A COMPARATIVE ANALYSIS of MACHINE LEARNING ALGORITHMS to PREDICT POLYCYSTIC OVARIAN SYNDROME (PCOS)

BY

FARIHA JANNAT ANANNA
ID: 171-15-9369

FATEMA TUZ ZOHORA SHEFA
ID: 171-15-8755

AND

SOMA ROY
ID: 171-15-8749

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

MS. SUBHENUR LATIF
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

MD. RIAZUR RAHMAN
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
9 SEPTEMBER, 2021

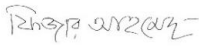
APPROVAL

This Project titled "**A Comparative Analysis of Machine Learning Algorithms to Predict Polycystic Ovarian Syndrome (PCOS)**", submitted by***Fariha Jannat Ananna*** *, ***Fatema Tuz Zohora Shefa****and ***Soma Roy**** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on ***09 SEPTEMBER, 2021****

BOARD OF EXAMINERS

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

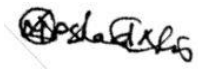


Internal Examiner

Dr. Fizar Ahmed
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Ms. Nazmun Nessa Moon
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



External Examiner

Dr. Md. Arshad Ali
Associate Professor
Department of CSE
Hazee Mohammad Danesh Science & Technology University, Dinajpur.

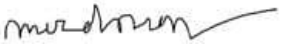
DECLARATION

We hereby declare that; this project has been done by us under the supervision of **Ms. Subhenur Latif, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

SUPERVISED BY:

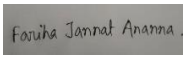
Ms. Subhenur Latif
Assistant Professor
Department of CSE
Daffodil International University

CO-SUPERVISED BY:

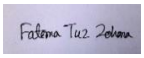


Md. Riazur Rahman
Assistant Professsor
Department of CSE
Daffodil International University

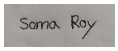
SUBMITTED BY:



Fariha Jannat Ananna
ID: 171-15-9369
Department of CSE
Daffodil International University



Fatema Tuz Zohora Shefa
ID: 171-15-875
Department of CSE
Daffodil International University



Soma Roy
ID: 171-15-8749
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project successfully

We really grateful and wish our profound our indebtedness to **Ms. Subhenur Latif, Assistant Professor,** Department of CSE, Daffodil International University,
Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Machine Learning" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan, Head, Department of CSE**, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

In health services, artificial intelligence is used for early diagnosis functions to manage huge quantities of medical studies with great accuracy and precision. PCOS is a hormonal illness that affects a woman's menstrual cycle when she reaches reproductive age. Women with PCOS have a menstrual cycle that lasts 21 days or longer. Women with PCOS might have less periods (less than eight in a year) or stop having feminine cycles inside and out. This can end in infertility including the appearance of cysts in the ovaries. Irregular menstruation cycles, weight gain, skin darkening, thinning hair on the scalp, diabetes, and high blood pressure are all symptoms of PCOS. It's better to have a diagnosis and treatment as quickly as possible. Assortment of indications and the presence of an assortment of gynecological disorders, PCOS is especially hard to analyze. The time and money spent on the many clinical testing and ovarian scans has become a burden for PCOS sufferers. To resolve this concern, this paper compares machine learning methods for the initial prediction of PCOS using an ideal and basic but promising clinical and metabolic parameter that serves as an early marker for the condition. To collect the data needed for this comparative analysis, a patient questionnaire of 280 women was conducted during doctor consultations and clinical examinations. Based on the significance of the 19 features from medical and physiological test results, 12 prospective points are listed. In the Jupyter Python IDE, PCOS is identified using various machine learning techniques such as logistic regression, K-Nearest Neighbor (KNN), Gaussian Naive Bayes, Random Forest Classifier, and Support Vector Machine (SVM). Random Forest Classifier (RFC) was discovered to be the most appropriate and effective methodology for PCOS prediction, with an accuracy of 100 percent.

Keywords – Polycystic Ovarian Syndrome, Machine Learning, PCOS, PCOS Predict.

TABLE OF CONTENTS

& Implication for Future Research

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

Technology and humanity working together can pave the way for good health and assistance services. Machine learning is artificial intelligence' branch that allows a system to learn and develop on its own without being predictive analytics. It primarily focuses on the development of algorithms that can obtain and use provided datasets for machine learning. Machine Learning systems, like detection, data prediction, and image recognition, have had a significant impact on the healthcare business. One of the most common hormonal disorders in sexually active women is polycystic ovarian syndrome (PCOS). This is a sophisticated endocrine condition that Miscarriage, polycystic ovaries, heart disease, type 2 diabetes, overweight, and other issues can result. Changes in progesterone, estrogen, FSH, and LH levels can occur in the absence of ovulation. PCOS is a chronic disease that affects roughly 12–21% of procreative women, with 70% of them suffering undetected. Clinical, biochemical, and radiological test results are used to make a diagnosis. Due to the lack of insight of its convoluted persuasive, PCOS is diagnosed by excluding insignificant signs or diagnostic testing. It has been observed that symptoms decrease with age and as women approach menopause. PCOS can be controlled to some level by the use of prescribed medications and changes in lifestyle. Contraceptive pills, diabetes, hormone treatments, anti-androgen meds, and monitoring strategies such as an ultrasound are all examples of this. If such methods fail, sophisticated product procedures such as surgical drilling of reproductive organs are used to enhance the ovary's ovulation capacity by lowering testosterone levels. Because of the variety of symptoms associated with this disease, medical practitioners are required to conduct a great number of medical reports and x - ray imaging techniques that are excessive. It is critical to identify and treat PCOS as early as possible, using as few tests and imaging procedures as possible because the situation caused ovary disorder, which increases the chance of infertility, pregnancy complications, or even obstetric tumors, as well as mental distress for patients as a result of money and resources wasted. Although research was conducted to diagnose PCOS using various machine

learning algorithms, there is still necessity improvement in terms of accuracy and precision based on medical data.

1.2 Motivation

Concerns about women's healthcare were outlined as a concentrate of our interest among the numerous complexities around us due to their importance in today's society. Polycystic Ovarian Syndrome, AKA PCOS, is not known to all levels of women. They are not aware of their reproductive organs as well as problems related to them. It causes infertility, uterus tumors, and ends up with cancer. Lots of medical tests and time can discourage a PCOS affected woman from curing herself properly. Besides, doctors can easily predict by analyzing one’s symptoms. This machine-learning algorithm could be a blessing for both doctors and patients. Many doctors are not familiar with these possibilities, so using them in their diagnosis process can be a milestone in their successful careers.

1.3 Objective

- To detect PCOS based on clinical symptoms related with the use of the popular machine learning algorithms on random data sets.
- To analyze the quality of various algorithms and select the best suitable algorithm from among them.
- To help doctors as well as PCOS patients so that they can detect PCOS in the very early stage and start medication or treatment as quick as possible.
- Better women's health can ensure better children for future generations, and women can be protected from life-threatening gynie diseases.

1.4 Rationale of the Study

There is less critical work has done beforehand with PCOS prediction in Bangladeshi point of view. That is the reason we are intrigued to work with it and machine learning

procedures. Artificial intelligence may be applied in healthcare systems for diagnostic purposes to manage huge quantities of clinical data with great accuracy and precision. Machine Learning is a part of artificial intelligence that includes a variety of factual, probabilistic, and optimization techniques to empower Machines to "learn" from previous versions and detect difficult-to-perceive patterns in complex, congested, or sophisticated data collections. Machine learning models are shown in a lot of areas, including classification and clustering. This is used for predicting malignant expansion, performing a basic audit of programming defect forecasting, discovering dermatological illnesses, and so on. These techniques are most often used to lead various kinds of systematic risk prediction. We assumed that since machine learning has a wide range of applications, we should use it for our prediction task. The correct diagnosis is the core of any effective treatment, and in this analysis, we used machine learning methods such as Logistic regression, KNearest neighbor (KNN), Gaussian Naive Bayes, Random Forest Classifier, Support Vector Machine (SVM) to detect PCOS based on patient clinical data.

1.5 Expected Outcome

Our research will provide a large informative database for detecting PCOS. People will get to know about the symptoms of PCOS patients of Bangladesh. This will be very helpful for women of this country. We will provide results of the feasibility by using existing techniques of algorithms of machine Learning. We will try to introduce new techniques or variation of existing techniques. We expect that, we will publish one or more articles in international conferences or journals.

1.6 Research Questions

- What measure of information do we gather and where do we gather them?
- What will our original data resemble?
- Will our information and machine learning be viable?
- Do we need to train our unique data to machine learning model?

- Would it be a good idea for us to utilize well-known ML procedures or utilize another one?
- Could some other procedure give a better result over ML?

1.7 Report Layout

This research paper contains the following Sixes contents as given below:

- **Chapter 1** explains the introduction of the research which highline the motivation, objective, rationale of the study, research questions, and expected outcome.
- **Chapter 2** discusses its background information with related works, research summary, the scope of the problem, and challenges respectively.
- **Chapter 3** contains with the detailed workflow of this research, data collection procedure, and statistical analysis and feature implementation.
- **Chapter 4** covers experimental result and some relevant discussions, the analytical outcome of research via numerically and graphically.
- **Chapter 5** covers this research impact on society, environment, sustainability and ethical aspects.
- **Chapter 6** contains a summarization of this research work along with the limitations, conclusions and future work.

CHAPTER 2

Background

2.1 Preliminaries

In this chapter, we will cover relevant research, research summaries, the scope of the problem, and challenges. In the related works section, we summarize various research articles, related works, fundamental methods, and correctness's that are connected with our work. In the research summary section, we will present a list of relevant related studies. The extent of the challenging area explains how we may contribute to the problem that we are attempting to solve. Finally, the Difficulties section includes a few comments on the difficulties we had while working.

2.2 Related Works

Amsy Denny [1], by Using Machine Learning algorithm to build a system to detect and predict PCOS. They have used [1] Classification and Regression Trees (CART), Random Forest Classifier, Naïve Bayes classifier, Support Vector Machine (SVM), K-Nearest neighbor (KNN), logistic regression. [1] To prepare the model they utilized 541 women who were diagnosed during expert consultations. At last, they showed up with great exactness of 89% when testing with the dataset.

Malik Mubasher Hassan, Tabasum Mirza [2], they identified PCOS by machine learning algorithms. They have used Support Vector Machine, CART, Naive Bayes Classification, Random Forest and Logistic Regression. For this research, an information base was made from 10 different hospitals across Kerala, India and is available on Kaggle site. When tested with the dataset, they exhibited improved accuracy of 92 percent.

Namrata Tanwani [3], by using Machine Learning Techniques they identified PCOS with some factors such as obesity, insulin resistance, blood pressure, depression, inflammation. They have used Logistic Regression and K-Nearest Neighbor for

identification. There were 39 parameters in total, for just 9 parameters, with the greatest weights considered for KNN and 10 parameters considered for Logistic Regression. A comparison was made between the two different classifiers, The F1 score helped to determine the best model between the two. The F1 score for KNN is 0.90 and for that of Logistic Regression is 0.92, hence, the model of Logistic Regression is selected to determine the absence or presence of PCOS.

Palvi Soni, Sheveta Vashisht [4], they identified PCOS by data mining techniques. They have used Support Vector Machine, Naïve Bayes Classifier, Decision Tree, which will be accurate in assessing PCOS, using these methodologies in the future will be very fruitful in anticipating this disease in an organized manner.

Xing-Zhong Zhang [5], They're working on creating a new machine-learning model to find new PCOS genotype. 233 PCOS candidates were used for this research. They aligned the computational properties of two genetic variants: known PCOS genotype and the remaining genotypes in the genome. They tested different classifiers, K-nearest neighbor (KNN), decision tree and SVM with different kernel functions. SVM with linear kernel achieved the best performance.

Palak Mehrotra [6], They portrayed a technique for detecting PCOS based on medical and physiological parameters. Their algorithm entails the creation of a [6] feature vector based on medical and physiological features, and statically meaningful features for distinguishing between normal and PCOS groups are chosen using a two-sample test [6]. Bayesian and Logistic Regression (LR) algorithms are used to describe the selected feature. The highest accuracy of the Bayesian classifier is 93.93 percent, comparison to 91.04 percent for logistic regression.

Vaidehi Thakre [7], They developed a system for the initial classification and prevention of PCOS treatment based on a maximal and a marginal set of parameters. To detect whether a woman is suffering from PCOS, 5 different machine learning classifiers like Random

Forest, SVM, Logistic Regression, Gaussian Naïve Bayes, K Neighbors have been used. It has been observed that the accuracy of Random Forest Classifier is the highest and the most reliable.

2.3 Comparative Analysis and Summary

There has already been some research involved on prediction and detection using the machine learning algorithm and data mining process. Nowadays, the use of machine learning technology has increased with the use of various disease detection. The comparison between these related works has shown in this part. Here, the comparison of different research works with their subject, methodology, and the outcome are given below in Table 2.3.1

Table 2.3.1: Comparative Analysis

Reference Number	Author	Methodology	Description	Outcome
[1]	Amsy Denny, Maneesh Ram C, Anita Raj, Ashi Ashok, Remya George [1]	CART, Random Forest Classifier, Naïve Bayes classifier method, Support Vector Machine (SVM), [1] K-Nearest neighbor (KNN), logistic regression [1]	Machine Learning based system to detect and predict PCOS	89% accuracy in Random Forest
[2]	Malik Mubasher Hassan, Tabasum Mirza [2]	Support Vector Machine, CART, Naive Bayes Classification, Random Forest and Logistic Regression [2]	Machine Learning based system to identify PCOS	96% accuracy in Random Forest
[3]	Namrata Tanwani [3]	Logistic Regression and K-Nearest Neighbor [3]	Identified PCOS with ML	92% accuracy in Logistic Regression
[4]	Palvi Soni, Sheveta Vashisht [4]	Support Vector Machine, Naïve Bayes Classifier, Decision Tree [4]	Discussion on data mining techniques to identify PCOS	Various data mining tasks and techniques
[5]	Xing-Zhong Zhang, Yan-Li Pang, Xian	[5] K-nearest neighbor (KNN), decision tree, SVML	a new machine-learning model to	80% accuracy in SVM linear

	Wang & Yan-Hui Li [5]		find new PCOS genotype	
[6]	Palak Mehrotra, Jyotirmoy Chatter-jee, Chandan Chakra-borty, Biswanath Ghoshdastidar, Sudarshan Ghoshdastidar [6]	Bayesian and Logistic Regression (LR) [6]	a technique for detecting PCOS based on medical and physiological parameters	93% accuracy in Bayesian classifier
[7]	Vaidehi Sunil Thakre, Shreyas Vedpathak [7]	Random Forest, SVM, Logistic Regression, Gaussian Naïve Bayes, K Neighbors [7]	a system for the initial classification and prevention of PCOS treatment based on a maximal and marginal set of parameters	90% accuracy in Random Forest

A fusion of machine learning, artificial intelligence, and deep learning is currently being investigated with new technologies that can be used in any type of prediction and detection model. Recently, various machine learning algorithms have been used to diagnose and detect material. KNN, SVM, logistic regression and many algorithms are popular for any detection model. From previous research, we can see that the SVM, random forest, naïve Bayes, and logistic regression algorithm's popularity and effectiveness for prediction or detection models are high. In our research, we have tried to implement KNN, SVM, logistic regression, naive Bayes, random forest to predict the risk of PCOS in Bangladesh's perspective and we have 100% accuracy in Random Forest.

2.4 Scope of the Problem

Past works in different datasets have distinguished by utilizing machine learning procedure. By dissecting and taking assistance from past explores and works, we chose to chip away at identification of PCOS patients by utilizing machine learning algorithms.

2.5 Challenges

We had a tough time collecting data. We experienced a few challenges while working on our project. Surprisingly, when it comes to polycystic ovarian syndrome, the great majority of people in Bangladesh are unfamiliar of it. What's more, the most testing thing is, during this Coronavirus pandemic period, we needed to gather information from various hospitals & clinics with a tension in our psyche of getting affected. We were additionally curious about anaconda, Jupyter notebook, and some new machine learning techniques. It took us some time to know and find out about it from the start, yet with the assistance of our supervisor and accomplishing more practice we snatch them without any problem. At that point we keep on taking care of our responsibility.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

Our research topic is to set up a comparative model for the prediction of polycystic ovarian syndrome which we are working with. The recognition Model is created dependent on the parameters by which PCOS can be detected and some other related data. The research field's methods are involved with speculative facts given to speak to obviously. Because one of the most important components is data, which contributes in the development of an innovative model that may be beneficial for medicinal researchers or experts attempting to predict the correct diagnosis of this condition. This method has been used in the circulation analysis to enhance the diagnosis of PCOS by using machine learning combining methods to look at the patient's symptoms. As a result, the sections in the accompanying segment summarize the examination techniques and methods. Questionnaire is the way we chose for collecting data and some internet resources helped us setting the questions. Lastly, updated and well configured hardware and software have been used. We will apply our gathered information to different calculations to see which calculations will perform well for our model. We utilize different machine learning calculations like [1]Logistic regression, K-Nearest neighbor (KNN), Decision Tree, Gaussian Naive Bayes, Random Forest Classifier, Linear Support Vector Machine (SVML), and Radial Support Vector Machine (SVMR) [1]. We utilized Python as a programming language and Anaconda, Jupyter notebook as an information mining apparatus and local folders as our dataset in our exploration work.

3.2 Data Collection Procedure

There are several ways to gather data, such as real-time data collection or data collection via repository sites such as [1] Kaggle and UCI machine learning repository, which is one of the most commonly used. For the data collection, in this pandemic situation, we had to go outside from one hospital to another for taking the accurate data of PCOS patients. We have taken data of the symptoms of PCOS patients by questionnaire form.

We firstly, identify the patients, then asked them about their symptoms. Then later on, we transferred it to a google form and converted it to a csv file.

3.3 Statistical Analysis

Gathered data from 280 people, who are of different physical conditions, different ages. Figure 3.3.1 shows that in our dataset how many PCOS suffered and non-PCOS people were. We had prepared our model based on data from 160 PCOS suffered people and 120 non-PCOS people.

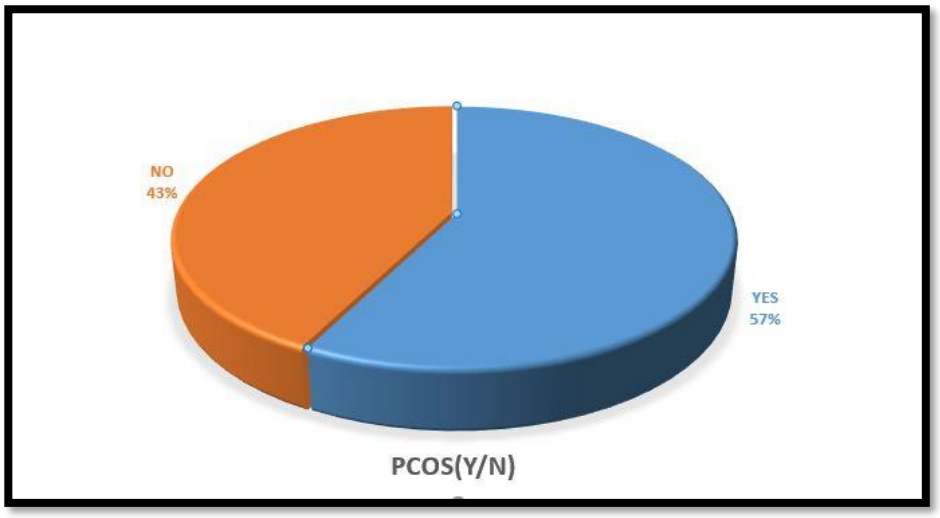


Figure 3.3.1: Pie Chart of PCOS.

Figure 3.3.2 shows that information from people of some ages. This picture shows we have information about how many people of any age. Most of the data we collected were about young people.

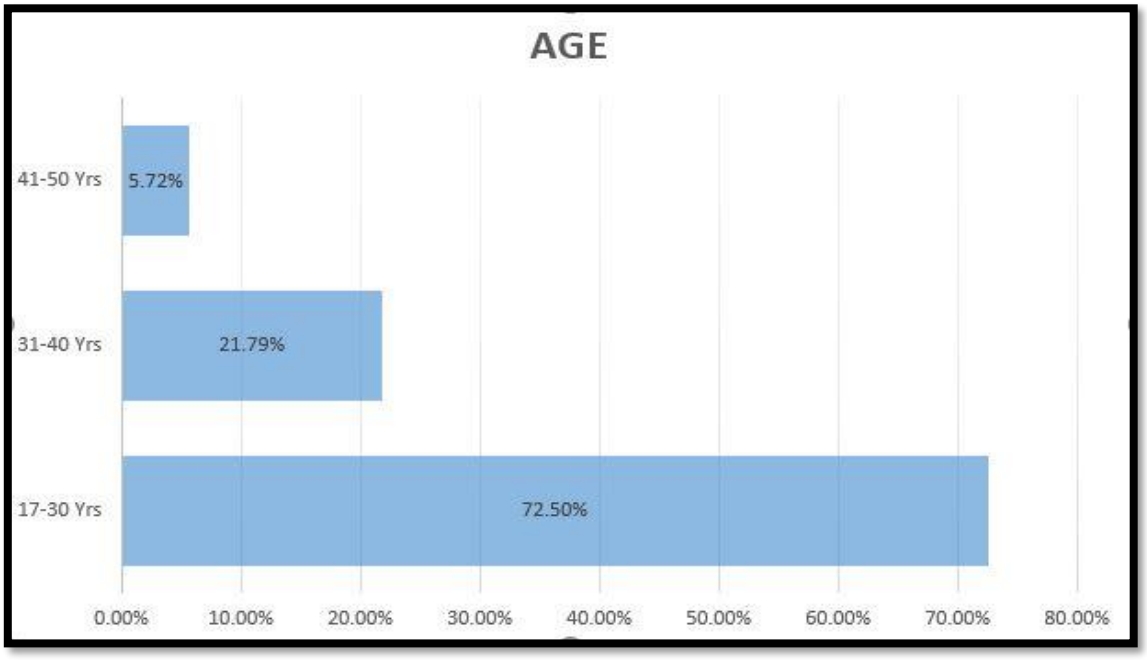


Figure 3.3.2: Bar Chart of Age.

Figure 3.3.3 shows the bar chart of the BMI range of the patients. We can assume that most of the patients are overweight than their ideal weight. That can be the prominent reasons of suffering from PCOS.

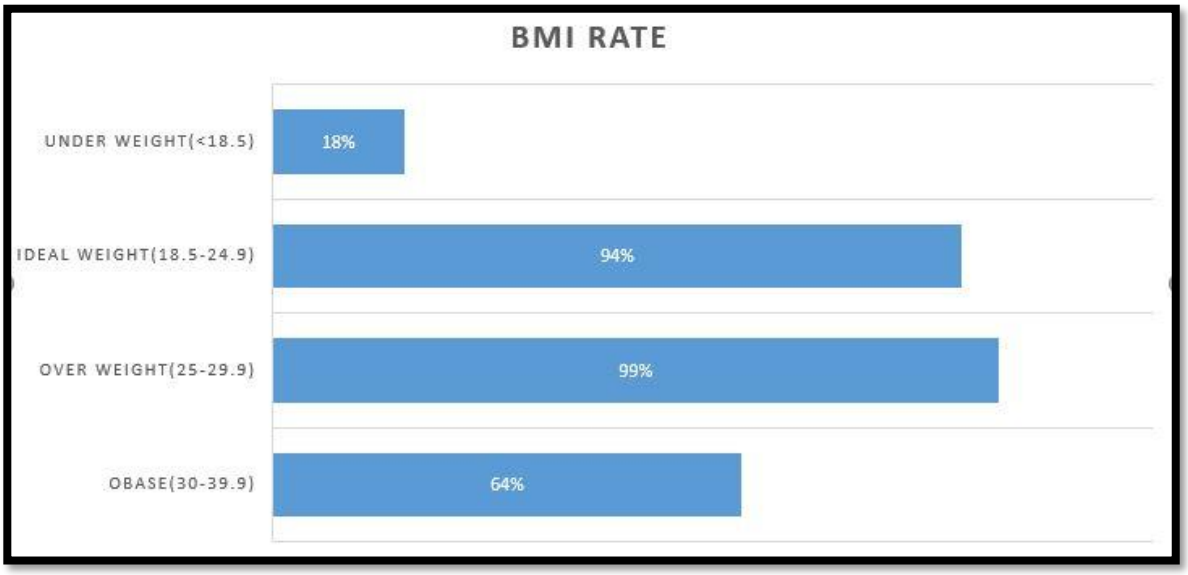


Figure 3.3.3: Bar Chart of BMI Rate.

Figure 3.3.4 shows that Most of the patients' period cycle is not regular and that is one of reasons of suffering from PCOS.

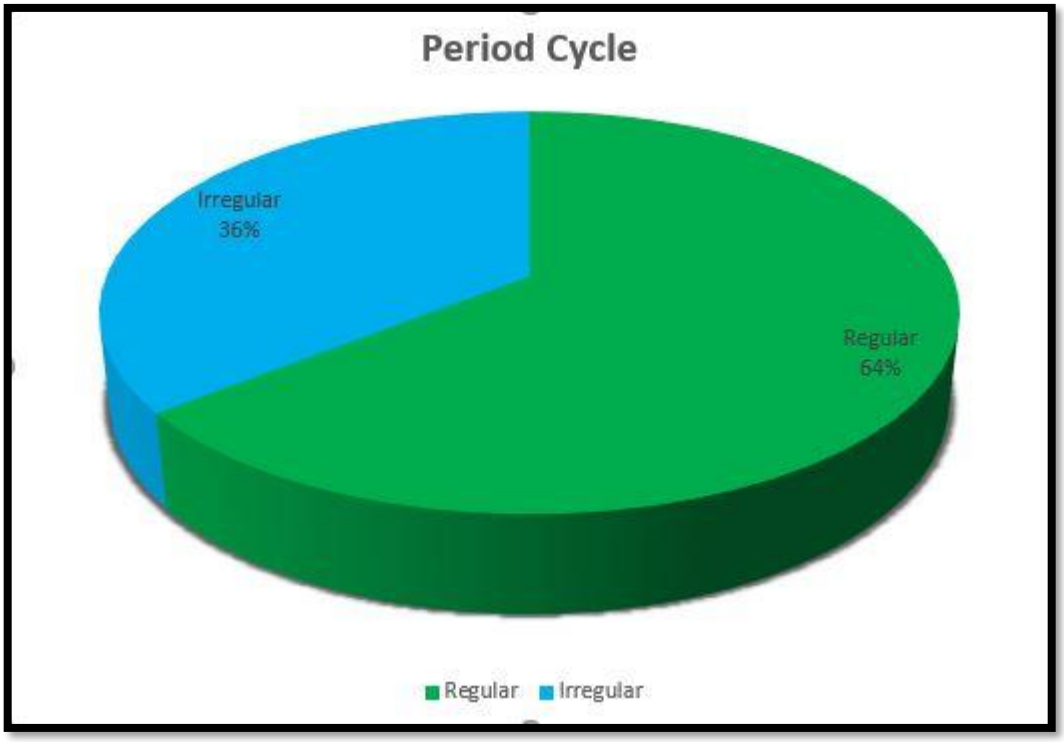


Figure 3.3.4: Pie Chart of PCOS Cycle.

Figure 3.3.5 shows the data of unwanted hair growth in the patients whose information is in dataset. It can be seen that most of the patients are facing the problem of this and it can be marked as a symptom of PCOS.

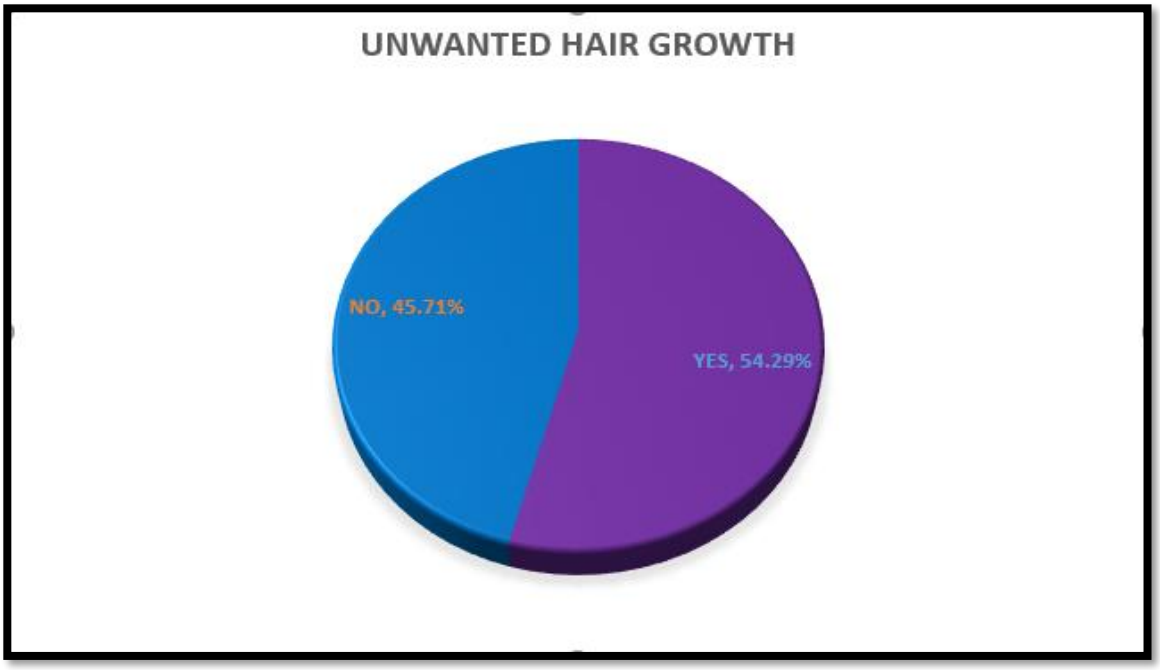


Figure 3.3.5: Pie Chart of Unwanted Hair Growth.

Figure 3.3.6 showing the statistical form of skin darkening of the patients from dataset. Majority of the people has noticed this problem on their skin which is one of the remarkable symptoms of PCOS.

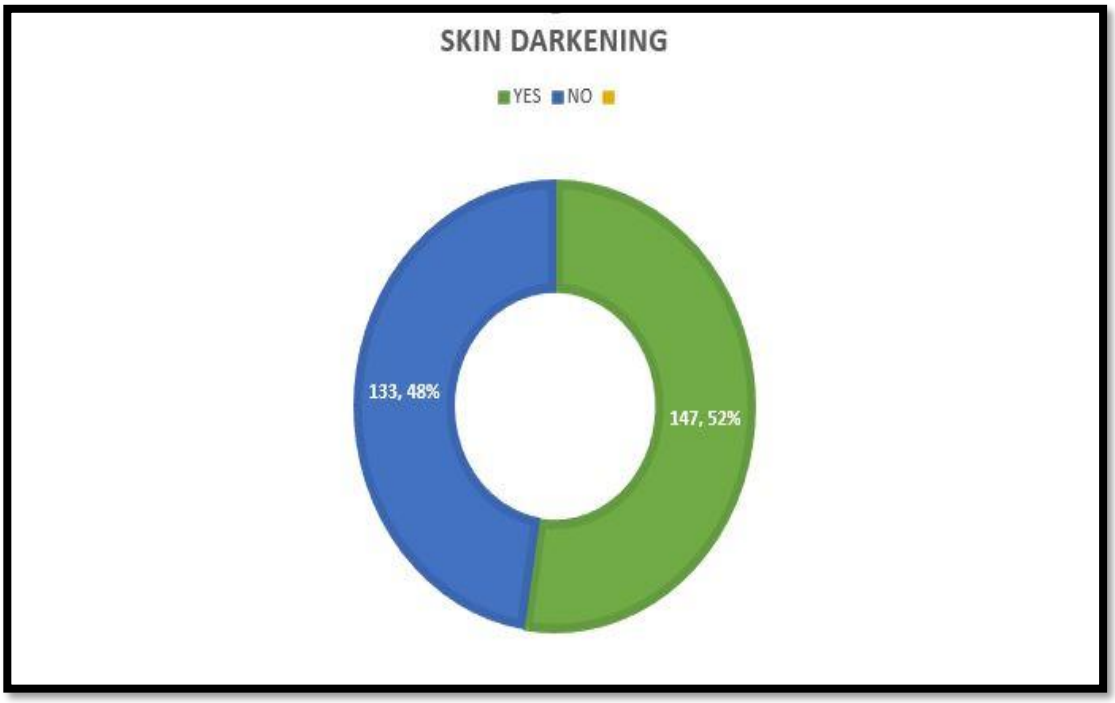


Figure 3.3.6: Pie Chart of Skin Darkening.

3.4 Proposed methodology

A dataset will be required for our research purpose. So, we collected the dataset from hospitals then checked the missing values. Then applied the imputation technique and made the data ready for pre-processing. Make it ready for the six algorithms (RF, SVML, SVMR, LR, GNB, KNN). After applying all the algorithms, we found the most accuracy by RF (Random Forest which is 100%. Here is the architecture of our proposed methodology:

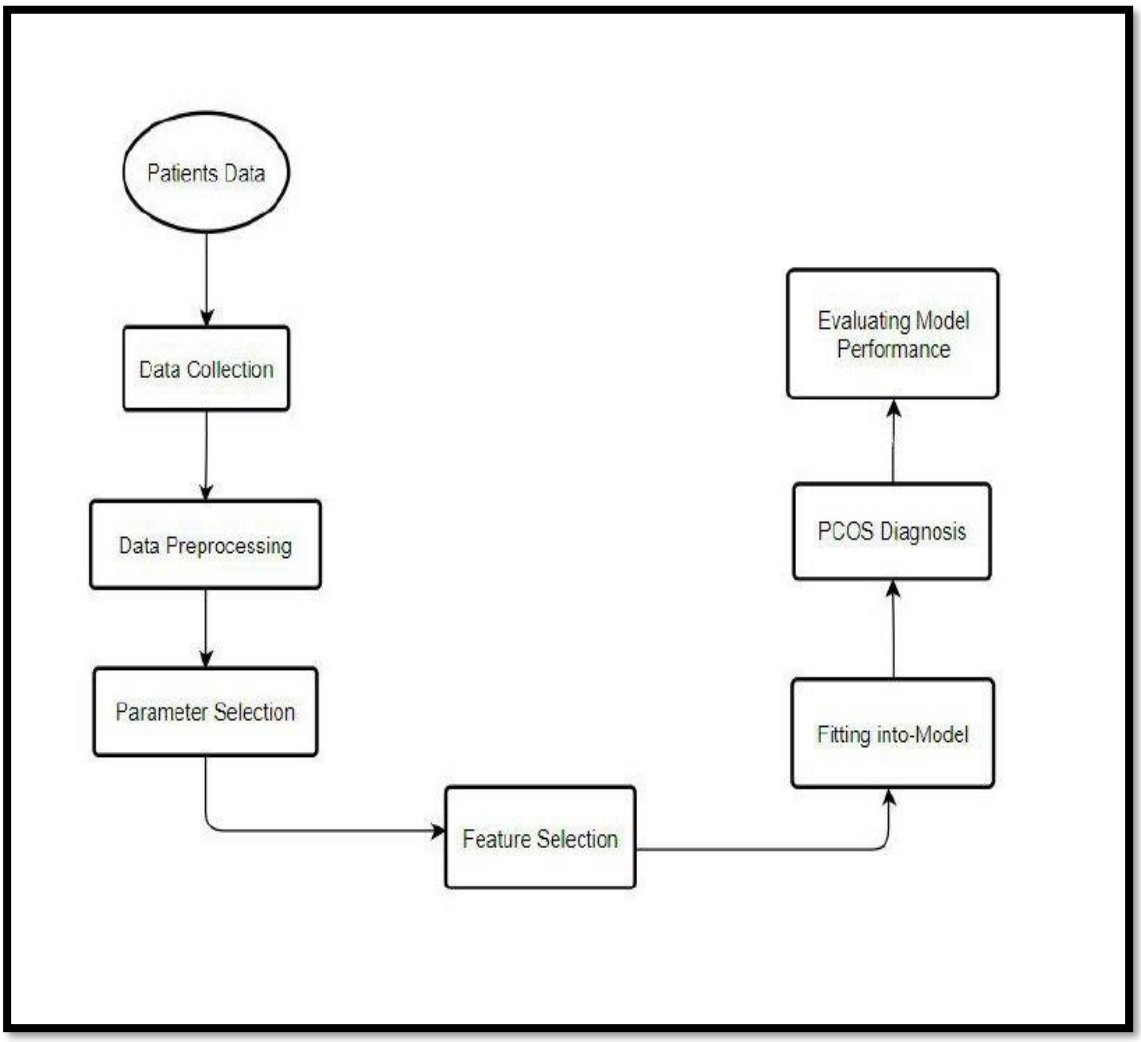


Figure 3.4: Proposed Methodology.

3.4.1 Data Preprocessing

Data preprocessing is the ability to manipulate information into a systematic order of information collection. We create some miss information, straight out information, mathematical information, and text information from the gathered information. At that moment, we determine that, through our information processing, we will make this information usable for analysis. These complex essential requirements dealing with invalid information, numerical values, function expanding, and feature extraction. The dataset's invalid values are indicated with 'NaN'. Overfitting can occur when a model is unable to read a variable. Overfitting can be avoided by lowering the set of possible combinations in the sample. This is done by reducing dimensionality and normalizing the data.

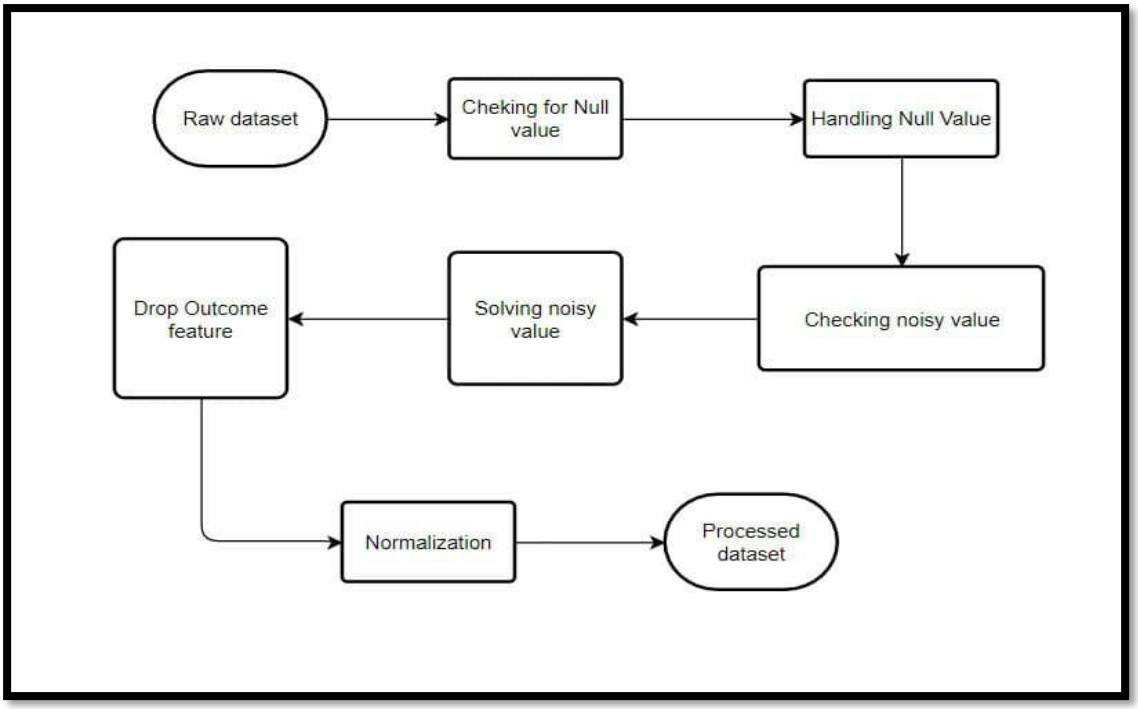


Figure 3.4.1: Data Preprocessing.

3.4.2 Parameter Selection

The data collected from the questionnaires was fed into the correlational method. The parameters were finalized with the help of trained opinions and after taking into account recent studies that had an impact on PCOS in some way. The features are incorporated

to remove any interrelationships between them that may have a harmful influence on the training set. Table:3.4.2 shows the completed set of parameters before modification. Biological, biochemical, and physicochemical properties are among these parameters.

Table 3.4.2: Parameter Selection

SL No	Parameters	Value
1	Age [1]	15-40
2	BMI [1]	< 18.5-Underweight 18.5 -24.9 - Ideal weight 25-29.9 -Overweight 30 - 39.9 - Obese
3	Cycle Duration	valid numeric value
4	Cycle Regularity	Regular/Irregular
5	Height(cm)	valid numeric value
6	Weight	valid numeric value
7	Unwanted Hair Growth	Yes(Y)-1 No(N)-0
8	Skin Darkening	Yes(Y)-1 No(N)-0 [1]
9	Pimples [1]	Yes(Y)-1 No(N)-0 [1]
10	High BP	Yes(Y)-1 No(N)-0
11	Diabetes	Yes(Y)-1 No(N)-0
12	Intake Fast Food [1]	Yes(Y)-1 No(N)-0 [1]
13	Regular Exercise [1]	Yes(Y)-1 No(N)-0 [1]
14	Loss of Hair [1]	Yes(Y)-1 No(N)-0 [1]
15	Marriage Status	Yes(Y)-1 No(N)-0
16	Pregnant	Yes(Y)-1 No(N)-0

17	Anxieties	Yes(Y)-1 No(N)-0
18	Weight Gaining	Yes(Y)-1 No(N)-0
19	Under Treatment	Yes(Y)-1 No(N)-0

.

3.4.3 Feature Selection

Only selected attributes of the samples are used as features to improve the model's performance and reduce the computational cost. As we have seen above, we have 19 features. We could use all of them but it could happen that all of them are not useful or there can be a chance of overfitting. Hence, we will use the Chi Square method to determine important features. Chi square method will calculate a score. The score calculated tells us how important that feature is. Top 12 most important features will used. Select KBest and chi-squared will used to find the feature importance.

Table: 3.4.3 displays the features and their correlation to the target, arranged in descending order. The more is the weight of the feature the more is its influence on the target, independently.

Table 3.4.3: Feature Selection.

Feature	**Score**
Weight (kgs)	314.327905
BMI	188.179570
Age (yrs.)	51.003770
Weight gain(Y/N)	22.075743
Marriage Status (Y/N)	16.612233
Pregnant(Y/N)	13.272180
High BP(Y/N)	13.007881
Unwanted Hair Growth(Y/N)	8.339211
Skin Darkening (Y/N)	7.415745

Height(cm)	6.608136
Cycle(R/I)	6.393862
Anxieties(Y/N)	6.15033

3.4.4 Fitting into Model

With the data cleaned and selected, it is now ready to be processed by the models. The six models will be used for supervised machine learning are 'Linear SVM', 'Radial SVM', 'Logistic Regression', 'Random Forest Classifier', 'K Neighbors Classifier', 'Gaussian Naive Bayes'.

3.4.5 PCOS Diagnosis

We built a predictive system where we inserted the values of those features selected by Chi square method to predict the disorder. We implemented all the six algorithms and found out that all of them are predicting well there is PCOS existing or not.

3.4.6 Evaluating Model Performance

We evaluate the performance of the models calculating their accuracy. Calculated the precision, recall, f1-score, roc-curve, and confusion matrix of each algorithm, in addition to their accuracy. Evaluation of that model is required for any model selection.

3.5 Implementation Requirements

For operational requirements, we require data mining tools, data handling equipment’s, and data storage devices. We collect data using questionnaire by google form. We created informative indexes on the PC's local drives. We used Anaconda navigator and Jupyter notebook for information preprocessing, data augmentation, and computation.

We have applied some algorithms like Random Forest algorithm, SVM, Logistic Regression, Gaussian Naive Bayes, K- nearest neighbor.

3.5.1 Random Forest (RF)

Familiar example of Clustering techniques is the Random Forest Algorithm. It integrates the outcomes of many decision trees to arrive at a conclusion. Used to solve issues based on both regression and classification.

3.5.2 Support Vector Machine (SVM)

Support Vector Machine algorithms are supervised machine learning algorithms that are used to solve issues such as regression, classification, and outlier detection. In SVM, the data is essentially shown as points in an n-dimensional space, where n is the number of features. The method seeks a hyperplane that may divide the plotted points into the needed or specified number of classes.

3.5.3 Logistic Regression (LR)

A classification algorithm is Logistic Regression. It's a machine learning algorithm that's been supervised. It performs its hypothesis using the sigmoid function. The calculated probability is the result of the hypothesis. It is expressed in binary terms, i.e., will it happen or will it not happen, with 1 or 0 being the answer.

3.5.4 Gaussian Naive Bayes (GNB)

Naive Bayes algorithms are those that employ the Bayes' theorem to compute the probability and determine which class the provided data belongs to. For our hypothesis,

we employed Gaussian Naive Bayes. It uses the Gaussian Normal Distribution, thus there will be no correlation between the features, and it handles continuous data.

3.5.5 KNeighbours Classifier

The KNN algorithm predicts that comparable data, if plotted, would exist nearby. We first load the data and specify how many classes we want the algorithm to classify it into. The method first calculates the distance between K neighbors using the distance formula, and then it selects the K closest neighbors based on the distance.

CHAPTER 4

Experimental Result and Discussion

4.1 Experimental Setup

In the past phase, we have talked regarding the dataset and dataset handling measures. The ready info is employed in sure calculations and therefore the consequences of the calculation are talked regarding during this chapter. 'Linear SVM', 'Radial SVM', 'Logistic Regression', 'Random Forest Classifier', 'K Neighbors Classifier', 'Gaussian Naive Bayes'.

4.2 Experimental Results & Analysis

We utilized six machine learning algorithms and contrasted them and every calculation figuring their accuracy, confusion matrix and ROC curve.

4.2.1 Prediction on PCOS Analysis

The purpose of our work is to predict PCOS mainly a patient has the disorder or not. So, we build a system where we will input the value of the features and will notice what the result has come out. All of the algorithms predicted very well and they are given below:

Logistic Regression:

```
input_data = (63,26.2,24,1,0,0,1,1,1,154.94,1,1)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have PCOS')
else:
  print('The Person has PCOS')
```

```
[1]
The Person has PCOS
```

Figure 4.2.1.1.1: PCOS Prediction using LR.

```
input_data = (59,23.8,23,0,0,0,0,1,0,157.48,1,0)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have PCOS')
else:
  print('The Person has PCOS')
```

```
[0]
The Person does not have PCOS
```

Figure 4.2.1.1.2: Not-PCOS Prediction using LR.

Random Forest:

```
input_data = (63,26.2,24,1,0,0,1,1,1,154.94,1,1)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have PCOS')
else:
  print('The Person has PCOS')
```

```
[1]
The Person has PCOS
```

Figure 4.2.1.2.1: PCOS Prediction using RF.

```
input_data = (59,23.8,23,0,0,0,0,1,0,157.48,1,0)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have PCOS')
else:
  print('The Person has PCOS')
```

```
[0]
The Person does not have PCOS
```

Figure 4.2.1.2.2: Not-PCOS Prediction using RF.

KNN:

```
input_data = (63,26.2,24,1,0,0,1,1,1,154.94,1,1)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have PCOS')
else:
  print('The Person has PCOS')
```

```
[1]
The Person has PCOS
```

Figure 4.2.1.3.1: PCOS Prediction using KNN.

```
input_data = (59,23.8,23,0,0,0,0,1,0,157.48,1,0)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have PCOS')
else:
  print('The Person has PCOS')
```

```
[0]
The Person does not have PCOS
```

Figure 4.2.1.3.2: Not-PCOS Prediction using KNN.

Gaussian Naive Bayes:

```
input_data = (63,26.2,24,1,0,0,1,1,1,154.94,1,1)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have PCOS')
else:
  print('The Person has PCOS')
```

```
[1]
The Person has PCOS
```

Figure 4.2.1.4.1: PCOS Prediction using GNB.

```
input_data = (59,23.8,23,0,0,0,0,1,0,157.48,1,0)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have PCOS')
else:
  print('The Person has PCOS')
```

```
[0]
The Person does not have PCOS
```

Figure 4.2.1.4.2: Not-PCOS Prediction Analysis using GNB

Linear SVM:

```
input_data = (63,26.2,24,1,0,0,1,1,1,154.94,1,1)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have PCOS')
else:
  print('The Person has PCOS')
```

```
[1]
The Person has PCOS
```

Figure 4.2.1.5.1: PCOS Prediction using SVML.

```
input_data = (59,23.8,23,0,0,0,0,1,0,157.48,1,0)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have PCOS')
else:
  print('The Person has PCOS')
```

```
[0]
The Person does not have PCOS
```

Figure 4.2.1.5.2: Not-PCOS Prediction using SVM

Radial SVM:

```
input_data = (63,26.2,24,1,0,0,1,1,1,154.94,1,1)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have PCOS')
else:
  print('The Person has PCOS')
```

```
[1]
The Person has PCOS
```

Figure 4.2.1.6.1: PCOS Prediction using SVMR.

```
input_data = (59,23.8,23,0,0,0,0,1,0,157.48,1,0)

# change the input data to a numpy array
input_data_as_numpy_array= np.asarray(input_data)

# reshape the numpy array as we are predicting for only on instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = model.predict(input_data_reshaped)
print(prediction)

if (prediction[0]== 0):
  print('The Person does not have PCOS')
else:
  print('The Person has PCOS')
```

```
[0]
The Person does not have PCOS
```

Figure 4.2.1.6.2: Not-PCOS Prediction using SVMR

4.2.2 Experimental Evaluation by their Accuracy

We applied six algorithms on our processed datasets where the number of the features were 19. Then we used the chi square method to determine important features. Chi square method calculated a score and tells us how important that feature is. Figure 4.2.1 shows the accuracy of six algorithms.

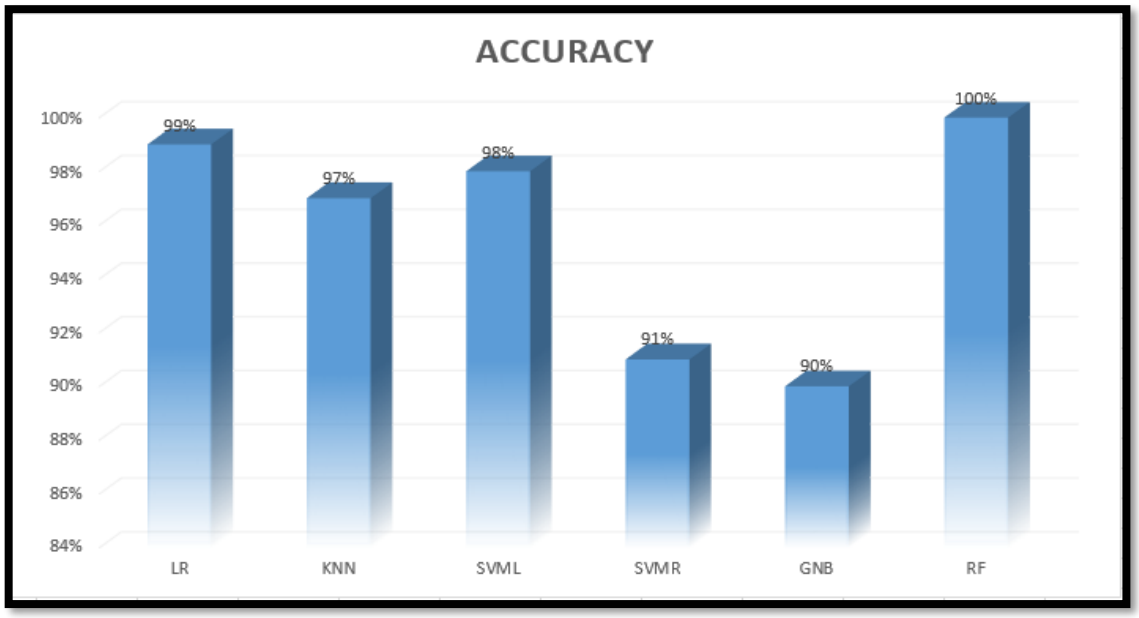


Figure 4.2.2: Accuracy of Six Algorithms.

From the bar chart, we can assume that Logistic Regression has obtained 99% accuracy, KNN has obtained 97% accuracy, Linear SVM has obtained 98% accuracy, Radial SVM 91% accuracy, Gaussian Naive Bayes 90% accuracy and lastly Random Forest 100% accuracy. Finally, we can make a decision that Random Forest showed its accuracy at its best.

4.2.3 Experimental Analysis by Confusion Matrix

The Confusion Matrix is obtained by leveraging the Confusion Matrix () library work. Using framework groups an undeniable origination regarding specific bad or impartial characteristics on an informational index. It is basically a table that may be created for a classifier on a paired informative collection and used to represent the classifier's presentation. False Positive (TP), False Negative (TN), False Positive (FP), and False Negative (FN) values are displayed there. It simply gives a proportion doing prescient qualities figuring from the informational index.

- **True Positive (TP)** = The number of instances is correctly classified which leads to PCOS.
- **False Positive (FP)** = The number of instances is wrongly detected which has PCOS.
- **True Negative (TN)** = The number of instances is currently level which referred to healthy conditions.
- **False Negative (FN)** = The number of instances incorrectly classified which leads to hygiene conditions.

The Confusion Matrix of the algorithms below:

Random Forest:

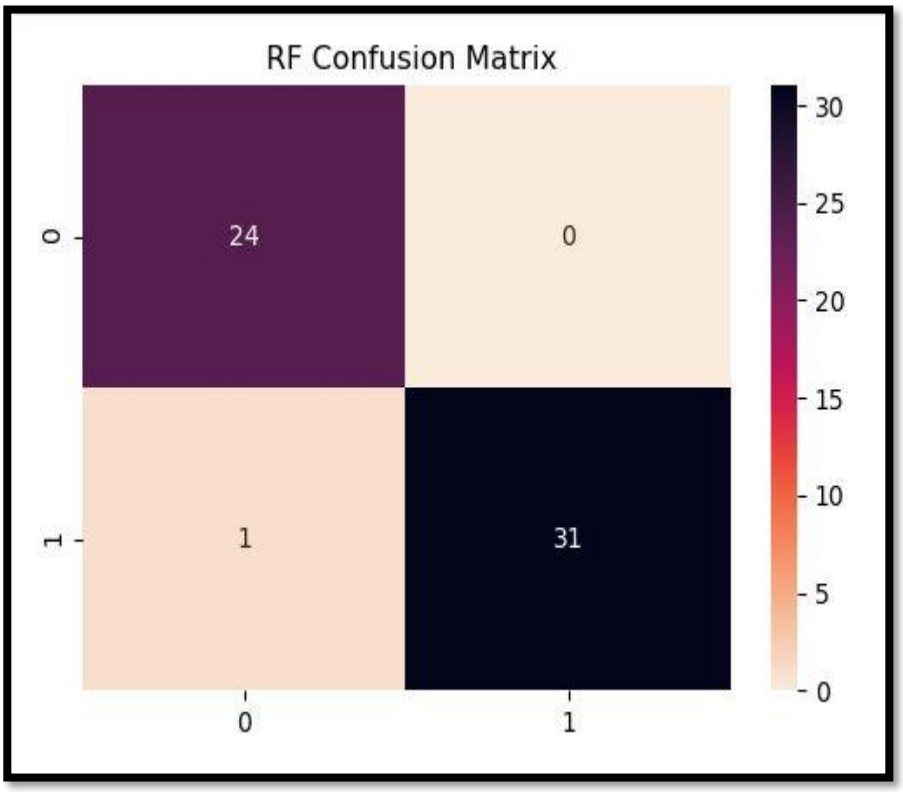


Figure 4.2.3.1: Confusion Matrix of RF.

SVML:

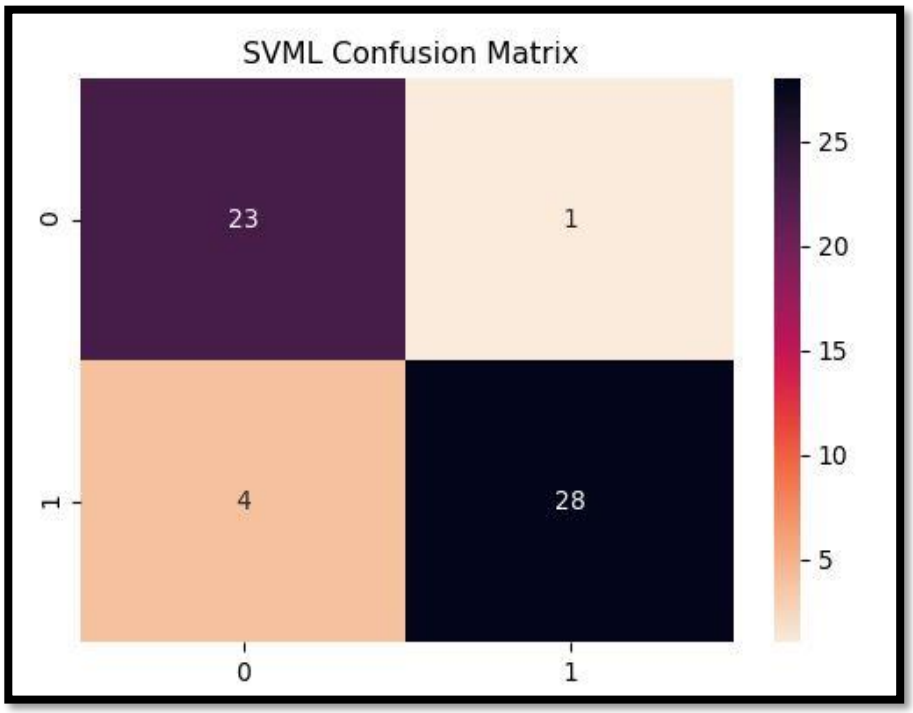


Figure 4.2.3.2: Confusion Matrix of SVML

Logistic Regression:

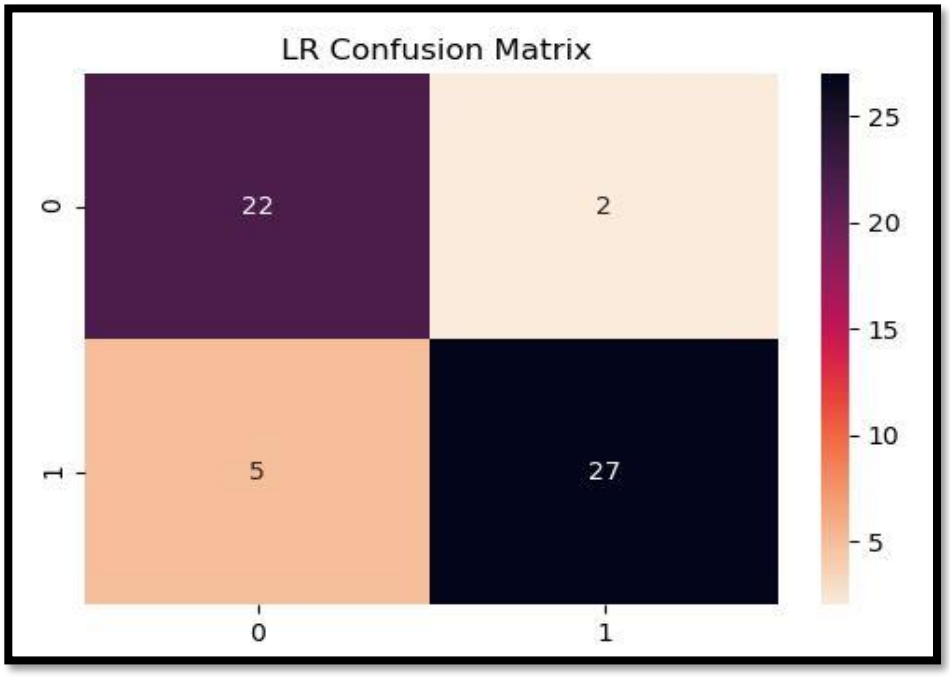


Figure 4.2.3.3: Confusion Matrix of LR.

KNN:

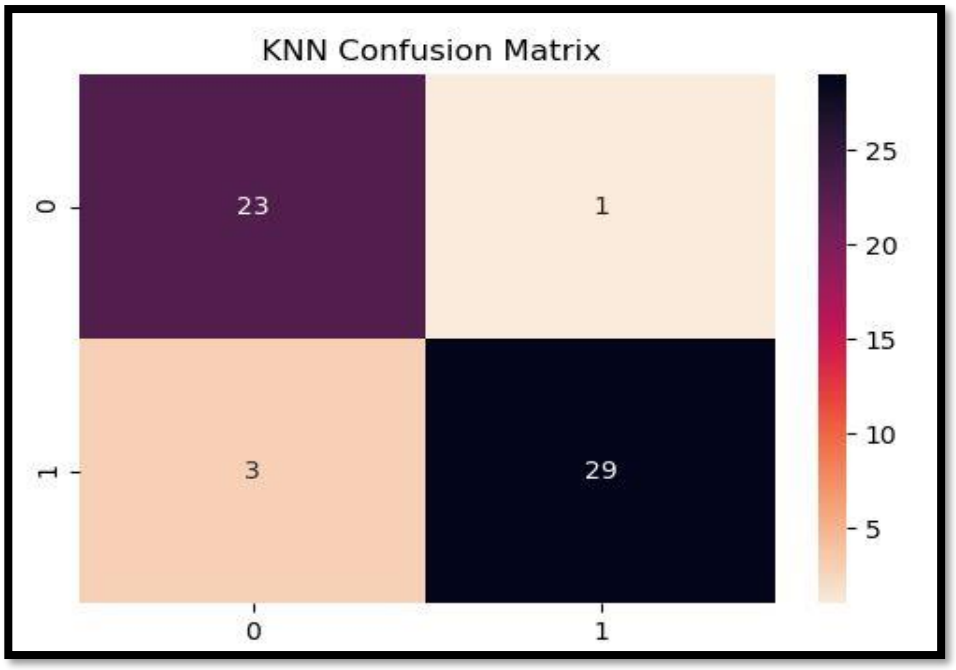


Figure 4.2.3.4: Confusion Matrix of KNN.

Gaussian Naive Bayes:

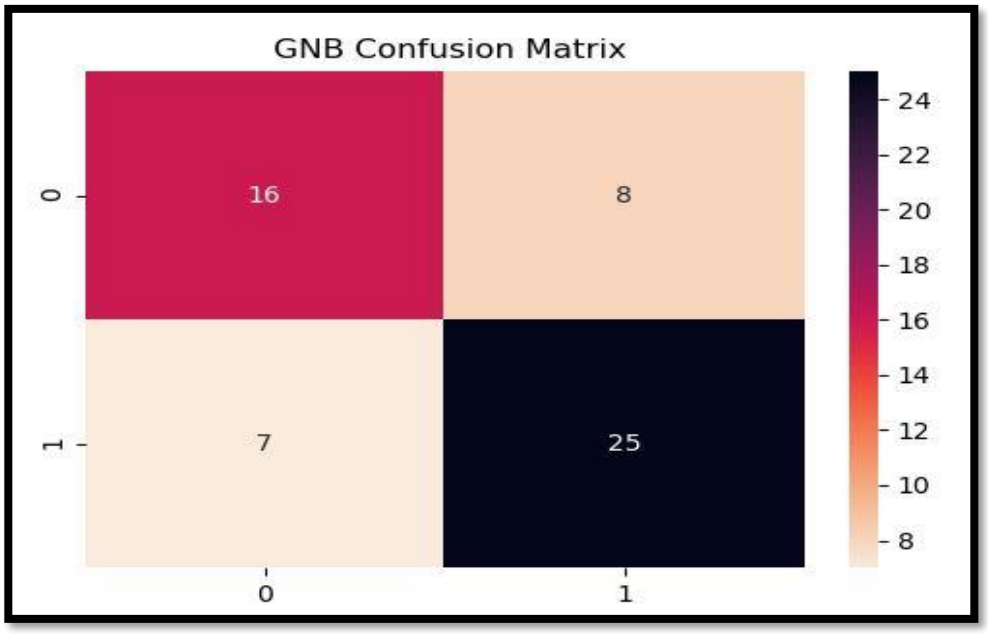


Figure 4.2.3.5: Confusion Matrix of GNB.

SVMR:

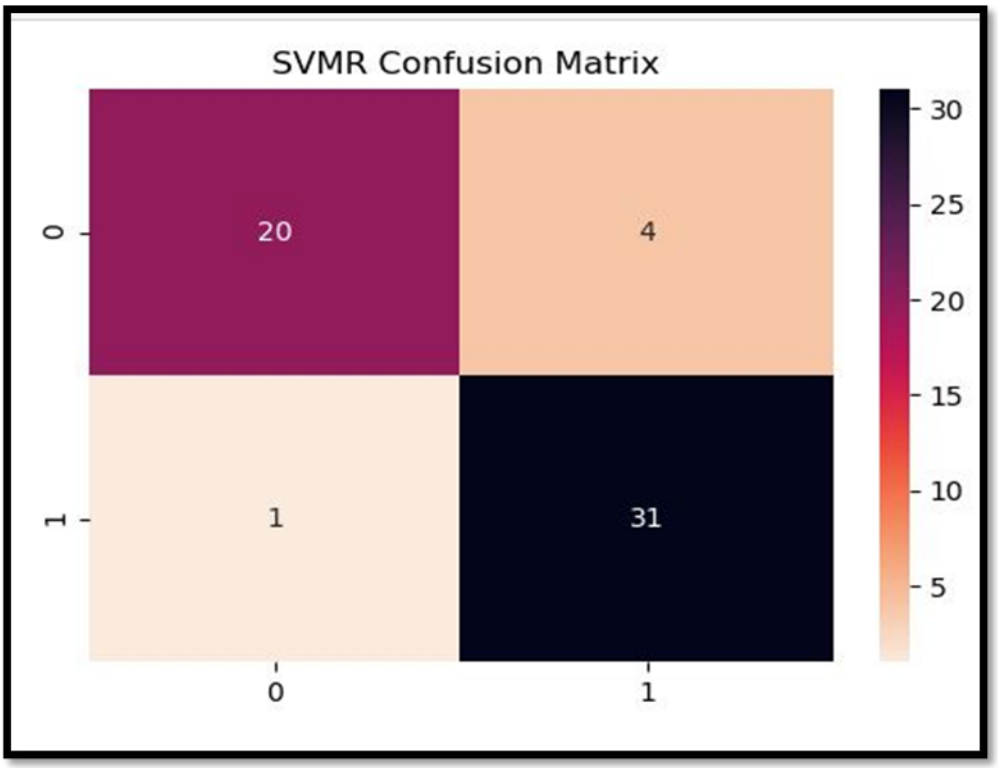


Figure 4.2.3.6: Confusion Matrix of SVMR

Now we will evaluate all the algorithms calculating precision, recall, F1 score according to these confusion matrices:

Precision is a unit of measurement for specificity. This is the proportion of true positive to predicted positive value.

$$Precision = \frac{TP}{TP + FP} \times 100\%$$

A **recall** is a metric for accuracy. This is the true positive rate positive to true positive value.

$$Recall = \frac{TP}{TP + FN} \times 100\%$$

The **F1 score** is a calculation of the harmonic mean of recall and precision. For calculation, it takes into account both false positive and false negative values.

$$F_1 \text{ score } = \frac{2 \ x \ precision \ x \ recall}{precision + recall} \times 100\%$$

Table 4.2.3: Performance Evaluation

Models	**Precision**	**Recall**	***F*$_1$-score**
Random Forest Classifier	100%	96%	98 %
K-Neighbors Classifier	96%	89 %	92 %
Linear kernel SVM	96 %	85%	90 %
Logistic Regression	96%	82%	88%
Radial kernel SVM	83 %	95 %	89 %
Gaussian Naive Bayes	67 %	70 %	68 %

From this table, this evaluation is showing that Random Forest working better than others.

4.2.4 Experimental Analysis by ROC Curve

Receiver operating characteristics (ROC) curves is very useful for visual comparison of classification models. ROC curve is made with a true positive rate and false-positive rate. The diagonal line is representing the random guessing. The curve of a model is close to random guessing, which is a less accurate model. Therefore, for an accurate model the curve will be far away from the random guessing line. The ROC curves of our using algorithms are given below:

Random Forest:

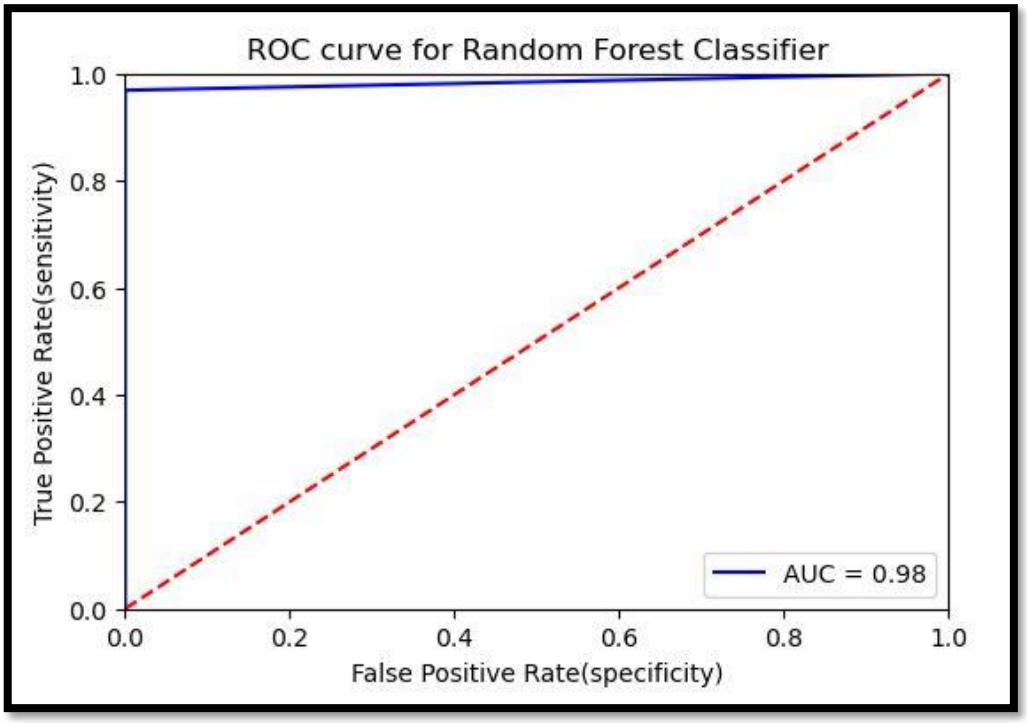


Figure 4.2.4.1: ROC curve of RF.

SVML:

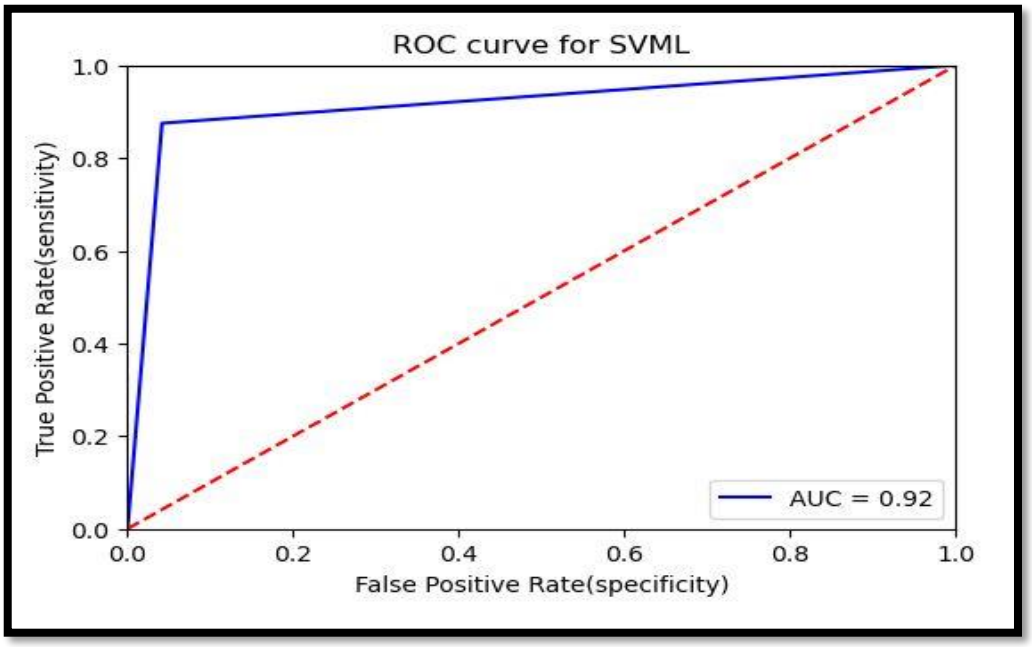


Figure 4.2.4.2: ROC curve of SVML.

Logistic Regression:

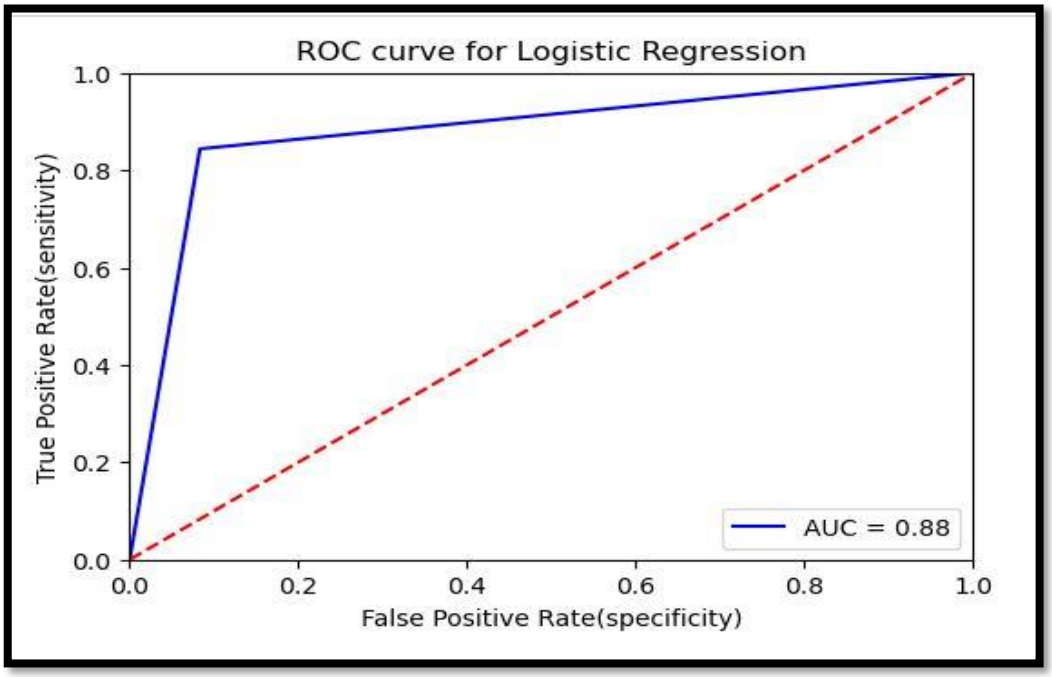


Figure 4.2.4.3: ROC curve of LR.

KNN:

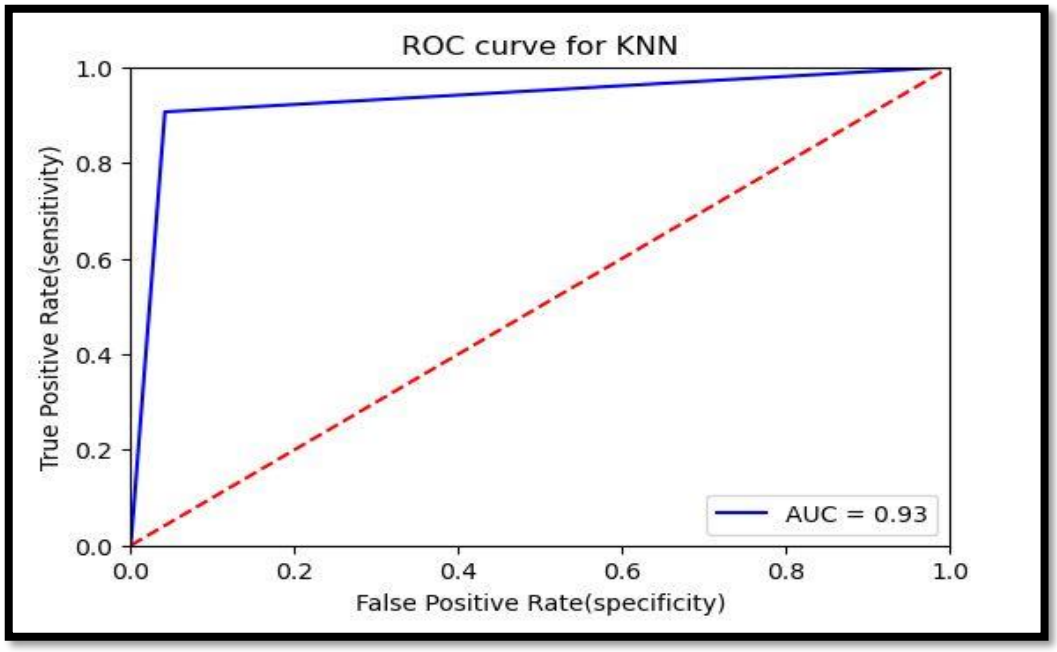


Figure 4.2.4.4: ROC curve of KNN.

Gaussian Naive Bayes:

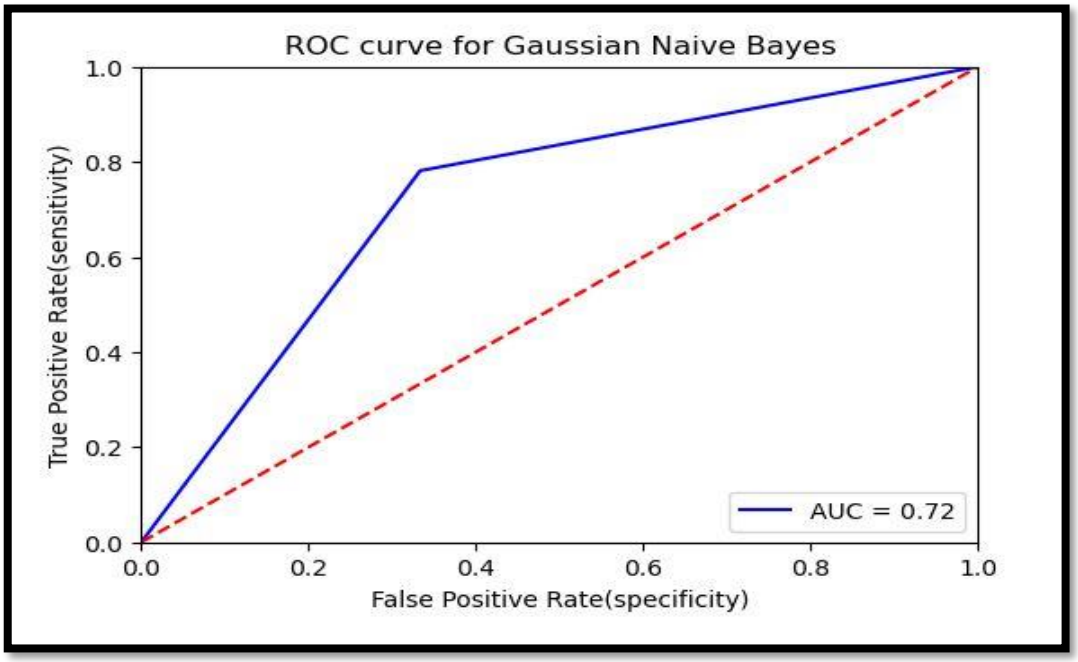


Figure 4.2.4.5: ROC curve of GNB.

SVMR:

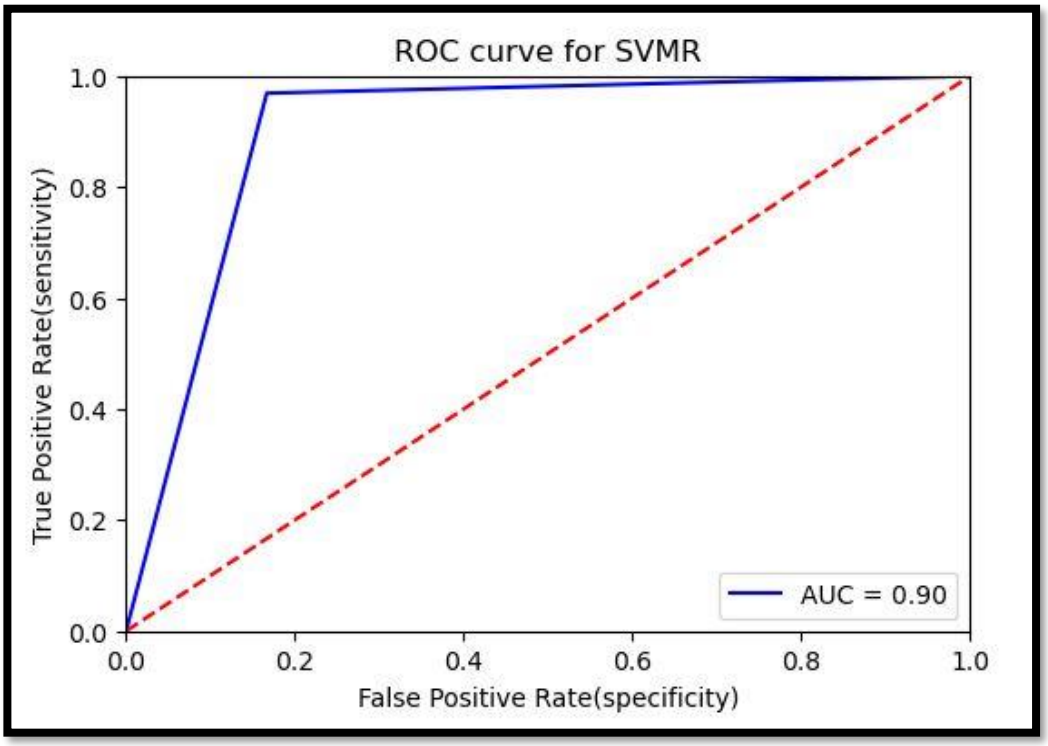


Figure 4.2.4.6: ROC curve of SVMR.

After evaluating all the algorithms from ROC curve, Random Forest Classifier worked better.

4.2.5 Comparative Analysis

The table below is showing the comparison better our work and some top works from our reference. The purpose of our work is to predict the gynecological disorder PCOS. In paper [1] ,they worked on 541 patients with 23 features and succeeded with 89% accuracy applying Random Forest. In paper [2] working on huge dataset with 42 features, they achieved 96% accuracy using Random Forest. In paper [3] 538 patients and 39 features were used to detect PCOS and Logistic Regression achieved 92% accuracy for the model. Authors of [5] worked on PCOS genotypes and SVML found out the best one after all analysis. In [6] , Bayesian Classifier worked well with 93% accuracy on 250

patient's datasets containing 9 features. Lastly, [7] showed a PCOS detection model in which Random Forest was chosen for its 90% accuracy.

After analyzing all the top papers from our reference and comparing them with our work we have found that, in our model Random Forest worked better and its accuracy was better than all other models. Finally, after all evaluation we can declare Random Forest the best one for our work.

Table 4.2.5: Comparative Evaluation.

Method/Work Done	**Sample size**	**Size of Feature set**	**Algorithm**	**Accuracy**
This work	280	19	Random Forest	100%
[1]Amsy Denny	541	23	Random Forest	89%
[2]Malik Hasan	1000	42	Random Forest	96%
[3] Namrata Tanwani	538	39	Logistic Regression	92%
[5] Xing-Zhong Zhang	306	25	SVM linear	80%
[6] Palak Mehrotra	250	9	Bayesian classifier	93%
[7] Vaidehi Thakre	540	30	Random Forest	90%

4.3 Discussion

Overall, our work considers a number of techniques to obtaining reliable anticipated results. However, the researchers used their method based on their understanding of the PCOS dataset. The predicted values, however, are not always given with the right values. However, our prediction method outperforms other current systems in terms of accuracy, with a 100 percent accuracy based on the dataset and this model can predict PCOS well. Finally, we defeat all sorts of hurdles to accurate prediction and arrive at a stage that aids in identifying the disease's status.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Polycystic Ovarian Syndrome detection by utilizing ML will positively affect women health as well as society. Day by day women are started to suffer on this disease, but they are not familiar to this term not even aware of this. Women health is as important as any other asset of a countries. So, in this context, we will identify the stage of the diseases so that people can aware of this before.

5.2 Impact on Environment

As earlier stated, there has been some recent work that can have an impact on the environment. The major goal of all of the effort was to improve the environment so that people might survive the sickness and raise awareness among them. If people are, we have a better society and a healthier environment. Specially Women who are giving birth to children and building a nation. It's really important to be careful with their health so that they can participate in a nation's development.

5.3 Ethical Aspects

Journal by Prof. Kohinoor Begum SSMC, in 2000 says that - 22 percent of women of reproductive age, suffering from PCOS in Bangladesh. Women health results in infertility, obesity. Mass people suffering much as they are not familiar to this term. Their ignorance can lead them to cancer and life threating as well.

5.4 Sustainability Plan

The ability to maintain a specific ratio or level is known as sustainability. The ability to exist indefinitely is referred to as sustainability. The manageability plan is divided into three sections: network, monetary, and authoritative. The Maintainability Plan provides us with a practical idea of any project run and preliminary plans for the project. This

model must be aimed to make it simple for people to change, and it is critical to remember that people do not suffer the negative impacts of mediocrity while using this model.

CHAPTER 6

Summary, Conclusion, Implication for Future Research

6.1 Summary

Our work can be partitioned into 3 significant parts-Information assortment, Strategy, and Test Results. There is no doubt that some examination word on this setting previously existed. Above all, we provided a model for this research project. We'll need to gather information at that moment. We gathered the data from hospitals as a questionnaire form. We pre-handled the dataset at that time for testing reasons. There are several features that are inadequate. After the dataset has been created, we perform six ML algorithms. We acquire better performance from the Random Forest (RF) and KNN computation, which is about 98 percent. Finally, we find our usual result, which is more accurate than the rest of the task-related work. We notice the early stages of diseases and then the rate of illnesses, which has a better outcome than other tasks.

6.2 Limitations

Though we have tried our level best to collect more classes of data, but during this situation, it wasn’t possible. So, this is a limitation. We collected few features of the PCOS which was not enough, more medical data should have added to the dataset. This is another limitation. And we have used few ML methods, we could try other more Machine Learning models or DL methods.

6.3 Conclusion

Polycystic Ovary Syndrome (PCOS) is one of the most common endocrine disorders in women of reproductive age. Infertility and anovulation may develop as a result of this. The clinical and metabolic indicators that serve as biomarkers for the disease are included in the diagnostic criteria. We created a method that detects PCOS based on a small number of possible indicators. We have used popular machine learning algorithms, i.e., SVM (linear & radial), Logistic Regression, K-Neighbors Classifier, Gaussian

Naive Bayes and Random Forest on the clinical data of patients diagnosed PCOS based on symptoms associated with the disease. The performance validation metrics recall, accuracy, precision, and F- statistics indicated best performance of the Random Forest and KNN algorithm in diagnosis of PCOS with an accuracy of 98%. Thus, it is concluded that the Random Forest algorithm and KNN both are the best suitable algorithm for diagnosis of PCOS on the given data. The future scope of the study can include use of different or large data sets for diagnosis of the disease.

6.4 Implication for Further Study

Nowadays, innovation and current science make our lives faster and easier. We need to use our model later in a product, web application, or Android application in our country. Later on, we will have the option of increasing the accuracy of our model by using a larger data set. Furthermore, by creating simple-to-use GUIs, the model's product may be accessible by individuals. In the future, we will add the medical diagnosis images with the current dataset and will make it more informative as much as we can. We can work with this research project for making a research paper. We can make it for the conference paper.

Appendices

Appendix A: Research Issues

Before this work, we had seen there was not that much work on PCOS patients from Bangladesh. Our exceptionally kind and accommodating supervisor assisted us with picking our exploration subject and gave the total rule expected to complete our examination. In this era, it’s important to make our majority of people, women safe and sound by health. So, we decided to work on this not familiar but important as all other diseases.

Appendix B: Related Issues

At the point when we began our examination, we neither thought about the data mining, Jupyter notebook, anaconda, nor did we have any involved involvement in Machine Learning calculations. We set aside a decent measure of effort to learn them all, rehearse and prepare them for work. To start with, we applied them on an alternate dataset just to test in the event that we are doing it right. At that point we began working with the primary dataset to see our normal results. One thing we have gained from our first exploration is, when you get the total image of your work in your mind, and right now realize how to do it, it won't take that long to wrap up.

References

[1] Amsy Denny, Maneesh Ram C, Anita Raj, Ashi Ashok, Remya George. [1] "i-HOPE: Detection and Prediction System for Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques" IEEE TENCON. Kochi, India, December 2019.

[2] Malik Mubasher Hassan, Tabasum Mirza. "Comparative Analysis of Machine Learning Algorithms in Diagnosis of Polycystic Ovarian Syndrome", September 2020, International Journal of Computer Applications (0975 – 8887) Volume 175 – No.17.

[3] Namrata Tanwani. "Detecting PCOS using Machine Learning", Issue:01 2020, IJMTES (International Journal of Modern Trends in Engineering and Science), Volume:07.

[4] Palvi Soni, Sheveta Vashisht. "Exploration on Polycystic Ovarian Syndrome and Data Mining Techniques" Proceedings of the International Conference on Communication and Electronics Systems (ICCES 2018).

[5] Xing-Zhong Zhang, Yan-Li Pang, Xian Wang & Yan-Hui Li. "Computational characterization and identification of human", 28 August 2018, www.nature.com/scientificreports.

[6] Palak Mehrotra, Jyotirmoy Chatterjee, Chandan Chakraborty, Biswanath Ghoshdastidar. Sudarshan Ghoshdastidar. "Automated Screening of Polycystic Ovary Syndrome using Machine Learning Techniques", December 2011, IEEE.

[7] Vaidehi Sunil Thakre, Shreyas Vedpathak. "PCOcare: PCOS Detection and Prediction using Machine Learning Algorithms", December 2020, Article in Bioscience Biotechnology Research Communications.

[8] V. Deepika. "Applications of Artificial Intelligence Techniques in Polycystic ovarian syndrome Diagnosis", November 2019, Journal of Advanced Research in Technology and Management Sciences, Volume: 01 Issue: 03 ISSN: 2582-3078.

[9] Asa Lindholm, Liselotte Andersson, Mats Eliasson, Marix Bixo, Inger Sundström-Poromaa. "Prevalence of symptoms associated with polycystic ovary syndrome". International Journal of Gynecology and Obstetrics (2008).

[10] Gautam N. Allahbadia, Rubina Merchant. “Polycystic ovary syndrome and impact on health”, Middle East Fertility Society Journal (2011).

[11] https://pcos.com/female-sex-hormones-and-pcos/

[12] https://www.webmd.com/women/treatment-pcos#2

[13] Legro RS,” Polycystic Ovary Syndrome and Cardiovascular Disease: premature association?, Endocrine Reviews, vol24, pp.302-312, 2003.

[14] Lin Li, Dongzi Yang, Xiaoli Chen, Yaxiao Chen, Shuying Feng, Liangan Wang. “Clinical and metabolic features of polycystic ovary syndrome”, (2007), International Journal of Gynecology and Obstetrics.

[15] Dumesic, D. A. et al. Scientifc Statement on the Diagnostic Criteria, Epidemiology, Pathophysiology, and Molecular Genetics of Polycystic Ovary Syndrome. Endocrine Reviews 36, 487–525, https://doi.org/10.1210/er.2015-1018 (2015).

[16] Mastorakos G, Lambrinoudaki I, Creatsas G. Polycystic ovary syndrome in adolescents: current and future treatment options. Paediatr Drugs 2006; 8:311–8.

[17] Fauser, BCJM & Chang, Jaehyuk & Azziz, Ricardo & Legro, Richard & Dewailly, Didier & Franks, Stephen & Tarlatzis, BC & Fauser, Bart & Balen, Adam & Bouchard, P & Dahlgren, Eva & Devoto, Luigi & Diamanti, E & Dunaif, A & Filicori, M & Homburg, Roy & Ibanez, L & Laven, Joop & Magoffin, Denis & Lobo, R. (2004). Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS). Human Reproduction. 19. 41-47. 10.1093/humrep/deh098.

Plagiarism Report

PCOS Report_Updated

ORIGINALITY REPORT

17%	12%	10%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	Palak Mehrotra, Jyotirmoy Chatterjee, Chandan Chakraborty, Biswanath Ghoshdastidar, Sudarshan Ghoshdastidar. "Automated screening of Polycystic Ovary Syndrome using machine learning techniques", 2011 Annual IEEE India Conference, 2011 Publication	1%
4	Submitted to University of New England Student Paper	1%
5	www.ijitee.org Internet Source	1%
6	Submitted to Universiti Teknologi MARA Student Paper	1%
7	www.nature.com Internet Source	1%

Daffodil International University Library
Daffodil Tower, (DT)-3, 3rd Floor
102/1, Shukrabad, Dhanmondi – 1207
Tel: 9116774 (Ext.-123, 150,151)

Library Confirmation Form					
Name of the Student	Fatema Tuz Zohora Shefa, Fariha Jannat Ananna, Soma Roy				
Student ID	171-15-8755, 171-15-9369, 171-15-8749				
Group		✓	Yes		No
Group IDs	Fall19D066				
1. Project Title	A Comparative Analysis of Machine Learning Algorithms to Predict Polycystic Ovarian Syndrome (PCOS)				
Submission of Soft Copy of Reports		✓	Yes		No
Name and Designation of the Project Supervisor	Signature of the Project Supervisor				
Ms. Subhenur Latif Assistant Professor					

This is for your kind information that the management of DIU has decided to receive students' Project/Thesis Reports by DIU Library through this email (projectreport@diu.edu.bd) to check Plagiarism by Turnitin Software before submitting to the departments. Students have to submit a plagiarism checking report provided by the DIU library with their Project Report/ Thesis to the respective departments.

Acceptable range of plagiarism at DIU has been settled as follows:

a) Project/ Thesis report of undergraduate students – 40%
b) Project/ Thesis report of Masters students – 30%

Only the acceptable reports will be submitted for further processing.
Actual plagiarism -17%

……
Name and Signature

Submission guidelines of Project/ Thesis/ Internship report

This document contains a guide on soft copy submission of student Project/ Thesis/ Internship Report/ Project Report to DIU library.

Project Report should be arranged as ordered below:

1. **Title page**
2. **Letter of approval /acceptance (with supervisor's signature)**
3. **Acknowledgment**
4. **Dedication**
5. **Abstract / Executive Summary**
6. **Table of Contents**
7. **List of Figures, Tables, Abbreviations, etc.**
8. **The main body or chapters:**
   a. Introduction
   b. Literature review/ Review of Related Literature
   c. Significance of the Study/ Scope of the Study (Optional)
   d. Methodology/ Experimental Details
   e. Analysis / Discussion / Findings / Recommendations
9. **Conclusions**
10. **Appendices**
11. **References (APA style)**
12. **Page Numbering:**
    a. Preliminary pages must be in lower case roman numerals e.g. i, ii, iii.
    b. All pages of the main body or from chapter one will be numbered in Arabic numerals e.g. 1, 2, 3.
    c. All pages have to be arranged according to the table of contents
13. **Format:**
    The report should be in ONE FILE and PDF/ Word format document.
14. **Copyright Note:**
    Write "©Daffodil International University" at footer
15. **Plagiarism checking**: Students' reports will not be accepted without plagiarism checking by Turnitin software.
16. **Submission:**
    Student may send the file to **projectreport@diu.edu.bd** or bring in softcopy in person (Pen Drive) to library project report section (3rd. Floor, Library Building, Daffodil Tower-03).

- - - - - - - - - - - - - - - - - - -

(Dr. Md. Milan Khan)
Librarian
Daffodil International University