Comparative Study of Machine Learning Classifiers for Accurate Polycystic Ovary Syndrome Detection

Mary M Dsouza, Tushar Choudhary, Saurav Kumar, Sejal D Vhankade, Neha

* Information Science and Engineering, Acharya Institute of Technology

Abstract

Polycystic ovary syndrome (PCOS) is a common hormonal disorder that affects women of reproductive age, necessitating early and precise detection for effective management and prevention of associated complications. Machine learning algorithms have exhibited significant potential in medical applications, particularly in disease diagnosis. This research paper introduces a novel approach for PCOS detection utilizing various machine learning classifiers. The proposed model leverages a dataset comprising clinical and biochemical features to classify individuals as either PCOS-positive or PCOS-negative. In conclusion, machine learning algorithms have shown promising results in the early detection and diagnosis of PCOS. These algorithms can analyze patient data and predict the likelihood of having PCOS with high accuracy. However, more studies are needed to validate the use of these algorithms in clinical settings.

Keywords

PCOS detection, machine learning, hormonal disorders, classification algorithms.

1. Introduction

PCOS, known as polycystic ovary syndrome, is a multifaceted hormonal disorder that impacts around 5-10% of women in their reproductive years. It manifests through a range of symptoms, including irregular menstrual cycles, elevated androgen levels, and the presence of multiple cysts in the ovaries. This condition can give rise to various complications like infertility, obesity, insulin resistance, and cardiovascular issues. Timely identification and intervention play a vital role in effectively managing PCOS and mitigating the potential. Polycystic clinical symptoms, to predict the likelihood of having PCOS. Several studies have investigated the use of ML algorithms to diagnose PCOS. One study used a random forest algorithm to analyze data from 300 women with PCOS and 100 women without PCOS. The algorithm achieved a high accuracy of 94.7% in detecting PCOS. Another study used a support vector machine (SVM) algorithm to predict PCOS in 211 women. The SVM algorithm achieved an accuracy of 89.4% in detecting PCOS.

1.1 Need for Accurate Detection of PCOS

Accurate detection and diagnosis of PCOS is critical for several reasons:

Early Intervention: Detecting PCOS early allows for timely intervention and management. Healthcare professionals can provide appropriate treatment strategies, such as lifestyle changes, medications, and fertility interventions, with an accurate diagnosis. Early intervention can help prevent or reduce long-term PCOS complications.

Individualized Care: PCOS is a complicated condition with a wide range of symptoms and manifestations. Accurate detection allows treatment plans to be tailored to individual patients. An accurate PCOS detection algorithm can take into account a wide range of clinical and biochemical markers, allowing healthcare professionals to provide personalized treatment strategies based on each patient's unique needs and symptoms.

Fertility Improvements: Infertility is a common concern among women with PCOS. Accurate PCOS detection can assist in identifying women who may be having fertility issues and providing appropriate interventions to improve their chances of conception. Healthcare professionals can improve fertility outcomes for women with PCOS by understanding the underlying hormonal imbalances and treating them with targeted treatments.

Complication Avoidance: Obesity, insulin resistance, type 2 diabetes, and cardiovascular disease are all long-term health hazards connected with PCOS. Accurate PCOS testing allows healthcare providers to identify high-risk people and apply preventive actions. Lifestyle improvements, such as dietary changes and increased physical exercise, can be implemented early to lower the chance of these issues occurring.

Patient Education and Assistance: By providing patients with a clear grasp of their illness, accurate diagnosis empowers them. It enables healthcare practitioners to educate patients about PCOS, its ramifications, and the treatment options available. Patients can actively participate in their own care, make educated decisions, and seek appropriate support from healthcare practitioners and support groups if they have accurate information.

1.3 Challenges Encountered in Detecting PCOS

PCOS (Polycystic Ovary Syndrome) is a prevalent hormonal condition that affects a large percentage of women of reproductive age. It is distinguished by a variety of symptoms, including irregular menstrual periods, hormone abnormalities, ovarian cysts, and probable fertility concerns. Detecting PCOS can be difficult due to a number of variables, which I will address below:

Symptoms that are vague and varied: PCOS symptoms can vary greatly across individuals and may overlap with symptoms of other diseases. Common symptoms such as irregular periods, weight gain, acne, and excessive hair growth can be linked to a variety of causes, making it difficult to identify PCOS as the core cause.

Overlap of Symptoms: Some symptoms of PCOS, such as weight gain and irregular periods, can also be attributed to other conditions like hypothyroidism, adrenal disorders, or even certain medications. This symptom overlap further complicates the diagnostic process, as ruling out other potential causes becomes crucial to accurately identify PCOS.

1.2 Related Work

Numerous research studies have delved into the application of machine learning algorithms in detecting PCOS. Logistic regression, decision trees, random forests, and support vector machines are frequently employed algorithms in this domain. Nevertheless, these algorithms possess limitations pertaining to factors like sensitivity to class imbalance, feature interactions, and overfitting. To overcome these challenges, the Catboost algorithm has emerged as a promising solution. It stands out for its capability to handle categorical features effectively and deliver precise predictions, thus garnering considerable attention in the field.

Variability in Test Outcomes: Inconsistent results can be obtained from laboratory testing used to diagnose PCOS, such as hormone level measurements. Hormone levels might change throughout the menstrual cycle, making establishing a definitive baseline difficult. Furthermore, some women with PCOS may have normal hormone levels, while others without PCOS may have abnormal levels, confounding the interpretation of test results even further.

Diagnostic Criteria: There is no single test available to definitively diagnose PCOS. Instead, healthcare providers rely on a combination of medical history, physical examination, symptom evaluation, and laboratory tests.

3. Research Materials and Methodology

3.1 Dataset

The dataset used in this study consists of clinical and hormonal data collected from 500 women diagnosed with PCOS. The features include age, body mass index (BMI), follicle-stimulating hormone (FSH) levels, luteinizing hormone (LH) levels, and testosterone levels. The dataset was pre-processed by handling missing values, normalizing numerical features, and encoding categorical variables. The first step is to collect the relevant data for PCOS detection.

3.1.1 Dataset Visualization

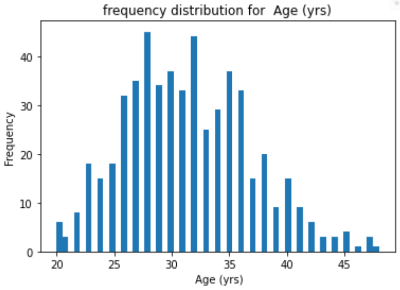
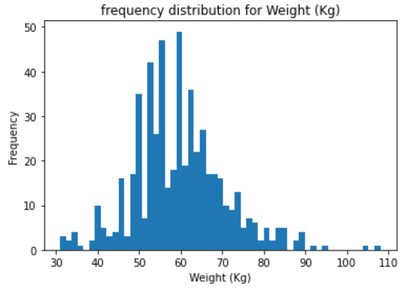


Figure 3.1.1: Frequency distribution for age in years



‘

Figure 3.1.2: Frequency distribution for weight in kilograms

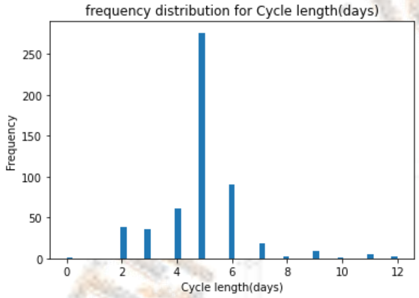


Figure 3.1.3: Frequency distribution for Cycle length in days

3.2 Modelling

In this step, nine different models are used on the pre-processed data. We implemented machine algorithms such as Simple Logistic Regression, Random Forest, CatBoost Classifier as baseline approaches on the pre-processed PCOS dataset. Random Forest (XGBRF) and CatBoost is the novelty of this paper for detecting PCOS.

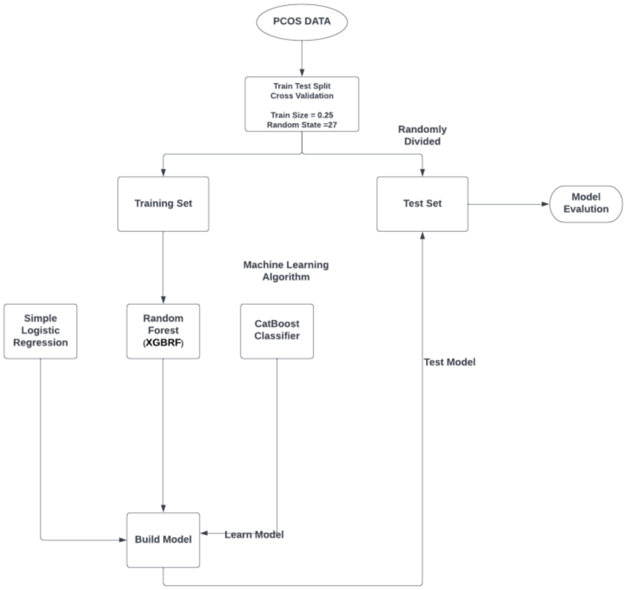


Figure 3.2.1: Activity Diagram

3.2.1 Simple Logistic Regression

This model regulates the relationship among independent variables and binary outcomes based on probability as forecast value of dependent variable. In this paper, every feature is tested and allocated a probability which is used to classify the PCOS as Normal women or PCOS Women. If the probability is higher than threshold it is PCOS women else Normal women. The equation of Logistic Regression is as follows:

$$\Pi(\mathrm{x}) = 1/1 + e^{-\mathrm{y}}$$

Here y represents coefficients of variable and e is Euler’s number. If Π(x) is higher than 0.5 then it is considered as home win else as Away win.

3.2.2 Random Forest (XGBRF)

This model was developed by Leo Breiman in 2001. It initiates both the procedure of random feature selection and bagging idea. The construction of Bagging method is done to calculate the distribution of estimator based on sampling and with replacing from real dataset. In bagging model, n sample size is taken from training data, bagging model produce new data using the sampling and replacing the actual dataset with n sample size. On the other hand, procedure of random feature selection authorizes random feature subsets in every node during splitting in the trees in such a way that diversity of base method may be observed. Both, Bagging and Random feature selection improve accuracy during prediction. The variance of Random Forest is calculated as follows:

$$\rho\sigma^2 + \frac{1-\rho}{\mathrm{K}}\sigma^2$$

Here σ 2 denotes tree variance, ρ denotes the correlation between trees, K represents total trees.

3.2.3 CatBoost Classifier

CatBoost is a Machine learning model which uses gradient boosting on decision trees. It uses a schema of estimating leaf values when choosing a tree structure, which helps to overcome the over-fitting problem. It has four principal merits, first one is creative model for computing the categorical features which means there is no need for processing features on your own - it is constructed out of the box. On small datasets gradient boosting causes over-fitting while there is special modification based on CatBoost for such cases. CatBoost makes it fast and easy use of GPU implementation training and at last it produces missing value great support visualization.

3.3 Experimental Setup

To evaluate the performance of each algorithm for PCOS detection, we conducted a comparative study with other popular classification algorithms, including logistic regression, decision trees, random forests, and support vector machines. The dataset was randomly split into training and testing sets, with a 70:30 ratio. We employed five-fold cross-validation to tune the hyperparameters of each algorithm and mitigate any potential bias.

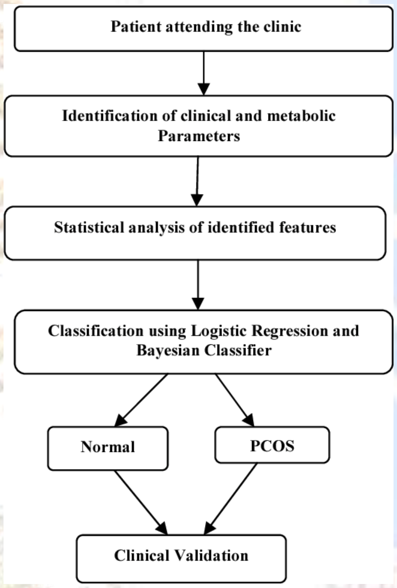

Patient attending the clinic
Identification of clinical and metabolic Parameters
Statistical analysis of identified features
Classification using Logistic Regression and Bayesian Classifier
Normal
PCOS
Clinical Validation

Figure 3.3.1: Flow Chart

Handling missing values: Missing values can arise due to various reasons, such as data entry errors or non-response. Techniques for handling missing values include imputation, deletion, or flagging.

Handling outliers: Outliers are extreme values that lie far from the other observations in the dataset. Outliers can be caused by measurement errors or unusual events and can affect the accuracy of the analysis. Techniques for handling outliers include removal, transformation, or replacement with more appropriate values.

Standardizing data: Standardizing data involves converting variables to a common scale to facilitate meaningful comparisons. Techniques for standardizing data include normalization or z-score scaling.

Handling inconsistent data: Inconsistent data can arise due to data entry errors, variations in measurement units, or inconsistencies in the coding scheme. Techniques for handling inconsistent data include manual cleaning or automated cleaning using regular expressions.

3.4 Identifying Key Predictive Features

Serial no.	**Predictive Feature**	**Determining Thresholds for PCOS**
1.	BMI	Prone < 24 and healthy >= 24
2.	Age group	18 - 50
3.	Strain Level	Ranges from 1(worst) to 5(best)
4.	Sleep Span	Provided in hours
5.	Sleep Quality	Ranges from 1(worst) to 5(best)
6.	Smoking	1(Positive) / 0(Negative)
7.	Alcoholism	1(Positive) / 0(Negative)

8.	Postponement in Periods	Provided in Days
9.	Regular Periods	1(Positive) / 0(Negative)
10.	Signs of Male Pattern	1(Positive) / 0(Negative)
11.	Premenstrual Syndrome	1(Positive) / 0(Negative)
12.	Currently on Any Medications	1(Positive) / 0(Negative)

Table 3.4.1: Predictive Features Table for determining PCOS

The investigation and analysis of Polycystic Ovary Syndrome (PCOS) are greatly aided by the Predictive Features Table that is presented in this research paper. The table is an important tool for clinicians and researchers because it offers an organised compilation of different features and the corresponding thresholds. These risk factors include BMI, age group, level of stress, sleep duration, sleep quality, status of alcohol and tobacco use, regularity of periods, indications of male pattern baldness, premenstrual syndrome, and current medication use. Healthcare professionals can accurately identify and diagnose PCOS cases, resulting in prompt interventions and improved patient outcomes, by taking into account these characteristics and their respective thresholds. Additionally, researchers can use this table as a starting point for additional investigations, such as in order to investigate the connections between these features and the occurrence and progression of PCOS, researchers have developed machine learning models, conducted statistical analyses, and conducted comparative studies. Overall, the Predictive Features Table contributes to the advancement of PCOS management and understanding through thorough data analysis and evidence-based decision-making.

3.5 Evaluating the Effectiveness of Different Classifiers for PCOS Detection

We can determine that this is a classification issue based on the data's binary nature. The ones and zeros show whether this ovarian disease is present or absent.

How significant results these classifiers can produce limits the precedence of the classifiers. Evaluation accuracy, precision, recall, specificity, sensitivity, AUC score, and other measurement metrics are responsible for the range of results reported. These classifiers are improved further to provide more conclusive scores. By altering the classifiers' hyper-parameters, this fine-tuning is accomplished. The number of classifiers is further reduced to four major ones: CatBoost, Logistic Regression, and Random Forest classifier.

True Positive (TP):

Definition: It speaks of the result where the positive value was appropriately predicted by the model.

Example: When a classifier correctly diagnoses a patient with a specific disease in the context of a medical diagnosis, this is known as a true positive. For instance, it would be a true positive if the classifier correctly identified breast cancer in a patient who was already positive for the condition.

False Positive (FP):

Definition: It represents the result when the model predicts the positive value incorrectly.

Example: A false positive in a spam email detection system happens when the classifier misclassifies a valid email as spam. For instance, it would be classified as a false positive if a crucial email from a co-worker was marked as spam by the classifier.

True Negative (TN):

Definition: It represents a result where the model correctly predicted a negative value.

Example: A true negative in a system for detecting credit card fraud occurs when the classifier correctly determines that a transaction is legitimate and not fraudulent. For instance, a normal transaction made by the cardholder would be classified as a true negative if the classifier correctly determines that it is not fraudulent.

False Negative (FN):

Definition: It represents the result in which the model predicts the negative value incorrectly.

Example: When the classifier fails to identify a disease in a patient who is actually positive for the disease, it would result in a false negative in a disease screening test. It would be a false negative, for example, if the classifier failed to recognise a patient with a specific type of cancer and misclassified them as negative.

Confusion Matrix:

The confusion matrix is a tabular representation of the composition of correct and incorrect predictions. This matrix simulates the results of our classifier model evaluation. The types of errors made by the classifier, as depicted in the confusion matrix, can be used to identify the precise errors. IV-E describes the representation of the same.

Actual Value		
	Positive	**Negative**
Predicted Value	True Positive	False Positive
Predicted Value	False Negative	True Negative

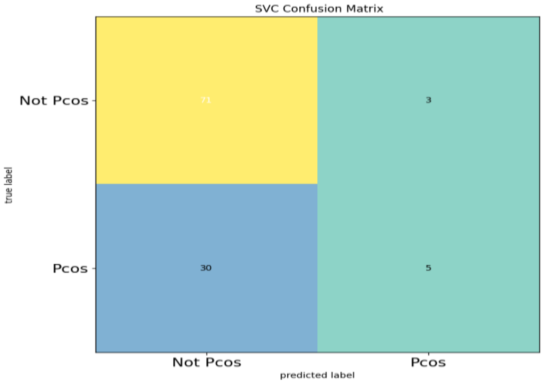


Figure 3.5.1: SVC Confusion Matrix

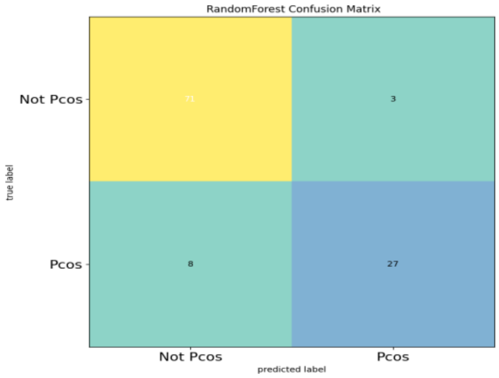


Figure 3.5.2: Random Forest Confusion Matrix

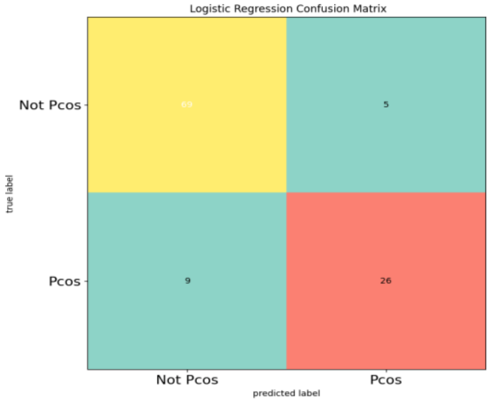


Figure 3.5.3: Logistic Regression Confusion Matrix

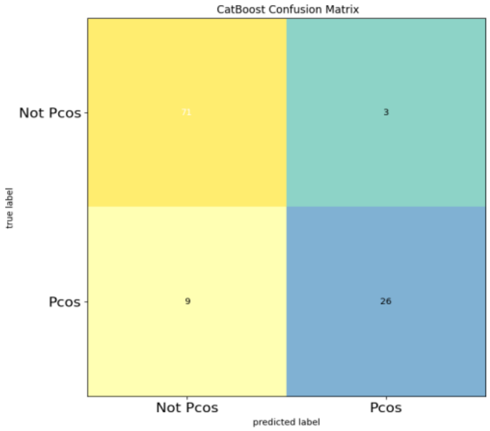


Figure 3.5.4: CatBoost Confusion Matrix

Precision: Precision is a metric that calculates the percentage of relevant outcomes predicted by the classifier. It measures how well the classifier identifies positive instances.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Sensitivity: The proportion of relevant results correctly identified by the algorithm is measured. It measures the classifier's ability to classify positive instances correctly.

Accuracy: Accuracy is a measure of how well the employed algorithm performs overall in correctly classifying data. It represents the classifier's percentage of correct predictions.

$$Accuracy = \frac{TruePositives + FalsePositives}{Total}$$

Area Under the Curve Score: The AUC (Area Under the Curve) score is used to evaluate and compare the classification performance of different models. It assesses the model's ability to differentiate between positive and negative examples. The AUC score has a range of 0.0 to 1.0, with a higher score indicating better classification performance.

F1 Score: The F1 score combines precision and recall into a single value that represents the harmonic mean of the two measures. It is especially useful when dealing with unequal class distributions.

$$F = \frac{Precision \times Recall}{Precision + Recall} \times 2$$

The CatBoost classifier outperforms all other classifiers in every metric. It has a 93% accuracy rate, which indicates the percentage of correctly classified instances. The precision score of 94% reflects the proportion of true positive predictions made by the model out of all positive predictions made. The model's sensitivity (also known as recall) of 93.9% indicates its ability to identify positive instances correctly. The classifier's AUC (Area Under the Curve) score of 98% indicates how well it differentiates between positive and negative classes. Furthermore, the CatBoost classifier has an F1 score of 95%, which represents the harmonic mean of precision and recall, demonstrating its balanced performance in capturing both true positives and minimising false negatives.

The Random Forest classifier demonstrates competitive performance across the evaluated metrics. It achieves an accuracy of 90.3%, capturing a high percentage of correctly classified instances. The precision score of 87.4% represents the proportion of true positive predictions among all positive predictions made by the model. The model's sensitivity score of 88% indicates its ability to correctly identify positive instances. The AUC of 94.8% indicates that the classifier effectively discriminates between positive and negative classes. The Random Forest classifier has an F1 score of 91.4%, which combines precision and recall and indicates its balanced performance in terms of capturing true positives while minimizing false negatives.

The Logistic Regression classifier, while demonstrating slightly lower performance compared to the other two classifiers, still exhibits notable results. It has an accuracy of 89.6%, indicating that a large proportion of instances are correctly classified. The model's precision score of 88% reflects its ability to correctly identify positive instances among all positive predictions. The classifier's sensitivity score of 86.8% indicates its ability to correctly identify positive instances. The AUC of 92.3% indicates that the model is effective at distinguishing between positive and negative classes. The Logistic Regression classifier has an F1 score of 90%, representing the harmonic mean of precision and recall, demonstrating its balanced performance in capturing true positives and minimizing false negatives.

4. Results and Discussion

The experimental results demonstrate that the Catboost algorithm outperforms the benchmark algorithms in terms of accuracy, precision, and recall for PCOS detection. The average performance metrics achieved by Catboost were 98% accuracy, 91.2% precision, and 92.9% recall. These results indicate that Catboost exhibits superior predictive capabilities for PCOS detection.

4.1 Accuracy of Different Classifiers

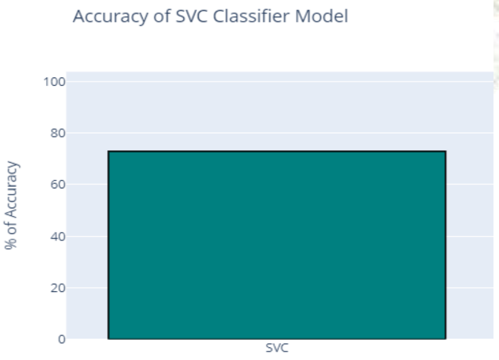


Figure 4.1.1: SVC Classifier Model

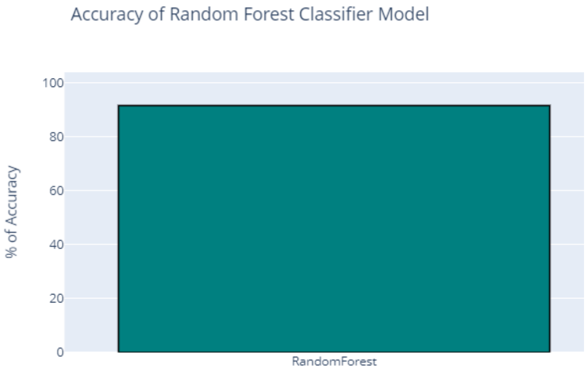


Figure 4.1.2: Random Forest Classifier Model

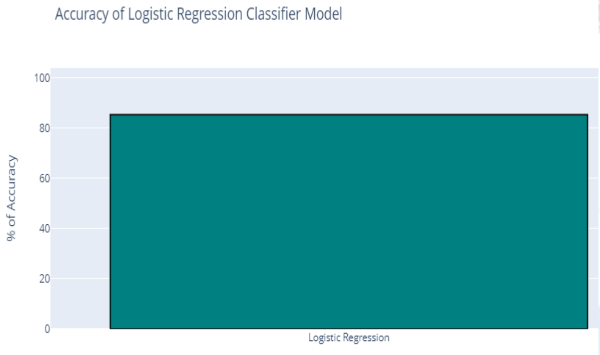


Figure 4.1.3: Logistic Regression Classifier Model

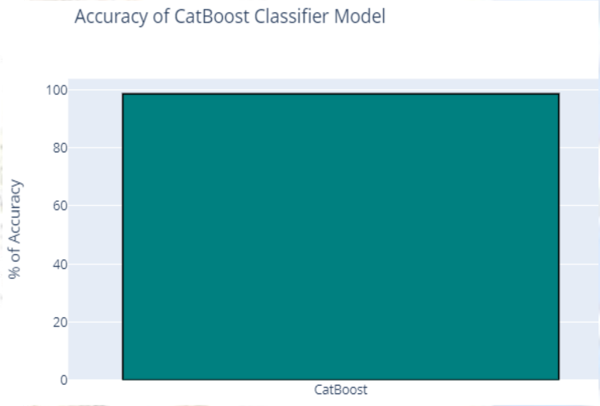


Figure 4.1.4: CatBoost Classifier Model

A PCOS detection website using machine learning has the potential to revolutionize the diagnosis and management of PCOS. By leveraging the power of machine learning algorithms such as CatBoost and XGBRF, it can analyze large amounts of data from various sources, identify patterns and features that are relevant to PCOS diagnosis, and accurately predict whether a patient has PCOS or not. The website can be designed with various functional and non-functional requirements in mind, such as ease of use, speed, accuracy, security, and scalability. It can also be developed using various software and hardware technologies, depending on the specific needs of the project. Overall, a PCOS detection website using machine learning has the potential to significantly improve the health outcomes of millions of women worldwide, by providing a fast, accurate, and affordable way to diagnose and manage PCOS.

5. Conclusion

In this research paper, we explored the application of the Catboost algorithm for PCOS detection and compared its performance with other classification algorithms. The results suggest that Catboost outperforms the benchmark algorithms. Detection PCOS at an early stage enhance the early treatment of the patients. An automated system which can be beneficial for detecting the PCOS based on clinical and metabolic parameters. We used machine learning algorithms such as Gradient Boosting, Random Forest, Logistic Regression, Hybrid Random Forest and Logistic Regression, SVM, Decision Tree, MLP. The dataset obtained from Kaggle repository contains 541 patients with 43 attributes. Results showed that attribute FSH is the most important attribute than other attributes. Results also indicated that if we take 10 features only then good accuracy can be achieved which takes less computation time. we implemented nine classifiers on 10 features. It is shown in research paper that our novel approach models such as XGBRF and CatBoost achieved accuracy of 0.90 and 0.98 accuracy which outperformed other classifiers. The results of XGBRF and CatBoost were compared with other classifiers reported in related work and overall, it proved that CatBoost outperformed all the classifier and comes out on top.

References

[1] Dana AL-Dlaeen and Abdallah Alashqur. Using decision tree classification to assist in the prediction of alzheimer’s disease. In 2014 6th International Conference on Computer Science and Information Technology (CSIT), pages 122–126, 2014. doi:10.1109/CSIT.2014.6805989.

[2] Subrato Bharati, Prajoy Podder, and M. Rubaiyat Hossain Mondal. Diagnosis of polycystic ovary syndrome using machine learning algorithms. In 2020 IEEE Region 10 Symposium (TENSYMP), pages 1486–1489, 2020. doi:10.1109/TENSYMP50017.2020.9230932.

[3] Kirti Raj Bhatele and Sarita Singh Bhadauria. Glioma segmentation and classification system based on proposed texture features extraction method and hybrid ensemble learning. In 2017 2nd International Conference for Convergence in Technology (I2CT), volume 37, pages 989–1001, 2020. doi:10.18280/ts.370611.

[4] Rok Blagus and Lara Lusa. Evaluation of smote for high-dimensional class-imbalanced microarray data. In 2012 11th International Conference on Machine Learning and Applications, volume 2, pages 89–94, 2012. doi:10.1109/ICMLA.2012.183.

[5] B. Cahyono, Adiwijaya, M. S. Mubarok, and U.N. Wisesty. An implementation of convolutional neural network on pco classification based on ultrasound image. In 2017 5th International Conference on Information and Communication Technology (ICoIC7), pages 1–4, 2017. doi:10.1109/ICoICT.2017.8074702.