$$

This means that the expected response given by the response function $E(y_i) = x_{\cdot i}\beta$ is just the probability that the response variable takes on the value 1. Since the response is binary, then error terms $\varepsilon_i$ can take only two values,
i.e.

$$\varepsilon_i = 1x_i'\beta \; when \; y_i = 1$$

$$\varepsilon_i = -x_i'\beta \; when \; y_i = 0$$

Thus, the error is not normal. Also, the variance of error is not constant,
i.e.

$$\sigma_{yi}^2 = E(y_i)[1 - E(y_i)]$$

since $E(y_i) = x_i'\beta = \pi_i$ This indicates that the variance of each observations is a function of mean. There arises an additional constraint on response function, because

$$0 \leq E(y_i) = \pi_i \leq 1$$

This restriction can cause serious problem with the choice of a **linear response function**, because it would be possible to fit a model to the data for which the predicted value of the response lies outside the interval [0, 1]. Generally, when the response variable is binary, there is a considerable empirical evidence indicating that the shape of the response function should be nonlinear. A monotonically increasing (or decreasing) S-shaped (or reverse S-shaped) function (shown in figure 3.1) is usually employed. This function is called the **logistic response function** and has the form,

$$E(y) = \frac{exp(y)}{1 + exp(y)} = \frac{exp(x'\beta)}{1 + exp(x'\beta)}$$

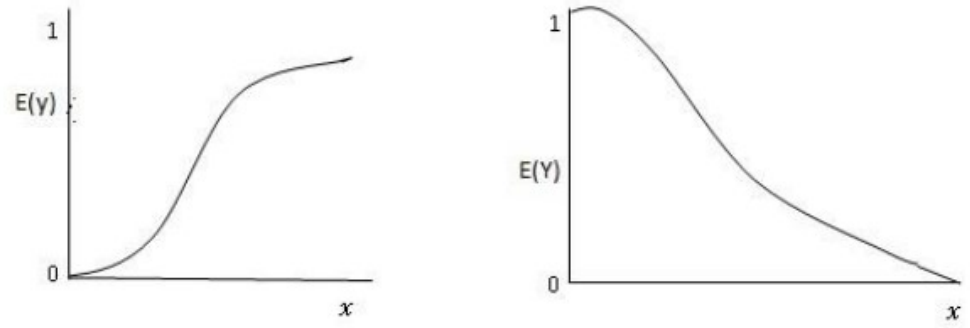



Figure 3.1:

### 3.1.2 Linear predicator and Link functions

The systematic component in $E(y)$ is the linear predictor and is denoted by $\eta_i = x_i'\beta, i = 1, 2, ..., n$. The link function in generalized linear model relates the linear predictor $\eta_i$ to the mean response say $\mu_i$. Thus $g(\mu_i) = \eta_i$ *or* $\mu_i = g^{-1}(\mu_i)$where g(.) is the linked function. In the usual linear models based on normality assumption, the link $g(\mu_i) = \mu_i$is used and is called as identity link. A link function maps the range of $\mu_i$ onto the whole real line provides good empirical approximation and carries meaningful interpretations in real applications.

In the case of logistic regression, the link function is defined as $= \log \frac{\pi}{1-\pi}$ This transformation is called logit transformation of probability $\pi$ and the ratio $\frac{\pi}{1-\pi}$is called odds in favor. The link $\eta$ is also called the log-odds. The link function is obtained as follows.

$$\pi = \frac{1}{1 + exp(-\eta)}$$

OR

$$\eta(1 + exp(-\eta)) = 1$$

OR

$$exp(-\eta) = \frac{1 - \pi}{\pi}$$

OR

$$\mu = \log \frac{\pi}{1 - \pi}$$

### 3.1.3 Maximum likelihood estimation of parameters

Consider the general form of the logistic regression model

$$y_i = E(y_i) + \varepsilon_i$$

where $y_i'$ are independent Bernoulli random variables with parameter $\pi$ with

$$E(y_i) = \pi_i = \frac{exp(x_i'\beta)}{1 + exp(x_i'\beta)}$$

The PDF of $y_i$ is

$$f_i(y_i) = \pi_i^{yi}(1 - \pi_i)^{1-y_i}, i = 1, 2, ...n; y_i = 0 \; or \; 1$$

The likelihood function is given by,

$$L(y, \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{yi}(1 - \pi_i)^{1-y_i}$$

The log likelihood is given by

$$\log L(y, \beta) = \Sigma_{i=1}^{n}[y_i \log \pi_i + (1 - y_i)]$$

$$= \Sigma_{i=1}^{n} y_i \log \frac{\pi_i}{1 - \pi_i} + \Sigma_{i=1}^{n} \log(1 - \pi_i)$$

$$Now \; since \; 1 - \pi_i = [1 + exp(x_i'\beta)]^{-1}$$

$$and \quad \eta_i = \log \frac{\pi_i}{1 - \pi_i} = x_i'\beta$$

The log likelihood function can be written as

$$\log L(y, \beta) = \sum_{i=1}^{n} y_i(x_i'\beta) - \sum_{i=1}^{n} \log[1 + exp(x_i'\beta)]$$

Often in logistic regression models, we have repeated observations on trials at each level of the $x$ variables. Let $y_i$ represent the number of one's observed for the $i^{th}$, observation and $n_i$ be the number of trials at each observation. Then the log likelihood function becomes

$$\log L(y, \beta) = \sum_{i=1}^{n} y_i\pi_i + \sum_{i=1}^{n} n_i \log(1 - \pi_i) - \sum_{i=1}^{n} y_i \log(1 - \pi_i)$$

The maximum likelihood estimate, $\hat{\beta}$ of $\beta$ can be obtained by numerical maximization. If the model assumptions are correct, then we can show that asymptotically $E(\hat{\beta}) = \beta$ and $V(\hat{\beta}) = (x'v^{-1}x)^{-1}$where $v$ is the dispersion matrix for the error term $\varepsilon$. Also, the estimated value of the linear predicator is $n_i = x_i'\hat{\beta}$ and the fitted value of the logistic regression model is given by

$$\hat{y_i} = \hat{\pi}_i = \frac{exp(\hat{\eta}_i)}{1 + exp(\hat{\eta}_i)}$$

$$\frac{exp(x_i'\hat{\beta})}{1 + exp(x_i'\hat{\beta})}$$

### 3.1.4 Interpretation of the parameters

Suppose that a linear predictor has only a single regressor, so that the fitted value of the model at a particular value of $x$, say $x_i$ is,

$$\hat{\eta}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

and the fitted value at $x_i + 1$ is,

$$\hat{\eta}(x_i + 1) = \hat{\beta}_0 + \hat{\beta}_1(x_i + 1)$$

The difference in the two predicted values is,

$$\hat{\eta}(x_i + 1) - \hat{\beta}(x_i) = \hat{\beta}_1$$

Now, $\hat{\eta}(x_i)$ is just the log-odds when the regressor variable is equal to $x_i$ and $\hat{\eta}(x_i + 1)$ is just the log-odds when the regressor is $x_i + 1$. Therefore, the difference in the two fitted values is,

$$\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) = \log(odds_{x_i+1}) - \log(odds_{x_i})$$

$$= \log \frac{odds_{x_i+1}}{odds_{x_i}}$$

$$= \hat{\beta}_i$$

or we obtain the odds ratio,

$$\hat{O}_R = e^{\hat{\beta}_1}$$

The odds ratio can be interpreted as the estimated increase in the probability of success associated with a one-unit change in the value of the predicator variable. In general, the estimated increase in the odds ratio associated with a change of $d$ units in the predicator variable is exp $d\hat{\beta}_1$

### 3.1.5 Statistical Inference on Model Parameters

Statistical inference in logistic regression is based on certain properties of maximum likelihood estimators and on likelihood ratio tests. These are large sample or asymptotic results.

Likelihood ratio test:

The likelihood ratio test procedure compares twice the logarithm of the value of the likelihood function for the full model $(FM)$ to twice the logarithm of the value of the likelihood function of the reduced model $(RM)$ to obtain a test statistic,

$$LR = 2ln\frac{L(FM)}{L(RM)} = 2[ln\ L(FM) - \ (RM)]$$

$$where \quad (FM) = \Sigma_{i=1}^{n} y_i ln\hat{\pi}_i + \Sigma_{i=1}^{n}(n_i - y_i)ln(1 - \hat{\pi}_i)$$

$$and \quad (RM) = yln(y) - (n - y)ln(n - y) - nln(n)$$

which asymptomatically follows chi-square distribution. If the test statistic LR exceeds the upper $\alpha$ - percentage point of this chi - square distribution, we would reject the claim that the reduced model is appropriate.

Testing goodness of fit:

The goodness of fit of the logistic regression model can also be assured using a likelihood ratio test procedure. This test compares the current model to a saturated model, where each observation is allowed to have its own parameters. The deviance is defined as twice the difference in log–likelihood between this saturated model and the full model that has been fit to the data with estimated probability of success

$$\pi_i = exp\{x_i'\hat{\beta}\}/[1 + exp\{x_i'\hat{\beta}\}]$$

i.e.

$$D = 2ln\frac{L(saturatedmodel)}{L(FM)} = 2\Sigma_{i=1}^{n}[y_i ln\frac{y_i}{n_i\hat{\pi}_i} + (n_i - y_i)ln\frac{n_i - y_i}{n_i(1-\hat{\pi}_i)}]$$

When the logistic regression model is an adequate fit to the data and the sample size is large, the deviance has a chi-square distribution with $(n - p)$ degrees of freedom, where p is the number of parameters in the model. Small values of the deviance or large p-value imply that the model provides a satisfactory fit to the data, while large values of the deviance imply that the fitted model is not adequate.

## 3.2 RANDOM FOREST

Random forest algorithm is a supervised classification and regression algorithm. As the name suggests, this algorithm randomly creates a forest with several trees. Generally, the more trees in the forest the more robust the forest looks like. Similarly, in the random forest classifier, the higher the number of trees in the forest, greater is the accuracy of the results.

In the random forest approach, a large number of decision trees are created. The forest it builds is a collection of Decision trees, trained with the bagged method. The idea of bagging reduces the variation in the predictions by combining the result of multiple Decision trees on different samples of the data set. Every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. A new observation is fed into all the trees and taking a majority vote for each classification model.

An error estimate is made for the cases which were not used while building the tree. That is called an OOB(Out-of-bag) error estimate which is mentioned as a percentage.

### 3.2.1 Random forest algorithm

Random forest models reduce the problem of overfitting by introducing randomness by:

STEP 1: Draw a random bootstrap sample of size n (randomly choose n samples

from training data).

STEP 2: Draw a decision tree from bootstrap sample. At each node of tree, randomly select d features.

STEP 3: Split the node using features(variables) that provide best split according to objective function. For instance, by maximizing the information gain.

STEP 4: Repeat steps 1 to step 2, k times (k is the number of trees you want to create using subset of samples).

STEP 5: Aggregate the prediction by each tree for a new data point to assign the class label by majority vote i.e. pick the group selected by the greatest number of trees and assign new data point to that group.

## 3.3 CONFUSION MATRIX

A confusion matrix is basically N x N matrix and preferred for evaluating the performance of a classification model, where N represent the number of target classes. The matrix compares the prediction done by machine learning model with the actual target values. This gives us a holistic view for the performance of our classification model and what sorts of errors it is making. In the matrix there is value of True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) is given

| | | Actual data | |
|---|---|---|---|
| | | NO | YES |
| Prediction | NO | TN | FN |
| | YES | FP | TP |

Table 3.1: Confusion matrix

- True Positive (TP)

  The positives that are identified as positives. In other words, the number of patients with PCOS that are tested positive.

- False Positive (FP)

  The negatives that are identifies as positives. That means the number of healthy (without PCOS) patients that are tested positive. (Also known as a "Type I error.")

- True Negative (TN)

The negatives that were identified negatives. It means that healthy patients are tested negatives.

- False Negative (FN)
  The positives that were identified positive. In this case, the patients with PCOS are tested negative. (Also known as a "Type II error.)"

Recall: It is the ability of the model to find all the positives. It is also called sensitivity and is defined, for each class, as the ratio of true positives and the sum of true positives and false negatives.

$$Recall = TP/(TP + FN)$$

Precision: It is the ability of the model not to classify a sample positive that is actually negative. For each class (0 or 1), it is defined as true positives divided by the sum of true positives and false positives.

$$Precision = TP/(TP + FP)$$

F1 score: It is defined as the harmonic mean of precision and recall, a measure of test's accuracy where 1 is its best score and 0 is its worst. It is not easy to comparing the two model with low precision and high recall or vice versa. So, make them comparable, we use F-score.

$$F1score = 2 * ((Precision * Recall)/(Precision + Recall))$$

Accuracy: Out of all the classes, what proportion we predicted correctly.

$$Accuracy = \frac{Number\ of\ object\ correctly\ classified}{Total\ No.\ of\ Objects\ in\ the\ test\ set}$$

# Chapter 4

# DATA ANALYSIS

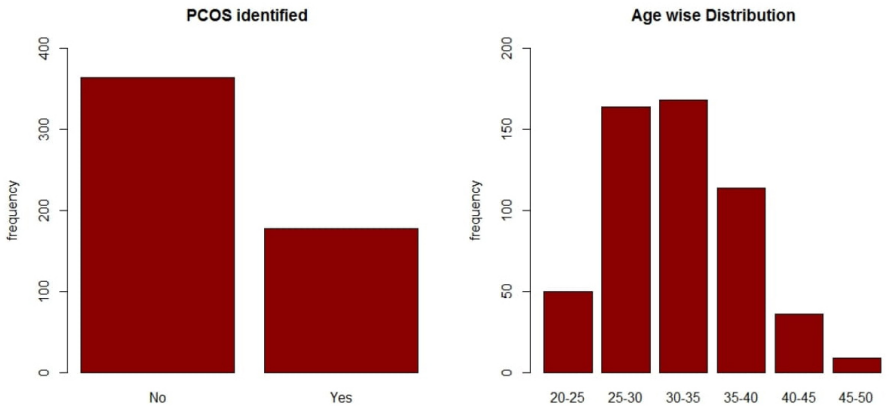
## 4.1 DESCRIPTION OF DATA SETS AND ATTRIBUTES

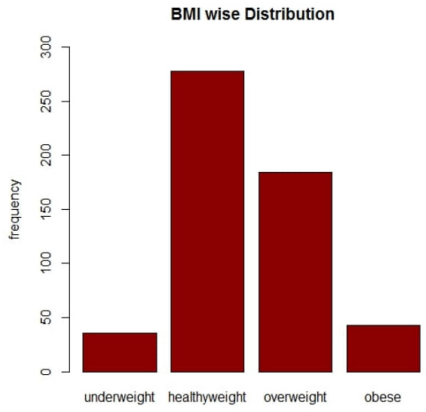
The PCOS dataset in this study is retrieved from Kaggle. The PCOS original dataset consist of 541 instances with 42 attributes in which one attributes as patient file number (not taken into consideration for data analysis). Finally, the total number of 41 attributes includes 40 as input attributes and PCOS as a class label [Positive (Yes) and Negative (No)]. The attributes are categorized as continuous, nominal, and ordinal.

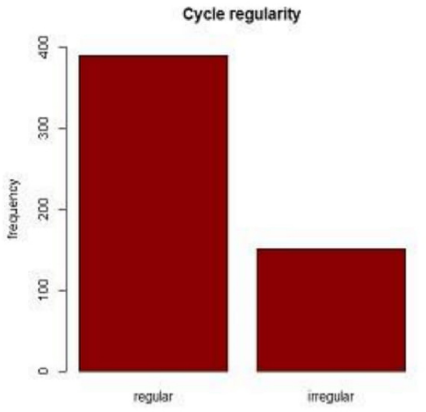
Attributes description and their units

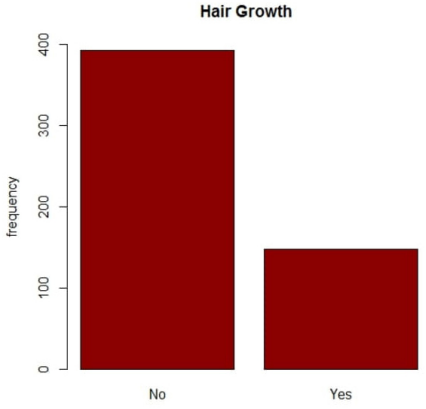
| No | Attributes | No | Attributes |
|---|---|---|---|
| 1 | Patient File number | 22 | Thyroid-Stimulating Hormone: TSH (mIU/L) |
| 2 | PCOS (class label) | 23 | Anti-Mullerian Hormone: AMH (ng/mL) |
| 3 | Age (Yrs) | 24 | Prolactin: PRL (ng/mL) |
| 4 | Weight (Kg) | 25 | Vit D3 (ng/mL) |
| 5 | Height (Cm) | 26 | Progesterone: PRG (ng/mL) |
| 6 | BMI: body mass index | 27 | BP-Systolic (mmHg) |
| 7 | Blood Group | 28 | Random Blood Sugar: RBS (mg/dl) |
| 8 | Pulse rate (bpm) | 29 | Weight gain (Y/N) |
| 9 | RR (breaths/min) | 30 | Hair growth (Y/N) |
| 10 | Haemoglobin: Hb (g/dl) | 31 | Skin darkening (Y/N) |
| 11 | Menstrual cycle: Cycle (R/I) | 32 | Hair loss (Y/N) |
| 12 | Cycle length (days) | 33 | Pimples (Y/N) |
| 13 | Marriage Status (Yrs) | 34 | Fast food (Y/N) |
| 14 | Pregnant (Y/N) | 35 | Reg. Exercise (Y/N) |
| 15 | No: of abortions | 36 | BP-Diastolic (mmHg) |
| 16 | Follicle stimulating hormone: FSH (mIU/mL) | 37 | Follicle No. (R) |
| 17 | LH (mIU/mL) | 38 | Follicle No. (L) |
| 18 | FSH/LH | 39 | Avg. F size (L) (mm) |
| 19 | Hip (inch) | 40 | Avg. F size (R) (mm) |
| 20 | Waist (inch) | 41 | Endometrium (mm) |

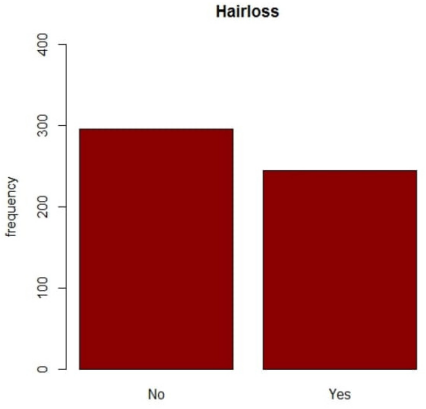
Description of the dataset In this data set, there are 541 samples of patients suspected and tested for PCOS from ten different hospitals across Kerala, India. For these 541 data bar diagrams are plotted for some of the attributes to give some insights to the data set and are given below.

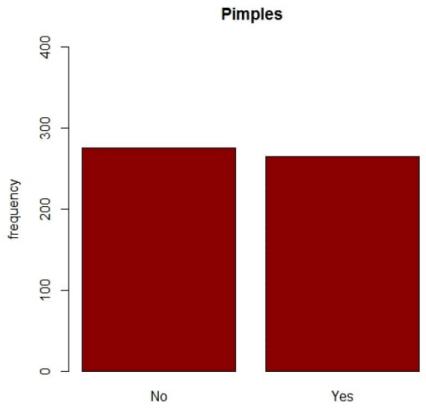


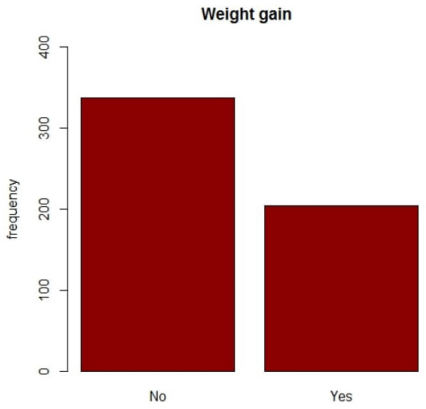

BMI wise Distribution
frequency
0
50
100
150
200
250
300
underweight
healthyweight
overweight
obese

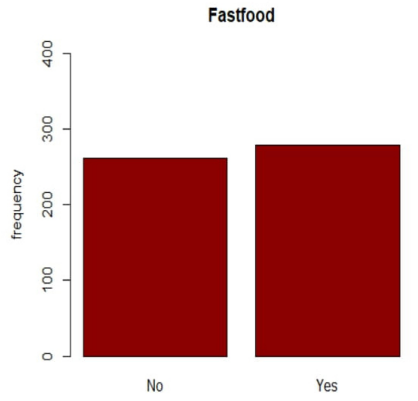

Cycle regularity
frequency
0
100
200
300
400
regular
irregular

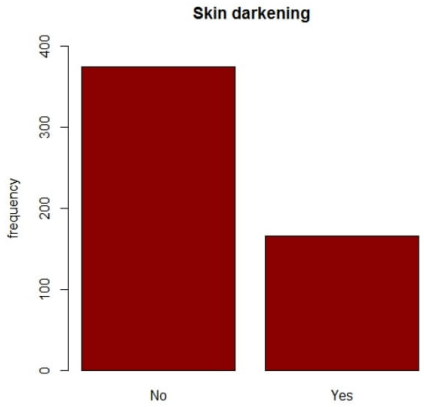

Hair Growth
frequency
0
100
200
300
400
No
Yes

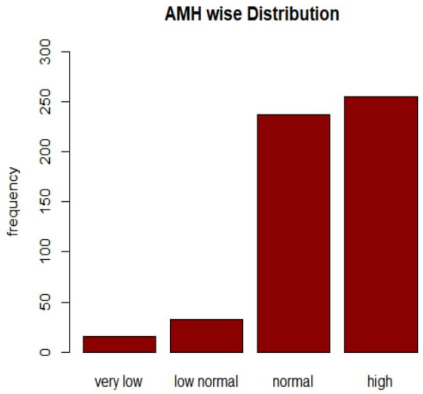

Hairloss
frequency
0
100
200
300
400
No
Yes

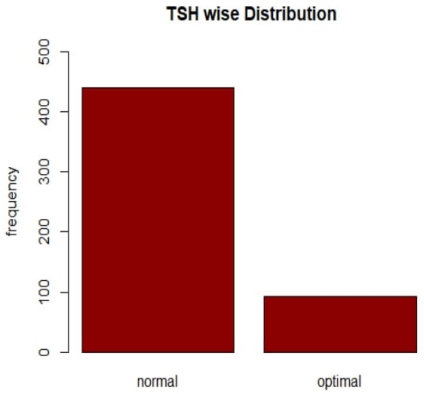

Pimples
frequency
0
100
200
300
400
No
Yes

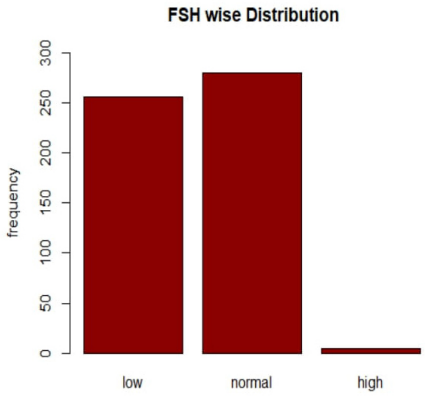

Weight gain
frequency
0
100
200
300
400
No
Yes

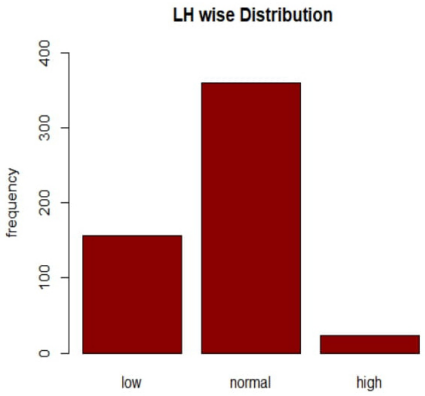

Fastfood
frequency
0
100
200
300
400
No
Yes


Skin darkening
frequency
0
100
200
300
400
No
Yes


AMH wise Distribution
frequency
0
50
100
150
200
250
300
very low
low normal
normal
high


TSH wise Distribution
frequency
0
100
200
300
400
500
normal
optimal


FSH wise Distribution
frequency
0
50
100
150
200
250
300
low
normal
high


LH wise Distribution
frequency
0
100
200
300
400
low
normal
high

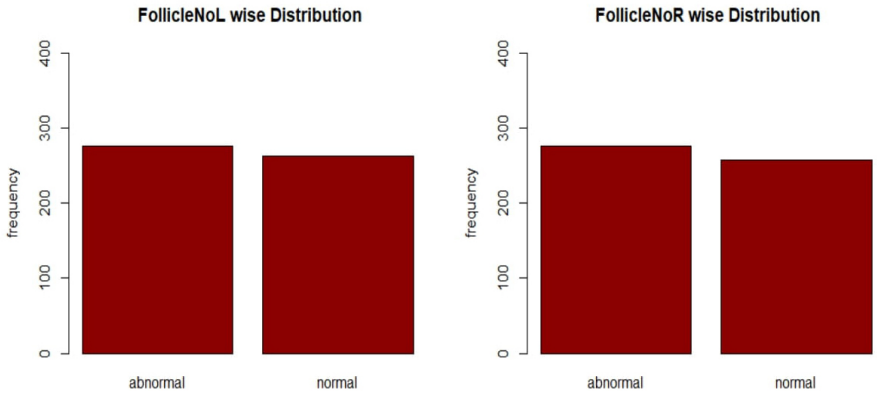



Figure 4.1: Bar diagrams of some of the attributes

The dataset was imported into the workshop of R-program software. The packages necessary for the smooth running of analysis was installed in the R console. Here, Logistic regression and random forest methods are used for the modeling the detection of PCOS. For that first split dataset into train and test with the ratio 80:20. Then construct a model for train data and predictions are made with test data.

## 4.2 VARIABLE SELECTION

In order to improve the performance of the model and reduce the computational cost, only selected attributes of the sample act as variables. This is done by doing a chi-square test for independence for categorical variables and a correlation test for numerical variable

Chi-square test for independence

Chi square test for independence is a statistical hypothesis test used to determine whether two categorical or nominal variables are likely to related or not. The null hypothesis of the independent assumption is to be rejected if the p-value of the following Chi-squared test statistic is less than a given significance level $\alpha$. The formula for the chi-square statistic used in the chi-square test is

$$X^2 = \sum_{n=1} \frac{(0_i - E_i)^2}{E_i}$$

Where $Oi$

is the observed value and $Ei$

is the expected value.

The function used for performing chi-square test is chisq.test().

**Contingency tables**

| | | Cycle | |
|---|---|---|---|
| | | IRREGULAR | IRREGULAR |
| PCOS | NO | 56 | 308 |
| | YES | 95 | 82 |

| | | Weight gain | |
|---|---|---|---|
| | | NO | YES |
| PCOS | NO | 281 | 83 |
| | YES | 56 | 121 |

| | | Hair growth | |
|---|---|---|---|
| | | NO | YES |
| PCOS | NO | 317 | 47 |
| | YES | 76 | 101 |

| | | Skin darkening | |
|---|---|---|---|
| | | NO | YES |
| PCOS | NO | 308 | 56 |
| | YES | 67 | 110 |

| | | Hair Loss | |
|---|---|---|---|
| | | NO | YES |
| PCOS | NO | 221 | 143 |
| | YES | 143 | 102 |

| | | Pimples | |
|---|---|---|---|
| | | NO | YES |
| PCOS | NO | 222 | 142 |
| | YES | 54 | 123 |

| | | Fast food | |
|---|---|---|---|
| | | NO | YES |
| PCOS | NO | 225 | 139 |
| | YES | 38 | 139 |

| | | Reg Exercise | |
|---|---|---|---|
| | | NO | YES |
| PCOS | NO | 281 | 83 |
| | YES | 126 | 51 |

| | | Pregnant | |
|---|---|---|---|
| | | NO | YES |
| PCOS | NO | 222 | 142 |
| | YES | 113 | 64 |

Chi square values and its significance of categorical variables:

| Categorical variables | Chi-square | P-value | Significance |
|---|---|---|---|
| PCOS & Cycle | 84.873 | 2.20e-16 | ** |
| PCOS &Pregnant | 0.29897 | 0.5845 | |
| PCOS &Weight gain | 103.31 | 2.20e-16 | ** |
| PCOS &Hair growth | 114.6 | 2.20e-16 | ** |
| PCOS &Skin darkening | 120.25 | 2.20e-16 | ** |
| PCOS &Hair loss | 15.437 | 8.53e-05 | ** |
| PCOS &Pimples | 43.064 | 5.30e-11 | ** |
| PCOS &Fast food | 74.963 | 2.20e-16 | ** |
| PCOS &Reg Exercise | 1.9982 | 0.1575 | |

Correlation test and its significance of numeric variables:

| Variables | P-value | Significance |
|---|---|---|
| BMI | 2.70e-06 | ** |
| Pulse rate | 0.03274 | |
| RR | 0.3913 | |
| Hb | 4.27e-02 | |
| Cycle length | 2.98e-05 | ** |
| No abortion | 0.1843 | |
| Beta HCG | 0.5215 | |
| II Beta HCG | 0.7671 | |
| FSH: LH Ratio | 0.6705 | |
| Waist: Hip Ratio | 0.7781 | |
| TSH | 0.814 | |
| AMH | 4.49e-10 | ** |
| PRL | 0.905 | |
| VitD3 | 0.04686 | |
| PRG | 0.3088 | |
| RBS | 0.256 | |
| BP Systolic | 0.8538 | |
| BP Diastolic | 0.3773 | |
| Follicle No L | 2.20e-16 | ** |
| Follicle No R | 2.20e-16 | ** |
| Avg F size L | 0.01936 | |
| Avg F size R | 0.02306 | |
| Endometrium | 0.01307 | |

## 4.3 ANALYSIS BY LOGISTIC REGRESSION

Variable selection for PCOS is done by selecting those variables that are proven in the literature as related to the response variable and also with respect to the given data set those variables that are associated with the response variable are taken as independent variables. The variable 'PCOS' is the response variable, 'age', 'cycle', 'BMI', 'AMH', 'weight gain', 'skin darkening', 'hair loss', 'hair growth', 'pimples', 'fast food', 'cycle length', 'Follicle No L ', 'Follicle No R' are the explanatory variables. The response variable is defined as

$$Y = \begin{cases} 0 & \text{If patient is not having pcos} \\ 1 & \text{if patient is having pcos} \end{cases}$$

We use the glm() function to create the logistic regression model.

Fitting the dataset by logistic regression model,

Deviance Residuals:

| Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|
| -2.9056 | -0.3082 | -0.1083 | 0.1480 | 3.4329 |

Table 4.1: five-point summary of residuals

| Parameters | Estimate Std. | Error | z value | Pr ($>\|z\|$) |
|---|---|---|---|---|
| Intercept | -4.10373 | 1.77848 | -2.307 | 0.021031 |
| Age | -0.05514 | 0.03382 | -1.631 | 0.102985 |
| BMI | -0.02236 | 0.05468 | -0.409 | 0.682595 |
| Cycle Regular | -1.20969 | 0.40648 | -2.976 | 0.002920 |
| Cycle length | -0.05005 | 0.11752 | -0.426 | 0.670178 |
| AMH | 0.04091 | 0.03299 | 1.240 | 0.214952 |
| Weight gain | 1.61742 | 0.44672 | 3.621 | 0.000294 |
| Hair growth | 1.63376 | 0.42130 | 3.878 | 0.000105 |
| Skin darkening | 1.19253 | 0.40622 | 2.936 | 0.003328 |
| Hair loss | 0.11383 | 0.40087 | 0.284 | 0.776442 |
| Pimples | 1.14223 | 0.38715 | 2.950 | 0.003174 |
| Fast food | 0.44725 | 0.40976 | 1.091 | 0.275055 |
| Follicle No L | 0.17558 | 0.06738 | 2.606 | 0.009164 |
| Follicle No R | 0.36551 | 0.06801 | 5.374 | 7.69e-08 |

Table 4.2: five-point summary of residuals

Null deviance : 575.46 with 454 degrees of freedom

Residual deviance : 205.93 with 441 degrees of freedom

AIC : 233.93

Model deviance : 205.9303

Chi-square value : 490.96

p-value : 1

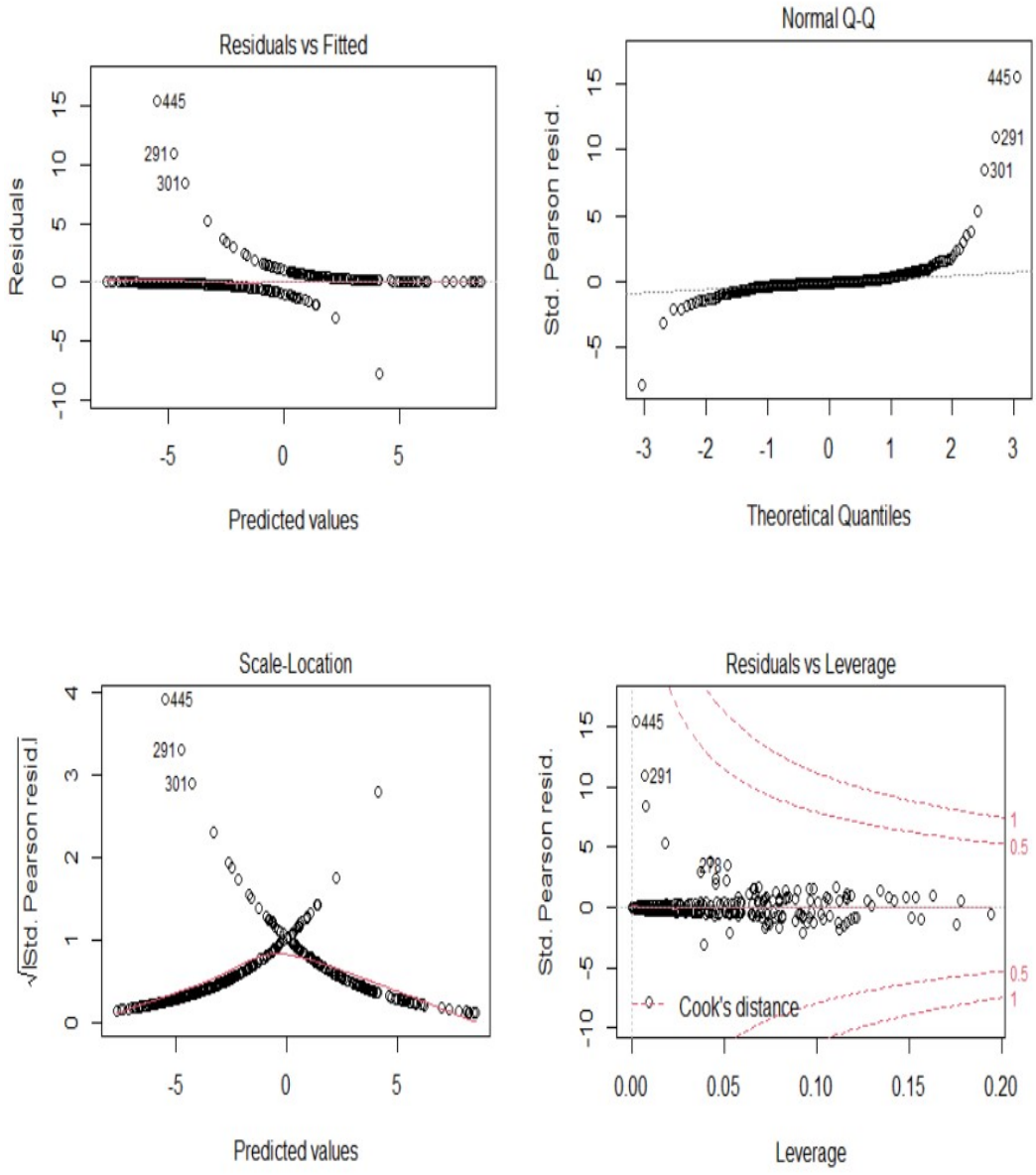



Figure 4.2: Residual plots

Model fit:

Here, the null hypothesis is

H0: The model is a good fit to the data

H1: The model is not a good fit to the data

The test statistic is model deviance and it asymptotically follows chi-square distribution with n-p degrees of freedom where n is sample size, p is the number of parameters in the model.

Model deviance, D=205.9303

Chi-square value=490.96

Since $D < x^2(\alpha)$

, we accept H0 that the model is a good fit and also from the p-value, which is $1, (> 0.05)$ leads us not to reject H0. Thus, the model is good fit to the data.

## 4.4 ANALYSIS BY RANDOM FOREST METHOD

The R package "randomForest" is used to create random forests. This package has the function randomForest () which is used to create and analyze random forests. The model automatically attempts to classify each of the samples in the Out-Of-Bag dataset and display a confusion matrix with the results.
The output obtained using R-Software is given below:
Type of random forest : classification
Number of trees : 500
No. of variables tried at each split : 3
OOB estimate of error rate : 10.88%

| | NO | YES | Class Error |
|---|---|---|---|
| NO | 276 | 16 | 0.05442177 |
| YES | 31 | 107 | 0.22463768 |

Table 4.3: Confusion matrix

So, we can conclude that the model has 10.88% error which means we can predict with 89.12% accuracy. Classification error in NO is 0.0544 and YES is 0.2246. In the confusion matrix shown in Table 4.3, out of 307 non PCOS patients, 276 are correctly classified and out of 123 PCOS patients, 107 are correctly classified. The plot of the model is given below:

Plotting the model will illustrate the error rate is stabilized with an increase in

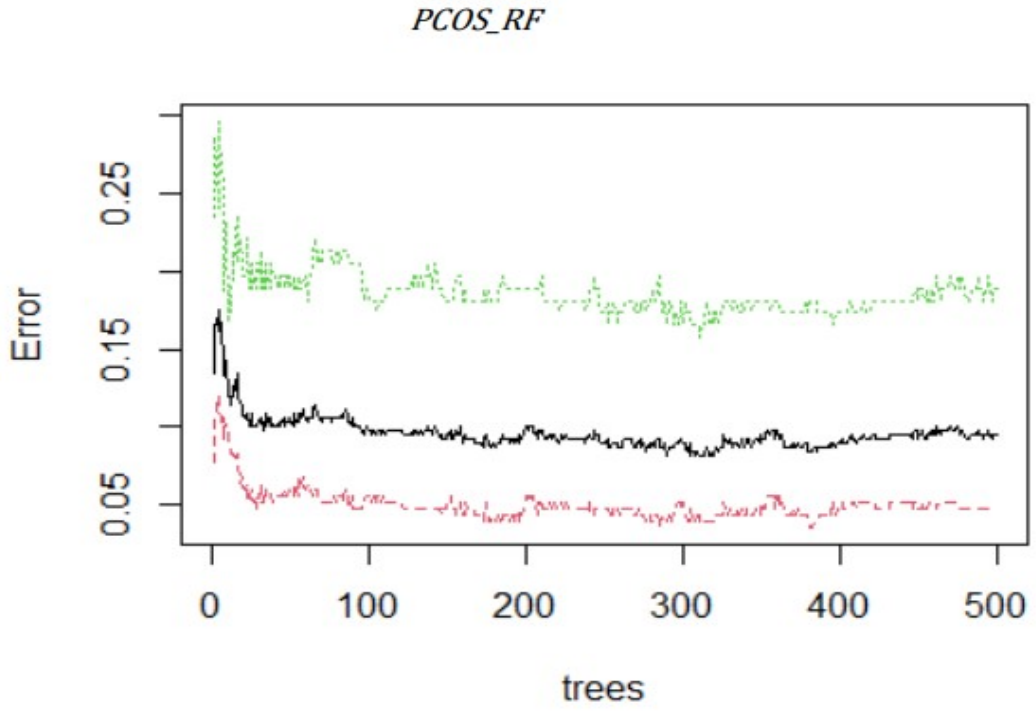



Figure 4.3: Plot of Random Forest

the number of trees.

NO YES Class Error

NO 276 16 0.05442177

YES 31 107 0.22463768

PCOS_RF

## 4.5 MAKING PREDICTION

Now using prepared model, predictions are made with test data using the R command "predict". Comparing the predicated models with the help of confusion matrix. The confusion matrix of logistic regression and random forest model are given below:

| | | Actual data | |
|---|---|---|---|
| | | NO | YES |
| Prediction | NO | 54 | 3 |
| | YES | 4 | 25 |

Table 4.4: Confusion matrix of Logistic regression out of 86 correctly predicted instances is 79 and incorrectly 7.

| | | Actual data | |
|---|---|---|---|
| | | NO | YES |
| Prediction | NO | 67 | 9 |
| | YES | 3 | 34 |

Table 4.5: Confusion matrix of Random forest, out of 113 correctly predicted instances is 101 and incorrectly 12.

Receiving operating characteristic (ROC) & Area under the curve (AUC)

The receiving operating characteristic (ROC) is a visual measure of classifier performance. Using the proportion of positive data points that are correctly considered as positive and the proportion of negative data points that are mistakenly considered as positive, we generate a graphic that shows the tradeoff between the rate at which you can correctly predict something with the rate of incorrectly predicting something. Ultimately, we're concerned about the area under the ROC curve, or AUC. That metric ranges from 0.50 to 1.00, and values above 0.80 indicate that the model does a good job in discriminating between the two categories which comprise our target variable. We can compare the ROC and AUC for models, which show a strong difference in performance.

To plot ROC curve, we use the "pROC" package. ROC plots of logistic regression and random forest are given below.

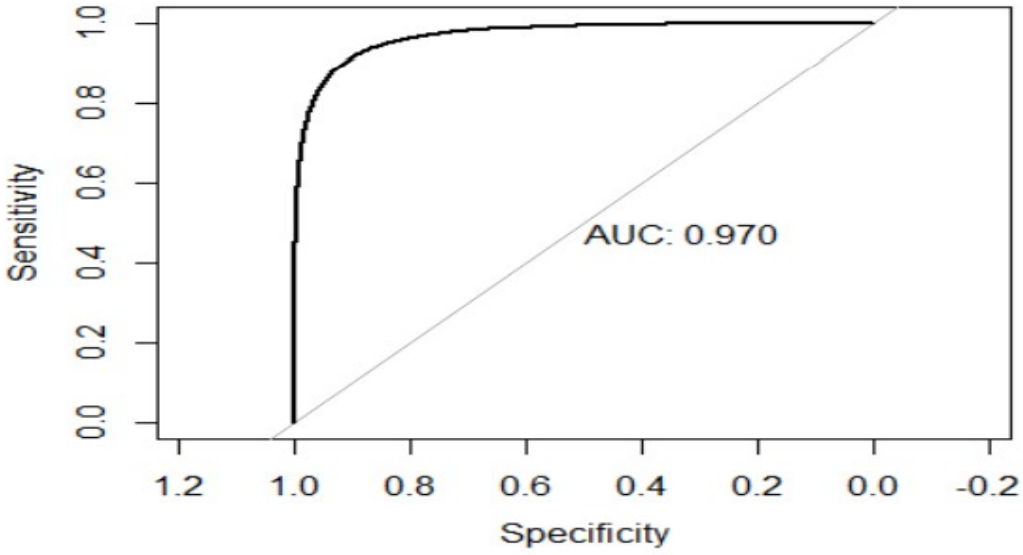



Figure 4.4: ROC Curve-Logistic Regression

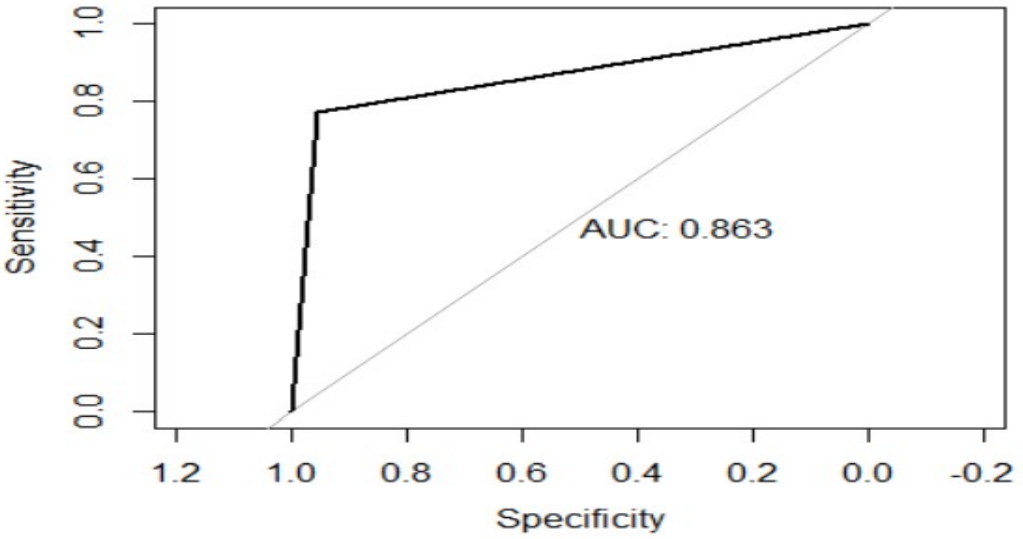



Figure 4.5: ROC Curve-Random Forest

From the confusion matrix tables 4.4 & 4.5, we can calculate Recall, Precision, Accuracy, F1 score & AUC

| S.N. | Algorithm | Accuracy | Recall | Precision | F1 | score AUC |
|---|---|---|---|---|---|---|
| 1 | Logistic regression | 0.92 | 0.89 | 0.86 | 0.88 | 0.97 |
| 2 | Random forest | 0.89 | 0.79 | 0.92 | 0.85 | 0.86 |

Table 4.6: Comparative Analysis of logistic regression and random forest.

Diagrammatic comparison is given below

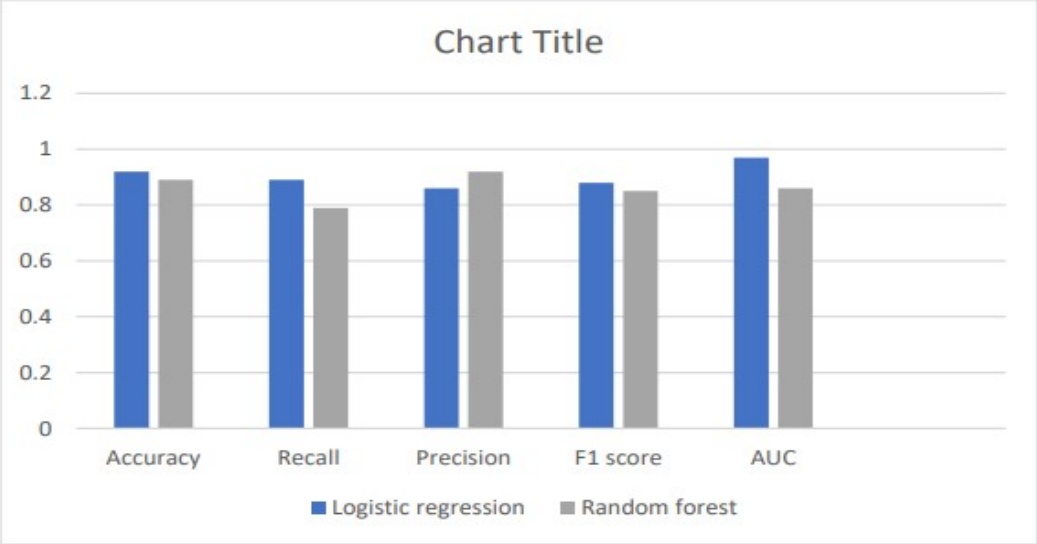



Figure 4.6: Comparative analysis of Logistic regression and Random forest

# Chapter 5

# CONCLUSION

Polycystic Ovary Syndrome or PCOS is an endocrine disorder that occurs in women of reproductive age. The condition once detected cannot be cured but treatment can help relieve its affects. The exact cause of PCOS is still unknown but there are certain factors that portray the risk of getting PCOS. Correct diagnosis is the baseline of any proper treatment and here using the statistical methodology we try to develop a risk diagnosing tool or screening tool to assess the condition of PCOD.

In this data set, there are 541 samples of patients from ten different hospitals across Kerala, India. Using the data, a model is prepared in order to accept causes or symptoms as features and outputs the presence or absence of this condition. Logistic regression and Random Forest techniques were used here in this study. 'age', 'cycle', 'BMI', 'AMH', 'weight gain', 'skin darkening', 'hair loss', 'hair growth', 'pimples', 'fast food', 'cycle length', 'Follicle No L ', 'Follicle No R' were used as explanatory variables which are found significantly associated with the PCOS variable.

The reason behind to build multiple models is to find out the best one for the given dataset. We compare the ROC curve and other comparison measures such as Accuracy, F1 score and AUC values. Higher these values better the model is and these values are higher for Logistic regression than the Random Forest classifier. Hence from our study with respect to the given data Logistic regression is the better model than Random Forest of detecting the absence or Presence of PCOS.

# REFERENCES


[1] Draper N.R and Harry Smith, Applied Regression Analysis, John Wiley & Sons, New York, 2003.

[2] Hassan, Malik & Mirza, Tabasum. (2020). Comparative Analysis of Machine Learning Algorithms in Diagnosis of Polycystic Ovarian Syndrome. International Journal of Computer Applications. Volume 175

[3] Montgomery D.C, Peck E.A and Vining G.G, Introduction to Linear Regression Analysis (Fifth edition), John Wiley & Sons, New York, 2012.

[4] Samprit Chatterjee, Ali S Hadi, Regression Analysis by Example, John Wiley & Sons, New York, 2006.

[5] S. Bharati, P. Podder and M. R. Hossain Mondal, "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms," 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 1486-1489.

[6] Tanwani, Namrata. (2020). Detecting PCOS using Machine Learning. 10.13140/RG.2.2.10265.24169. International Journal of Modern Trends in Engineering and Science.

[7] https://www.kaggle.com/prasoonkottarathil/polycystic-ovarysyndrome-pco