Comparative Analysis of Machine Learning Models for PCOS Diagnosis

Nitipat Wuttisasiwat Saowaluk Jirapornsirikul

Abstract—This study presents a comparative analysis of different machine learning models for the diagnosis of Polycystic Ovary Syndrome (PCOS). Using a dataset from Kaggle containing features such as age, BMI, menstrual irregularity, testosterone levels, and antral follicle count, we trained and evaluated three machine learning algorithms: Logistic Regression, Random Forest, and Decision Tree. Our evaluation using 5-fold cross-validation showed that Random Forest and Decision Tree classifiers performed exceptionally well on this dataset, suggesting potential applications in clinical decision support systems for PCOS diagnosis.

Keywords—machine learning, PCOS diagnosis, Random Forest, Decision Tree, Logistic Regression, healthcare analytics

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is one of the most common endocrine disorders affecting women of reproductive age, with prevalence estimates ranging from 6% to 10% worldwide. Despite its prevalence, PCOS remains challenging to diagnose due to its heterogeneous presentation and the lack of a single diagnostic test. Current diagnostic approaches rely on a combination of clinical symptoms, hormonal assessments, and ultrasound findings, which can lead to delayed diagnosis and treatment.

Machine learning techniques have demonstrated potential in a number of medical diagnostic applications by spotting complicated patterns in patient data that clinicians might not notice right away.

This project aims to evaluate the performance of various machine learning algorithms in diagnosing PCOS based on available clinical parameters, with the goal of developing a reliable tool that could assist healthcare providers in making more accurate and timely diagnoses.

II. EXPERIMENTAL SETUP

A. Dataset Description

We used a PCOS dataset from Kaggle for this research. The dataset consists of records from female patients with the following features:

1. Age (Integer): Patient's age in years
2. BMI (Decimal): Body Mass Index, calculated as weight (kg) divided by height squared (m^2)
3. Menstrual Irregularity (Boolean): Presence (1) or absence (0) of irregular menstrual cycles
4. Testosterone Level (Decimal): Serum testosterone level in ng/dL
5. Antral Follicle Count (Integer): Number of antral follicles as observed in ultrasound
6. PCOS Diagnosis (Boolean): Target variable indicating confirmed PCOS diagnosis (1) or absence (0)

The dataset includes a balanced distribution of positive and negative PCOS cases, with approximately 1000 total records.

B. Machine Learning Models

We implemented and compared three different machine learning models:

1. Logistic Regression: A baseline linear model commonly used for binary classification problems.
2. Random Forest: An ensemble learning technique that builds several decision trees during training and produces the class that represents the average of the classes of the individual trees.
3. Decision Tree: A non-parametric supervised learning method that creates a model predicting the target variable value by learning simple decision rules from the data features.

C. Evaluation Methodology

To evaluate the models, we used 5-fold cross-validation on the training set, which involves

1. Separating the training data into five equal sections, or folds
2. Using four folds to train the model and the remaining fold for validation
3. Repeating this five times, using one validation set for each fold
4. Averaging the performance metrics across all five iterations

Performance metrics collected during cross-validation included

1. Accuracy: The proportion of correct predictions based on all cases evaluated
2. Precision: The proportion of true positive predictions to all positive predictions
3. F1-Score: The harmonic mean of precision and recall

After cross-validation, we trained each model on the entire training set and evaluated them on the held-out test set to obtain the final performance metrics and generate confusion matrices.

All experiments were executed using Python 3.12.7 with scikit-learn 1.6.1 on a machine with an Apple M1 processor and 16 GB RAM.

III. EXPERIMENTAL RESULTS

A. Cross-Validation Results

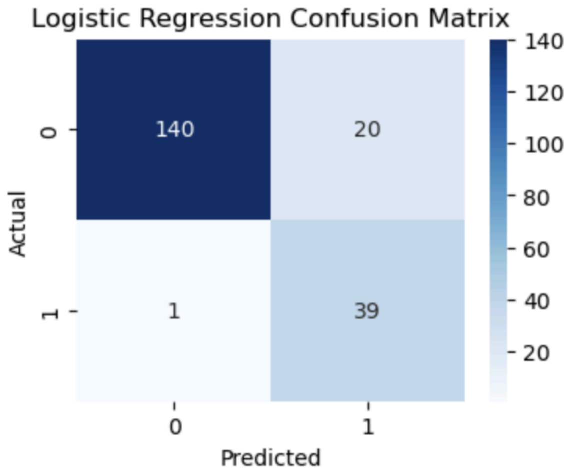
Model	Accuracy	Precision	F1 Score
Logistic Regression	0.914172	0.892309	0.913949
Random Forest	0.997659	1.000000	0.997653
Decision Tree	0.998437	1.000000	0.998431

Table I presents the average performance metrics from 5-fold cross-validation for each model.

The cross-validation results indicate that both tree-based models (Random Forest and Decision Tree) significantly outperformed the baseline Logistic Regression model across all metrics. Decision Tree achieved the highest overall performance with an accuracy of 0.998437 and an F1-score of 0.998431.

B. Confusion Matrices

Confusion matrices provide each model's performance by showing the counts of true positives, false positives, true negatives, and false negatives.



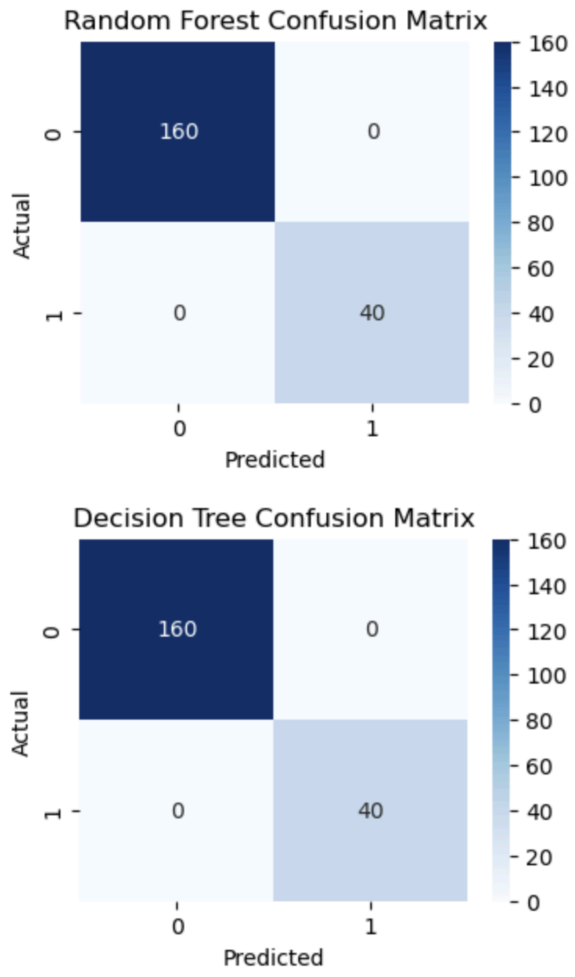


Fig. 1. Confusion matrices for (a) Logistic Regression, (b) Random Forest, and (c) Decision Tree models on the test set.

The confusion matrices revealed that Random Forest and Decision Tree had the lowest number of false negatives (0), making it particularly valuable in a medical diagnostic context where missing positive cases can have serious consequences Logistic Regression had the highest number of both false positives (1) and false negatives (20)

IV. DISCUSSION

A. Comparison of Model Performance

The experimental results clearly demonstrate that tree-based models outperform Logistic Regression for PCOS diagnosis based on the given clinical parameters. The excellent performance of Random Forest and Decision Tree can be attributed to

1. Its ability to describe complex, non-linear connections between features and the target variable
2. Inherent feature selection capabilities that give appropriate weight to the most informative features
3. Robustness to different types of features without requiring complicated preprocessing

B. Limitations

Several limitations of this study should be acknowledged

1. The dataset size is relatively modest, and results might vary with larger, more diverse patient populations
2. The dataset only comes from one source (Kaggle), which might restrict its applicability to various patient demographics and clinical contexts.
3. Our study did not include some potentially relevant features such as insulin resistance markers, luteinizing hormone to follicle-stimulating hormone ratio, and family history
4. We did not fine-tune all of the models' hyperparameters, which could have enhanced their performance.

C. Conclusion

This study shows how machine learning techniques, in particular tree-based models like Random Forest and Decision Tree, may assist in PCOS diagnosis. These models achieved high accuracy and precision rates in identifying PCOS cases based on available clinical parameters.

The effective use of these models may result in the creation of clinical decision support tools that assist clinicians in diagnosing PCOS more quickly and accurately, potentially reducing the current delays in diagnosis and treatment that many patients experience.

REFERENCES

[1] A. Yasmin, S. Roychoudhury, A. P. Choudhury, A. B. F. Ahmed, S. Dutta, F. Mottola, V. Verma, J. C. Kalita, D. Kumar, P. Sengupta, and A. Kolesarova, “Polycystic ovary syndrome: An updated overview foregrounding impacts of ethnicities and geographic variations,” Biomedicines, vol. 10, no. 12, p. 3184, 2022.

[2] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[3] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[4] Kottarathil, P. (2020). Polycystic Ovary Syndrome (PCOS) [Data set]. Kaggle. https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos

[5] Kennaruk. (2024). PCOS ML Benchmark [Source code]. GitHub. https://github.com/kennaruk/pcos-ml-benchmark