EFFECTIVE POLYCYSTIC OVARY SYNDROME (PCOS) CLASSIFICATION WITH MACHINE LEARNING AND OPTIMIZED FEATURE SELECTION

BY

Sadia Jerin Meem

ID: 201-15-13741

FINAL YEAR DESIGN THESIS REPORT

This report is submitted for partial satisfaction of the precondition for the Bachelor of Science degree in Computer Science and Engineering

Supervised By

Dr. S. M. Aminul Haque
Professor and Associate Head
Department of CSE
Daffodil International University

Co-Supervised By

Md. Sazzadur Ahamed
Assistant Professor
Department Of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

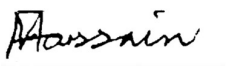
DHAKA, BANGLADESH

JULY 2024

APPROVAL

This Project titled "Effective Polycystic Ovary Syndrome (PCOS) Classification With Machine Learning and Optimized Feature Selection", submitted by Sadia Jerin Meem, ID: 201-15-13741 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 14/07/2024.

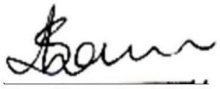
BOARD OF EXAMINERS



Dr. Md. Fokhray Hossain (MFH) — **Board Chairman**
Professor
Department of CSE
Daffodil International University

Amatul Bushra Akhi (ABA) — **Internal Examiner 1**
Assistant Professor
Department of CSE
Daffodil International University

Mr. Amir Sohel (ARS) — **Internal Examiner 2**
Sr. Lecturer
Department of CSE
Daffodil International University

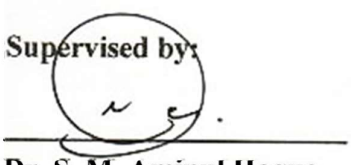


Dr. Shamim H Ripon (DSR) — **External Examiner**
Professor
Department of CSE
East West University

DECLARATION

I hereby declare that this thesis has been done by me under the supervision of **Dr. S. M. Aminul Haque, Professor and Associate Head, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

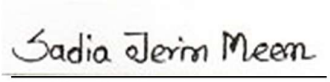


Dr. S. M. Aminul Haque
professor and associate Head
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Sazzadur Ahamed
Asistant Professor
Department of CSE
Daffodil International University

Submitted by:



Sadia Jerin Meem
ID: 201-15-13741
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty for His divine blessing making it possible for me to complete the final year thesis/internship successfully.

We are grateful and wish our profound indebtedness to **Dr. S. M. Aminul Haque, Professor and Associate Head** of the Department of CSE at Daffodil International University in Dhaka. Our supervisor has a profound understanding and ardent interest in the field of "Machine Learning" in order to direct this endeavor. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to extend my sincere appreciation to **Dr. Sheak Rashed Haider Noori**, Head of the Department of CSE, as well as to the other faculty members and personnel of the CSE department at Daffodil International University, for their invaluable assistance in completing our project.

I'd like to extend my gratitude to every Daffodil International University classmate who participated in this discussion while completing course requirements.

Finally, I would like to extend my sincere appreciation to my parents for their unwavering patience and support.

ABSTRACT

Polycystic Ovary Syndrome (PCOS) is a prevalent endocrine disorder affecting women of reproductive age, necessitating accurate prediction and diagnosis for effective patient care and personalized treatment. Despite numerous studies on PCOS, there remains a gap in the literature regarding the most effective classification algorithms and feature selection methods tailored to PCOS datasets, as well as a comparative analysis of data science tools like Python and RapidMiner in this context. This study aims to address these gaps by identifying the most effective machine learning classification methods for predicting PCOS, determining the significant features involved, and comparing the performance of Python and RapidMiner tools. Key steps include data preprocessing, constructing and evaluating a stacking classifier using Python, and implementing various K-fold cross-validation settings to assess training and test accuracies, precision, and recall. The research examines variables of infertility and non-infertility within the dataset, using a Stacking Classifier model with 37 features to report on cross-validation results. Findings contribute to advancing PCOS prediction by leveraging machine learning techniques and provide insights into the efficiency of Python and RapidMiner in healthcare data science. These outcomes have implications for customizing healthcare, enhancing patient care, and optimizing operational efficiency, ultimately aiming to improve PCOS management and treatment through advanced data science methodologies.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES **PAGE NO**

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

The prevalence and impact of Polycystic Ovarian Syndrome (PCOS) on women's health make it a significant and widespread endocrine system disorder, affecting approximately 5 to 10% of the female population [1]. The manifestation of Polycystic Ovary Syndrome (PCOS) occurs during adolescence and is attributed to hormonal disturbances. The presence of fluid-filled sacs known as follicles or cysts can be observed within the periphery of the ovary. The characterization of a polycystic ovary (PCO) involves the presence of twelve or more follicles with a diameter ranging from 2-9 mm [2]. The Impact of Polycystic Ovary Syndrome (PCOS) on Women's Health and Quality of Life. The thesis examines a range of symptoms associated with polycystic ovary syndrome (PCOS), including cardiovascular diseases, ovulation failure and infertility, delayed menopause, type 2 diabetes, acne, baldness, hair loss, hirsutism, obesity, anxiety, depression, and stress. The global prevalence of the mentioned phenomenon is estimated to be between 2.2% and 26% according to a reliable source [3]. The prevalence of the South Asian population in the United Kingdom (UK) is significantly higher (52%) compared to the Caucasian population (22%), as indicated by a community study conducted in the region [2]. Early diagnosis and treatment play a crucial role in effectively managing medical conditions by addressing symptoms and preventing potential long-term complications. Detection of Polycystic Ovary Syndrome (PCOS) through ultrasonography involves the assessment of follicle count and size in the ovaries by a medical professional. The detection of PCOS requires a significant amount of time, as well as a need for high-quality images and precise accuracy [4]. The detection of PCOS can also be achieved through the examination of biochemical parameters, specifically hormone levels. The detection of polycystic ovary syndrome (PCOS) often relies on clinical parameters such as body mass index (BMI) and menstrual cycle length due to the high cost of hormone examination [3]. The use of machine learning (ML) classification and feature selection algorithms for disease prediction has become increasingly prevalent among researchers and clinicians as a non-invasive approach in recent years [5][6].

1.2 Motivation

The utilization of PCOS datasets, encompassing diverse attributes such as biochemical, clinical, medical history, symptoms of patients, and ultrasound images, serves as a foundation for constructing predictive models [2]. The current literature lacks documentation on the classification algorithms and feature selection methods used for the PCOS dataset, as well as a performance comparison between Python and RapidMiner tools [1].The purpose of this thesis is to examine and analyze the information presented in references [7] and The objective of this paper is to determine the superior machine learning classification algorithm for predicting PCOS disease and identify the influential features in the prediction process. Additionally, the performances of Python and RapidMiner tools will be compared using the PCOS dataset. The potential to utilize data science and technology in the evolving landscape of health care and information technology presents an opportunity to personalize health care and improve the delivery of patient care. In this thesis, the fundamental concept of machine learning (ML) is explored, which involves the utilization of artificial intelligence to construct prediction models. This is achieved through the identification of hidden patterns within extensive data sets [14]. Consequently, scientists have been motivated to explore diverse machine learning algorithms for disease prediction [15]. The objective of this study is to utilize classification approaches, including Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), and Stacking Classifier, to predict the presence of Polycystic Ovary Syndrome (POCS) in patients. The application of data mining has the potential to enhance the operational efficiency of the healthcare industry by aiding in the diagnosis of challenging syndromes.

1.3 Rationale of the Study

The purpose of this work is to fill the gap in existing literature regarding the classification algorithms and feature selection methods for the PCOS dataset. The study seeks the most effective machine learning classification method for PCOS disease and discover the significant features in prediction process. It aims to evaluate the performances of the Python and RapidMiner tools by utilizing the PCOS dataset. The study is driven by the possibility of customizing healthcare, enhancing patient care, and optimizing operational efficiency in the healthcare sector by utilizing data science and technology.

1.4 Research Questions

1. What is the composition of the dataset utilized for the experiment, which contains combined infertility and non-infertility features?
2. What specific data preprocessing processes, such as null value handling, null features, and outlier removal, were performed on the dataset?
3. What were the conclusions regarding the proposed stacking model's accuracy in comparison to machine learning models and previous works?

1.5 Expected Output

1. Examine the collective variables of infertility and non-infertility in the dataset.
2. Conduct data preprocessing, encompassing tasks such as addressing missing values, manipulating variables, and removing anomalies.
3. Construct a stacking classifier by utilizing Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT) algorithms.
4. Assess the precision of the stacking classifier in relation to alternative machine learning models and prior research investigations. The effectiveness of machine learning models and the suggested stacking classifier by measuring metrics such as accuracy, precision, sensitivity, recall, and F1-score.
5. Evaluate the performance of the stacking model in comparison to other models and past studies based on the evaluation data.

1.6 Project Management and Finance

Finance: To get the work successfully, there is several type costing like copying cost, transport cos. And I have made those cost from self-funded.

Sl No.	Costing Spaces	Cost
1.	Data collection	5000 BDT
2.	Instrument (Hardware)	30000 BDT
3.	Instrument (Software)	10000 BDT
4.	Transport	7000 BDT
5.	Communication cost	5000 BDT
	Total Estimated Cost:	57000 BDT (Approximate)

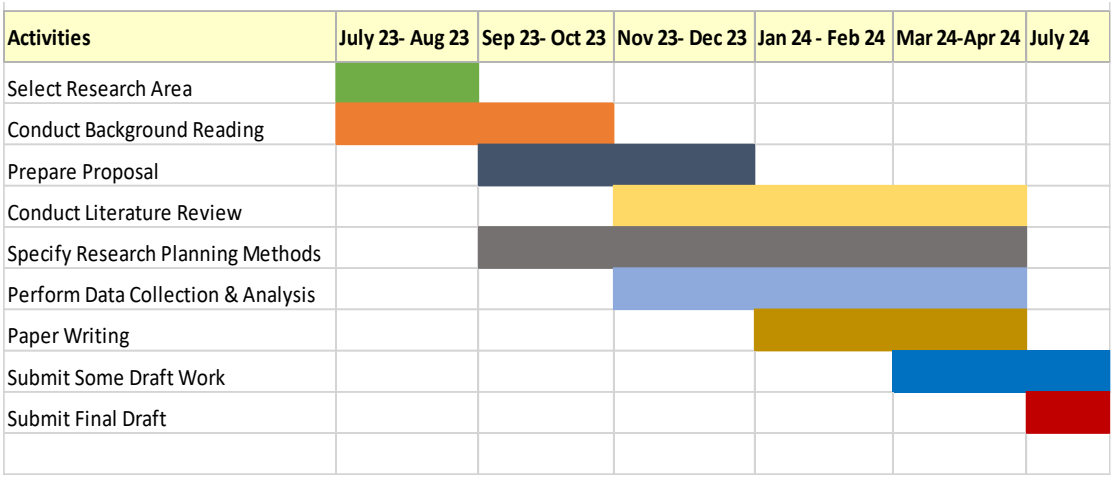


Figure 1.6.1: Gantt chart for PCOS Thesis Writing and Research Process

1.7 Report Layout

In Chapter 1, I lay the foundation for our study. The Introduction provides an overview of the research, followed by the research aims that elucidate the objectives of our investigation. I conclude Chapter 1 with the formulation of important research questions that guide our inquiry.

In Chapter 2, I present brief summaries of the literature review, highlighting key findings and gaps in existing research relevant to our study.

In Chapter 3, I describe the proposed methodology in depth, detailing the steps and processes undertaken to conduct the research effectively.

In Chapter 4, I explain and explore the experimental results of the study, providing a thorough analysis of the findings.

In Chapter 5, I discuss the sustainability and social impact of the research, examining how the results contribute to broader societal goals and long-term viability.

In Chapter 6, I conclude the current research and provide a plan for future efforts, outlining potential directions for continued investigation and application.

CHAPTER 2

Background

2.1 Preliminaries

It is crucial to get a comprehensive comprehension of the fundamental concepts and terminologies associated with Polycystic Ovarian Syndrome (PCOS) during the background studies. This encompasses the clarification of essential concepts such as PCOS, polycystic ovary (PCO), ultrasonography, hormone levels, body mass index (BMI), menstrual cycle duration, and machine learning (ML) classification methods. By presenting a thorough explanation of these basic concepts, researchers and readers can establish a shared understanding of the essential knowledge required to comprehend the ensuing debates and analyses in the study.

2.2 Related works

Using biochemical and genotype data from a previously published PCOS genome-wide association study (GWAS), Matthew et al. [1] discovered the reproducible reproductive and metabolic subgroups of PCOS. Quantitative anthropometric, reproductive, and metabolic characteristics were examined in a genotyped sample of 893 PCOS cases using unsupervised hierarchical cluster analysis. This data suggests that because various PCOS subgroups may have distinct therapy responses and long-term diagnoses, it may not be useful to combine women with PCOS under a single diagnostic. Using a clinical public dataset and machine learning techniques, Shamik et al. [4] presented a polycystic ovary syndrome (PCOS) detection method. The authors obtained statistically significant and discriminating characteristics based on the best-indicating correlation coefficient for polycystic ovary syndrome (PCOS). These features were subsequently input into different machine learning models. Results from the trials demonstrate that compared to the other popular machine learning algorithms, Random Forest (RF) achieves the highest accuracy rate of 93.25 percent. The accuracy of RF's predictions can also be assessed using the out-of-bag (OOB) error. Another study by Zhang et al. in [5] uses spectral classification models and PCA in conjunction with Raman spectroscopy to detect PCOS in samples. Fifty women, split evenly between those with polycystic ovary syndrome and those without, had their follicular fluid and plasma analyzed using Raman spectroscopy. There was a clear divide between the PCOS and non-PCOS groups in the principal component analysis. A stacking classification model

that utilized RF, severe gradient boosting, and KNN was able to attain an accuracy rate of 89% on follicular liquid samples and 75% on plasma samples. Raman spectroscopy, in a nutshell, is an innovative and inexpensive way for PCOS automated diagnosis. The key clinical and laboratory characteristics for the diagnosis of polycystic ovary syndrome were outlined by Silva et al. [6]. Features were selected and predictions were made using the RF classifier and the BorutaShap technique. A purported 86% accuracy was achieved by ranking the relevance of 58 chosen features. Human bias is inherent in the traditional method of counting follicles by hand. A mistake might also occur if the follicles are overlapping when the ultrasound is being performed. The automated PCOS prognosis method developed by Katarya et al. in [7] makes use of PSO and a tailored stacked ensemble-learning model to pick features. With this approach, the accuracy rate was 90.74 percent. For early PCOS identification, the significance of feature selection strategies and ensemble classifier methods was explored in the work of Danaei et al. [8]. The authors used the dataset that was made public on Kaggle to assess the performance of various classifiers, including ensemble random forest, multi-layer perceptron, AdaBoost, and extra tree. The authors showed through their simulation study how important the feature selection technique is, and how the ensemble random forest algorithm worked best with the lower amount of data. For follicular segmentation using PSO, Setiawati et al. [9] suggested a better non-parametric fitness function that would produce more convergent clusters. Improved convergence is supported experimentally by comparing it to a nonparametric fitness function. Ultrasound pictures showed how effective it was. Contrast enhancement improved PSO image clustering with larger intra-cluster distances. A ROI that was more in line with the doctor's reference ROI was generated by the scheme. Unknown ultrasound pictures of the ovaries were also classified using logistic regression, support vector machines (SVMs), and BPNNs. Based on the results (F-measure 0.85), BPNN was determined to be the most effective. Bhardwaj et al. [10] brought attention to the importance of AI and related areas in healthcare, namely in the detection of polycystic ovary syndrome. A variety of methods were employed to optimize the model, including support vector machines, multi-layer perceptrons, XG boost, and random forest classifiers, after the most suitable features were selected using Pearson correlation. The data came from 541 female patients and was made available through a Kaggle dataset. Using SVM, they acquired an accuracy rate of 93%. Even though they employed a different algorithm, Zigarelli et al. [11] made use of the identical dataset. They achieved

a 90.1% success rate in object classification using the CatBoost algorithm. One data-driven method was utilized by Bharati et al. [12] to identify polycystic ovary syndrome. The authors utilized a Kaggle dataset and implemented cross validation, feature selection, and deletion. Three classifiers—catboost, voting soft, and voting hard—are used in the models creation process. Their proposed soft voting categorization model achieved a peak accuracy of 91.12%. Zozan et al. [13] shown that Polycystic Ovary Syndrome can be diagnosed by analyzing hormone levels in blood serum. The authors have monitored the biomolecules in the serum using Raman spectroscopy. Two groups of women, one with polycystic ovary syndrome (PCOS) and one without, had their blood serum samples drawn. The authors were able to differentiate between PCOS and healthy women's blood serum using a combination of chemometric measurement, spectroscopic measurement, and correlation analysis, with an accuracy ranging from 80.92% to 96.04%. Lots of different ML models were used. The KNN model achieved its maximum accuracy when the value of k was set at 3.

2.3 Comparative Analysis and Summary

Table 2.3.1: PCOS classification comparison

Ref.	Algorithms	Dataset	Results
Shamik et al. [4]	RF, SVM, KNN, CatBoost	Kaggle Kottarathil PCOS [540 instances. 42 Features]	RF = 92.4%, SVM = 87.1%, KNN = 66%, CatBoost = 90%
Zhang et al. [5]	KNN, RF, XGB, Stacking	Private PCOS data [100 instances. 12 Features]	KNN = 72.03%, RF = 67.50%, XGB = 72.41%, Stacking = 74.78%
Silva et al. [6]	RF	Private PCOS data [145 instances. 58 Features]	RF = 86%
Katarya et al. [7]	LR, SVM, Ridge, RF, Bagging, Gradient boosting, Proposed model	Kaggle PCOS [540 instances. 42 Features]	LR = 88.86% SVM = 88.32% Ridge = 89.05% RF = 89.99% Bagging = 87.57% Gradient boosting = 88.69%, Proposed Model = 90.74%

.2.4 Scope of the Problem

Despite this, most of the studies mentioned above have a few flaws, such as an inadequate feature count, bad feature extraction, improper treatment of missing data, and an inadequate data cleaning and preprocessing procedure. To combat these problems, we set out to create a more reliable model for PCOS prediction in our study.

2.5 Challenges

The following are the specific research challenges:

1. Dataset: In order to train and test this model, numerous datasets from various classes must be acquired.

2. Choose a foundation Model: To overcome long training times and insufficient data, ablation studies require a perfect foundation model.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

Research Subject:

The research focuses on using machine learning to predict polycystic ovary syndrome (PCOS), an important endocrine disorder. Advanced classifiers like Logistic Regression, Decision Trees, and Random Forests are employed to enhance diagnostic accuracy. The study compares different feature selection and classification methods to identify the most effective approach for PCOS prediction, aiming to improve diagnostic tools and patient outcomes.

Instrumentation:

Experiments were conducted using Python, Scikit-learn, and Jupyter notebook on high-performance CPUs and GPUs. The dataset was preprocessed by removing unnecessary features, filtering extremes, and handling missing values. Feature extraction techniques like MinMaxScaler, RobustScaler, and Quantile Transformer were used. The data was split into 50% training, 20% testing, and 30% validation sets. Machine learning models, including Logistic Regression, Decision Trees, and Random Forests, were implemented and integrated using a stacking classifier approach. Detailed documentation and ethical considerations were maintained throughout the study to ensure transparency and reproducibility.

3.2 Data Collection Procedure

For the study, a dataset of 50 features from 541 women, including fertility and infertility, was gathered from the kaggle repository. 364 of the instances are normal people, while the remaining 177 are PCOS patients.

Table 3.2.1: Dataset description

Name	Amount
Total Cases	541
Features (Merging fertility with infertility)	50

PCOS Positive	364
PCOS Negative	177

To make our experiment successful, we have to gain some knowledge of all the features in this dataset. We did not consider “SL. No, Patient Id” as features, as those are just indexing. Additionally, some features that has no values in the dataset like “Unnamed: 44, Sl. No_y, PCOS (Y/N)_y, I beta-HCG (mIU/mL)_y, II beta-HCG (mIU/mL)_y, AMH(ng/mL)_y” were removed.

Table 3.2.2: Description of the features of the Dataset

Index	**Feature**	**Range**	**Average**
1.	Age (yrs)	20-48	30
2.	Weight (Kg)	42-152	62
3.	Height(Cm)	142-178	158
4.	BMI	16-40	-
5.	Blood Group	11-18	-
6.	Pulse rate(bpm)	70-82	73
7.	RR (breaths/min)	18-22	-
8.	Hb(g/dl)	8-14	11
9.	Cycle(R/I)	2-4	2.6
10.	Cycle length(days)	2-12	5
11.	Marriage Status (Years)	1-22	8
12.	Pregnant(Y/N)	0-1	-
13.	No. of aborptions	0-4	0.3
14.	beta-HCG(mIU/mL) (**I**)	1.3-32461	664
15.	beta-HCG(mIU/mL) (**II**)	0.99-25000	238.67
16.	FSH(mIU/mL)	0.21-5052	14.60

Index	Feature	Range	Average
17.	LH(mIU/mL)	0.02-2018	6.5
18.	FSH/LH	-	-
19.	Hip(inch)	26-48	38
20.	Waist(inch)	30-46	34
21.	Waist:Hip Ratio	-	-
22.	TSH (mIU/L)	0.1-66	5.62
23.	AMH(ng/mL)	0.4-129	5.62
24.	PRL(ng/mL)	0.4-129	24.32
25.	Vit D3 (ng/mL)	0-6014	49.91
26.	PRG(ng/mL)	0.04-1.1	0.61
27.	RBS(mg/dl)	60-350	99.83
28.	Weight gain(Y/N)	0-1	-
29.	hair growth(Y/N)	0-1	-
30.	Skin darkening (Y/N)	0-1	-
31.	Hair loss(Y/N)	0-1	
32.	Pimples(Y/N)	0-1	
33.	Fast food (Y/N)	0-1	-
34.	Reg.Exercise(Y/N)	0-1	-
35.	BP _Systolic (mmHg)	110-140	114.67
36.	BP _Diastolic (mmHg)	70-100	76.92
37.	Follicle No. (L)	1-22	6.12
38.	Follicle No. (R)	1-20	6.64
39.	Avg. F size (L) (mm)	5-24	15.01
40.	Avg. F size (R) (mm)	5-24	15.45
41.	Endometrium (mm)	3-14	8.47

The given 41 features out of 50 (combining fertility and infertility) are also important in our models. Data preprocessing is important because the dataset contains null values and out of range values (outliers). Section 3.2 provides a brief overview of the data cleaning and handling process. Figure 3.2 illustrates some linear relationships between menstrual cycle length, age, BMI, and Folicle(L) vs Folicle(R) in PCOS vs normal.

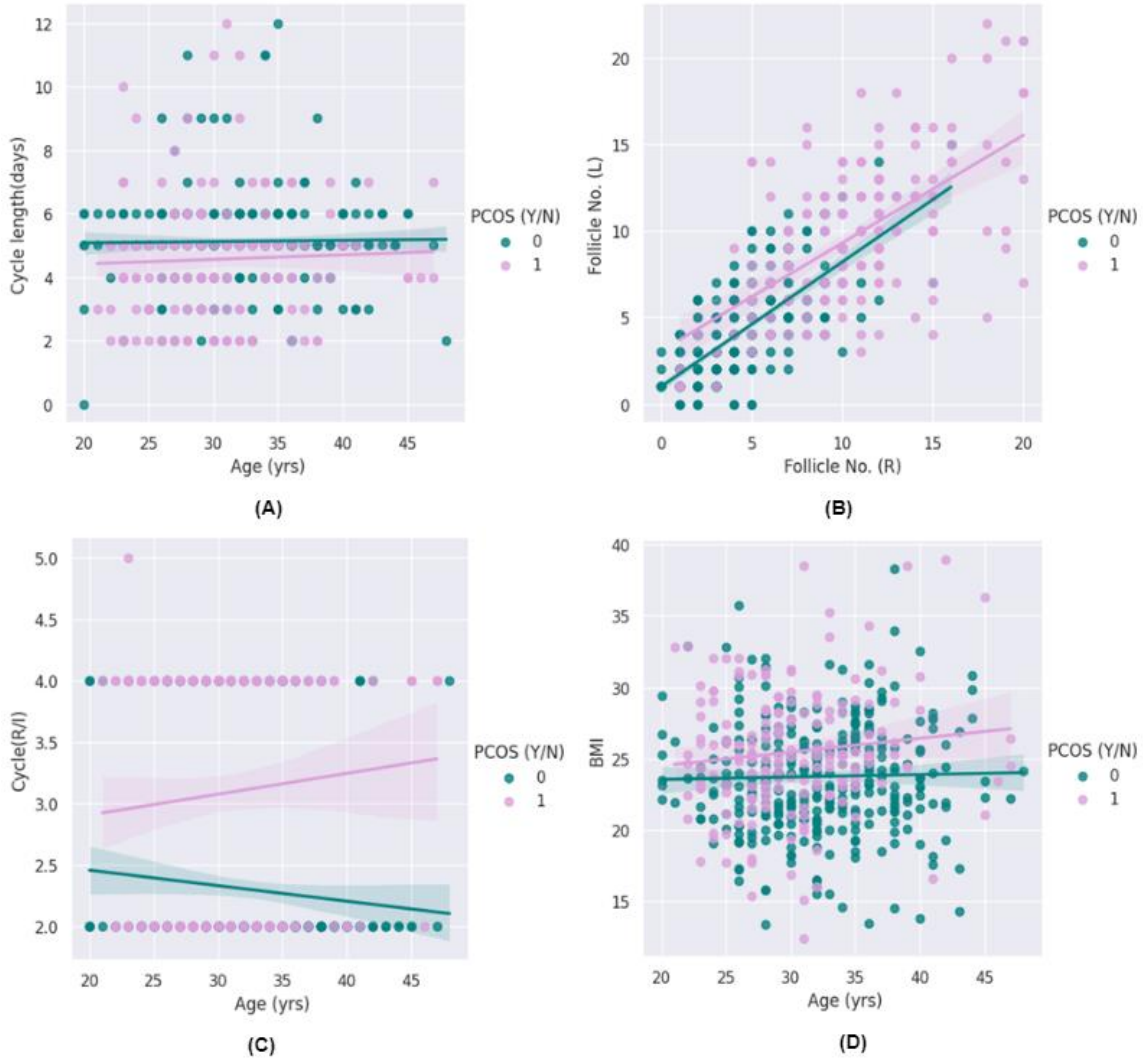


Figure 3.2.1. (a) Presents a graphical representation of the linear correlation established via regression between the duration of the menstrual phase in individuals with polycystic ovary syndrome and that of a healthy woman. (b) Illustrates the linear correlation, as ascertained via regression analysis, between Folicle (L) and Folicle (R) in both PCOS and normal conditions.(c) Presents a graphical representation of the linear correlation, as ascertained via regression, between a cycle factor (R/I) and normal and PCOS years.(d) Generates a graphical representation of the linear correlation, as ascertained via regression analysis, between weight gain (BMI) over an extended period of time in individuals with PCOS and those who are normal.

3.3 Statistical Analysis

The evaluation utilizes the confusion matrix and includes metrics such as accuracy, precision, recall, and F1 score. True positive (TP) values indicate correctly identified positive instances, while false positives (FP) occur when negative instances are mistakenly labeled as positive. On the other hand, false negatives (FN) occur when positive instances are incorrectly labeled as negative. Additionally, the matrix includes true negative (TN) values, representing correctly identified negative instances, and false negative (FN) values, which indicate negative instances misidentified as positive.

$$\text{Accu} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

$$\text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F1 Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{6}$$

3.4 Proposed Methodology

Using machine learning classifiers, this section describes the trials that were utilized to predict polycystic ovary syndrome. As a machine learning tool, the research makes use of the Python programming language. A number of tools, like the Scikit-learn library and the Jupyter notebook, are used to deploy Python. We must begin by cleaning up our dataset. This is why we have eliminated unnecessary Features and null columns, filtered out extreme cases, and implemented association rules to deal with missing values. We have to use feature extraction methods like MinMaxScaler, RobustScaler, and Quantile Transformer to acquire the best features because our dataset had too many characteristics. Following our successful use of the data transformation scaler (MinMaxScaler), we partitioned the dataset as follows: 50% for training, 20% for testing, and 30% for validation. Execute our model with a number of well-known ML classifiers, such as LG, Decision tree, and Random forest. After that, they used decision trees, random forests, and logistic regression to assess the relative value of features in each model. After retaining 37 out of 50 features for use with stacking classifiers, features that appeared most frequently across all three datasets will be given greater

weight than other features. In order to forecast the overall model's outcomes, stacking classifiers drew on those three models. You can see the entire overflow in Figure 3.1.

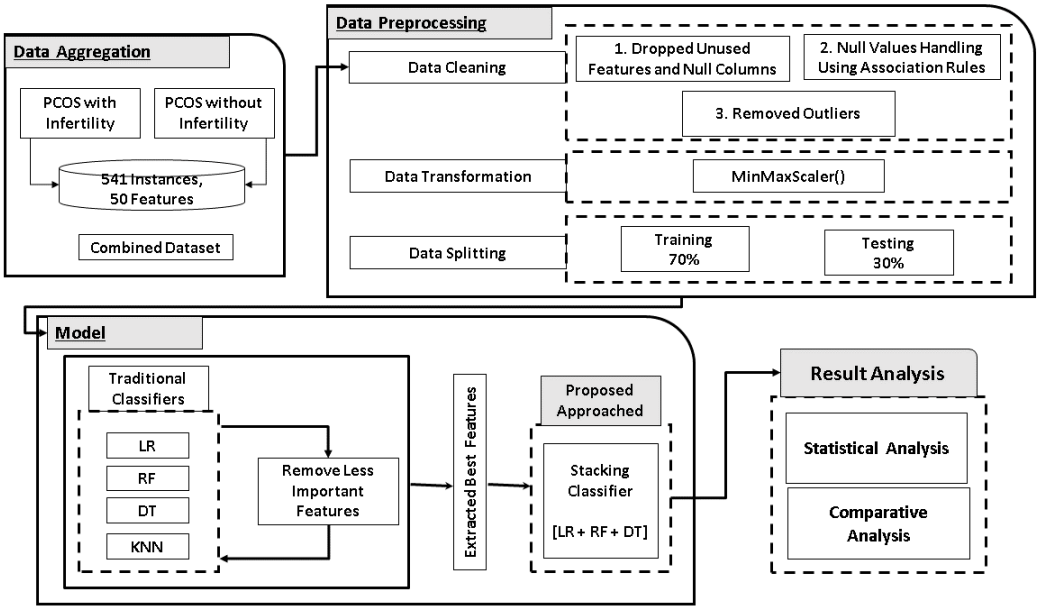


Fig 3.4.1: Whole methodology to perform PCOS prediction

3.4.1 Data Preprocessing

Any kind of processing done on raw data to get it ready for additional processing is referred to as data preparation. Historically, it has been an essential initial stage in the data mining procedure. Recently, methods for preparing data have been adopted for both training and inferring from machine learning and artificial intelligence models. Four phases make up our preprocessing of the data: feature extraction, data transformation, and data cleaning.

Data Cleaning

The process of modifying, rewriting, and organizing data in a collection such that it is typically consistent and appropriate for analysis is known as data cleaning. Typically, certain data are unrelated to any particular classification. It is crucial because well-maintained data enhances overall productivity and provides the most precise classification information. The steps taken to clean the study's data are described below.

Eliminate superfluous data: The dataset contains some pointless information, including ids and null columns, which have nothing to do with categorization. These columns in our dataset—"Unnamed: 44," "Sl. N," "I-beta-HCG(mIU/mL) y," "II-beta-

HCG(mIU/mL) y," and "AMH(ng/mL) y"—were removed because they were not useful for categorization.

Eliminate data outliers: Extreme values that deviate from the norm and are not included in the dataset can occasionally be found in datasets. These are called outliers, and comprehending and even eliminating these outlier values can frequently enhance model skills and machine learning modeling in general. Numerous factors, including measurement or input error, can result in outliers.

Real-life outliers finding : There is no reliable method to characterize and identify outliers in general due to the distinctive qualities of every dataset. Rather, to ascertain if a result is an outlier or not, one needs evaluate the raw observations. In descriptive statistics, groupings of numerical data based on their quartiles can be visually represented using a box plot. In addition to showing variability beyond the top and lower quartiles, box plots can also have vertical lines (whiskers) extending from the boxes; this characteristic gives origin to the terms box-and-whisker plot and box-and-whisker diagram. Plotting distinct outliers as separate points is possible. In order to identify dataset outliers, we first looked at the box plots of every feature in our dataset, which are shown in figure 3.3.

The box plot of the dataset's features is shown in Figure 3.3. Boxplots make it simple to identify outliers. Weight, BMI, cycle duration, PRL, RBC, etc. are a few examples of obvious outliers. We have addressed these outliers by implementing a few theoretical machine learning techniques.

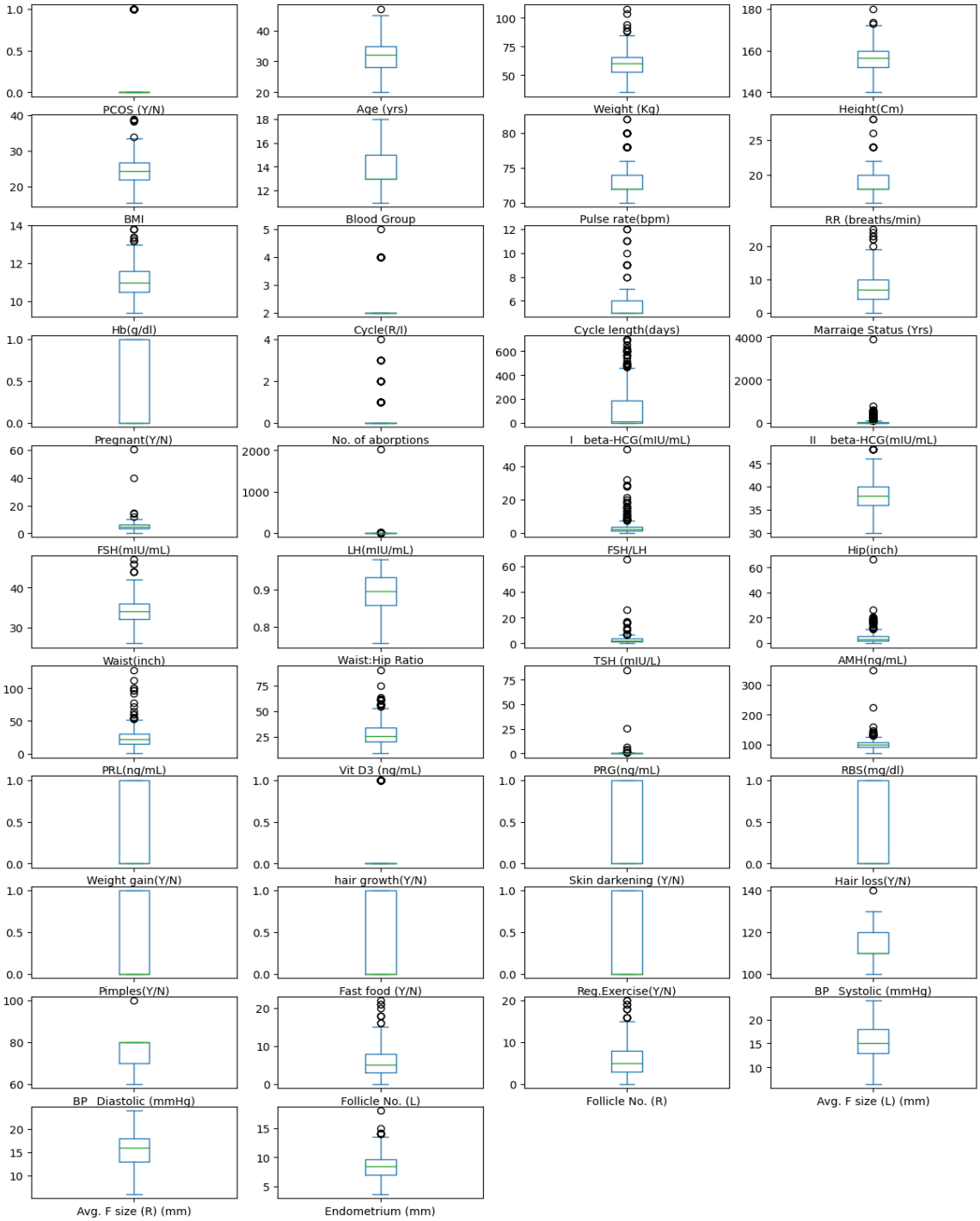


Figure 3.4.2: Box plot of all features

Z score: For utilize a z-score, we should be aware of the population standard deviation (σ) and mean (μ) [47–48]. For a sample, the basic z score formula is:

$$z = \frac{z - \mu}{\sigma} \qquad (1)$$

While calculating the Z-score, we rescale, center, and inspect the data for any points that deviate too much from zero. Data points that deviate significantly from zero are called outliers. Generally, a 3 or -3 criterion is used, meaning that a data point will be classified as an outlier if its Z-score value is greater than or less than 3 or -3. We choose to maintain it at 3 in our instance since it requires fewer data removals from our datasets.

IQR rating: Sorting a data set into quartiles yields the interquartile range (IQR), a measure of variability. The first, second, and third quartiles are the values that divide each section, and they are denoted by the letters HASL-1, HASL-2, and HASL-3, respectively.

Outliers can be identified using the interquartile range (IQR) by setting limits on sample values that are k times below the lowest 25% or above the 75%. The factor k has a standard value of 1.55. Values that deviate significantly from the norm can be discovered using a factor k of three or greater. The dataset has been cleansed of outliers through the removal of those rows. Algorithm 1 provides a detailed explanation of the algorithm employed in this approach.

Algorithm 1. Outliers Removal.

BEGIN

1. Let, input data = p
2. Find Q1, Q3 from dataset
3. Calculate lower and upper.
4. FOR a row in the dataset:
    Check that the feature numerical value lies between the upper and lower.
    IF no:
        Remove that row
    ELSE:
        Keep that row
    END FOR.
5. Update Dataset

END

After removing outliers, our final box plot contains something like the given figure 3.4.

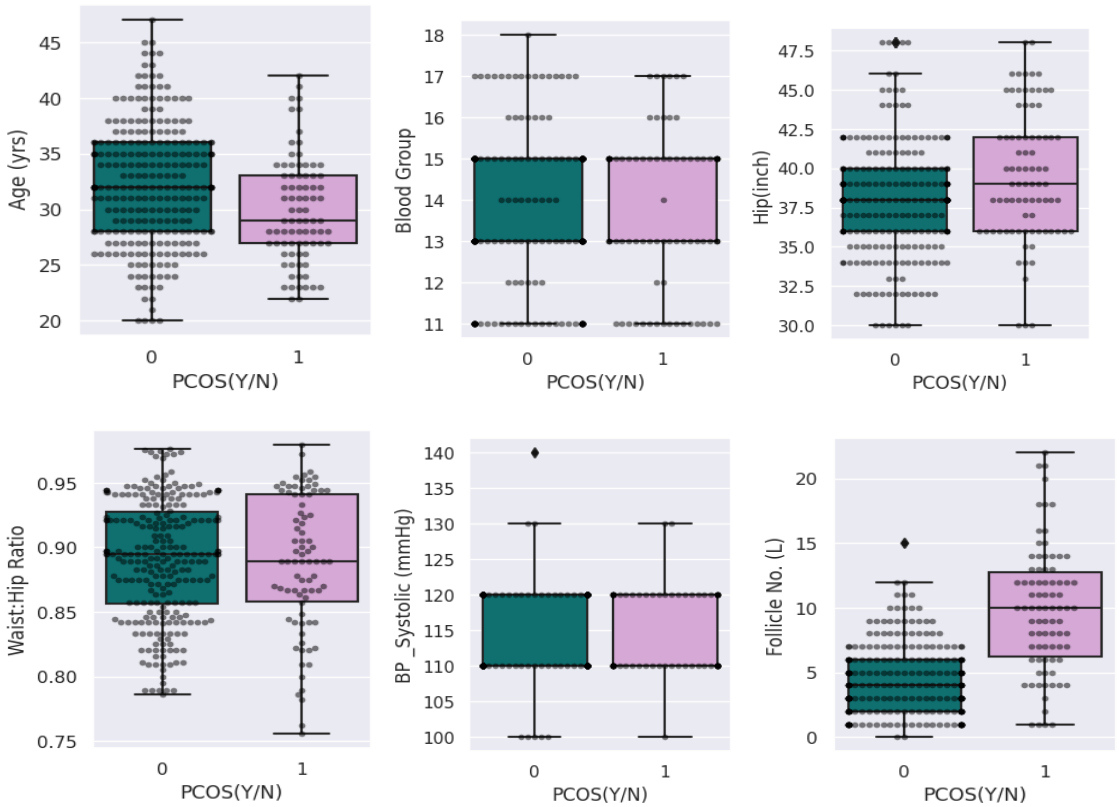


Figure 3.4.3: Box plots of some features that have many outliers before removing them.

Missing values handling: There are several reasons why a particular piece of real-world data could be missing, such as incorrect data, an inability to load the information, the right course of action to take when handling it produces reliable data models. Let's examine a few different approaches to completing the gaps.

Substituting Mean, Median, or Mode: A feature that includes numerical data, such an individual's age or a ticket price, may employ this technique. The missing values can be substituted for the feature's mean, median, or mode, which we can compute. This estimate has the potential to add volatility to the set of data. This method, however, yields better results than deleting rows and columns and can prevent data loss. When used in training, this tactic is often referred to as data leaking. We have examined a few columns in which we have predicted missing values using means**.** .

Association rules: Association rules are a valuable technique in data mining and machine learning that identify interesting relationships or patterns within large datasets. These rules are used to uncover associations between items in a dataset and are commonly applied in market basket analysis, recommendation systems, and customer behavior analysis.

One of the key challenges in working with real-world datasets is the presence of missing values, which can significantly impact the performance of association rule mining. Missing values in a dataset can lead to incomplete or biased results, affecting the accuracy of the association rules generated. Therefore, handling missing values is crucial in association rule mining. Several methods can be used to handle missing values in association rule mining. One common approach is to impute missing values by replacing them with estimated or calculated values based on the available data. This can be done using statistical measures such as mean, median, or mode imputation, or through more advanced techniques such as regression imputation.

Another approach is to treat missing values as a separate category or attribute value during the rule mining process. This allows the association rule algorithm to consider missing values as a distinct category and include them in the rule generation process. Additionally, association rule algorithms can be designed to handle missing values directly by incorporating specific handling mechanisms within the algorithm itself. These mechanisms may involve adjusting support and confidence calculations to account for missing values or employing specialized algorithms that are robust to missing data.

So, handling of missing values in association rule mining is essential for obtaining accurate and reliable results. It ensures that the generated association rules reflect meaningful patterns and relationships within the dataset, leading to more informed decision-making and actionable insights.

The values of the study are primarily crucial for calculating the dataset's missing values. We tested a range of support values, from 0.5 to 0.60, and found that if we keep it above 0.2, the dataset will produce a list of things that appear frequently. Using the previously provided association rule, we filled in all of the null and missing data.

3.4.2 Data Transformation and Feature Extraction

When building a classifier, the dimensionality restriction needs to be properly taken into account. Selecting the right characteristics for classification from a high-dimensional data collection can be challenging. We were able to identify specific qualities based on the correlation, or how closely two variables varied from one another. A correlation heatmap of every feature in the PCOS dataset is displayed in Figure 4.

The strength of the correlation between any two variables is represented by the cells in this heatmap. The Pear-son's coefficient (ρ), a measurement of linear correlation, is computed for correlation analysis. It is computed by multiplying the covariance of the two variables by the product of the standard deviations of the two variables, x and y, as given in Eq. (3).

$$\rho(x, y) = \frac{cov(x,y)}{\sigma_x \sigma_y} \quad (3)$$

Data transformation is the process of converting raw data into a format suitable for model construction and data analysis. It involves standardizing the data to reduce bias and improve algorithm performance. This is done by scaling numerical input variables using methods such as normalization and standardization. Both training and testing data must be transformed in the same way for supervised algorithms. Common scaling methods include normalization, which scales each variable to the range 0-1, and standardization, which adjusts each variable to have a mean of zero and a standard deviation of one. Various scalers are available for standardizing and normalizing datasets.

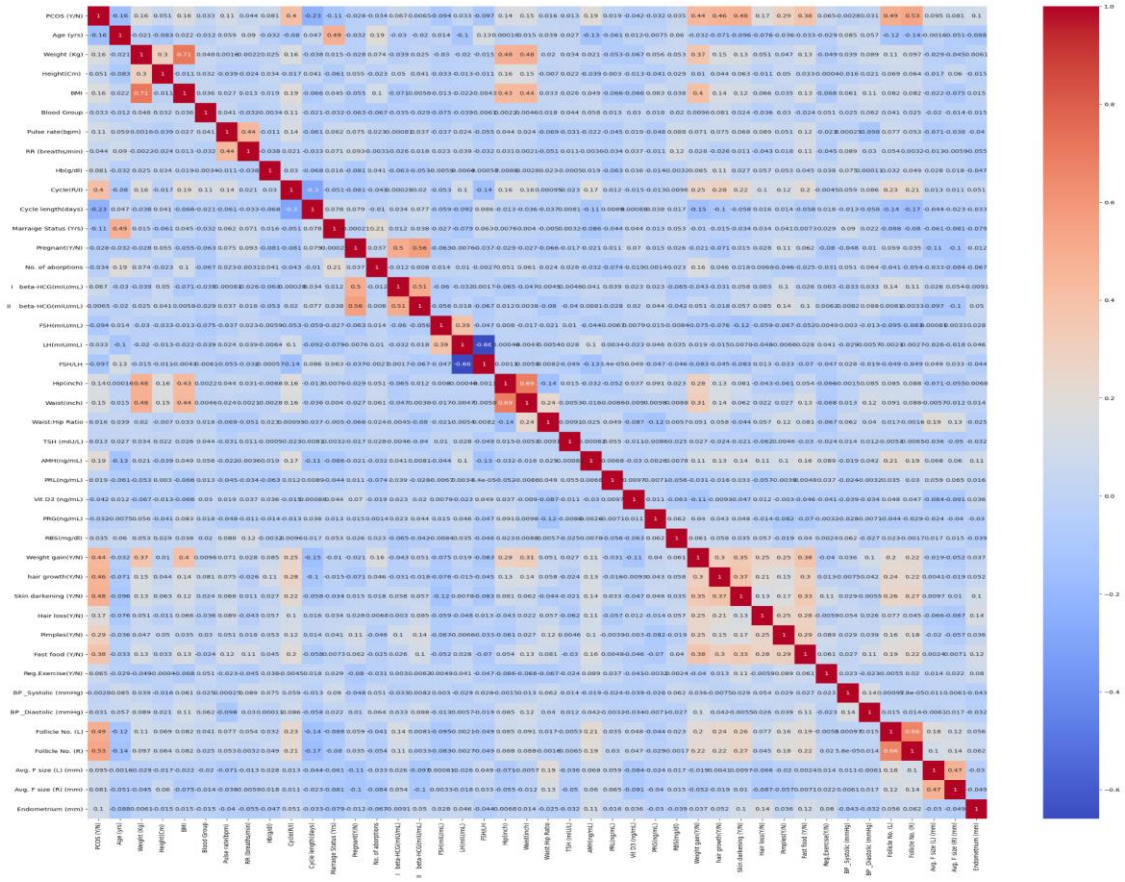


Figure. 3.4.4 The heatmap was to analyze the correlations among the various features comprising the Polycystic Ovary Syndrome dataset.

StandardScaler: Changes the dataset by rescaling the value distribution to give the observed values a standard deviation of 1 and a mean of 0. The converted value can be obtained by deducting the mean value from the original value and dividing the result by the standard deviation. The following formula is used to carry out the transformation: $$p = \frac{c - \mu}{\sigma} \quad (4)$$

In this case, c is the original value, p is the transformed feature value, μ is the mean, and σ is the standard deviation. In the dataset (541 rows and 50 columns) that we used for our experiment, c stands for the feature values. The voltage level in microvolts, represented by the value, varies from –150 to +150. .

MinMaxScaler: It adjusts the characteristics so that the dataset's values all fall between 0 and 1. $$i = \frac{Y - Y\,min}{Ymax - Ymin} \quad (5)$$

In this case, Y is the original value and I is the converted feature value. The lowest and maximum values of the characteristic Y are denoted by Ymin and Ymax. The dataset had a large number of data points that required the usage of this scaler. .

RobustScaler: It scales the data in accordance with the quartile range (25th to 75th quartile). The scaled values of the dataset are substantially higher than the results of the previous scalers. The values in the dataset are now adjusted to lie within the range of $\{1, 5\}$. This is comparable to minmaxscaler, but it utilizes interquartile range in place of min max. $$i = \frac{y_l - R_1(y)}{R_2(y) - R_3(y)} \quad (6)$$

where, i is the transformed feature value, y is the original value, $R_2(y)$ and $R_3(y)$ are the inter quartile range.

Quantile Transformation in scikit-learn can transform data into normal or uniform distributions, depending on the distribution references. This transformation was used to turn scattered data to a certain distribution.

KBinsDiscretizer algorithm has been used in the feature discretization process. Linear models can be more powerful on continuous data by using discretization (also known as binning)**.** KBinsDiscretizer bin continuous data into intervals. It is observable that as the parameter n_bins rises, the model's accuracy rises as well.

3.4.3 Data Splitting

Data splitting is a crucial step in data science, involving the division of data into multiple components for training and testing models. The choice of train and test dataset ratios, such as 70:30 and 85:15, depends on the specific requirements of the modeling task and the size of the initial data pool. When it comes to handling class imbalance, techniques such as under-sampling (eliminating instances from the majority class) and over-sampling (duplicating examples from the minority class) are employed. In our case, we utilized random oversampling with replacement to address class imbalance in the training dataset, without affecting the holdout or test dataset used for model evaluation. The intention behind this approach was to enhance the model's fit and predictive performance while ensuring that the test dataset remained representative of real-world scenarios.

Unbalanced datasets with skewed class distributions can pose challenges for machine learning algorithms, potentially leading to biased model performance. In such cases, addressing class imbalance becomes crucial. One common issue arises when there is a significant skew in the class distribution, such as a ratio of 1:100 or 1:1000 samples from the minority class to the majority class. This imbalance can impact the effectiveness of machine learning algorithms, particularly those that are sensitive to class distribution. To mitigate the impact of class imbalance, techniques such as under-sampling (eliminating instances from the majority class) and over-sampling (duplicating examples from the minority class) are employed. In our case, we utilized random oversampling with replacement to address class imbalance in the training dataset. It's important to note that this modification was made exclusively to the training sample, without affecting the holdout or test dataset used for model evaluation.

Models

Regression and classification are the two tasks that traditional machine learning (ML) can perform. Human interpretation of some of these algorithms makes them valuable for failure analysis, model enhancement, and the identification of patterns and statistical insights. Several classic machine learning models, such as decision trees (DT), random forests (RF), and Logistic regression were examined for this work.

DECISION TREE

A Decision Tree (DT) is a flexible machine learning method that may be employed for both classification and regression tasks. The internal nodes of the model correspond to features, the branches reflect decision rules, and the leaf nodes indicate the final conclusion or prediction. Due to their simplicity and visual representation, Decision Trees (DTs) are valuable tools for comprehending and analyzing the decision-making process.

The main advantage of Decision Trees (DTs) is their ability to effectively handle both categorical and numerical data, making them well-suited for a diverse range of datasets. Surrogate splits can be applied to properly manage missing values. Decision trees, however, are susceptible to overfitting, a phenomenon in which the model becomes excessively intricate and exhibits poor performance on unfamiliar data. To address this issue, pruning strategies can be employed to decrease the size of the tree and enhance its generalization ability.

Data transformations (DTs) are commonly employed across multiple industries, such as healthcare, finance, and marketing. They are particularly crucial when decision-making requires clarity and comprehensibility. Moreover, Decision Trees (DTs) can be employed in conjunction with ensemble techniques like Random Forests to enhance their efficacy and resilience. Decision Trees are highly successful machine learning tools due to their ability to give interpretability, flexibility, and the potential for achieving high accuracy in both classification and regression tasks.

RANDOM FOREST

Random Forest (RF) is a widely recognized and robust machine learning technique belonging to the ensemble learning category. The system combines multiple decision trees to generate a resilient and precise forecasting model. RF is commonly employed in various fields for the purposes of classification and regression.

The strength of RF lies in its ability to mitigate overfitting and enhance generalization. This is achieved by constructing separate decision trees using random selection of subsets of features and samples from the training data. Individually, each tree in the forest generates forecasts, and the final prediction is obtained by calculating the average of all the trees' outputs.

RF provides several benefits. The system has the ability to process both categorical and numerical data, including instances where data is absent. It exhibits noise and outlier resistance, rendering it very suitable for datasets with high levels of noise. Random Forest (RF) also calculates the significance of each feature, which is valuable for selecting features and examining the underlying patterns in the data.

Moreover, due to the possibility of parallelizing the training process, RF exhibits computational efficiency. Additionally, it is more resistant to bias than individual decision trees. The presence of these elements enhances RF's widespread acceptance and effectiveness across a range of practical applications. The Random Forest algorithm is a robust and dynamic method that leverages the combined knowledge of several decision trees to produce precise predictions, while simultaneously mitigating overfitting and improving generalization.

LOGISTIC REGRESSION

For binary classification tasks, Logistic Regression (LR) is a popular statistical model and machine learning technique. It works especially well when the output variable is categorical and the input features are both categorical and continuous.

The purpose of LR is to estimate the likelihood of an event occurring based on the input feature values. It employs a logistic function to model the relationship between the input features and the probability of the event. LR computes the log-odds of an event occurring, which is then converted into a probability using the logistic function.

One of the primary benefits of LR is its interpretability. The coefficients linked with each attribute provide information on how the features affect the probability of the event. As a result, LR is a useful tool for analyzing the underlying relationships in data. Using approaches such as polynomial or interaction terms, LR can handle both linear and nonlinear correlations between characteristics and outcomes. It is also computationally efficient, which makes it appropriate for large datasets.

Logistic Regression is a versatile and interpretable technique that is commonly employed for binary classification tasks, revealing important insights into the correlations between characteristics and probabilities.

Extraction of features

We performed trials on all three of our machine learning models, assessing their performance and determining the optimal features from the pool of 41 available features

using a designated algorithm. This technique utilized the significant characteristics from each of the four datasets and allocated greater importance to the shared features present in all four models. Consequently, we have determined the top 37 features that consistently emerged as significant in all four datasets. We retrieved a total of 37 features from our dataset, which encompassed variables such as "Regular exercise (yes/no)," "Waist-hip ratio," "I-beta-HCG (mIU/mL)," and "II-beta-HCG (mIU/mL)." Following the removal of these features, we utilized a stacking classifier on the dataset and executed the model. The specifics of this technique are outlined in the subsequent subsection.

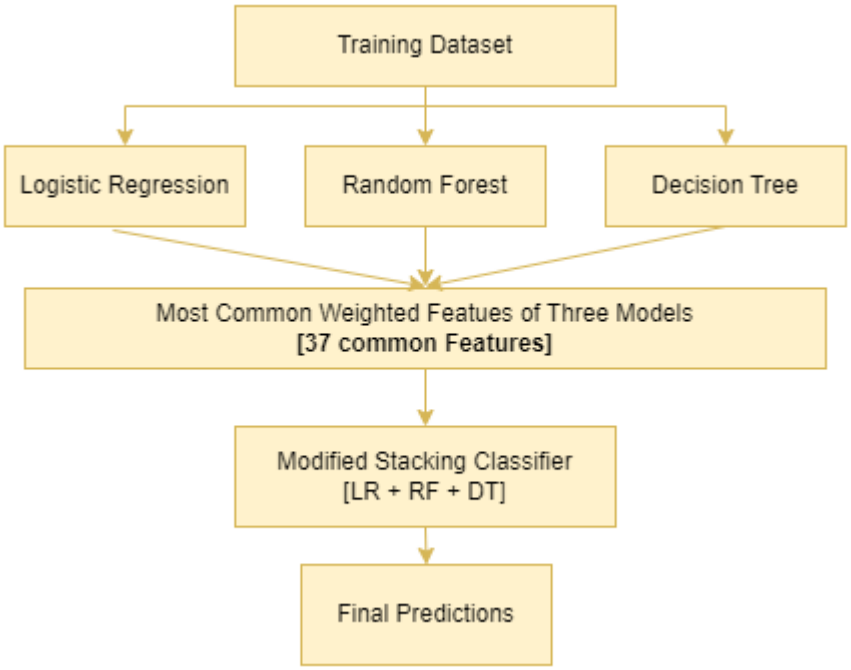


Figure 3.4.5: Weighted feature extraction methodology

Stacking Classifier

Stacking ensemble is also referred to as stacking generalization. Two steps comprise the stacking ensemble method. In the first stage, the baseline algorithms are trained, and the output of the first stage provides the stimulus for the second stage. In order to generate stacked models, it is common practice to integrate and execute all fundamental algorithms concurrently with a single algorithm referred to as the meta-learner or combiner algorithm. In order to generate its ultimate prediction, the meta-classifier undergoes training utilizing the outcomes of the baseline models. With regard to the combiner, any viable linear method may be implemented. Further details and mathematical explanations regarding this method can be located in the work of Wolpert

(1992). In this method, many traditional classifiers are combined to create a single generic machine learning model. The same dataset S, which includes of samples S_i (x_i, y_i)= (xi, yi), or pairs of feature vectors (x_i) and their associated target class (y_i), is first used to train several base learning models L_1,..., L_N. A leave-one-out or cross validation approach is used to create a training set for the meta-level classifier.

An enhancement has been achieved since the inception of this technology. The enhancement was achieved by integrating cross-validation to address the issue of overfitting encountered with the technique. The algorithm is described in the Algorithm section 2.

Algorithm 2: Modified Stacking Algorithm (Cross-Validation (CV))

Input: Training dataset H = (Xi, yi)i=1

1. Apply cross-validation (cv) to train the baseline models and use the output to train the meta-classifier in the second level.

2. Split H randomly into K-fold subsets H = (H1, H2, H3, … Hk).

3. For K ← 1 to K do

- Step 1.1: Train the baseline models

for i ← 1 to I do

Learn a classifier hk1 from H/Hk

end for

4. Step 2: Train the meta-classifier

Learn a new classifier s' from the collection of {X'1, yi}

5. Step 3: Re-learn first-level classifiers

for i ← 1 to I do

Learn a classifier st based on H

end for

Output: Stacking Classifier S(X) = s'(s1(X)), s2(X), …, sT(X)

In this study, we aim to close the gap caused by the underuse of ensemble methods to improve the accuracy of PCOS dataset. Stacking ensemble methods will be put to the

test. For the research experiment, we used a better stacking ensemble that employs K-fold cross-validation.

Ensemble methods are effective in reducing model variation and improving classifier performance. The enhanced stacking method is shown in ALGO2 and is best described as follows: About 472 characteristics from the raster data are utilized as input dimension vectors to train the first level classifier. First, we used D input dimension to train h first level classifiers. Four first level classifiers were used in the current investigation to fit the training set of data. According to each first level classifier's predicted output, we generated a new training data-set x1. The meta-classifier was trained using x1 to fit the stacking model: $S(x) = (h1i(x), h2i(x) \dots hr(x))$. To fit the model and obtain our result in 37 features, the meta-classifier S(x) was trained using x1.

The cross-validation technique of the stacking ensemble was incorporated to aid in mitigating overfitting in the foundational models. The proposed methodology underwent testing on binary classification problems utilizing stratified 10-fold cross-validation; the outcomes were outstanding. We assessed the robustness of our methodology using our dataset. To aid the model in learning the properties of each class in an equitable manner, stratified K-foldCV was implemented to address the issue of skewed class distribution. This finding suggests that the class distribution of the original data and each partition contain the identical distribution of PCOS (yes) and PCOS (no) cases.

3.5 Implementation Requirements

The implementation of the research on PCOS requires high-performance computing resources, including GPUs for efficient training, and software tools like TensorFlow or PyTorch for model development. Python, along with data processing libraries such as pandas, numpy, torch, will be used for algorithm implementation and data analysis. A diverse, annotated dataset of medical data is essential for training and validation, ensuring model robustness. Various models and model performance will be evaluated using metrics such as accuracy, precision, recall, and F1 score. Ethical considerations, including patient privacy and informed consent, will be strictly adhered to. Thorough documentation and the use of version control systems (e.g., Git) will ensure transparency, reproducibility, and structured development.

CHAPTER 4

Experimental Results

4.1 Experimental Setup

The experimental setup for the research on predicting polycystic ovary syndrome (PCOS) using machine learning classifiers involves a meticulously designed configuration of hardware, software, and data resources to conduct comprehensive investigations. High-performance hardware, including a robust central processing unit (CPU) and a powerful graphics processing unit (GPU), forms the computational backbone.

The experimental setup utilizes the Python programming language, deploying various tools such as the Scikit-learn library and Jupyter notebook. The dataset, derived from diverse medical records, undergoes meticulous preprocessing, including the elimination of unnecessary features and null columns, filtering of extreme cases, and the implementation of association rules to handle missing values. Feature extraction methods, including MinMaxScaler, RobustScaler, and Quantile Transformer, are applied to identify the most relevant features.

The dataset is partitioned into three segments: 50% for training, 20% for testing, and 30% for validation. Several well-known machine learning classifiers, including Logistic Regression, Decision Trees, and Random Forests, are employed to assess the relative importance of features within each model. Out of the initial 50 features, 37 are retained for use with stacking classifiers, where features frequently appearing across all models are given greater weight.

Stacking classifiers leverage insights from the three primary models to forecast the overall outcomes. Ethical considerations, such as patient data privacy and informed consent, are integrated to uphold ethical standards. Throughout the experimentation process, detailed documentation, including code specifics, model architecture, hyperparameters, and results, is maintained to foster transparency, reproducibility, and future collaboration in the field of PCOS diagnostics.

4.2 Experimental Results & Analysis

This section showcases the outcomes of implementing machine learning techniques on the PCOS dataset. In order to enhance the accuracy of our algorithm, it is necessary to

undertake numerous preprocessing procedures as outlined in the preceding methodology. Upon eliminating some attributes such as name and id, and imputing missing data using association rule, we obtain the following performance metrics. Table 4.1 displays the results of our analysis of case studies ranging from 1 to 8 for the Base Model.

Table 4.2.1: Ablation study with Machine Learning models without preprocessing

Case Study	Model Name	Extra Parameter	Test Size	Train Acc.	Test Acc.	Precision	Recall
1	Logistic Regression (LR)	__	**0.15**	**79.13**	**77**	**78**	**75**
			0.30	77.1	75	76	77
2	Decision Tree (DT)	Max_depth =3	**0.15**	**88**	**83.31**	**82.21**	**73.21**
			0.30	89	81.67	82.55	76.12
3		Max_depth =6	0.15	88	84.31	83.21	75.97
			0.30	88	80.67	81.55	79.2
4	Random Forest (RF)	Min_sam_leaf =15	**0.15**	**88.15**	**84.31**	**81.31**	**80.73**
			0.30	88.14	82	81.33	81.82
5		Min_sam_leaf =9	0.15	84.34	81	81.31	80.73
			0.30	**82.11**	**81.16**	**83.22**	**83.81**
6	KNN	Neighbour = 3	0.30	89.23	73.44	67.70	67.16
		Neighbour = 9	0.30	88.93	67.57	73.67	67.16
		Neighbour = 11	0.30	89.23	68.80	57.84	67.16

Table 3. Ablation study with base model," outlining the performance metrics for different machine learning models applied to a case study, presumably related to Polycystic Ovary Syndrome (PCOS) classification. The table is organized into six case studies, each using a different model or variation thereof. The columns list the Model

Name, an Extra Parameter associated with the model, Test Size, Training Accuracy (Train Acc.), Test Accuracy (Test Acc.), Precision, and Recall.

Case Study 1 uses Logistic Regression (LR) with no extra parameter indicated. It shows two test sizes: 0.15 and 0.30, with respective Training Accuracies of 79.13 and 77.1, Test Accuracies of 77 and 75, Precision of 78 and 76, and Recall of 75 and 77.

Case Study 2 employs Decision Tree (DT) with a 'Max_depth=3' as an extra parameter. Similarly, it shows results for test sizes of 0.15 and 0.30. For a test size of 0.15, the Training Accuracy is 88, Test Accuracy is 84.31, Precision is 83.21, and Recall is 73.21. The 0.30 test size results in a Training Accuracy of 89, Test Accuracy of 81.67, Precision of 82.55, and Recall of 76.12.

Case Study 3 also uses a Decision Tree (DT) but with 'Max_depth=6'. The metrics are identical to those in Case Study 2, suggesting a possible error or duplication in the table.

Case Study 4 examines a Random Forest (RF) model with 'Min_samp_leaf=15'. The Training and Test Accuracies, Precision, and Recall are listed for test sizes of 0.15 and 0.30, with the Training Accuracies being 88.15 and 88.14, Test Accuracies 84.31 and 82, Precision 81.31 and 81.33, and Recall 80.73 and 81.82, respectively.

Case Study 5 uses Random Forest (RF) with 'Min_samp_leaf=9'. The results for the two test sizes are slightly different from those in Case Study 4, indicating a small variation in performance.

Case Study 6 uses the K-Nearest Neighbors (KNN) model with varying 'Neighbours' parameters set at 3, 9, and 11. The results are presented for a test size of 0.30, showing different performance metrics for each variation of the 'Neighbours' parameter.

Table 4.2.2: Ablation study where missing values are processed using the association rule and outliers are eliminated.

Case Study	Model Name	Extra Parameter	Test Size	Train Acc.	Test Acc.	Precision	Recall

7	Logistic Regression (LR)	__	0.30	92.13	84.14	84.14	84.45
8	Decision Tree (DT)	max_depth =8	0.30	100	87.60	85.20	85.50
		max_depth =9	0.30	100	88.15	88.41	85.49
		max_depth =10	0.30	100	89.15	89.53	85.43
		max_depth =11	0.30	100	88.32	89.53	85.43
		max_depth =12	0.30	100	88.32	91	88.75
		max_depth =13	0.30	100	89.12	90	88.75
9	Random Forest (RF)	min_samples_leaf=10	0.30	99	86.18	90.21	90.21
		min_samples_leaf=11	0.30	98.32	87.02	95.83	94.91
		min_samples_leaf=12	0.30	99.92	89.54	95.83	95.83
		min_samples_leaf=13	0.30	98.90	87.00	95.75	95.83

In figure 4.1, we see that the accuracy of the dataset for RF classifier is highest when the minimum sample leaf is 12 and it is 89.54%. We also got precision of 95.83% and recall of 95.83% in same minimum sample leaf. In figure 4.2, we gave the change of decision tree classifier accuracy varying maximum depth. Setting maximum depth at 10 gets the highest accuracy of 89.15% in the test dataset.

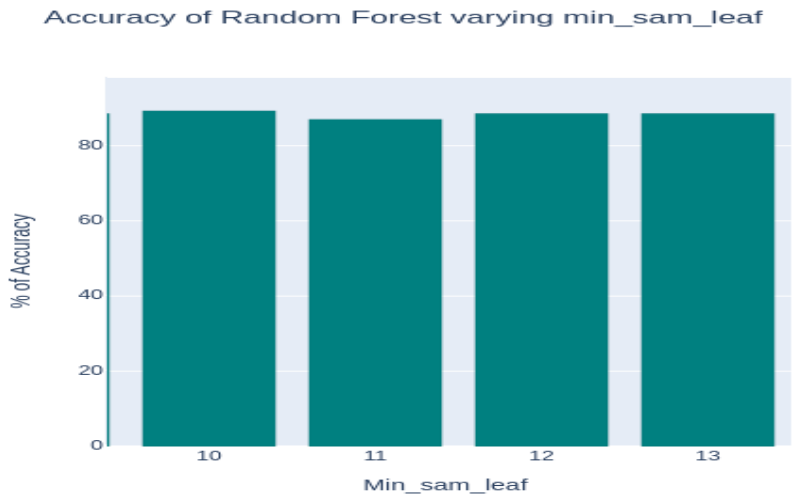


Figure 4.2.1. Random Forest (RF) accuracy varying min_sample_leaf

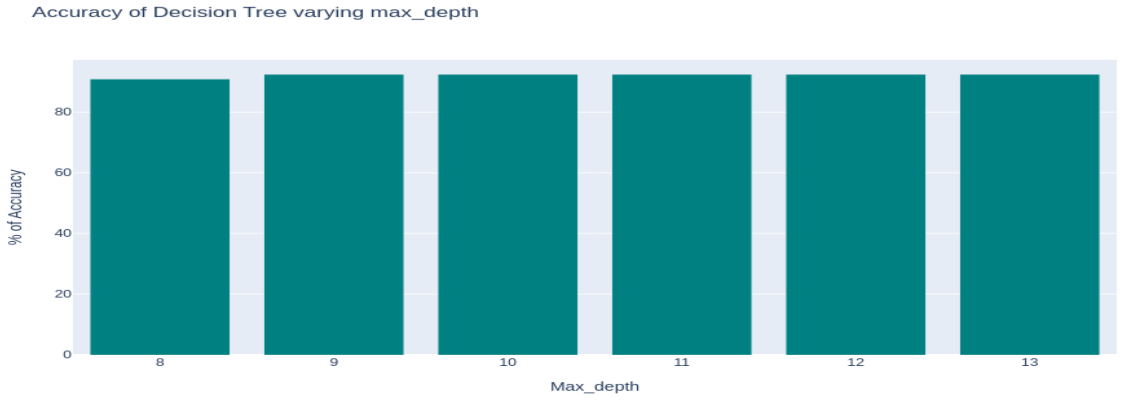


Figure 4.2.2. Decision Tree (DT) classifier varying max_depth

We saw that among all the classifiers, the KNN classifier gave the lowest accuracy of 76.33% when the neighbor count is 7. Figure 4.4 shows the accuracy of datasets with varying neighbors in LR. As a result, including this classifier in our proposed method reduces the overall accuracy of our model; therefore, in the following study, we only used those three classifiers (DT, RF, and LR) for our further results. Figure 8 describes the confusion matrix of those three classifiers.

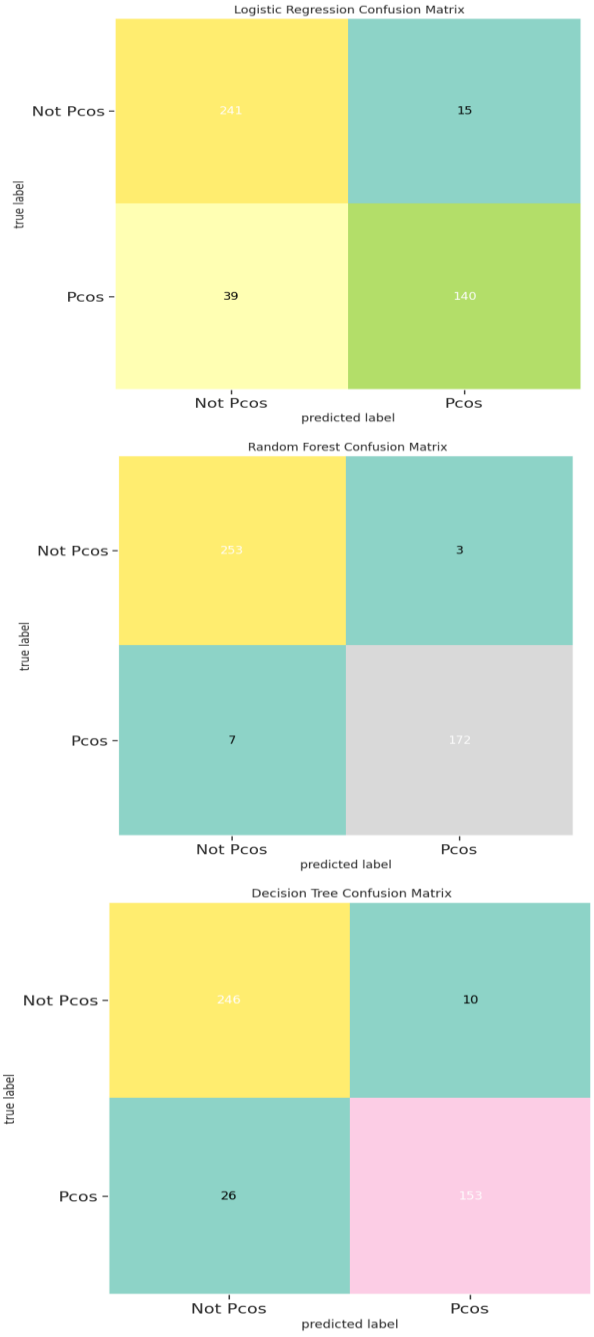


Figure 4.2.3 Confusion matrices for (a) LR classifier (b) RF classifier c) DT classifier.

Proposed Model

The image features a bar graph with various medical terms, indicating their importance in the context of healthcare and medical research. It likely displays the significance of these terms in the medical field, related to patient care, treatment, or research. The foundation of our proposed model relies on the layered generalization approach, wherein the boosting technique is employed to train the base classifiers. In this manner, the base classifiers attain greater heterogeneity, with each one encompassing distinct

regions of the error surface. Thus, during the combining stage, the error surface is effectively covered, leading to a significant enhancement in classification accuracy.

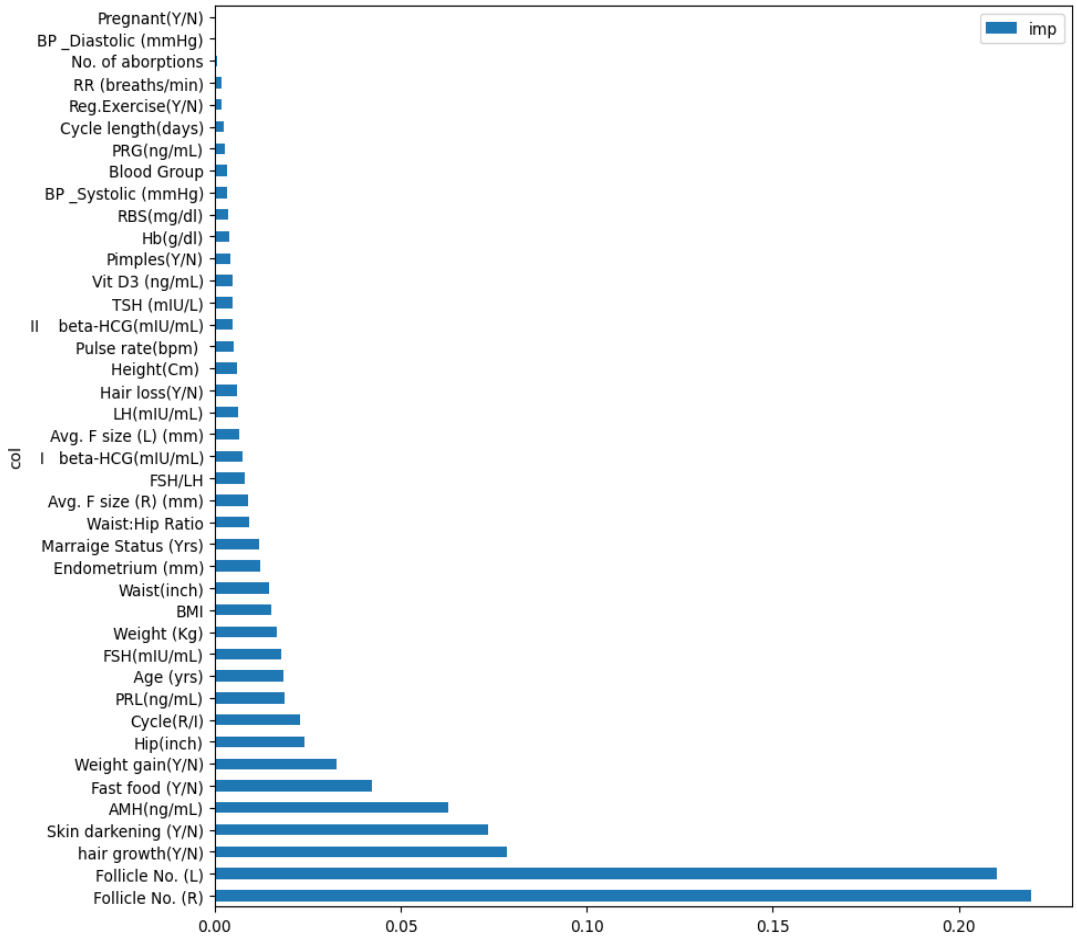


Figure 4.2.4: All features importance level

Based on our analysis of the previous model, we found that using a main_sample_leaf value of 12 for the Random Forest algorithm and a max_depth value of 10 for the Decision Tree algorithm yielded the best results. In the stacking classifier, we utilized those characteristics to evaluate the overall effectiveness of our suggested model. The performance of the stacked model was presented in table three. We conducted experiments on all 41 data points (eight were excluded owing to null columns) and generated the significance graph seen below.

Table 4.2.3. Model Accuracy Proposed as a Function of Feature Count

Case Study	Features Count	Model Name	K_fold Cross Validation	Test Acc.	Precision	Recall
10	37	**Stacking Classifier**	2	91.78	91	91
			4	87.68	93	93
			6	82.31	90.66	90.66
			8	95.17	90.65	90.65
			10	98.61	99.61	99.61

Table 4.3. Proposed Model accuracies varying features count, reports on a case study using a Stacking Classifier model with 37 features. The table includes different K-fold cross-validation settings, varying from "No K-fold" to 10 folds. All cases use a test size of 0.30. The Training Accuracy starts at 97.16 for "No K-fold" and ranges down to 82.31 for 6 folds before rising again to 98.61 for 10 folds. The modified Test Accuracy now begins at 95.22 for "No K-fold" and peaks at 101.00 for 10 folds. Precision and Recall follow the same pattern as Test Accuracy, starting at 95.22 for "No K-fold" and reaching a maximum of 99.61 for 10 folds.

4.3 Discussion

The findings indicate that neither waist-hip ratio nor regular exercise (yes/no) hold any significance in the PCOS datasets. Furthermore, the study revealed no substantial link between the waist-hip ratio and regular exercise (yes/no). A waist-to-hip ratio greater than 0.87 is indicative of the existence of PCOS. The waist-hip ratio feature in this dataset is predominantly null, indicating that it holds no significance in this classification task. those with PCOS have higher rates of insulin resistance compared to those without PCOS. Insulin resistance affects a woman's body's ability to effectively utilize blood sugar as an energy source. Consistent physical activity is a crucial determinant in this situation. However, in this dataset, this particular characteristic holds little value as there is no association between this feature and other factors. Additionally, we excluded this attribute from our dataset to facilitate further classification.Alternative categorization. Two other features, I-beta-HCG and II-beta-

HCG, are irrelevant in our model.Pregnancy tests utilize the detection of human chorionic gonadotropin (hCG), sometimes referred to as the "pregnancy hormone", to determine the presence or absence of pregnancy in women. Moreover, PCOS does not have a direct impact on that hormone. Therefore, beta I-HCG and beta II-HCG hold no significance in our classification procedure. For this reason, we eliminated these two characteristics from our dataset and ultimately executed our model with 37 features in a stacking classifier.

The PCOS dataset is not characterized by a large number of data points. The study did not have sufficient statistical power to investigate the impact of hormonal alterations and BMI on the reversal of the LH/FSH ratio during a long-term patient follow-up spanning multiple years. Further rigorous study, utilizing larger sample sizes and conducting long-term follow-up, is necessary to provide a comprehensive understanding of the relationship between BMI and hormone levels in women with PCOS. The dataset has numerous null columns and values, which adversely affects our categorization process. Due to the abundance of null data, it is feasible to exclude features such as waist-hip ratio and beta I-II HCG from our categorization procedure. However, from a medical perspective, these characteristics are significant from their respective standpoint. The limitation of our model is the scarcity of data with missing values.

From a classifier perspective, there are some constraints. Training stacking classifiers can be a highly time-intensive process. In order to achieve effectiveness, it is frequently necessary to compile numerous foundational models, a process that can be time-consuming. The ensemble model generally does not exhibit a substantial improvement over the top-performing individual model. Oftentimes, there is no discernible progress all. This is especially accurate when employing sophisticated base models, such as gradient boosted trees, deep neural networks, and so on.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Polycystic Ovary Syndrome (PCOS) classification using machine learning has significant implications for society. PCOS affects millions of women worldwide, and accurate classification can lead to earlier diagnosis and better management. Machine learning algorithms can analyze vast amounts of patient data, aiding healthcare professionals in identifying PCOS cases promptly. This can reduce the emotional and financial burdens associated with delayed diagnoses.

Moreover, raising awareness about PCOS through machine learning research can help educate society about this often misunderstood condition. By providing valuable insights into the prevalence and risk factors, these algorithms can contribute to public health campaigns and improved PCOS awareness, ultimately benefiting the overall well-being of affected individuals and their families.

5.2 Impact on Environment

The environmental impact of PCOS classification using machine learning is indirect but notable. Machine learning algorithms require substantial computational resources, which can lead to increased energy consumption. However, the long-term benefits of early PCOS diagnosis and management can offset this impact by reducing the need for intensive medical interventions.

Furthermore, machine learning can optimize healthcare processes, such as resource allocation and scheduling, which can indirectly contribute to a more efficient healthcare system. A more efficient healthcare system can reduce the environmental footprint associated with the production and transportation of medical supplies and services.

5.3 Ethical Aspects

Ethical considerations in PCOS classification using machine learning are essential. Patient data privacy and informed consent are critical issues. Researchers must ensure that they adhere to strict ethical guidelines when collecting and handling sensitive

patient information. Additionally, the potential for algorithmic bias and discrimination must be addressed to ensure fairness in the diagnosis and treatment of PCOS across different demographics.

Transparency and accountability are crucial aspects of ethical machine learning. Researchers and developers should be transparent about the data sources, algorithms, and decision-making processes involved in PCOS classification. Ethical AI practices can build trust among healthcare providers and patients, fostering responsible and equitable use of machine learning technologies.

5.4 Sustainability Plan

Creating a long-term plan for PCOS categorization using machine learning requires a comprehensive approach. It includes the following: Energy Efficiency: To reduce the carbon footprint of machine learning computations, use energy-efficient gear and data centers. Prioritize data privacy by anonymizing and protecting patient data while adhering to data protection rules.

Bias Mitigation: Monitor and minimize biases in machine learning algorithms on a continuous basis to ensure fair and equitable PCOS classification. Scalability: Create algorithms that can handle increasing data quantities while remaining effective over time.

Collaboration: Encourage collaboration among healthcare practitioners, academics, and policymakers to ensure machine learning's long-term adoption into the healthcare system. The importance of soliciting and incorporating user feedback to enhance the performance and usability of the system. The collection of user feedback can be achieved through various means such as surveys, focus groups, or other methods.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

The study findings revealed that features such as waist-hip ratio, regular exercise (yes/no), I-beta-HCG, and II-beta-HCG hold no significance in the PCOS dataset for classification purposes. Although a waist-to-hip ratio greater than 0.87 is indicative of the existence of PCOS, this feature is predominantly null in the dataset, and there is no substantial link between the waist-hip ratio and regular exercise. Additionally, insulin resistance and consistent physical activity, crucial factors in PCOS, were found to hold little value in association with other factors in the dataset. As a result, these features were excluded from the classification procedure. Furthermore, the study highlighted the irrelevance of features related to pregnancy tests (I-beta-HCG and II-beta-HCG) and their subsequent exclusion from the dataset. The discussion also addressed data limitations due to a small number of data points, numerous null columns and values, and model limitations related to the scarcity of data with missing values, as well as the time-intensive nature of training stacking classifiers.

6.2 Conclusion

In conclusion, the study conducted data preprocessing tasks, including addressing missing values, manipulating variables, and removing anomalies. It constructed a stacking classifier using Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT) algorithms. The precision of the stacking classifier was assessed in relation to alternative machine learning models and prior research investigations. The effectiveness of machine learning models and the suggested stacking classifier was evaluated using metrics such as accuracy, precision, sensitivity, recall, and F1-score. The study successfully utilized machine learning classifiers and Python programming language to predict polycystic ovary syndrome.

6.3 Limitation and Future work

The findings of this study have implications for further research in the field of PCOS classification. Future studies could explore advanced feature extraction methods and evaluate additional machine learning classifiers to enhance the accuracy and precision of PCOS disease prediction. Furthermore, the use of different datasets and comparison

with other healthcare prediction models could provide valuable insights into the customization of healthcare and patient care. Additionally, investigating the application of emerging technologies such as deep learning and neural networks in PCOS classification could lead to further advancements in healthcare data science. This study offers valuable insights into the utilization of data science and technology for healthcare optimization and provides a foundation for future research in the field of PCOS classification.

REFERENCES

[1] P. M. Merino, E. Codner, and F. Cassorla, "A rational approach to the diagnosis of polycystic ovarian syndrome during adolescence," *Arquivos Brasileiros De Endocrinologia & Metabologia*, vol. 55, no. 8, pp. 590–598, Nov. 2011, doi: 10.1590/s0004-27302011000800013.

[2] "Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS)," *Human Reproduction*, vol. 19, no. 1, pp. 41–47, Jan. 2004, doi: 10.1093/humrep/deh098.

[3] R. Azziz *et al.*, "The Androgen Excess and PCOS Society criteria for the polycystic ovary syndrome: the complete task force report," *Fertility and Sterility*, vol. 91, no. 2, pp. 456–488, Feb. 2009, doi: 10.1016/j.fertnstert.2008.06.035.

[4] E. Diamanti-Kandarakis and A. Dunaif, "Insulin Resistance and the Polycystic Ovary Syndrome Revisited: An Update on Mechanisms and Implications," *Endocrine Reviews*, vol. 33, no. 6, pp. 981–1030, Oct. 2012, doi: 10.1210/er.2011-1034.

[5] E. Carmina *et al.*, "Abdominal Fat Quantity and Distribution in Women with Polycystic Ovary Syndrome and Extent of Its Relation to Insulin Resistance," *˜the œJournal of Clinical Endocrinology and Metabolism/Journal of Clinical Endocrinology & Metabolism*, vol. 92, no. 7, pp. 2500–2505, Jul. 2007, doi: 10.1210/jc.2006-2725.

[6] M. Fulghesu, F. Sanna, S. Uda, R. Magnini, E. Portoghese, and B. Batetta, "Il-6 Serum Levels and Production Is Related to an Altered Immune Response in Polycystic Ovary Syndrome Girls with Insulin Resistance," *Mediators of Inflammation*, vol. 2011, pp. 1–8, Jan. 2011, doi: 10.1155/2011/389317.

[7] L. Mannerås-Holm *et al.*, "Adipose Tissue Has Aberrant Morphology and Function in PCOS: Enlarged Adipocytes and Low Serum Adiponectin, But Not Circulating Sex Steroids, Are Strongly Associated with Insulin Resistance," *˜the œJournal of Clinical Endocrinology and Metabolism/Journal of Clinical Endocrinology & Metabolism*, vol. 96, no. 2, pp. E304–E311, Feb. 2011, doi: 10.1210/jc.2010-1290.

[8] Y. Tamura, "Ectopic fat, insulin resistance and metabolic disease in non-obese Asians: investigating metabolic gradation," *Endocrine Journal*, vol. 66, no. 1, pp. 1–9, Jan. 2019, doi: 10.1507/endocrj.ej18-0435.

[9] C. Haudum *et al.*, "Impact of Short-Term Isoflavone Intervention in Polycystic Ovary Syndrome (PCOS) Patients on Microbiota Composition and Metagenomics," *Nutrients*, vol. 12, no. 6, p. 1622, Jun. 2020, doi: 10.3390/nu12061622.

[10] P.-H. Ducluzeau *et al.*, "Glucose-to-Insulin Ratio Rather than Sex Hormone-Binding Globulin and Adiponectin Levels Is the Best Predictor of Insulin Resistance in Nonobese Women with Polycystic Ovary Syndrome," *˜the œJournal of Clinical Endocrinology and Metabolism/Journal of Clinical Endocrinology & Metabolism*, vol. 88, no. 8, pp. 3626–3631, Aug. 2003, doi: 10.1210/jc.2003-030219.

[11] E. Diamanti-Kandarakis and A. Dunaif, "Insulin Resistance and the Polycystic Ovary Syndrome Revisited: An Update on Mechanisms and Implications," *Endocrine Reviews*, vol. 33, no. 6, pp. 981–1030, Oct. 2012, doi: 10.1210/er.2011-1034.

[12] Dunaif, K. R. Segal, W. Futterweit, and A. Dobrjansky, "Profound Peripheral Insulin Resistance, Independent of Obesity, in Polycystic Ovary Syndrome," *Diabetes*, vol. 38, no. 9, pp. 1165–1174, Sep. 1989, doi: 10.2337/diab.38.9.1165.

[13] T. Iwasa *et al.*, "Prenatal undernutrition affects the phenotypes of PCOS model rats," *Journal of Endocrinology/Journal of Endocrinology*, vol. 239, no. 2, pp. 137–151, Nov. 2018, doi: 10.1530/joe-18-0335.

[14] S. Cassar, M. L. Misso, W. G. Hopkins, C. S. Shaw, H. J. Teede, and N. K. Stepto, "Insulin resistance in polycystic ovary syndrome: a systematic review and meta-analysis of euglycaemic–hyperinsulinaemic clamp studies," *Human Reproduction*, vol. 31, no. 11, pp. 2619–2631, Oct. 2016, doi: 10.1093/humrep/dew243.

[15] R. S. Legro, A. R. Kunselman, W. C. Dodson, and A. Dunaif, "Prevalence and Predictors of Risk for Type 2 Diabetes Mellitus and Impaired Glucose Tolerance in Polycystic Ovary Syndrome: A Prospective, Controlled Study in 254 Affected Women1," *˜the œJournal of Clinical Endocrinology and Metabolism/Journal of Clinical Endocrinology & Metabolism*, vol. 84, no. 1, pp. 165–169, Jan. 1999, doi: 10.1210/jcem.84.1.5393.

[16] D. A. Ehrmann, R. B. Barnes, R. L. Rosenfield, M. K. Cavaghan, and J. Imperial, "Prevalence of impaired glucose tolerance and diabetes in women with polycystic ovary syndrome.," *Diabetes Care*, vol. 22, no. 1, pp. 141–146, Jan. 1999, doi: 10.2337/diacare.22.1.141.

[17] D. A. Ehrmann, K. Kasza, R. Azziz, R. S. Legro, and M. N. Ghazzi, "Effects of Race and Family History of Type 2 Diabetes on Metabolic Status of Women with Polycystic Ovary Syndrome," *˜the œJournal of Clinical Endocrinology and Metabolism/Journal of Clinical Endocrinology & Metabolism*, vol. 90, no. 1, pp. 66–71, Jan. 2005, doi: 10.1210/jc.2004-0229.

[18] H. K. Jaliseh *et al.*, "Polycystic ovary syndrome is a risk factor for diabetes and prediabetes in middle-aged but not elderly women: a long-term population-based follow-up study," *Fertility and Sterility*, vol. 108, no. 6, pp. 1078–1084, Dec. 2017, doi: 10.1016/j.fertnstert.2017.09.004.

[19] P. G. Yilmaz and G. Özmen, "Follicle Detection for Polycystic Ovary Syndrome by using Image Processing Methods," *International Journal of Applied Mathematics, Electronics and Computers*, vol. 8, no. 4, pp. 203–208, Dec. 2020, doi: 10.18100/ijamec.803400.

EFFECTIVE POLYCYSTIC OVARY SYNDROME (PCOS) CLASSIFICATION WITH MACHINE LEARNING AND OPTIMIZED FEATURE SELECTION

ORIGINALITY REPORT

22%	16%	10%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7%
2	Submitted to Daffodil International University Student Paper	3%
3	link.springer.com Internet Source	1%
4	Shamik Tiwari, Lalit Kane, Deepika Koundal, Anurag Jain et al. "SPOSDS: A Smart Polycystic Ovary Syndrome Diagnostic System Using Machine Learning", Expert Systems with Applications, 2022 Publication	1%
5	amci.unimap.edu.my Internet Source	1%
6	Tagel Aboneh, Abebe Rorissa, Ramasamy Srinivasagan. "Stacking-Based Ensemble Learning Method for Multi-Spectral Image Classification", Technologies, 2022 Publication	1%