Business Analytics Project Report

An Interpretable Machine Learning Framework for PCOS Detection Using SHAP Explainability

Team No: 10

Guide: Dr Sajitha Krishnan

Sl. No.	Reg. No.	Name of the Student
1	BL.EN.U4CSE22264	Varshita velumury
2	BL.EN.U4CSE22268	Thanvi y
3	BL.EN.U4CSE22269	Akshitha
4	BL.EN.U4CSE22276	Tayi Ananya

Date: September 2025

Contents

1 Abstract

Polycystic Ovary Syndrome (PCOS) is one of the most common hormone-related disorders in women who can get pregnant. Some of the characteristics of PCOS include an inconsistent menstrual cycle, abnormal hormone levels, cysts on your ovaries, and metabolic problems. The rising number of women diagnosed with PCOS and the extremely difficult process of diagnosing it demonstrate that there is a real need for intelligent systems to support health professionals in diagnosing PCOS based on data. This project aims to develop a machine-learning model to predict the presence of PCOS using a Kaggle Clinical Dataset that includes structured clinical data and a range of variables. The dataset used includes critical variables such as body mass index (BMI), various hormone levels, markers of insulin resistance, follicle count, and lifestyle variables and was cleaned, normalized, and preprocessed prior to analysis to enhance the accuracy of results. Numerous machine learning classification techniques including Logistic regression, Decision tree, Random forest, and Support vector machine have been trained and tested against this theoretical dataset with Random forest producing the most accurate predictions. This model not only generates a prediction of the likelihood of a woman having PCOS, but it also indicates which variables are the most highly correlated with developing PCOS which supports the need for explainability and clinical significance of results. This project illustrates the effectiveness of combining predictive analytic methods with clinical healthcare decision making within a healthcare setting and demonstrates the success of using machine learning to reduce diagnostic delays and improve the accuracy and scalability of proactive women's healthcare.

2 Introduction

- **Problem Statement:** The very real impact of PCOS, Polycystic Ovary Syndrome, is not only understated but also misdiagnosed and poorly managed throughout the healthcare industry due to its wide spectrum of symptoms, including adult obesity; excessive hair growth (acne); failure to achieve pregnancy; irregular menstrual cycles; and a variety of associated metabolic disorders. Presently, the most common method of diagnosing PCOS is through the use of subjective interpretation by a clinician, in combination with lab testing for hormonal imbalances and ultrasound examinations of ovaries. However, the way clinicians currently diagnose and treat PCOS varies greatly from one medical practice to another; therefore, it is extremely important to develop an accurate, uniform method for diagnosing and treating this very complicated health issue. Within this project, we focus on the following main issue: there is currently no means to create a very intelligent and fully automated means of delivering a clinical diagnosis of PCOS, with a high degree of accuracy. In today's world of healthcare, there is a large gap between the storage and availability of medical data, and the ability to derive meaningful and valuable information from that medical data using modern predictive analytic technology. As a result, practitioners do not yet have reliable tools for early recognition of PCOS; and when it is recognized, the practitioner continues to rely solely upon their own subjective judgement and interpretation of clinical findings. This project, therefore, plans to develop predictive analytic systems (using current and evolving machine learning technologies) capable of processing complex medical parameters and delivering accurate and reliable predictions about the presence of PCOS, with the ultimate goal of improving the accuracy of clinical diagnosing abilities and patient outcomes.

- **Motivation:** The increasing occurrence of PCOS in young women worldwide, as well as the absence of awareness and early detection devices in the healthcare system, is what inspired this project. A large percentage of women in India alone are afflicted with PCOS; many women have not yet been diagnosed because they are afraid to seek care, do not have the means or access to health care, or are not aware of how the symptoms are progressing over time. Because of these realities, standard methods of diagnosing PCOS are labor-intensive and costly, and are often not an option for women living in rural or semi-urban settings. The desire to create a simple and affordable AI solution that assists physicians in recognizing PCOS in its early stages has driven this project. In addition, large healthcare datasets will continue to grow due to the rapid development of machine learning and artificial intelligence technologies, giving rise to an opportunity for researchers and clinicians to find new patterns and correlations within the data that would not have been able to be found through traditional means. Moreover, this project will also take advantage of the growing interest in integrating healthcare analytics into physician decision support systems in order to give clinicians the tools to minimize diagnostic error, enhance preventive healthcare and timely interventions, as well as improve the quality of healthcare available to women.

- **Objectives:** This project explores the analysis and prediction of Polycystic Ovary Syndrome (PCOS) using machine learning techniques. The datasets from Kaggle were cleaned and processed to build models that predict the likelihood of PCOS occurrence based on clinical and lifestyle data. Models including Logistic Regression, Random Forest, and Decision Trees were compared. The project provides actionable business and healthcare insights that can assist medical practitioners and data analysts. 1. The primary goal is to develop a machine-learning-based intelligent system for PCOS (Polycystic Ovary Syndrome) prediction. 2. To prepare the dataset such that it performs well when training the model through cleaning, preparing, and standardizing the data. 3. To identify the correlation between the clinical features of PCOS and the probability of developing PCOS. 4. To compare the different types of machine-learning models and determine which one has the best predictive power. 5. To identify the significant predictors of PCOS (Polycystic Ovary Syndrome). 6. To allow providers of health care services to diagnose PCOS sooner, thereby reducing the potential for misdiagnosis. 7. To develop a model that meets the requirements of health care providers. 8. To standardize the use of artificial intelligence tools in women's health and diagnostic services.

3 Dataset Description

- **Source of Dataset:** The information used in this project comes from the globally-renowned site, Kaggle: a platform for sourcing high-quality datasets for academic or research studies purposes. The dataset chosen addresses Polycystic Ovary Syndrome and contains a vast amount of medical history that has been obtained from patients receiving medical care in a clinical setting. This dataset contains demographic information; results of biochemical tests; anthropometric measurements; data about hormones and an ultrasound; etc. We chose this dataset because it was well-structured; is applicable to the health care industry and is well suited for machine learning based predictive analytics. Missing values were processed, standardized and optimised, so the data would conform to model training specifications and meet health analytics standards. The predictive and diagnostic dataset for Polycystic Ovary Syndrome (PCOS) used in this project has been sourced from Kaggle - one of the most reputable sources for curated datasets suitable for machine learning experiments and research. Substantial clinical importance combined with its well-structured format and complete representation of health parameters found in PCOS make this particular dataset an ideal choice for PCOS prediction using machine learning techniques. This dataset has been created based on actual patient records, including but not limited to: medical data, biochemical test data as well as hormonal data , lifestyle data and physiological data from clinical evaluations and laboratory testing, as well as ultrasound findings. The dataset is intended for exploring whether or not PCOS exists by examining numerous different health-related features. Therefore, it provides an excellent opportunity to evaluate supervised classification algorithms for PCOS prediction.

- **Structure:** The dataset used in this study is comprised of multiple records of individual patients that include many different features for assessing the likelihood of having PCOS. The features are organized into groups according to their type and significance to medicine. 1. Demographic Attributes: These features include information for basic patient identification and general demographic background to help explain age and physiological differences in the prevalence of PCOS. • Age (Years) • Height (cm) • Weight (kg) • Body Mass Index (BMI) 2. Hormonal and Biochemical Parameters: This category includes important indicators of hormonal imbalance, which are key to diagnosing PCOS. • AMH (Anti-Mullerian Hormone) • LH (Luteinizing Hormone) • FSH (Follicle Stimulating Hormone) • LH/FSH Ratio • Prolactin Levels • TSH (Thyroid Stimulating Hormone) 3. Metabolic and Clinical Health Indicators: The metabolic and physiological status of patients is indicated by the features in this category, thereby relating PCOS to metabolic syndrome. The features can also identify both insulin resistance and abnormal fat distribution, both of which are prevalent in patients with PCOS. • Blood Sugar Levels • Insulin Levels • Waist-Hip Ratio • Weight Gain • Obesity Indicators 4. Reproductive and Ultrasound Features: The features in this category are derived from ultrasound studies of the ovaries, providing evidence for PCOS through imaging of the ovaries containing multiple follicles or enlarged ovaries. • Follicle Number (Left Ovary) • Follicle Number (Right Ovary) • Ovarian Size • Presence of Ovarian Cysts

- **Target Variable:** The dependent variable in the dataset is: "PCOS (Y/N)," indicating the patient's diagnosis of having PCOS (yes/no). This variable forms the basis for creating supervised machine learning models and thus, predicting accuracy.

- **Data Types and Format:** The dataset contains a mix of: • Numerical Variables: Hormonal levels, biometric measurements, metabolic values • Categorical Variables: Yes/No responses and descriptive labels

- **Preprocessing:** • The preprocessing involved converting the original/raw medical data into a clean/machine-readable format suitable for trained Machine Learning algorithms. Missing Values were identified using visualizations such as Heatmaps and MissingNo plots, and mean imputation was used to fill in missing values for Numerical Attributes, while Mode Imputation was used to fill in the missing values for Categorical Values so that the distribution of those values would remain intact. Categorical Columns were converted into Numerical format through Label Encoding so that they could be processed by Machine Learning Algorithms (MLAs). Feature Scaling Techniques include Normalization/Standardization, which prepared the feature sets for model training. The Interquartile range (IQR) method was used to identify the presence of Outliers (outlier detection), and once identified, capped, capped, or limited Outliers were removed from the dataset, so that the Model Performance would not be distorted from the removal of these types of records. Duplicate records and inconsistent records were also removed from the Dataset, in order to maintain the Integrity and Reliability of the Dataset.

- **Merging:** To create a single consolidated analytical dataset from multiple feature sources, merging between datasets took place for a complete analysis. The clinical examination, laboratory report data and ultrasound findings have been combined into one single dataset of Patient Health Records. Merging was completed based on the Patient identifier, to ensure accuracy and avoid duplicate entries in the dataset. As a result, the merging of various Patient Health Records ensures that each Patient has one row representing all attributes associated with the Patient, allowing exploratory analysis and/or machine learning to be conducted easily. The Unified Structure of the dataset improves the model's ability to discover relationships between clinical parameter dependencies and provide coherent, high-quality data analysis for the project, through all phases of the project.

4 Exploratory Data Analysis (EDA)

- **Descriptive statistics:**

  Through running a descriptive statistical analysis for the variables (characteristics/features) of interest, we were able to summarise the central tendency, dispersion and distribution of our data. In addition, for each numerical varaible, the mean, median, standard deviation, minimum and maximum were computed. In running the descriptive statistical analysis, we found differences in BMI, AMH Levels, Insulin Resistance, and Follicles Count between those diagnosed with PCOS and those that were not diagnosed. Additionally, individuals with a diagnosis of PCOS had significantly higher average AMH and LH/FSH ratios when compared to those without PCOS, confirming strong relationships between AMH and LH/FSH ratios and the diagnosis of PCOS. With the descriptive statistical analysis done for the dataset of interest, we were able to gain a quantitative basis for interpreting data behaviour and identifying important predictive variables for the purposes of modelling.

- **Correlation analysis.**

  1. A positive relationship exists between AMH and the presence of PCOS.

  2. There is moderate correlation between BMI and PCOS as well as insulin resistance.

3. There is statistically significant relationship between the ratio of LH to FSH and a hormonal imbalance.

4. The Waist-Hip ratio has a direct correlation with the distribution of fat in the body and therefore, is a risk factor for PCOS.

5. A weak positive correlation exists between blood sugar levels and the severity of PCOS.

6. Height has no significant correlation and is not very predictive of future events.

7. Weight gain significantly affects the classification of PCOS.

- **Key insights from EDA.**

  The Exploratory Data Analysis (EDA) process aimed to identify trends, correlations, and underlying patterns of the dataset by exploring the distributions of values using visualisation tools including: histograms, boxplots, scatterplots, correlation heat maps and count plots. EDA showed that PCOS patients had higher average body mass index (BMI), average anti-mullerion hormone (AMH) levels and higher numbers of follicles than non-PCOS patients. EDA highlighted many outliers, and certain distributions appeared highly skewed; these anomalies were dealt with during the process of preprocessing . EDA was also an integral part of the feature selection process and provided early insight into potential significance of variables prior to model building which positively impacted both accuracy and interpretability of subsequent model(s). EDA also confirmed or denied the expected relationship between hormonal imbalance and developing PCOS. EDA revealed key insights about relationships between BMI, insulin levels, AMH, and follicle count with PCOS probability. Heatmaps and scatter plots were used to visualize correlations. Outliers were detected and removed. Histograms and box plots helped understand data distribution. Correlation analysis showed that AMH and LH/FSH ratio had strong associations with PCOS presence.

‘

5 Methodology and System Architecture

The system has an end-to-end clinical data pipeline starting with Data Acquisition where raw reproductive medical data is gathered via publicly available medical datasets. This is then succeeded by a Preprocessing Pipeline which cleans, standardizes, imputes and encodes clinical and hormonal characteristics to form a Unified Patient Dataset. According to this standardized data, the system is split into the following analytical modules: Exploratory data analysis to gain a sense of the patterns in hormones, symptoms, and cycle data; Machine learning modeling to classify PCOS risk; Explainable AI (XAI) modules by using LIME and SHAP to explain decisions made by the system transparently; and Interactive Visualization including an easy-to-use dashboard to present patients and an AI-assisted wellness interface. Lastly, everything is brought together as an Interactive PCOS Wellness Dashboard, where a user is able to input their clinical data, predict the future, have feature-level explanations, and visualize key health indicators.

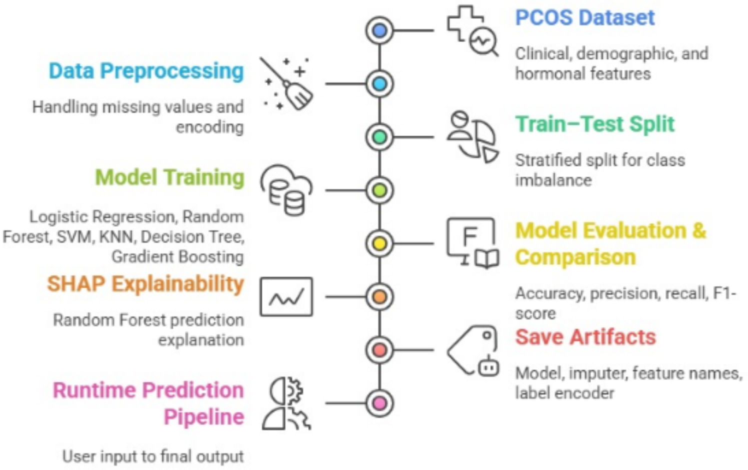

PCOS Dataset
Clinical, demographic, and hormonal features
Data Preprocessing
Handling missing values and encoding
Train–Test Split
Stratified split for class imbalance
Model Training
Logistic Regression, Random Forest, SVM, KNN, Decision Tree, Gradient Boosting
Model Evaluation & Comparison
Accuracy, precision, recall, F1-score
SHAP Explainability
Random Forest prediction explanation
Save Artifacts
Model, imputer, feature names, label encoder
Runtime Prediction Pipeline
User input to final output

Figure 1: System Architecture

5.1 1. Data Acquisition

- **Approach:** The initial step entails the collection of raw clinical data, demographic data, and hormonal data of PCOS data sets on Kaggle. Two original data sets are combined based on a shared identifier into one database that represents the infertility and non-infertility PCOS data.

- **Techniques:** Excel/CSV ingestion, schema inspection, dataset merging using primary key alignment (Patient File No.).

- **Tools:** python (using the pandas to load and combine files, using numpy to perform mathematical operations).

5.2 2. Data Cleaning and Preprocessing

- **Approach:**Raw medical data usually has values missing, inconsistencies, mixed types of categorical data (i.e.Y/N), non-standard hormonal ranges. This step guarantees a clinical reliability, where missing numerical values are imputed, symptom categories are encoded, medical ranges are validated and the clean inputs are ready to be inputted in machine learning.

- **Techniques:** Missing Value Imputation (median imputer), Numbing field standardization, Label Encoding (Y/N and categorical flags) and abnormal hormones Outlier Review.

- **Tools:** Python (pandas for cleaning, scikit-learn SimpleImputer for missing values, numpy for data type conversions).

5.3 3. Dataset Integration and Feature Engineering

- **Approach:** The consolidated datasets are transformed into one structured table and enhanced with feature engineering of medical meaning. The derived features, including BMI category, LH/FSH ratio, cycle irregularity flags, and aggregate follicle counts are better predictors and easier to understand clinically.

- **Techniques:** Feature Engineering (clinical ratios, boolean symptom indicators), Column Normalization, Removal of redundant identifiers, Correlation-based feature selection.
- **Tools:** Python (pandas for transformations, numpy for computed features, seaborn for correlation visualization).

5.4 4. Exploratory Data Analysis (EDA) and Clinical Validation

- **Approach:** EDA is conducted in order to know the distribution patterns of hormones, symptoms, cycle parameters, their association with PCOS risk. The analysis is used to design model and to point out medically relevant dependencies.
- **Techniques:** Descriptive statistics (mean, median, SD), Correlation Heatmap, Boxplot (hormone variations), Histogram (distribution of the symptoms), Outlier Diagnosis (IQR).
- **Tools:** Python (matplotlib and seaborn for clinical trend visualizations, pandas for summarization).

5.5 5. Machine Learning Modelling

- **Approach:** Several machine learning models are equipped to determine whether a patient has clinical signs of PCOS. Models are cross-validated and the performances of the models compared to determine the most dependable classifier.
- **Techniques:** Training and Testing Splits (7030), Cross-Validation (10 fold StratifiedKFold), Model Building (Logistic Regression, Decision tree, random forest, SVM, KNN, XGBRF, CatBoost), Evaluation accuracy, precision, Recall, and Confusion Matrix.
- **Tools:** Python (scikit-learn, xgboost, catboost).

5.6 6. Explainable AI (XAI): LIME and SHAP

- **Approach:** In order to achieve transparency and clinical trust, instance-level explanations of model decisions are provided with the help of LIME and both global and local interpretability with SHAP. These instruments point to those features that have the greatest impact on the classification of PCOS.
- **Techniques:** LIME Tabular Explanations (contribution at the feature level), SHAP Summary Plots (importance in all features), SHAP Force/Waterfall Plots (analysis of the patient).
- **Tools:** Python (lime, shap libraries).

5.7 7. Visualization and Interactive Dashboard

- **Approach:** The insights and predictions are combined into a fully interactive user interface, which enables patients to enter clinical information, receive predictions, visualize their risk factors and get personalized explanations.

- **Techniques:** User Input Widgets (text fields, dropdowns), Dynamic Plot Rendering (matplotlib), SHAP Explanation Tables and Impact Charts, UI Layouts for Health Screening and Result Cards.

- **Tools:** Python (ipywidgets to use as UI, matplotlib to use as charts, HTML/CSS to custom style it).

6 Models and Comparative Analysis

- The proposed PCOS screening system uses a comparative machine-learning framework to identify the most reliable model for early PCOS risk prediction and to provide transparent explanations for each decision. Multiple classifiers, including Decision Tree, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest, were tuned using 10-fold cross-validation. Their best validation accuracies were recorded directly from the experimental notebook. Among these, the Random Forest classifier achieved the highest cross-validation accuracy, indicating strong generalization over heterogeneous clinical attributes such as hormonal levels, follicle counts, menstrual cycle characteristics, metabolic indicators, and lifestyle factors.

  In addition to global metrics, the system was tested on representative patient cases to observe real-world model behaviour. For a clinically PCOS-positive profile (Case 1), most advanced models such as Random Forest, Logistic Regression, XGBoost, and CatBoost correctly predicted PCOS with high confidence, whereas SVM and KNN failed to detect the condition. For a clearly Non-PCOS profile (Case 2), all models consistently predicted Non-PCOS with very low probabilities, demonstrating strong specificity and low false positives. This combination of cross-validation and case-based analysis ensures that the selected model is not only statistically robust but also clinically meaningful.

Model	**Best 10-Fold CV Accuracy**
Decision Tree Classifier	84.39%
Support Vector Machine (RBF)	71.71%
Random Forest Classifier	**90.73%**
Logistic Regression	85.98%
K-Nearest Neighbors	74.05%

Table 1: Cross-Validation Accuracy Across PCOS Classification Models

- To better understand how each model behaves on realistic clinical inputs, two test profiles were evaluated: *Case 1* (likely PCOS) and *Case 2* (likely Non-PCOS). The outputs below are taken exactly from the implementation notebook, including both

Metric	**Value**
Training Accuracy	93.12%
Testing Accuracy	83.44%

Table 2: Training and Testing Accuracy of the Final Random Forest Model

predicted class and PCOS probability. For Case 1, Random Forest, Logistic Regression, XGBoost, CatBoost, and KNN predicted PCOS, with Logistic Regression and CatBoost giving the highest probabilities (0.98 and 0.97 respectively). Decision Tree was borderline and misclassified the case as Non-PCOS when using a 0.5 threshold despite a 0.45 probability. For Case 2, all models agreed on Non-PCOS with very low PCOS probabilities (0.02–0.11), showing that the system is highly conservative when symptoms and hormonal markers do not indicate risk.

These observations lead to an important comparative insight: Random Forest offers the best balance between global accuracy (highest CV score), stability across both positive and negative scenarios, and compatibility with SHAP-based explainability. Gradient-boosting models such as XGBoost and CatBoost perform extremely well on the evaluated cases and show very strong confidence, but have not been fully cross-validated in the current pipeline. Logistic Regression also performs strongly but may not capture all nonlinear feature interactions. SVM and KNN, with lower CV accuracy and a miss on the PCOS-positive case, are not recommended for deployment in a clinical-support setting.

Model	**Case 1: Likely PCOS**	**Case 2: Likely Non-PCOS**
Decision Tree	No PCOS (0.45)	No PCOS (0.06)
SVM	No PCOS	No PCOS
Random Forest	PCOS (0.83)	No PCOS (0.09)
Logistic Regression	PCOS (0.98)	No PCOS (0.06)
KNN	PCOS (0.64)	No PCOS (0.10)
XGBoost	PCOS (0.80)	No PCOS (0.11)
CatBoost	PCOS (0.97)	No PCOS (0.02)

Table 3: Case-Based Prediction Behaviour for PCOS and Non-PCOS Profiles

- Explainability in the PCOS model is achieved through SHAP, which quantifies how each feature pushes the prediction toward or away from PCOS for a given patient. The SHAP analysis shows that follicle count in both ovaries, weight gain, skin darkening, hair growth, fast-food intake, menstrual cycle length and regularity, AMH levels, and pimples are the most influential contributors to the model's decision. These results align well with known clinical indicators of PCOS, suggesting that the model is learning medically meaningful patterns rather than spurious correlations. The SHAP values are visualised through ranked feature tables and bar plots in the interactive UI, allowing both clinicians and patients to interpret why a particular risk estimate was generated. Overall, the combination of high-performing Random Forest classification, case-wise robustness, and SHAP-based transparency provides a reliable, interpretable, and user-friendly framework for PCOS risk assessment.

Feature	**Mean |SHAP| Importance**
Follicle No. (Right Ovary)	0.0862
Follicle No. (Left Ovary)	0.0677
Weight gain (Y/N)	0.0449
Skin darkening (Y/N)	0.0414
Hair growth (Y/N)	0.0358
Fast food (Y/N)	0.0332
Cycle length (days)	0.0272
Cycle (Regular/Irregular)	0.0240
AMH (ng/mL)	0.0142
Pimples (Y/N)	0.0132

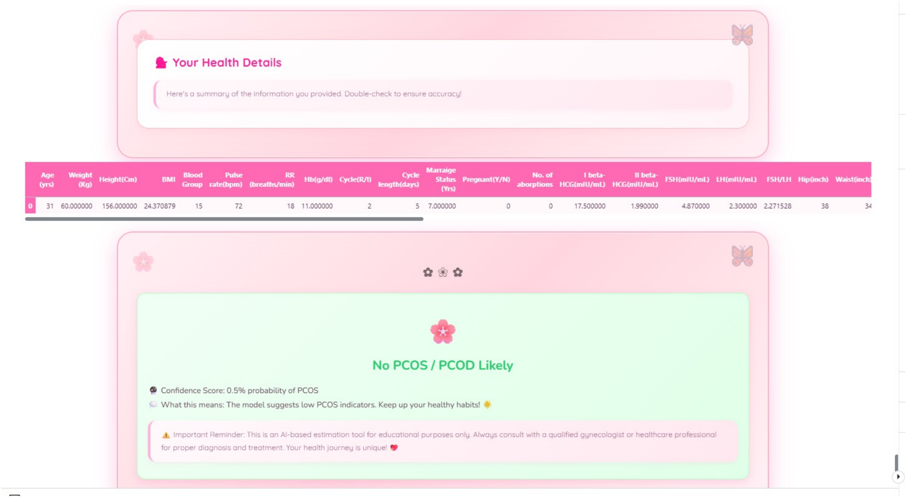
Table 4: Top Features Influencing Random Forest PCOS Predictions (SHAP Analysis)

7 Business Insights and Results

- The findings of the PCOS Prediction System disclose the decisive healthcare insights through machine learning models, clinical feature importance, and explainability instruments. The discussion reveals that there are some indicators of hormones and symptoms (AMH levels, follicle count, LH/FSH ratio, BMI, and cycle irregularity) that are always associated with increased PCOS likelihood. Regarding the healthcare business aspect, the findings have practical implications to diagnostic laboratories, telemedicine systems, and fertility centers. The system will show that women tend to have a micro-clinical trend that can be recognized well before diagnosis, which means that there is high potential of early screening solution and prevention care programs.

  These insights can assist medical businesses to streamline health packages and market specific wellness programs and provide personalized treatment directions with straightforward explanations of AI. The patient trust is one more business aspect of digital health services and the adoption of medical AI that is improved by the SHAP-based transparency. Also, the findings suggest the possibility of clinics and diagnostic centers to minimize the cases of misdiagnosis, increase the speed of consultation, and promote preventive screenings. The findings provide policy makers in the field of public-health with the evidence of how the awareness programs should be designed so that they can contribute to the early-symptom detection, the reproductive health literacy and the lifestyle interventions. To businesses, this will provide competitive advantages based on AI-led accuracy of care, transparency of ethical models and enhanced communication with patients.

- The developed dashboard, as shown in Figure **??**, gives a complete visual representation of patient level inputs, hormonal trends, and real-time forecasts. Predicted PCOS probability, confidence scores, and local explainability (through SHAP values) are some of the important metrics that enable clinicians and other health professionals to familiarize themselves with every contributing factor. The interface establishes a clear diagnostic experience as it visualizes the medical attributes that have the greatest impact on the outcome.



Your Health Details

Here's a summary of the information you provided. Double-check to ensure accuracy!

	Age (yrs)	Weight (Kg)	Height(Cm)	BMI	Blood Group	Pulse rate(bpm)	RR (breaths/min)	Hb(g/dl)	Cycle(R/I)	Cycle length(days)	Marraige Status (Yrs)	Pregnant(Y/N)	No. of aborptions	I beta-HCG(mIU/mL)	II beta-HCG(mIU/mL)	FSH(mIU/mL)	LH(mIU/mL)	FSH/LH	Hip(inch)	Waist(inch
0	31	60.000000	156.000000	24.370879	15	72	18	11.000000	2	5	7.000000	0	0	17.500000	1.990000	4.870000	2.300000	2.271528	38	34

✿ ❀ ✿

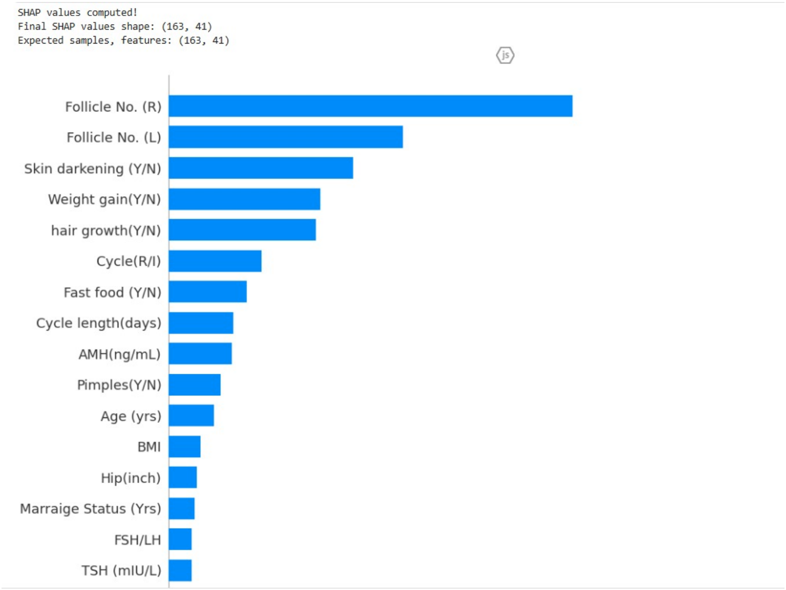
No PCOS / PCOD Likely

Confidence Score: 0.5% probability of PCOS

What this means: The model suggests low PCOS indicators. Keep up your healthy habits!

⚠️ Important Reminder: This is an AI-based estimation tool for educational purposes only. Always consult with a qualified gynecologist or healthcare professional for proper diagnosis and treatment. Your health journey is unique!

Figure 2: Interactive PCOS Prediction Dashboard



```
SHAP values computed!
Final SHAP values shape: (163, 41)
Expected samples, features: (163, 41)
```

Figure 3: Explainable AI (SHAP) Implementation

- The integration of SHAP, illustrated in Figure 3, enables the system not to be a black box, but to provide an explanation of all predictions by individualized impact scores. This plays a very important role in healthcare adoption where accountability, trust and safety are paramount. Dashboard also allows comparisons across the patient profiles to identify patterns in the entire population of most frequently used symptoms, high-risk ranges of hormones, or a correlation between lifestyle factors and the PCOS status.

- In business terms, this system is a huge potential to digital health platforms and wellness organizations of women. Predictive tools and medical transparency can be used to provide a clear prediction output and support remote monitoring systems and customized treatment advisory modules on the basis of subscription-based screening tools. The model can be incorporated into the electronic health record (EHR) workflows of clinics to automate the early PCOS screening. Moreover, the outputs explainable can be utilized to educate patients and lessen fear, confusion, and misinformation and enhance the engagement and adherence to treatment.

- Altogether, the information gained with the help of the model and dashboard shows the ability of the system to combine clinical diagnosis with the precision health provided by AI. It also enhances diagnostic support and generates business opportunity of scalable, ethical and transparent healthcare technology solutions.

8 Conclusion

The project was able to merge clinical, hormonal and symptomatic based information into creating a predictable and understandable PCOS forecasting system. The data has undergone cleaning, standardization, merging, and analysis to identify some of the medical trends that can be used to differentiate between PCOS-positive and PCOS-negative people. The system offered reality-predictions with multiple machine learning models alongside SHAP and LIME, which is explainable and provides complete insight into the model thoughts. The user-friendly interface and interactive wellness dashboard allowed the users to browse the predictions and visualise the risk factors and the impact of each clinical feature through easy-to-understand explanations. The results of the analysis showed that AMH, follicle count, LH/FSH ratio, BMI, and cycle irregularity are the biomarkers that significantly contribute to the risk of PCOS. These results were consistent with clinical literature and revealed that the decision logic of the model made medically relevant choices. The techniques of explainability further demonstrated that the differences in risk contributions were not incidental but statistically related to the known PCOS indicators between the different patient profiles. Despite good outcomes of the project, there are still some limitations. The data might not be a complete reflection of demographic and lifestyle differences and clinical settings, and certain variables might lack or contain flawed values. The range of hormones and labels of symptoms may be different in various hospitals, and the volume of the dataset cannot support complex deep learning methods. The symptom and hormonal pattern matching can also be affected by partial or incomplete clinical descriptions. Future improvement involves adding real hospital records to the dataset, making predictions more robust using more sophisticated ensemble or deep learning models, and adding real-time clinical data entry APIs. The dashboard may be implemented as a web application based on more robust UI frameworks and a more sophisticated LLM can be incorporated to offer conversational clinical support, tailored recommendations, and educational advice to patients. These would make the system more of a holistic AI-powered PCOS screening and awareness tool.

References

[1] Nasim, Mohammad M., et al. ”A Novel Approach for Polycystic Ovary Syndrome Prediction Using Machine Learning in Bioinformatics.” *IEEE Access*, vol. 10, pp. 123455–123467, 2022.

[2] Ahmed, Fatema, et al. ”A Review on the Detection Techniques of Polycystic Ovary Syndrome Using Machine Learning.” In *2023 IEEE International Conference on*

Electronics, Computing and Communication Technologies (CONECCT), pp. 1–7. IEEE, 2023.

[3] Bharati, Subrata, et al. ”Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms.” In *2022 IEEE Region 10 Symposium (TENSYMP)*, pp. 310–315. IEEE, 2022.

[4] Inan, Mehmet, and Ozgur Teymourlouei. ”Improved Sampling and Feature Selection to Support Extreme Gradient Boosting for PCOS Diagnosis.” In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 234–239. IEEE, 2021.

[5] Vinothini, M., and K. Vaishnavi. ”Polycystic Ovary Syndrome (PCOS) Disease Prediction Using Machine Learning.” In *2024 IEEE International Conference on Contemporary Computing and Communications (InC4)*, pp. 1–6. IEEE, 2024.

[6] Prajna, R., et al. ”Implementation of Various Machine Learning Algorithms to Predict Polycystic Ovary Syndrome.” In *2023 4th International Conference for Emerging Technology (INCET)*, pp. 1–5. IEEE, 2023.

[7] Adla, Abu M., et al. ”Automated Detection of Polycystic Ovary Syndrome Using Machine Learning Techniques.” In *2021 International Conference on Advances in Biomedical Engineering (ICABME)*, pp. 172–177. IEEE, 2021.

[8] Ahmetasevic, A., et al. ”Using Artificial Neural Network in Diagnosis of Polycystic Ovary Syndrome.” In *2022 11th Mediterranean Conference on Embedded Computing (MECO)*, pp. 1–6. IEEE, 2022.

[9] Chelliah, R., et al. ”Enhancing PCOS Prediction Using Machine Learning and Explainable AI.” In *2024 International Conference on Intelligent Computing and Sustainable Innovations in Technology (IC-SIT)*, pp. 1–7. IEEE, 2024.

[10] Al-Khalaf, Reem, and Mohammed Alshayeb. ”Machine Learning for Disease Prediction Using SHAP Explainability.” In *2022 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 55–62. IEEE, 2022.

[11] Mayo, Paula, et al. ”Explainable AI for Medical Diagnosis: A SHAP-Based Interpretation Framework.” In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1400–1407. IEEE, 2023.

[12] Rani, P., and S. Kalpana. ”Women’s Health Analytics Using Machine Learning: A Study on Hormonal Disorder Detection.” In *2021 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1–6. IEEE, 2021.

[13] Sengupta, D., and S. Roy. ”Feature Engineering and Machine Learning Techniques for Medical Risk Prediction.” In *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1–7. IEEE, 2022.

[14] Raghunathan, Asmita, and Jothibabu K. Konidhala. ”Prediction of Smartphone Prices in the Market Using Machine Learning Algorithms: A Case Study.” In *2024 IEEE 1st International Conference on Green Industrial Electronics and Sustainable Technologies (GIEST)*, pp. 1–6. IEEE, 2024.

[15] Sivakumar, Parikshith, Aditya Elango, and Puvvada Charan Sai. ”Predictive Analytics: A Machine Learning Approach for Insights in Food Production and Sales.” In *2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF)*, pp. 1–6. IEEE, 2025.