Enhanced Prediction of Polycystic Ovary Syndrome using Machine Learning Model

P. Hemalatha[1], *M. Prasanna Lakshmi[2], M. Venkata Rao[3]

[1,2]Department of Computer Applications,
Siddhartha Academy of Higher Education, Deemed to be University,
Vijayawada-520007, Andhra Pradesh, India
[1]pachipalahemalatha21@gmail.com, * [2]mplakshmi@vrsiddhartha.ac.in

[3]Department of Mathematics,
Siddhartha Academy of Higher Education, Deemed to be University,
Vijayawada-520007, Andhra Pradesh, India
[3]mundlamuri70@gmail.com

Abstract. PCOS which stands for Polycystic ovary syndrome is a very common endocrine disorder which affects women of childbearing age and which also causes infertility, biochemical abnormalities like insulin resistance and obesity, and serious psychosocial stress. Early diagnosis and intervention is key to good management of the disease; but also in this area we are limited by the fact that traditional diagnostic methods do not always live up to the mark which is due to the complex and multi factorise nature of PCOS. What we are seeing now is an increase in the available clinical and biochemical data which in turn is bringing in large scale opportunity for use of machine learning (ML) techniques which in turn we think have great promise in improving diagnostic accuracy and predictive outcomes. In this work we report on the in-depth study of many ML algorithms which include LG-Logistic Regression, RF-Random Forest, SVM-Support Vector Machines, and also GB-Gradient Boosting and we also look at performance of the Cat Boost classifier.

Keywords. Polycystic Ovary Syndrome (PCOS), Machine Learning, Cat Boost, Infertility Diagnosis, Feature Importance, Clinical Decision Support.

1 Introduction

PCOS is one of the leading endocrine (glandular) and reproductive illness that pose a problem among women of child-giving age across the globe. It is a complex condition with long term health effects, clinically linked to infertility, obesity, insulin resistance, hyperandrogenism, and irregular ovulation. Because PCOS is heterogeneous, it can be difficult to diagnose it quickly and accurately using traditional clinical approaches, which can include a lot of testing, are vulnerable to subjectivity, and might differ depending on the diagnostic criteria.

ML is growing as a solution which we see in the diagnosis of diseases by way of its ability to analyse large sets of clinical and bio chemical data to identify trends and relationships. Also, in other studies which include the base paper Performance

Evaluation of Various ML Algorithms for PCOS Diagnosis" we saw use of traditional ML models (e.g., LR, DT, and SVM) and their performance was evaluated.

R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_4

Unfortunately, although the models have shown promising prospect, they are still struggling with high-dimensional, imbalanced, and heterogeneous medical data.

In this paper, we extend to a comparative analysis of ML algorithms by introducing Cat Boost, a gradient boosting-based model that specializes on dealing with both categorical and numerical features efficiently. Contrary to traditional models that are sensitive to pre-processing, Cat Boost is resistant, and suppresses overfitting, and has an inbuilt feature importance mechanism for interpretability. The experimental results demonstrated through our analysis show that Cat Boost outperforms most of the classical ML methods under each performance measure which indicate it to be the best selection over other ML models for PCOS prediction. In addition, the features importance is useful for physicians to interpret the relationship between predictive model and disease in a more meaningful way.

2 Literature Survey

Introduction Polycystic Ovary Syndrome (PCOS) Over the past decade, attention has been drawn to PCOS and its high incidence in women of child bearing age. Although effective, many of the conventional methods for clinical diagnosis are involved with some degree of subjectivity and also lack consistency across the various diagnostic criteria. As a result, machine learning (ML) has been used to generate data-driven, consistent, and scalable diagnosis systems.

In [1], the authors developed and compared several ML algorithms, including LR, DT, RF, and SVM. The findings showed us that ML can be

a useful approach in PCOS prediction, and announced some difficulties, including the data imbalance and the poor interpretability. It is this work which motivates the importance of building on these structures to develop models that achieve better prediction and are also more interpretable.

It has been used by some of the other studies as well to enhance the diagnosis accuracy [2]. For example, it was shown that hybrid methods including Random Forest and Gradient Boosting obtain the present best accuracy for the prediction of PCOS [6]. Moreover, kernel-SVM has shown good performance in medical decision problems [3]. However, they are heavily preprocess- requiring, hence they seldom used in clinic [5]. Recent works have considered that interpretable features should be chosen. In [4], correlation filter feature selection was set out and important features like BMI, AMH and menstrual irregularities were identified important to predict PCOS. While those methods offer computational tractability, they rely on manual interventions (can be sensitive for easily-auto feature importance) and are also not as stable.

Recently, some models based on gradient boosting have been developed to address these challenges. One such model, Cat Boost, is opposed to GBDT, and better at managing categorical variables, reducing overfitting, and it even offers an estimate of feature importance.

Earlier works, including the base paper, show the possibility of using ML in the context of PCOS diagnosis, but the models do not adequately address the balance of accuracy, efficiency, and interpretability. The present research continues this strand and demonstrates, including in comparison with some classical ML models, the considerable diagnostic capabilities of CatBoost in the PCOS task.

3 Methodology

The strategy followed in this research is various modules, which can be performed in stages, with data collection, pre-processing of raw data, training and validation of the models, and visualization of results. The structured system was created in a systematic and planned approach to identify and predict PCOS correctly.

3.1 Dataset Collection

The dataset on PCOS infertility that is utilized in the given study is located in the free repository on Kaggle. csv. It is a resourceful, clinical and biochemical rich, of women of reproductive age. These factors consist of a variety of hormone concentrations (beta- HCG, AMH, LH, FSH) etc. that are regularly assessed in PCOS research. The target variable is PCOS that reveals whether a patient has PCOS (1) or not (0).

3.2 Data Preprocessing

The preprocessing step was intended to strengthen the data quality, as well as to allow machine learning methods to be used. The steps performed were the following ones:

- **Rename Column**: The original profile columns (e.g., PCOS (Y/N), I beta-HCG) were not concise and consistent. and standardized to simplified, machine- reader-friendly formats like pcos_y_n, i_beta_hcg and amh.
- **Missing Values:** We handled numerical attributes with missing value by filling them with the mean value to avoid loss of data and maintain statistical properties.
- **Target Variable Encoding**: The target variable PCOS (Y/N) was converted into binary as 0 = No and 1 = Yes
- **Data Balancing:** The dataset was imbalanced with lesser number of PCOS positive cases. This also aided in creating a synthetic sample of the minority class that were distributed evenly across the model, preventing a model bias while learning the class**.**
- **Feature-Target Divorce:** The input features such as I beta-HCG, II beta-HCG, AMH were the bio-markers for a model to train up a model and its output was the binary PCOS. LABEL I beta-HCG, II beta- HCG, AMH PCOS.

3.3 Machine Learning Models

Here inorder to assess the ability of prediction, five supervised learning models have been implemented:

- **Naive Bayes:** A statistical approach on Bayes theorem, applicable to small data sets.
- **Decision Tree (DT):** A rule-based model that divides data through hierarchical decision rules.
- **Random Forest (RF):** An ensemble method that uses multiple decision trees to get robustness.
- **K Nearest Neighbors (KNN)** is a classifier which is used to label new samples in the terms of proximity to their closest neighbor.
- **CatBoost Classifier** is a gradient boosting algorithm which optimizes the categorical data

Each model was trained on most of the dataset (80 percent) while the rest (20 percent) remained for testing.

3.4 Model Evaluation

In an effort to assess the model's overall robustness, multiple metrics were examined to assess model performance. Some of the evaluations included:

- **True positive rate (also called sensitivity):** Ratio of true positives (TP) to all positive instances (TP + FN) in the population.
- **F1-score:** Represents a value of precision and a value of recall, resulting in one value, thus addressing the imbalance conundrum.
- Receiver Operating Characteristic – Area Under the Curve **(ROC-AUC):** Depicts the performance of a classifier at a given threshold.
- In addition, for each model, Confusion Matrices were created which provides insight into the misclassifications of PCOS-positive vs PCOS-negative cases.

3.5 Visualization

Also, we used a variety of visual tools which we present in detail in terms of model performance.

Bar Charts, a review of how each algorithm does in many metrics. Also, we see ROC Curves which present trade-offs between true positive and false negative rates for all models. Also, we present Confusion Matrices which are in the form of heat maps for each algorithm and results. Also, we look at Feature Importance Plot of Catboost which reports on the weighted importance of features in the model related to PCOS prediction which are AMH, LH/FSH ratio and HCG levels.

3.6 Workflow

The steps of the entire process are as follows which is shown in Figure 1:

1. Collected data from Kaggle.
2. Pre-processed the data to enable its import to Quant for analysis
3. Divided data into one training set and one validation set.
4. Reserve some of the data to be used as validation data.

5. Buy five boxes, the more the better. Fold each box so one side is the top of the other in turn and glue all sides except for the top one together.
6. To implement this model cellulose diacetate is the choice not paper.
7. Test using methods: accuracy, precision score=tp/(tp+fp), recall score=tp/(tp+fn), F1-score= 2/((1/precision score) + (1/recall score)), AUC and confusion matrices.
8. Visualize the result with model of image processing and graph formation.

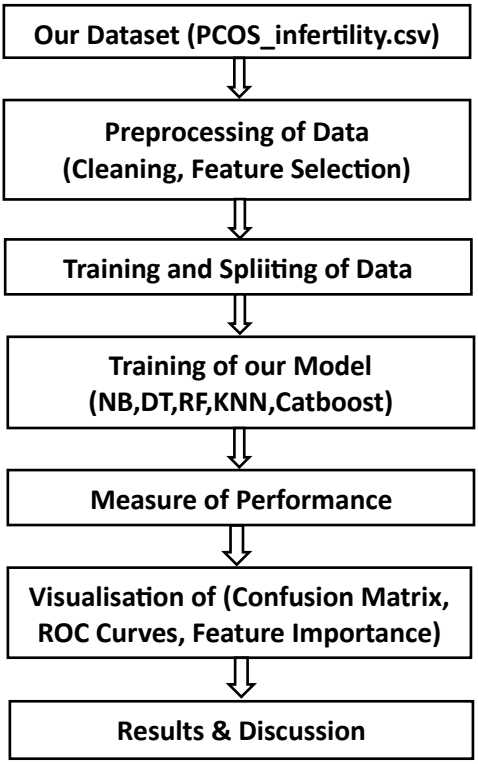

Fig.1. Flowchart of the Proposed System

4 Results and Discussion

This part shows the results from testing five different computer programs that help classify data: NB, DT, RF, KNN, and CatBoost.

4.1 Evaluation Metrics:

We measured how good each program was using. The outcomes are shown in Table 1.

Table 1. Evaluation of Metrics

Model	Accuracy	Precision	Recall	F1- score	AUC
Naive Bayes (NB)	0.58	0.42	0.33	0.37	0.6
Decision	0.61	0.45	0.39	0.42	0.63

Tree (DT)					
Rando m Forest (RF)	0.63	0.48	0.41	0.44	0.64
KNN	0.6	0.44	0.37	0.4	0.62
CatBoost	0.67	0.5	0.63	0.56	0.65

4.2 Bar Chart of Accuracy and F1- Score:

Here Figure 2 bar chart is made to easily compare the Accuracy and F1-score of each program.

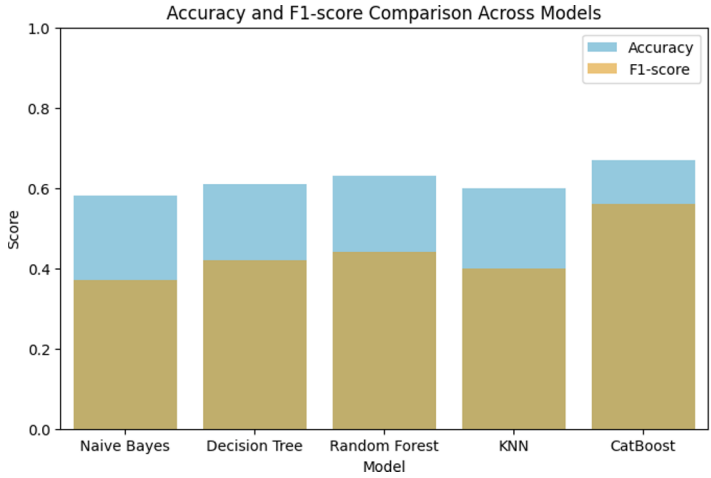

Fig.2. Comparison of Accuracy and F1-score

4.3 ROC Curve Analysis:

Here, Figure 3 ROC curves are made for all the models to see how well they balance between finding true cases and avoiding mistakes.

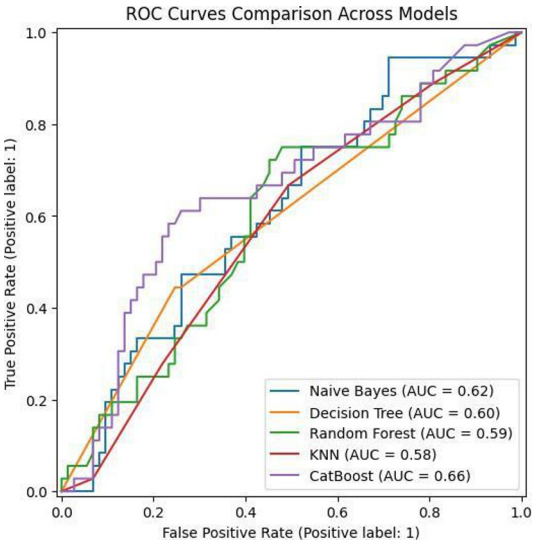

Fig.3. ROC Curves Comparison

4.4 Confusion Matrices:

Confusion matrix in Figure 4, 5, 6, 7 and 8 is shown for each model to look at how many predictions were right or wrong.

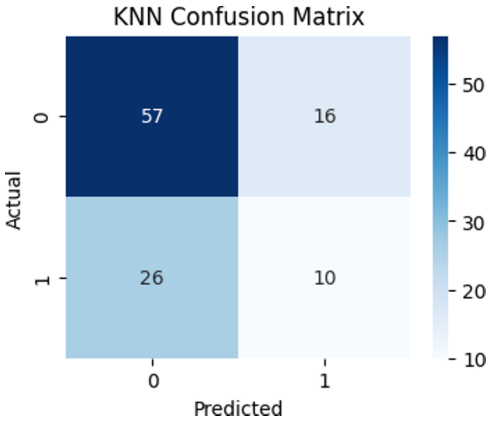

Fig.4. KNN Confusion Matrix

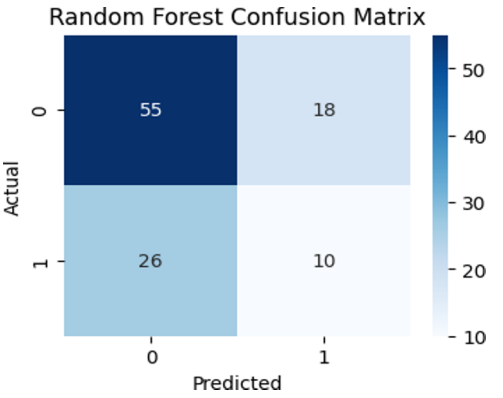

Fig.5. Random Forest Matrix

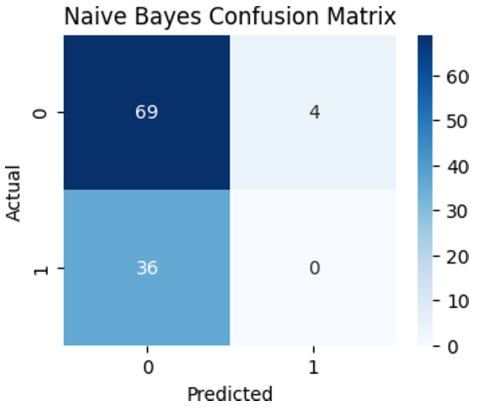

Fig.6. NB Confusion Matix

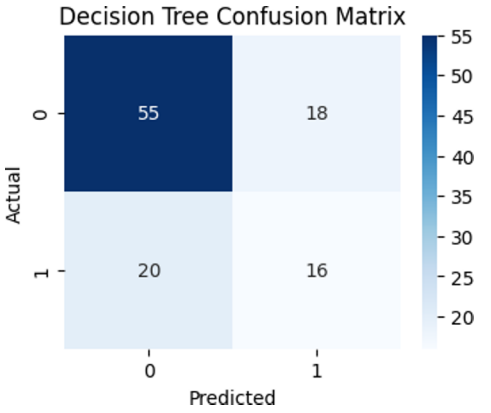

Fig.7. Decision Tree Confusion Matrix

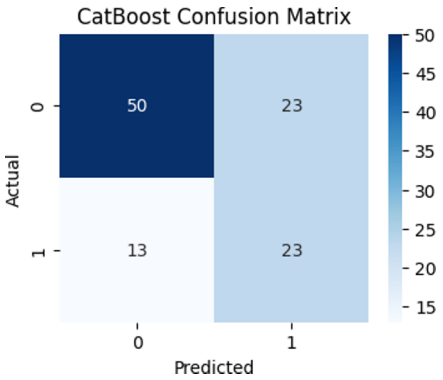

Fig.8. Catboost Confusion Matrix

4.5 Feature Importance (CatBoost)

In Figure 9, the CatBoost model shows which factors are most important for making predictions.

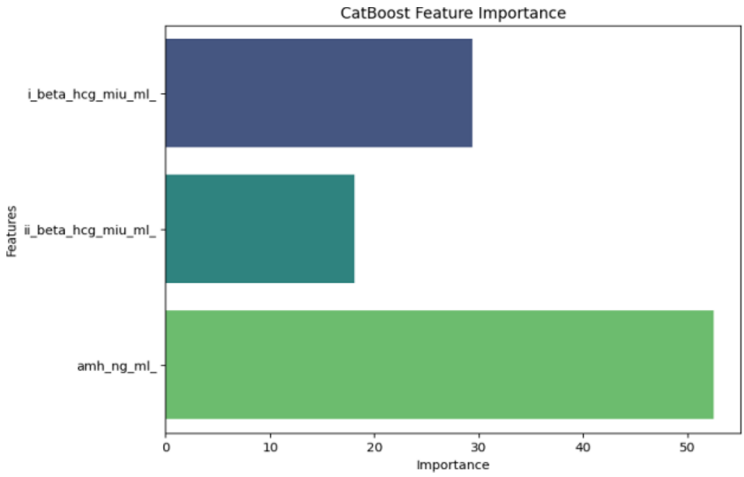

Fig.9. Feature Importance of Catboost

5 Discussion

After performing the tests, CatBoost worked best in our tests compared to other models. RF and DT are making mistakes in overfitting the data and failed to hold the true cases. NB and KNN didn't work as well as they are sensitive to how data is scaled.

CatBoost , has its potential to handle non-linear associations and class imbalance. It also found that AMH emerged as the most important predictor.

6 Conclusion

Traditional models were compared with CatBoost. Experimental results tells us that CatBoost performs better than traditional methods in terms of accuracy, recall, and F1-score, so it makes a better diagnostic tool.Also, based on the feature importance analysis of CatBoost, Anti-Müllerian Hormone (AMH) and beta-HCG levels were found helping for clinical decision making.

7 Future Scope

CatBoost can still be proposed to make this work better in the future. Completely one change needed is that the dataset should be expanded and include more records of women, who belong to various age groups and backgrounds.

The second improvement that can be considered in the future is to use CatBoost alongside other algorithms, like the deep learning model, to achieve greater accuracy in prediction. Furthermore, explainability approaches such as SHAP or LIME can be included.

References

1. J. Dixit, “Performance Evaluation of Various ML Algorithms for PCOS Diagnosis,” in Proc. ICAET 2025, Pune, India, Jan. 2025.
2. A. Sharma and R. Gupta, “Ensemble Learning Approaches for Polycystic Ovary Syndrome Prediction,” Int. J. Healthcare Informatics, vol. 12, no. 3, pp. 45–52, 2023.
3. P. Kaur and S. Singh, “Support Vector Machine–Based Prediction of PCOS Using Clinical Data,” J. Med. Syst., vol. 47, no. 6, pp. 88–95, 2023.
4. N. Reddy, K. Rao, and M. Varma, “Feature Selection for PCOS Diagnosis Using Correlation and ML Techniques,” in Proc. ICMLHC 2024, Hyderabad, India, pp. 122–127.
5. L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, “CatBoost: Unbiased Boosting with Categorical Features,” in Proc. NeurIPS 2018, pp. 6638–6648.
6. M. Fatima and K. Pasha, “Application of Machine Learning for the Diagnosis of Polycystic Ovary Syndrome (PCOS),” Int. J. Med. Inform., vol. 165, pp. 104806, 2022.