

Polycystic Ovary Syndrome (PCOS) Prediction System Using Machine Learning: An Age-Wise Analysis with SHAP-Based Feature Interpretation

Shraddha Rangsubhe¹ and Ms. Yogita kadbane²

¹ Student, Master of Computer Applications

² Assistant Professor, Master of Computer Application

Sadhu Vaswani Institute of Management Studies for Girls, Pune
Savitribai Phule Pune University, Pune, Maharashtra, India.

Abstract: Polycystic Ovary Syndrome or PCOS is a hormonal disorder that commonly affects women of reproductive age and is associated with many reproductive, metabolic and physical health problems. The symptoms and signs can differ a lot from one person to another, so the task of predicting PCOS early is hard. Machine learning techniques give a data-driven way to spot patterns in clinical data, hormones, lifestyle habits and symptom reports that may be linked to PCOS. In this study, a machine learning framework was built using 2,000 records. Four classification algorithms - Logistic Regression, Random Forest, Decision Tree and XGBoost - were. Compare to find which one predicts PCOS best. Besides comparing models, an age 1 analysis was done to see how common PCOS is and how symptoms spread across age groups. SHAP or Shapley Additive Explanations was used on the XGBoost model to explain which features drive its predictions. Combining machine learning prediction, age 1 analysis and SHAP feature explanation gives a way to understand PCOS patterns and help with data-driven PCOS prediction.

Keywords: Polycystic Ovary Syndrome (PCOS), Machine Learning, XGBoost, Random Forest, Decision Tree, Logistic Regression, PCOS Prediction, Age-wise Analysis, SHAP, Feature Importance..

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a hormonal disorder that affects women who are able to have children. This condition is linked to problems involving reproduction, metabolism and the body. Typical signs of PCOS are periods gaining weight too much hair on the body losing hair on the head, pimples and dark areas on the skin. The signs and how bad they are can be different for each person. This makes it hard to find PCOS early and correctly. Usually doctors look at the history the way someone feels check hormone levels and do other medical tests. These ways are helpful.. When many different things are involved, like medical information, hormone levels, lifestyle and symptoms the process of guessing who has PCOS can be very complex.



Polycystic ovary syndrome (PCOS)

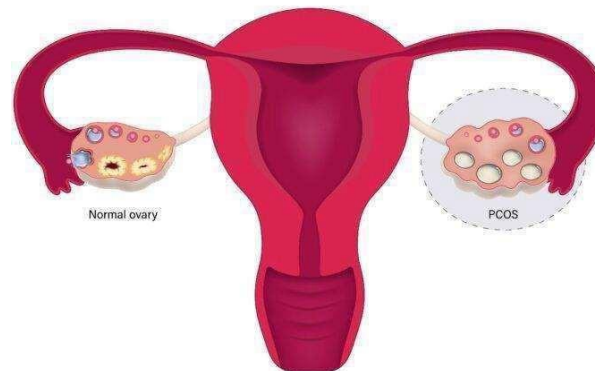


Figure 1. PCOS and Normal Ovary (ref by-<https://istockphoto.com/illustrations/pcos>)

Machine learning-based prediction can help find people who might be more likely to have PCOS and could help doctors take action quickly. Different ways of classifying data can give results based on the data that is used. So looking at machine learning models is important to choose the best one for predicting PCOS.

In this study a machine learning approach is created using data from 2,000 patients with details about their history, hormone levels, lifestyle and symptoms. Four ways of classifying data called Logistic Regression, Random Forest, Decision Tree and XGBoost are. Compared based on how well they predict PCOS. The models had accuracy rates of 93.72%, 98.49% and 98.99%. Among all the models tested XGBoost had the accuracy of 98.99%.

Along with comparing the models the study also looks at age groups to see how common PCOS is and how the symptoms are spread across different ages. The patients are put into groups like 20-24 years old 25-29 years old 30-34 years old 35-39 years old and 40 years and older. This part of the study gives an idea of how often PCOS happens and how symptoms like gaining weight too much hair, dark skin losing hair and pimples change with age.

Also SHAP is used to explain the predictions made by the XGBoost model. SHAP helps find out which parts of the data affect the predictions the most and helps understand what factors are important in predicting PCOS. So, the study uses comparing machine learning models looking at age groups and explaining predictions, with SHAP to create a way of predicting PCOS using patient data.

II. LITERATURE REVIEW

Previous Research has shown that machine learning methods can be useful for predicting and finding Polycystic Ovary Syndrome (PCOS) using information from patients like their history, age, hormone levels and symptoms. Different types of learning methods have been tested to find patterns that are linked to PCOS and to make predictions more accurate. Some studies have found results when using methods like Logistic Regression, Decision Tree, Random Forest and boosting methods.

Random Forest and Decision Tree models are often used because they can deal with different patient features and find complex patterns. Logistic Regression is easy to use. Gives clear results for classifying patients. XGBoost has become popular because it works well with data that is organized in tables. More recent studies have focused on techniques that explain how artificial intelligence makes decisions. These techniques help understand what factors are important, in the predictions made by machine learning.

Based on those results this study looks at Logistic Regression, Random Forest, Decision Tree and XGBoost for predicting PCOS using information. The study also looks at how important each feature's how symptoms change with age and uses SHAP analysis to make the models easier to understand. This combined way of working tries to give predictions and also a better understanding of the features that are linked to PCOS.



III. RESEARCH GAP

Previous research has used machine learning techniques for PCOS prediction. Some aspects remain less explored. Many studies focus mainly on prediction accuracy while limited attention is given to understanding the features influencing the predictions. Age-wise variations in PCOS-related symptoms are also frequently analyzed. In addition the combination of machine learning models with feature importance and SHAP-based explainability is limited. Therefore this study aims to address these gaps by combining model comparison feature importance analysis, age-wise symptom analysis and SHAP analysis, for PCOS prediction.

IV. METHODOLOGY

(A) Data Description

The dataset used in this study consists of 2,000 patient records with 44 features related to Polycystic Ovary Syndrome (PCOS). The dataset contains clinical, hormonal, lifestyle, physical and symptom-related information. The PCOS(Y/N) attribute is considered the target variable for predicting whether a patient is affected by PCOS. The diverse set of features provides information, about factors associated with PCOS and is used for developing and evaluating machine learning based prediction models. The study was implemented using Python in Jupyter Notebook, with Pandas, NumPy, Scikit-learn, XGBoost, Matplotlib, Seaborn and SHAP used for data processing, visualizations, model development, evaluation and interpretability.

Table 1. Dataset Features

Feature No.	Feature Name	Feature No.	Feature Name	Feature No.	Feature Name	
0	1	SL No	16	No. of abortions	31	Weight gain(Y/N)
1	2	Patient File No.	17	I beta-HCG(mIU/mL)	32	hair growth(Y/N)
2	3	PCOS (Y/N)	18	II beta-HCG(mIU/mL)	33	Skin darkening (Y/N)
3	4	Age (yrs)	19	FSH(mIU/mL)	34	Hair loss(Y/N)
4	5	Weight (Kg)	20	LH(mIU/mL)	35	Pimples(Y/N)
5	6	Height(Cm)	21	FSH/LH	36	Fast food (Y/N)
6	7	BMI	22	Hip(inch)	37	Reg.Exercise(Y/N)
7	8	Blood Group	23	Waist(inch)	38	BP _Systolic (mmHg)
8	9	Pulse rate(bpm)	24	Waist:Hip Ratio	39	BP _Diastolic (mmHg)
9	10	RR (breaths/min)	25	TSH (mIU/L)	40	Follicle No: (L)
10	11	Hb(g/dl)	26	AMH(ng/mL)	41	Follicle No: (R)
11	12	Cycle(R/L)	27	PRL(ng/mL)	42	Avg. F size (L) (mm)
12	13	Cycle length(days)	28	Vit D3 (ng/mL)	43	Avg. F size (R) (mm)
13	14	Marriage Status (Yrs)	29	PRG(ng/mL)	44	Endometrium (mm)
14	15	Pregnant(Y/N)	30	RBS(mg/dl)	45	

Table 1. Dataset Features Used in the Study



(B) Data Preparation

The dataset was first looked at to find values, duplicate records check data types and remove any attributes that were not needed. It had 2,000 records and 44 columns with information on clinical factors, hormones, lifestyle, reproductive history and symptoms. There were three values in the marriage Status (Yrs) feature but no duplicate records were found. The dataset was then. Prepared to make sure it was ready for machine learning development.

```
Original Dataset Shape: (2000, 44)

----- MISSING VALUES BEFORE CLEANING -----
Marriage Status (Yrs)    3
dtype: int64

Total Missing Values: 3

----- DUPLICATES -----
Duplicate Rows: 0

----- AFTER CLEANING -----
Final Dataset Shape: (1997, 44)
Missing Values: 0
Null Values: 0
None Values: 0
Duplicate Rows: 0

----- PCOS COUNT -----
PCOS (Y/N)
0    1392
1     605
Name: count, dtype: int64
```

Figure 3. Dataset Information and Data Quality Check

The SL.. Patient File No fields were removed because they were identification details and not useful for prediction. The PCOS(Y IN) variable was set as the target variable. All the other features-clinical, hormonal, lifestyle, reproductive and symptom-related-were used as input variables, for prediction.

Any string-based numerical values were changed into numerical formats before the model was trained and tested.

(B) Train-Test Split

After cleaning the data the final dataset contained 1,989 patient records, with 44 features and no missing or duplicate values. The dataset was split into 80% training data. 20% Testing data so that we could see how well the machine learning models work. A total of 1,591 records were used for training and 398 records were used for testing. The training set contained 1,111 -PCOS records and 480 PCOS records while the testing set contained 278 non-PCOS records and 120 PCOS records.



```

===== PCOS DISTRIBUTION =====
PCOS No : 1389
PCOS Yes : 600

===== TRAIN TEST SPLIT =====
Total Records : 1989
Training Records : 1591
Testing Records : 398

Training Shape: (1591, 43)
Testing Shape : (398, 43)

Training PCOS Distribution:
PCOS (Y/N)
0    1111
1     480
Name: count, dtype: int64

Testing PCOS Distribution:
PCOS (Y/N)
0     278
1     120
Name: count, dtype: int64

```

(C) Machine Learning Models

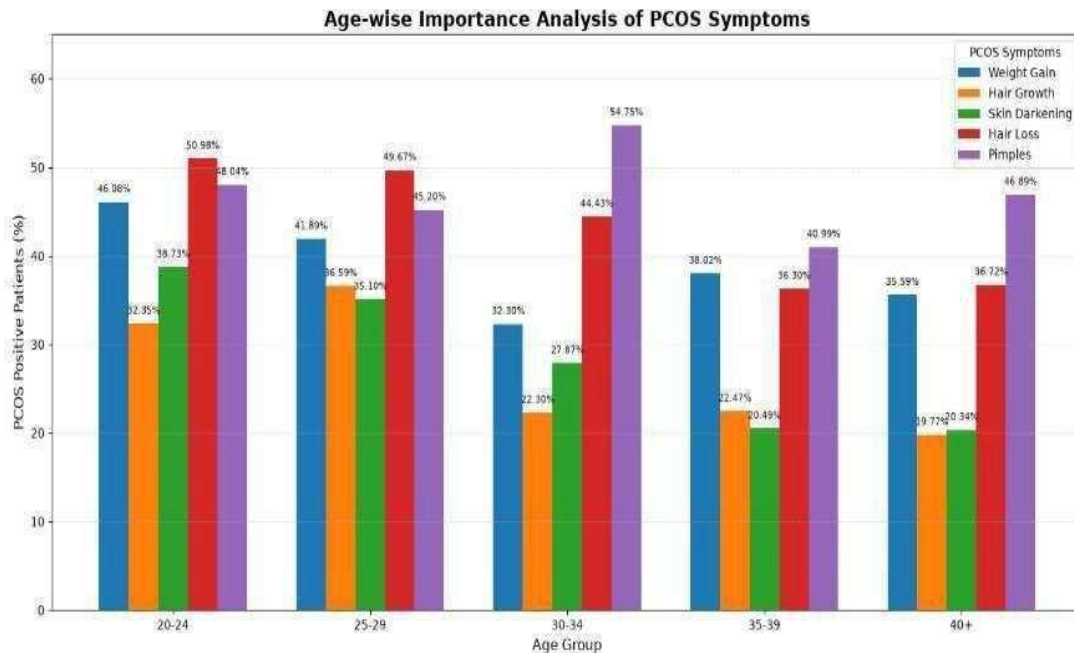
Four supervised machine learning models were used for PCOS prediction: Logistic Regression, Random Forest, Decision Tree, and XGBoost. These models represent linear, tree-based, and ensemble learning approaches for binary classification. All models were trained on the training dataset and evaluated on the testing dataset using Accuracy, Precision, Recall, F1 -Score, and ROC-AUC. The model performances were compared to identifying the best-performing model for PCOS prediction.

	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
0	Logistic Regression	93.72	92.79	85.83	89.18	0.9721
1	Random Forest	98.49	100.00	95.00	97.44	0.9999
2	Decision Tree	98.24	100.00	94.17	97.00	0.9708
3	XGBoost	98.99	100.00	96.67	98.31	0.9995

(D) Age-wise Importance Analysis

Age groups were examined to see how PCOS and its symptoms spread across ages. The patients were sorted into five age groups: 20-24 25-29 30-34 35-39 and 40 years and older. The study looked for differences in PCOS symptoms such as weight gain, hair growth, darker skin, hair loss and pimples, among the age groups.





(E) Evaluation

The measures used to check how well the machine learning models worked in this study are Accuracy, Precision, Recall and F-Score. To properly check how well the models predicted, four basic parts are looked at: Positive (TP) True Negative (TN) False Positive (FP) and False Negative (FN). True Positive (TP) is when the model correctly says a patient has PCOS. True Negative (TN) is when the model correctly says a patient does not have PCOS. False Positive (FP) is when the model says a patient has PCOS but they do not. False Negative (FN) is when the model says a patient does not have PCOS but they do. These parts are used to calculate the measures and check how well the models can tell the difference between PCOS cases and non-PCOS cases.

The measure terms are then used to build the performance evaluation metrics:

Accuracy: Accuracy shows how many of all the cases were classified correctly. This includes both PCOS and - PCOS cases.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Precision: Precision shows how many of the cases the model said were PCOS were actually PCOS. Higher precision means wrong predictions of PCOS.

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

Recall: Recall shows how well the model found patients who really had PCOS. Higher recall means the model missed PCOS cases.

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

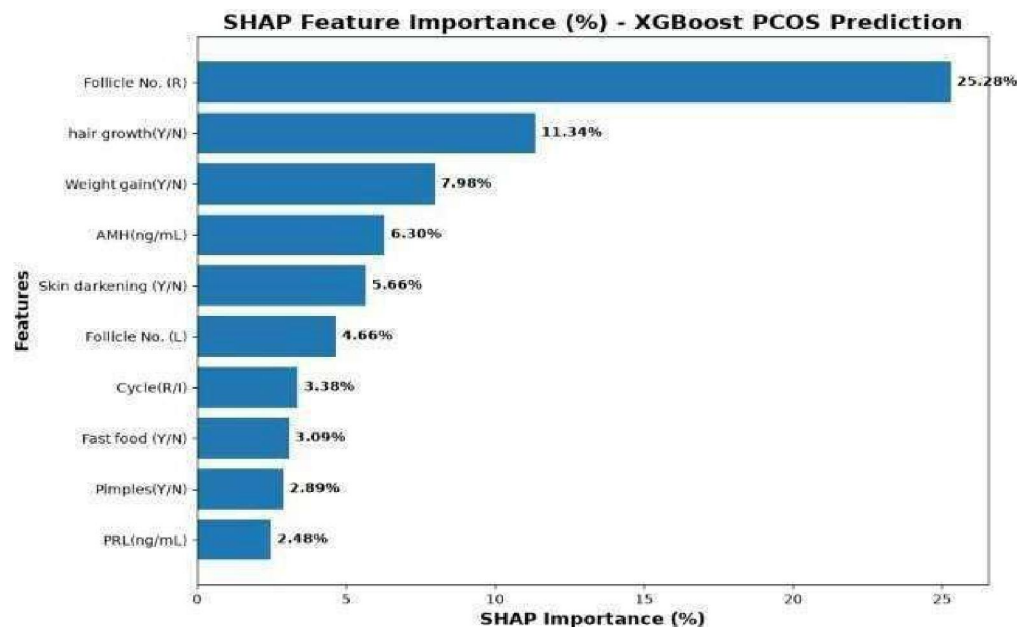
F-Score: F-Score is the average of Precision and Recall. It gives a look, at how well the model works when both precision and recall are important.

$$\text{F-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

(F) SHAP Analysis

SHAP (Shapley Additive explanations) analysis was used on the XGBoost model to see why the model says a person has PCOS. I used SHAP values to find which clinical and demographic details matter most for PCOS. When a SHAP value is large in terms that detail pulls the prediction strongly. A positive SHAP value nudges the prediction toward PCOS while a negative SHAP value pulls it away, from PCOS. This work makes the XGBoost model clearer and more open.





V. FINDINGS AND RESULTS

I found that the feature importance analysis highlighted Follicle No. (R) Follicle No. (L) Cycle (R/I) AMH (ng/mL) and Cycle Length (days) as the five important features for PCOS prediction. Among these follicle-related features had the importance showing that follicle-related features play a major role in the models predictions. The age-wise analysis revealed that PCOS-related symptoms varied across age groups. Hair Loss was most common in the 20 and 2 age groups whereas Pimples were most frequent in the 30 and 40+ age groups. Overall Hair Loss and Pimples were noticeable, in age groups. SHAP analysis added clarity to the XGBoost model by showing how each feature contributed to XGBoost predictions. Out of the four models tested XGBoost achieved the performance with 98.99% accuracy 100% precision, 96.67% recall, 98.31% F1-Score and 0.9995 ROC-AUC. These results show that machine learning can spot patterns linked to PCOS and offers a way to predict PCOS.

VI. LIMITATIONS

This work is based on a Kaggle dataset containing 2,000 patient records so the results may not represent all populations or clinical settings. I believe the model performance may also vary when applied to data collected from sources. The analysis was evaluated using a train-test split and validation on independent datasets would provide stronger evidence of model reliability. In SHAP the model explains its predictions. Does not establish a direct clinical relationship, between the features and PCOS.

VII. CONCLUSION

From the analysis done in this study machine learning showed potential for predicting PCOS using clinical and demographic features. Among the four models XGBoost gave the overall performance on the testing data. Feature importance and SHAP analysis showed that follicle-related cycle-related and hormonal features played a role in the predictions. The age-wise analysis also showed differences in PCOS-related symptoms across age groups. The results suggest that machine learning can help in identifying patterns related to PCOS and support prediction. The developed model can be useful as a tool for further clinical evaluation. However it should not be considered a replacement, for medical diagnosis.



REFERENCES

1. CM-037 Divy PCOS Dataset, Kaggle. Kaggle: PCOS Dataset
2. LS. Silva et al., "Polycystic ovary syndrome: clinical and laboratory variables related to new phenotypes using machine-learning models," Journal of Endocrinological Investigation, vol. 45, pp. 497-505, 2022
3. "Machine-Aided Self-Diagnostic Prediction Models for Polycystic Ovary Syndrome: Observational Study," JMIR Medical Informatics, 2022.
4. A. M. K. et al., "Machine Learning-Based Prediction of Polycystic Ovary Syndrome Using Clinical and Biochemical Features," 2019.
5. "SPOSDS: A Smart Polycystic Ovary Syndrome Diagnostic System Using Machine Learning, 11 Expert Systems with Applications, 2022.
6. "Exploring the Dominant Features and Data-Driven Detection of Polycystic Ovary Syndrome through Modified Stacking Ensemble Machine Learning technique," 11 Heliyon, 2023.
7. Yao, X. Yu, Y. Zhang, J. Chen, X. Zhu, C. Zhang, and T. Jijun, "Development and validation of an explainable machine learning and nomogram model for early detection and risk stratification of polycystic ovary syndrome: a multicenter study," Frontiers in Endocrinology, vol. 16, 1719631, 2025.

