© 2024 ADYPJIET, Volume 1(1), Online ISSN: 3049-1517, September 2024, pp. 24-28

Ajeenkya DY Patil Journal of Innovation in Engineering & Technology

Journal Homepage: https://www.adypsoe.in/adypjiet

# Empowering Early Detection: A Web-Based Machine Learning Approach for PCOS Prediction

[1]Poonam Musmade
Ajeenkya D Y Patil School of Engineering, Pune, INDIA
[2]Sachin Rajas
Ajeenkya D Y Patil Innovative University, Pune, INDIA
[3]S.M. Khairnar
Ajeenkya D Y Patil School of Engineering, Pune, INDIA
[4]S.V. Rupanar
Ajeenkya D Y Patil School of Engineering, Pune, INDIA

***Article History:***

***Received:*** *12-09-2024*

***Revised:*** *16-09-2024*

***Accepted:*** *20-09-2024*

**Abstract:**
Polycystic Ovary Syndrome (PCOS) is a prevalent hormonal disorder impacting many women globally, leading to a range of complications from menstrual irregularities to infertility. The rise in PCOS incidence highlights the critical need for early detection and effective management strategies. This study explores a web-based machine learning approach to predict PCOS using a dataset of 541 patient records. Various machine learning models, including Logistic Regression (LR), Decision Tree (DT), Ada Boost (AB), Random Forest (RF), and Support Vector Machine (SVM), are employed to uncover patterns and predict PCOS. Feature selection is performed using the Mutual Information model, resulting in the highest accuracy of 94% achieved by both AB and RF models. The integration of machine learning techniques into a user-friendly web interface aims to enhance early PCOS detection**.**

**Keywords**: Machine Learning, Mutual Information, Django, Polycystic Ovary Syndrome, Early Detectiontopics of the paper.

## 1. Introduction

Polycystic Ovary Syndrome (PCOS) is a complex endocrine disorder that affects a significant portion of women of reproductive age. It is characterized by hormonal imbalances that can result in irregular menstrual cycles, ovulation problems, hirsutism, acne, and weight gain. Long-term risks associated with PCOS include type 2 diabetes, cardiovascular diseases, and infertility.

Despite its high prevalence, PCOS remains underdiagnosed due to the variability in symptom presentation and lack of awareness. First described by Stein and Leventhal in 1935, PCOS affects approximately 5-10% of women aged 12-45 [1]. The condition's prevalence varies by ethnicity, with studies showing a higher incidence in certain populations, such as 15.3% among Indian women [2]. Early diagnosis is crucial for effective management and prevention of severe complications.

Recent advancements in artificial intelligence (AI) and machine learning (ML) offer promising methods for early detection of PCOS. This study proposes a web-based system that integrates ML models for PCOS prediction, aiming to provide a binary classification of PCOS status based on clinical features.

## 2. Literature review

Machine learning methods have shown significant potential in improving diagnostic accuracy for PCOS. The following studies illustrate the application of various ML techniques. Silva et al. [2] applied the Boruta Shap technique with a Random Forest model, achieving an accuracy of 86% [18]. Khanna et al. [3] utilized a multi-stacking ML approach on a dataset from Kerala, India, achieving 98% accuracy. Bharati et al. [4] employed a hybrid RFLR model, reaching 91.01% accuracy. Abu Adla et al, [5] achieved 91.6% accuracy with SVM using a hybrid feature selection strategy . Hassan et al, [6] reported a 96% accuracy with the Random Forest algorithm. Denny et al. and Mehr et al. (2023) highlighted Random Forest's efficacy with accuracy rates of 89.02% and 98.89%, respectively [7,8].

These studies demonstrate the efficacy of ML models in PCOS prediction, though results vary based on models and datasets. This study aims to improve accuracy and usability by combining advanced feature selection and ML models into a web-based interface.

## 3. **Methodology.**

**The proposed methodology for PCOS prediction consists of several key steps:**

### 3.1 Data Preprocessing

The dataset, sourced from Kaggle, undergoes comprehensive preprocessing:

Handling Null Values: Missing data is addressed using imputation techniques.

Removing Duplicates: Duplicate records are eliminated to ensure data integrity.

Categorical to Numerical Conversion: Categorical variables are converted to numerical format for ML algorithms.

### 3.2 Feature Selection

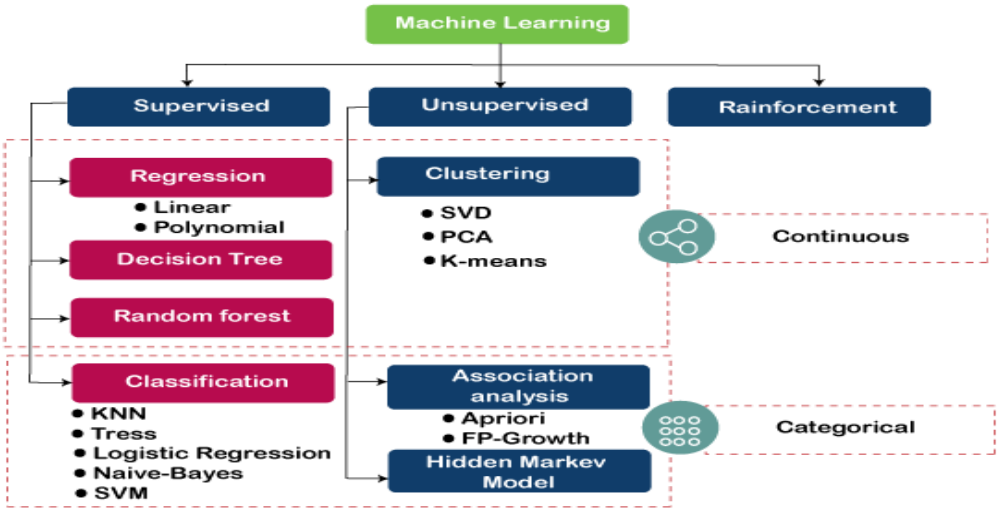


**Figure 1: Model Performance Comparison**

The Mutual Information (MI) model is employed for feature selection:

Mathematical Foundation: **:** Mutual information measures the dependency between variables. For feature $X_i$ and target Y, MI is calculated as:

$$I(X_i ; Y) = \sum_{x_i \in X} \sum_{y \in Y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)}$$

Where $p(x_i, y)$ is the joint probability of $x_i$ and y, and $p(x_i)$ and $p(y)$ are marginal probabilities.
Selection Process: Features are ranked based on their MI score, and the top features are selected for model training.

### 3.3 Model Training and Evaluation:

Various ML models are trained and evaluated:

Models Used: Logistic Regression (LR), Gaussian Naive Bayes (GNB), Decision Tree (DT), Ada Boost (AB), Random Forest (RF), Support Vector Machine (SVM).

Evaluation Metrics: Models are assessed using accuracy, precision, F1 score, and ROC-AUC. The ROC-AUC score is calculated as:

$$\text{ROC-AUC} = \int_0^1 TPR(t)\, dFPR(t)$$

where TPR is the True Positive Rate and FPR is the False Positive Rate.

### 3.4 Web Interface Development

A Django-based web interface is developed:

Functionality: Users can input clinical data to receive real-time PCOS predictions.

Integration: The best-performing models are integrated into the interface for accurate and timely predictions**.**

## 4. Results and Analysis

The performance of different ML models is evaluated:

Accuracy: Random Forest and Ada Boost achieved the highest accuracy of 94%.

Feature Selection Impact: Mutual Information significantly improved model performance by selecting relevant features.

Web Interface: The Django-based system provides a user-friendly platform for early PCOS detection, incorporating the top-performing models.

**Model Performance Diagram (Table 1)**

| Model | Accuracy (%) | Precision | F1 Score | ROC-AUC |
|---|---|---|---|---|
| Logistic Regression (LR) | 87 | 0.85 | 0.86 | 0.89 |
| Gaussian Naive Bayes (GNB) | 84 | 0.82 | 0.83 | 0.85 |
| Decision Tree (DT) | 90 | 0.89 | 0.9 | 0.92 |
| Ada Boost (AB) | 94 | 0.93 | 0.94 | 0.96 |
| Random Forest (RF) | 94 | 0.93 | 0.94 | 0.96 |
| Support Vector Machine (SVM) | 88 | 0.87 | 0.88 | 0.9 |

**Table 1** illustrates the accuracy, precision, F1 score, and ROC-AUC of various ML models used for PCOS prediction.

## 5. Conclusion

This study demonstrates the effectiveness of machine learning in early PCOS detection. By utilizing advanced algorithms and feature selection techniques, the proposed web-based system provides a practical tool for predicting PCOS with high accuracy. Future research could focus on expanding the dataset, incorporating additional features, and exploring more sophisticated ML techniques to further enhance prediction performance.

**References:**

1. Stein, I. F., & Leventhal, L. J. (1935). Polycystic Ovary Syndrome: A Clinical Study. Journal of Clinical Endocrinology.

2. Silva, (2023). BorutaShap and Random Forest for PCOS Prediction, Journal of Endocrinological Investigation].

3. Khanna, V. V., (2023). Multi-Stacking ML for PCOS Detection.

4. Bharati, S., et al. (2023). Hybrid RFLR Model for PCOS Diagnosis. [orman R. J., Dewailly D., Legro R.S., Hickey T.E. Polycystic ovary syndrome. Lancet. 2007; 370:685–697. doi: 10.1016/S0140-6736(07)61345-2.

5. Yasmine A. Abu Adla, Dalia G. Raydan, Mohammad-Zafer J. Charaf, Roua A. Saad, Jad Nasreddine, Mohammad O. Diab,(2021) Automated Detection of Polycystic Ovary Syndrome Using Machine Learning Techniques IEEE *Xplore*, 2021

6. Hassan, M. M., (2023). Random Forest for PCOS Detection.

7. Amsy Denny,Anita Raj, Ashi Ashok, C Maneesh Ram, Remya George (2019) i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques, TENCON, IEEE.

8. Homay Danaei Mehr, H. Polat Health technology,12,137-150, (2021), Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques.