SHORT ARTICLE

Machine Learning Models for the Development of a Probabilistic Screening Tool for Polycystic Ovary Syndrome

Hanumanth Narni[1], Vasudeva Rao Ananthasetty[2], SD Jilani[3], P Sree Sailaja[4]

[1,2,3]Department of Statistics, Acharya Nagarjuna University, Guntur, Andhra Pradesh

[4]Department of Obs. & Gynaecology, GITAM Institute of Medical Sciences and Research, Visakhapatnam, Andhra Pradesh

CORRESPONDING AUTHOR

Hanumanth Narni, Department of Statistics, Acharya Nagarjuna University, Guntur, Andhra Pradesh 522510

Email: hanumanth.narni@gmail.com

CITATION

Narni H, Ananthasetty VR, Jilani SD, Sailaja PS. Machine Learning Models for the Development of a Probabilistic Screening Tool for Polycystic Ovary Syndrome. Indian J Comm Health. 2025;37(2):339-342.

https://doi.org/10.47203/IJCH.2025.v37i02.027

ARTICLE CYCLE

Received: 16/10/2024; Accepted: 03/04/2025; Published: 30/04/2025

ABSTRACT

Background: Polycystic ovary syndrome (PCOS) is a common hormonal disorder in women of reproductive age that can lead to infertility and other long-term health problems. Early detection using simple, non-invasive tools is important to support timely intervention and improve outcomes. **Objective:** The study aimed to compare the performance of decision tree and naive Bayes models in predicting the likelihood of PCOS using non-invasive clinical features. **Methodology:** The study included 100 diagnosed cases of PCOS and 100 controls based on ultrasonographic findings. Clinical and lifestyle information was collected through a structured questionnaire. The models were evaluated using accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve. Five-fold cross-validation was used for validation. **Results** The decision tree model had high training accuracy but lower test accuracy, indicating overfitting. The naive Bayes model showed more consistent performance with 81 percent test accuracy and an F1 score of 0.81. **Conclusion:** The naive Bayes model shows promise as a simple, non-invasive screening tool for early identification of PCOS, particularly in primary care and low-resource settings.

KEYWORDS

Polycystic Ovary Syndrome; Screening; Machine Learning; Decision Tree; Naive Bayes

INTRODUCTION

Preventive medicine identifies high-risk individuals early, reducing disease burden through timely intervention. Machine learning (ML) models, capable of handling vast data, offer precise predictions, particularly in screening for complex conditions like polycystic ovary syndrome (PCOS). Affecting 6–10% of reproductive-aged women, PCOS leads to reproductive, metabolic, and psychological complications if untreated.(1,2) Traditional diagnostics often involve invasive tests like hormonal assays and imaging, which are impractical in low-resource settings.(3) This necessitates the development of cost-effective, non-invasive screening tools. ML models utilizing features such as body mass index (BMI), menstrual irregularities, and family history can address this gap.(4)

This study aim is to evaluates decision tree (DT) and naive Bayes (NB) models for predicting PCOS. Both models are simple and interpretable for clinical use. DT models offer clear decision paths, while NB models excel in handling categorical data through probabilistic frameworks.(5,6)

MATERIAL & METHODS

Study Design: This research employs a case-control study design to investigate the characteristics of Polycystic Ovary Syndrome (PCOS). Participants are categorized into two groups: cases and controls. Cases are women who have been diagnosed with PCOS, while controls are women without a PCOS diagnosis. Ultrasound scans are utilized as a diagnostic tool to determine the presence or absence of PCOS in each participant.

Sample Size Calculation: Sample size was estimated using a precision-based formula for the area under the ROC curve (AUC).(7) Based on an anticipated AUC of 0.8 and a desired margin of error of 0.1, the minimum required sample size was calculated to be 46 cases and 46 controls. To improve the statistical power and precision of the findings, the study included a sample of 100 cases and 100 controls.

Data collection: A structured validated questionnaire was administered to gather demographic and clinical data from participants. The dependent variable in this study was PCOS status (Positive or Negative), while the independent variables included abnormal body mass index (BMI), weight gain, physical inactivity, irregular menstrual cycles, acne, abnormal waist-hip ratio (WHR), stress, hirsutism, skin darkening, hair loss, and fast-food consumption.

Statistical Analysis: Pre-processing and Feature selection: Chi-square filter method was employed to select the most relevant features. Out of the eleven independent variables, ten were found to be significantly associated with PCOS at a 5% level of significance. These significant features included Abnormal BMI, Abnormal WHR, Physical Inactivity, Weight Gain, Hair Loss, Hirsutism, Irregular Cycles, Stress, Skin Darkening, and Fast-Food Consumption, with the exception of Acne. To address the potential issue of multicollinearity among the selected features, a Variance Inflation Factor (VIF) analysis was conducted. The results indicated that there was no significant multicollinearity present, as all VIF values were below 5.

Machine Learning models: Two machine learning models, namely Decision Tree (DT) and Random Forest (RF), were employed for the prediction of PCOS.

Training and Testing: To evaluate the performance of the machine learning models, the dataset was divided into training and testing sets. Eighty percent of the data was used to train the models, while the remaining twenty percent was reserved for testing and assessing their predictive accuracy.

Model Evaluation Metrics: The comparative analysis employs a comprehensive set of performance metrics, including Accuracy, Precision, Recall, F1 Score, and Area Under the Curve (AUC) from the Receiver Operating Characteristic (ROC) curve. To enhance the reliability of the model evaluations, 5-fold cross-validation is utilized, dividing the dataset into five subsets. This allows each subset to serve as a testing set while the remaining four are used for training, thereby reducing the potential for overfitting and providing a more accurate assessment of the models' generalization capabilities.

RESULTS

The age distributions between the PCOS-negative and PCOS-positive groups are well-matched, with similar means (26.76 vs. 26.51) and standard deviations (2.48 vs. 2.50). Confirming that age matching was effectively managed in this case-control study.

Table 1. Association of Independent variables with PCOS risk

Independent Variable	COR (95% CI)	P-value (Chi-square)
Hirsutism	8.6 (4.4 - 16.7)	<0.001
Weight gain	8.6 (4.5 - 16.2)	<0.001
WHR	8.5 (3.6 - 20.2)	<0.001
Irregular cycles	8.1 (4.2 - 15.6)	<0.001
Hair loss	4.9 (2.5 - 9.4)	<0.001
BMI	4.2 (2.3 - 7.7)	<0.001
Fast food	3.3 (1.8 - 6.1)	<0.001
Stress	2.6 (1.4 - 4.5)	0.001
No regular exercise	2.1 (1.2 - 3.7)	0.013
Skin darkening	2.0 (1.1 - 3.5)	0.021
Acne	1.0 (0.6 - 1.8)	0.886

Table 1. shows the association between various independent variables and the risk of developing PCOS, represented by crude odds ratios (COR). Hirsutism and weight gain are the strongest predictors, both with an OR of 8.6, indicating that women with these conditions are over 8 times more likely to have PCOS compared to those without. Other significant predictors include waist-to-hip ratio (WHR) (OR: 8.5), irregular cycles (OR: 8.1), and hair loss (OR: 4.9). Factors like BMI (OR: 4.2), fast food consumption (OR: 3.3), and stress (OR: 2.6) also show a notable increase in PCOS risk. Less significant but still relevant predictors include lack of regular exercise (OR: 2.1) and skin darkening (OR: 2.0). Acne, however, does not show a significant association with PCOS (OR: 1.0, p = 0.886). Overall, Hirsutism, weight gain, abnormal WHR, and irregular cycles showed strong associations with PCOS, each with odds ratios above 8. These features reflect key hormonal and metabolic imbalances and are commonly observed in clinical practice. Their strong predictive value and ease of identification make them important for early screening. Emphasizing such clinically relevant, non-invasive indicators can support timely diagnosis, especially in primary care settings.

Table 2. Average 5-fold cross validated evaluation metrics of DT, NB of train and test data

Data	ML model	Accuracy	Precision	Recall	F1 score	AUC
Train	DT	0.95	1.00	0.91	0.95	0.95
	NB	0.80	0.79	0.80	0.80	0.80
Test	DT	0.77	0.79	0.73	0.75	0.77
	NB	0.81	0.80	0.82	0.81	0.80

Table 2. represents that the performance metrics for the decision tree (DT) and naive Bayes (NB) models indicate differing levels of effectiveness in predicting outcomes across training and testing datasets. The DT model shows high accuracy on the training set (0.95) with perfect precision (1.00) and a recall of 0.91, resulting in an F1 score of 0.95 and an AUC of 0.95, suggesting it captures the positive class very well in the training phase. However, its performance declines on the test set, with accuracy dropping to 0.77, precision at 0.79, recall at 0.73, and an F1 score of 0.75, indicating potential overfitting. In contrast, the NB model performs relatively well on the test set, achieving an accuracy of 0.81, a precision of 0.80, a recall of 0.82, and an F1 score of 0.81, with an AUC of 0.80, suggesting a more balanced performance across both datasets. Overall, while the DT model performs better in training, the NB model demonstrates consistent predictive ability on the test set.

Discussion

In comparing the results of the current study with findings from relevant studies, we observe notable differences and similarities in the performance of machine learning models for PCOS prediction. For instance, a study(8) utilized decision trees and achieved an accuracy of 0.92, which aligns closely with the high accuracy of 0.95 found in our training set. However, their decision tree model showed a drop to 0.72 in test accuracy, comparable to our DT results (0.77), suggesting a consistent issue with overfitting in decision tree models when applied to unseen data. Conversely, the naive Bayes model in our study had a test accuracy of 0.81, which is slightly higher than the 0.75 reported by another study,(4) indicating that the NB model may provide better generalization compared to decision trees in some contexts. Furthermore, studies (9-10) highlighted the efficacy of ensemble methods, reporting accuracies of 0.85 and 0.84, respectively, suggesting that combining multiple models may yield better predictive performance than single models like DT and NB. In contrast, a recent investigation (3) achieved a test accuracy of 0.79 using logistic regression, reinforcing our findings that while naive Bayes performs well, more sophisticated models or ensemble approaches may be required for improved accuracy.

Conclusion

In comparing the performance of the decision tree (DT) and naive Bayes (NB) models for predicting PCOS, the naive Bayes model proves to be more effective for practical applications. While the DT model demonstrates high accuracy and F1 scores in the training phase, it shows a decline in accuracy on the test set. In contrast, the NB model achieves a higher accuracy of 0.81 on the test set, making it the preferable choice for reliable PCOS prediction.

These models can be integrated into mobile health applications to enable PCOS risk screening using simple, non-invasive inputs. Such apps can guide self-assessment or assist primary healthcare workers in identifying high-risk individuals. This approach is especially useful in remote areas with limited access to diagnostic facilities. It supports early detection, timely referrals, and improved reproductive healthcare delivery.

Recommendation

Future research could focus on validating the model in larger and more diverse populations to enhance its generalizability. Additionally, incorporating other relevant features such as family history and exploring ensemble machine learning models may further improve predictive accuracy and clinical utility.

Limitation of the Study

The study was hospital-based which may limit the generalizability of the findings to the broader population.

Relevance of the Study

This study offers a non-invasive, machine learning-driven approach for the preliminary screening of individuals at risk for PCOS

Authors Contribution

All authors have contributed equally.

Financial Support and Sponsorship

Nil

Conflict of Interest

There are no conflicts of interest.

ACKNOWLEDGEMENT

We gratefully acknowledge the participants for their valuable time and cooperation in this study

DECLARATION OF GENERATIVE AI AND AI ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The authors haven't used any generative AI/AI assisted technologies in the writing process.

REFERENCES

1. Teede HJ, Misso ML, Costello MF, et al. Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. Hum Reprod. 2018;33(9):1602–1618.
2. Escobar-Morreale HF. Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment. Nat Rev Endocrinol. 2018;14(5):270–284.
3. Zhang J, Li M, Chen Q, et al. Predictive models for polycystic ovary syndrome based on machine learning techniques. J Biomed Inform. 2019;93:103149.
4. Nandi A, Chen Z, Patel R, et al. Machine learning-based identification of clinical phenotypes and biomarkers for PCOS diagnosis using non-invasive features. J Clin Endocrinol Metab. 2020;105(5):1728–1736.
5. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1):81–106.
6. Lewis DD. Naive (Bayes) at forty: The independence assumption in information retrieval. In: European Conference on Machine Learning. Springer; 1998:4–15.
7. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform. 2014;48:193–204.
8. Jin Z, Wang Y, Zhang X, et al. Role of machine learning in the diagnosis and prediction of reproductive disorders: A comprehensive review. Front Endocrinol (Lausanne). 2021;12:697962.
9. Roy KK, Kumar N, Saxena A, et al. Predictive model for PCOS using machine learning techniques with non-invasive data. J Obstet Gynaecol Res. 2021;47(2):760–766.
10. Liang B, Liu X, Wang Y, et al. A machine learning model for polycystic ovary syndrome diagnosis based on phenotypic and genetic data. Endocr Connect. 2021;10(7):817–827.