PCOS Dataset

Data Cleaning after Merging:

So some columns look like numbers but they are actually stored as strings because of how data was stored in Excel or CSV, so we convert them back to numeric using functions and if there are any other strange values it converts them to NaN. Some column names have extra space at the start or end, we are removing thar as well. Then we are also giving easy names without spaces for easy coding.

Missing values:

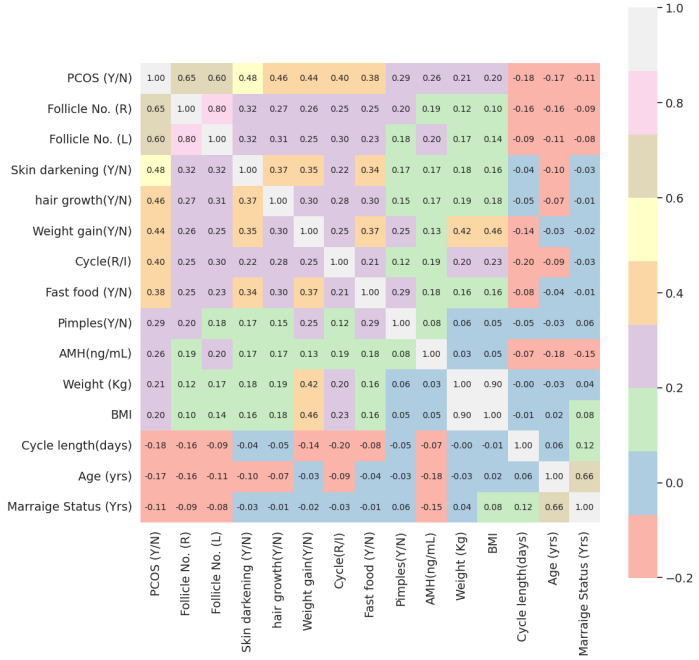
Then we are dealing with missing values by taking the median of that column and placing them there. We are using median because it is robust to outliers.

Then we do a summary also to see how everything is.

Correlation and Heatmap:

The `data.corr()` function calculates pairwise correlation coefficients between all numeric columns in your dataset. These coefficients range from -1 to 1, where values close to 1 indicate a strong positive correlation, values near -1 indicate a strong negative correlation, and values around 0 suggest no linear relationship.

Then we try to check which feature has the strongest relationship with the target variable. This kind of analysis is particularly useful in feature selection and understanding which medical or lifestyle indicators are more influential in predicting or diagnosing PCOS.



Interpretation:

We can see that Follicle No. ® and Follicle No. (L) have the highest positive correlations with PCOS at 0.65 and 0.60 respectively. This suggests that a higher number of follicles in

either ovary is strongly linked to PCOS, which makes clinical sense since PCOS is characterized by the presence of multiple ovarian follicles or cysts.

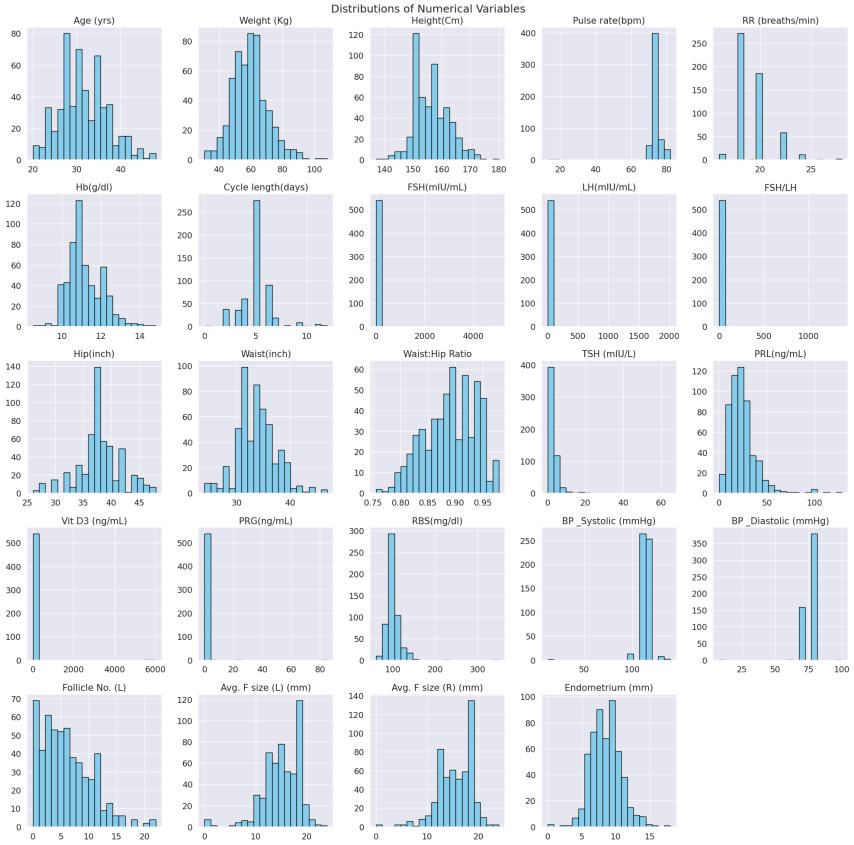
Skin darkening, hair growth and weight gain all show moderate positive correlations (ranging from ~0.40 to 0.48). These are symptoms or manifestations commonly observed in PCOS patients due to hormonal imbalances.

The heatmap also gives insight into multicollinearity—like the very strong correlation between `Follicle No. (L)` and `Follicle No. (R)` (0.80), which implies they tend to rise and fall together. So, we think we can remove either one of the features, particularly the one that shows the highest correlation with the target variable. Similarly, `Weight (Kg)` and `BMI` show a high correlation (0.92), which is expected because BMI is calculated using weight and height. So we can remove either one of them from the data.

So we will remove the Right Follicle Number and BMI from the data.

<u>Exploratory Data Analysis:</u>

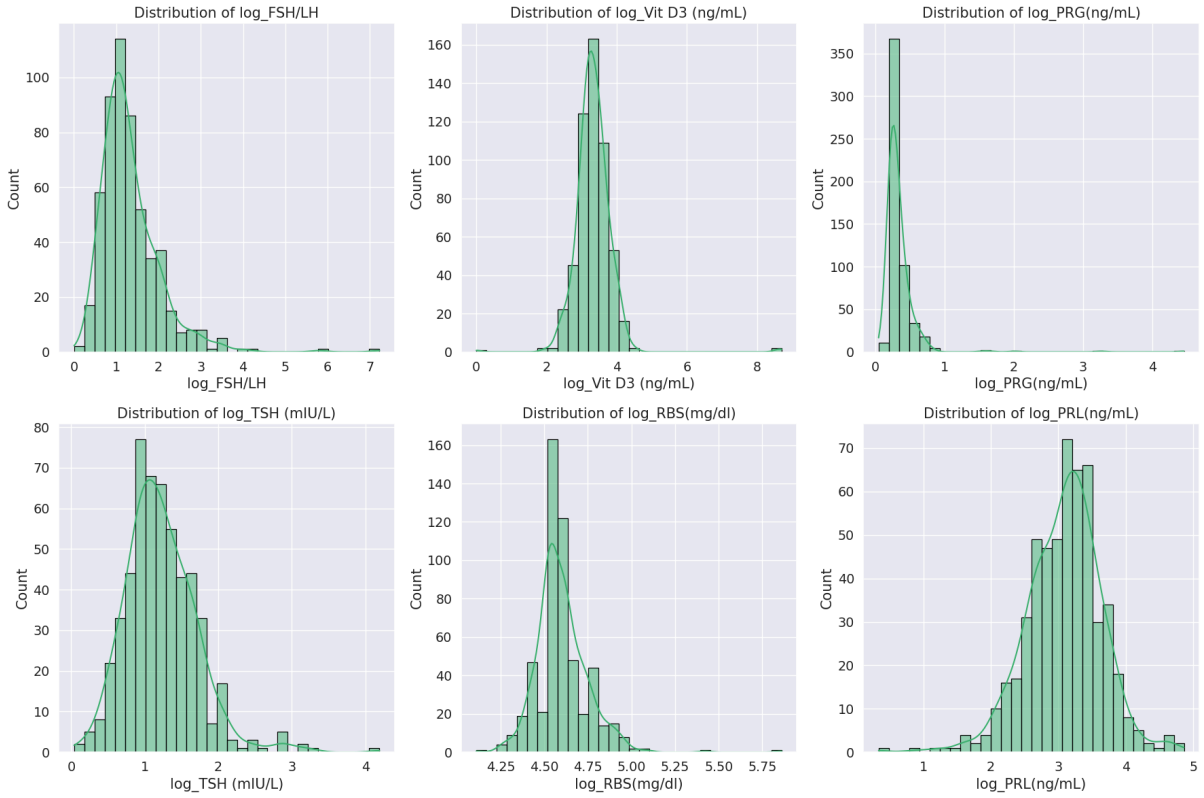
So first we wanted to see the distributions of all numerical variables through a simple histogram and then take the EDA from there.



We feel that something seems off with these columns like Cycle Length, FSH, LSH, FSH/LH, PRG and Vit D3 (Extremely right skewed). So before doing any further analysis we might go for inspecting what actually is happening here.

So the Cycle Length feature seems to have a certain number of days, which actually needs to be discretized if thought of logically. So, we discretize that. Also the FSH and LH have a ratio column together. We can use that instead of having two separate columns. It would reduce the

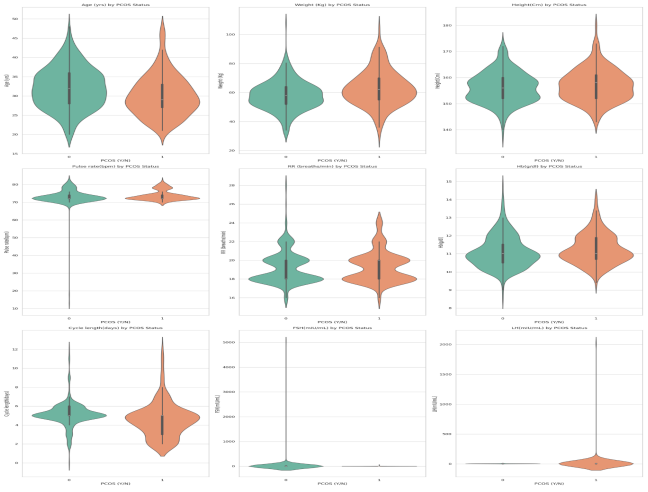
dimensionality. I think we will go for log transformations with the other highly skewed columns.



Now after log transformations, we get:

Which we are okay with. We still think that Systolic BP and Diastolic BP might be related and we might need only one of them, But let's keep them.

So, then we do a Violin Plot which shows the density of each feature, but grouped by PCOS Status. We can see that Slightly younger women tend to be PCOS positive.From the visualizations, we can see that women with PCOS generally tend to have higher body weight compared to those without the condition. Their menstrual cycle lengths also tend to be longer and more irregular, which aligns with known symptoms of PCOS.



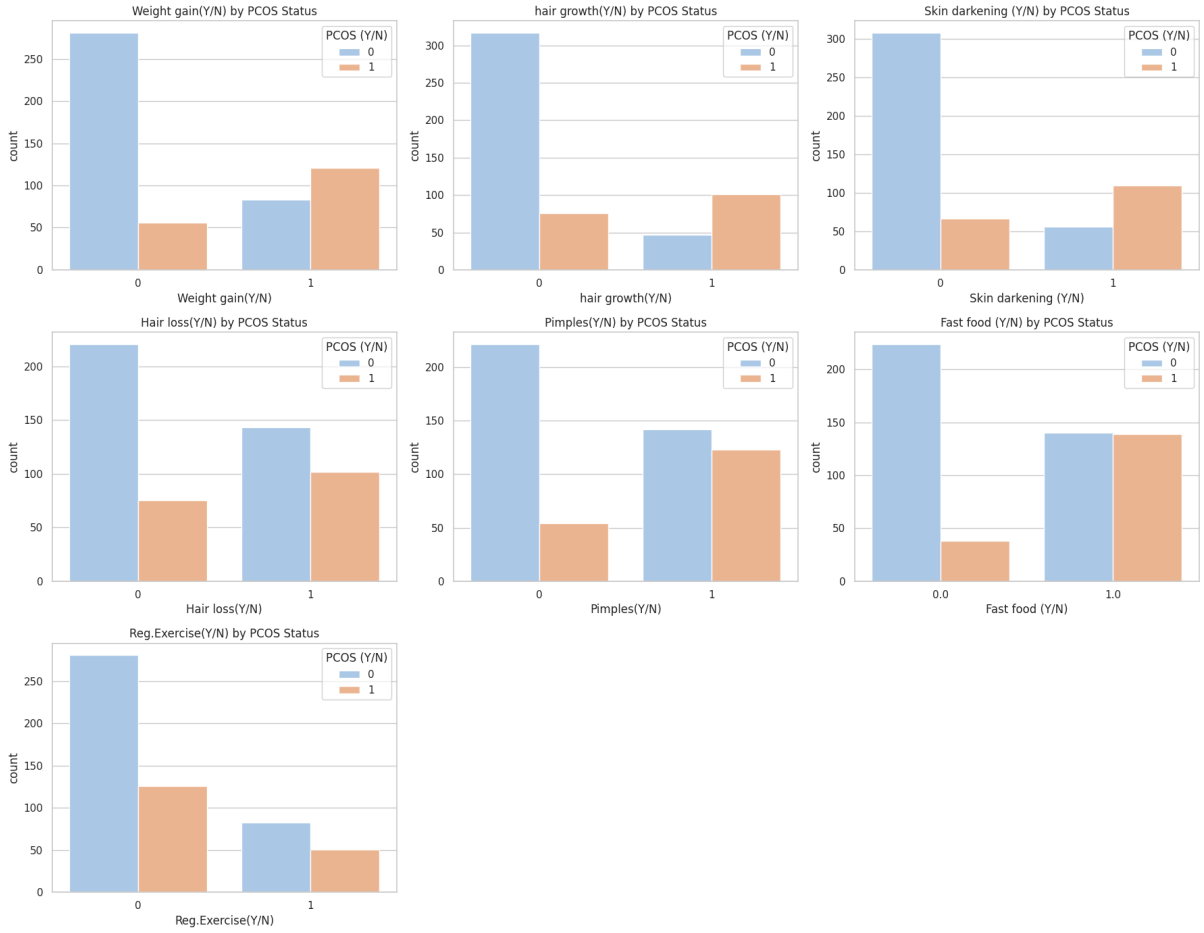
The age and height distributions are fairly similar between both groups, suggesting that these features alone may not strongly indicate PCOS. Pulse rate and respiratory rate also show some differences, but the patterns are less clear

and might be affected by measurement inconsistencies.

Hemoglobin levels appear slightly lower in women with PCOS, though the variation isn't drastic. When it comes to hormonal levels like FSH and LH, the plots are dominated by a few very high outlier values, making it hard to interpret meaningful differences.

Overall, weight and cycle length seem to be the most clearly distinguishable features between the two groups, while others may require more processing to reveal their value.

This set of bar plots compares the frequency of certain symptoms and lifestyle factors between women with and without PCOS. From the charts, it's clear that features such as weight gain, hair growth, skin darkening, hair loss, and pimples are reported more frequently among women with PCOS compared to those without. For example, significantly more women with PCOS report experiencing weight gain and increased hair growth, which are known physical symptoms linked to hormonal imbalance. Similarly, skin darkening and hair loss appear more common in the PCOS group.



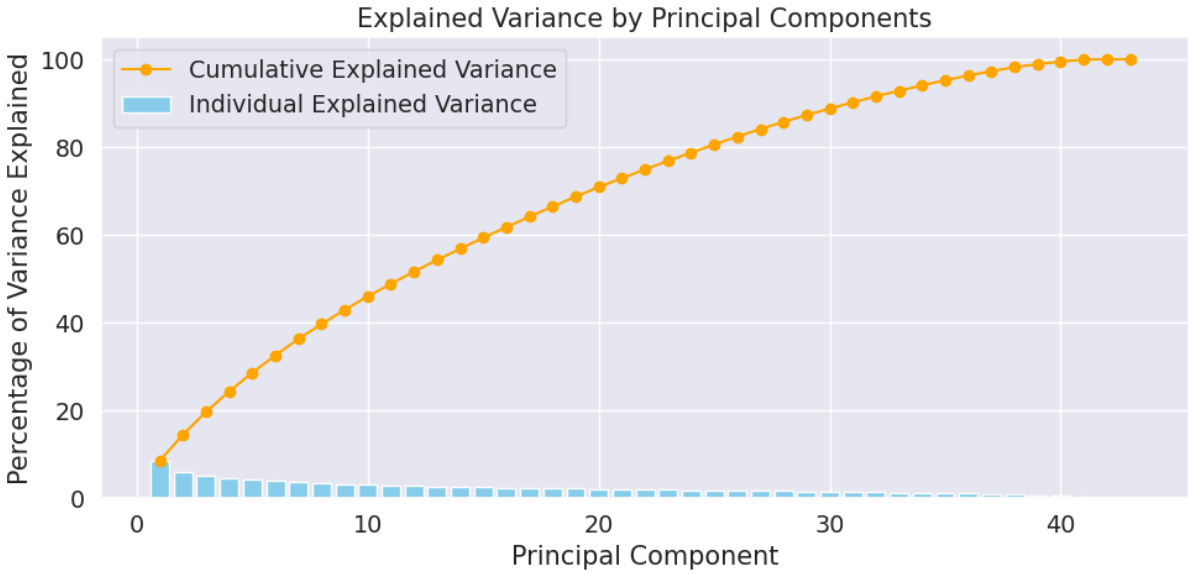
On the other hand, lifestyle-related factors like fast food consumption and regular exercise do not show major differences between the two groups. The distribution is almost even in both PCOS-positive and PCOS-negative individuals for these behaviors, suggesting that while

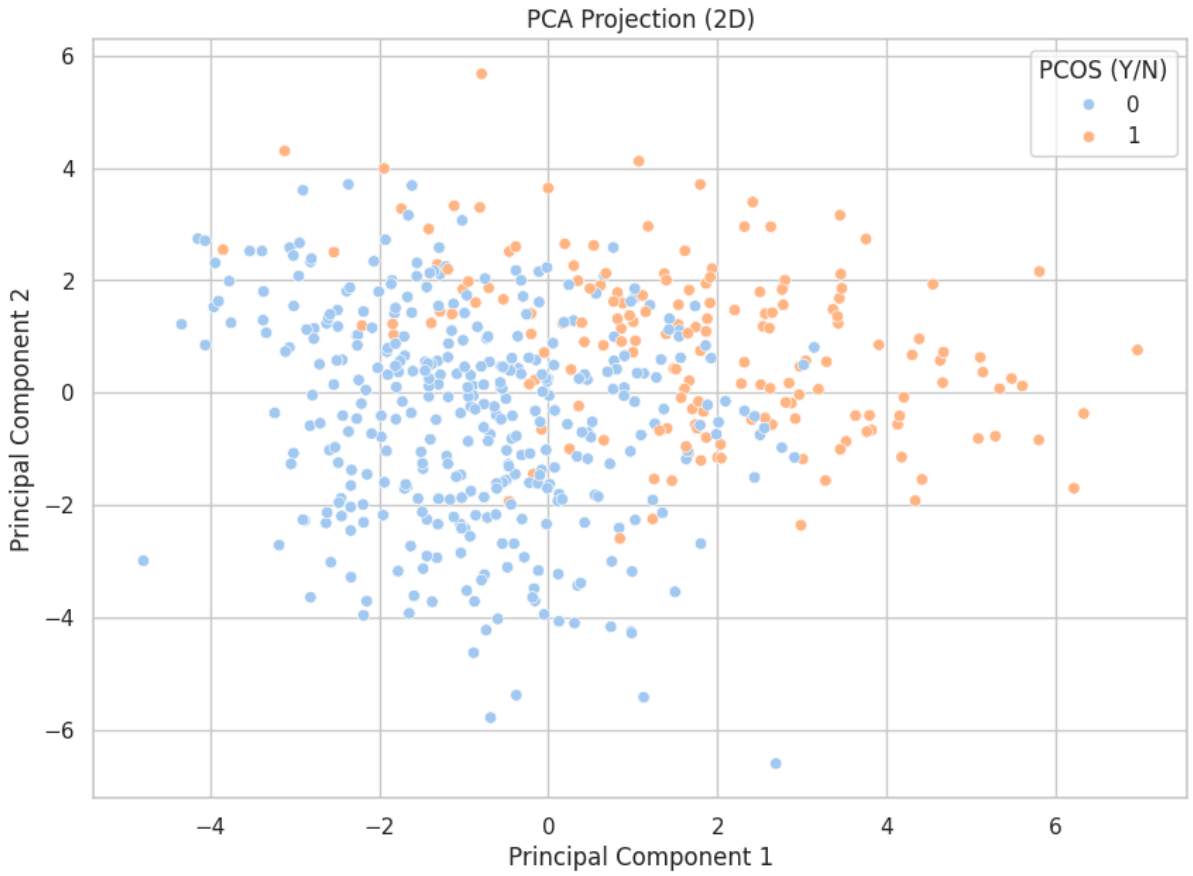
lifestyle may play a role in health overall, it may not be as distinct a feature for identifying PCOS in this dataset.

Now we think that the preprocessing and cleaning is almost done to the best of our knowledge. Now let us go to model building.

Dimensionality Reduction:

PCA is a linear algebra-based technique that works only on numerical data, assuming continuous, scaled input features. So, we try to do Principal Component Analysis and see how it is. We keep the components to n=2.The bars show how much variance each individual principal component explains. As we move to the right (higher components), each one contributes less to the overall variance. The orange line with dots shows the cumulative variance explained as more components are added.





This tells us that while PCA has captured some of the structure that distinguishes PCOS vs non-PCOS, the separation is not perfectly clear in 2 dimensions. Higher-dimensional structure may still be important. Again it reduces and says that approximately 20 components are required to explain the variance.

Moving forward we would not use this because we need 20 components and as the components are linear combinations of others it would reduce the interpretability of the model.

Model Building:

Random Forest Classifier:

In this block of code, we are building a Random Forest classification model to predict whether a patient has PCOS based on their medical features. You start by defining the input features `X` by removing the target column (`PCOS (Y/N)`) and non-informative identifiers (`Sl. No`, `Patient File No.`). The target `y` is the PCOS diagnosis. We then split the dataset into training and test sets using a 70-30 ratio.

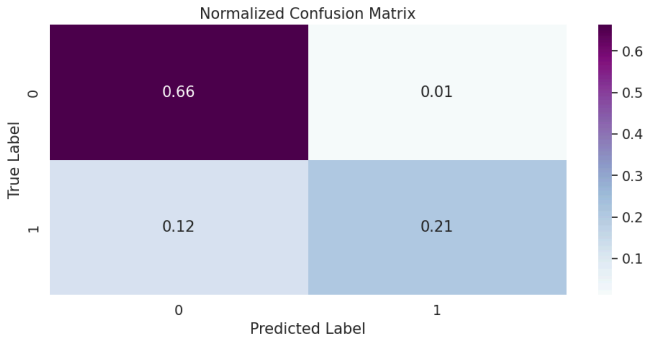
We begin with a basic Random Forest model to get an initial accuracy score. After that, we improve model performance by running a GridSearchCV, which tests different combinations of hyperparameters (like the number of trees, maximum depth, and splitting criteria) using

7-fold cross-validation. The grid search returns the best-performing parameter set based on training accuracy.

Once we identify the optimal parameters, we retrain the Random Forest model (`rfc_best`) using those settings and evaluate it on the test set. The final test accuracy is printed, and a classification report is generated to show precision, recall, F1-score, and support for each class.

Finally, we plot a normalized confusion matrix to visualize how well we model distinguishes between patients with and without PCOS. The values are normalized so that the proportions (not raw counts) can be interpreted more easily, making it clear where the model is performing well and where it's making errors.

Interpretation of the Model:



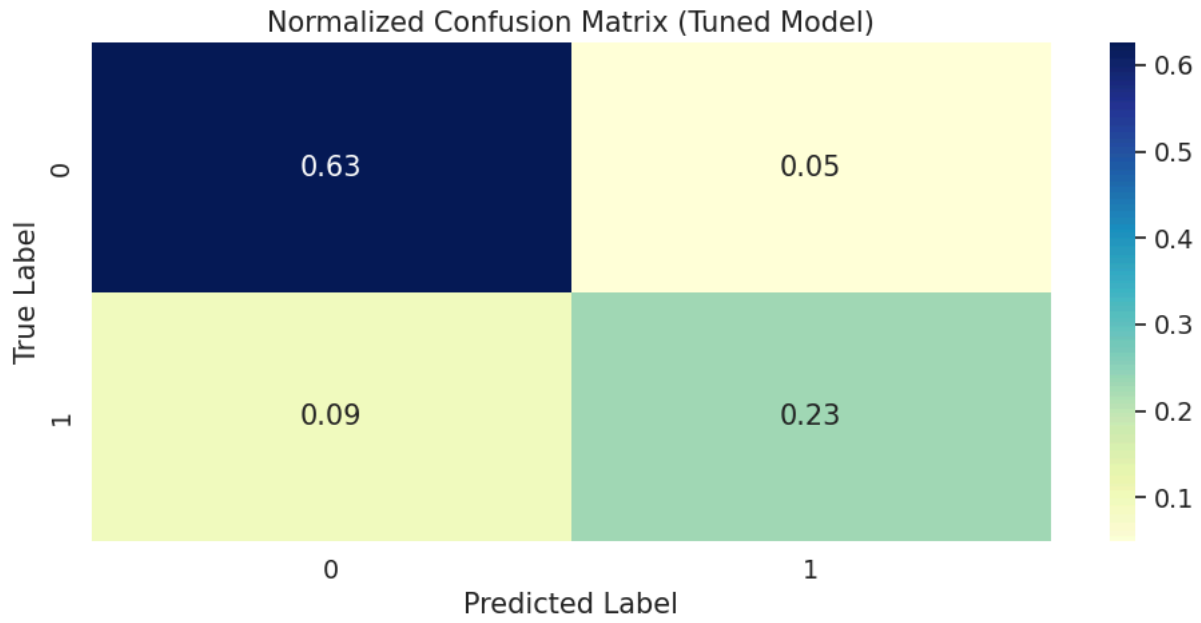
Best parameters are max depth is 8, maximum features is square root and n_estimators are 150, the final test accuracy is 87%

We need to focus on Recall (also called Sensitivity or True Positive Rate) tells you what percentage of actual PCOS cases your model successfully detects.

In the results, recall for PCOS = 1 is only 64%, meaning 36% of real PCOS cases are going undetected. So we go on for the next model

Neural Network Model:

The data splitting is the same as the Random Forest Classifier model. We use Standard scaling. This code performs hyperparameter tuning on a neural network model using KerasTuner's RandomSearch. It starts by preprocessing the PCOS dataset (encoding categorical variables and scaling features), then defines a flexible neural network architecture where the number of layers, units, activation functions, and optimizers are tuned. The tuner explores different combinations of these settings to find the best-performing model based on validation accuracy. Once the best model is found, it is evaluated on the test set using a classification report and a normalized confusion matrix to assess how well it distinguishes between PCOS and non-PCOS cases.



After hyperparameter tuning, the neural network achieved an overall test accuracy of 86%, showing good predictive performance. More importantly, the model's ability to identify actual PCOS cases — measured by recall for class 1 (PCOS-positive) — improved to 72%, compared to 64% in earlier models. This means the model is now successfully identifying nearly 3 out of every 4 PCOS patients, a significant gain for a healthcare-related task where missing a diagnosis is risky.

The precision for PCOS-positive predictions is also solid at 83%, indicating that most positive predictions are indeed correct. The F1-score (which balances precision and recall) is 0.77, reflecting a well-balanced classifier for the minority class. Overall, the tuned neural network does a much better job of minimizing false negatives, which is crucial for detecting PCOS in real-world screening or diagnosis scenarios.

Support Vector Machine (Extra Trees Classifier and ADASYN):

We start by preparing the data for modeling. We remove unnecessary columns like IDs and target labels, and convert any text-based or categorical features into numeric format using one-hot encoding. This step ensures that machine learning models can process the data correctly.

Next, we use a technique called Extra Trees Classifier to rank which features are most important for predicting PCOS. We select the top 9 features and use only those for building our final model, which helps reduce noise and improve performance.

Adaptive Synthetic Sampling Approach for Imbalanced Learning is an oversampling technique that generates synthetic data points for the minority class (e.g., PCOS-positive

cases) to balance the dataset. Unlike SMOTE, it adapts by creating more synthetic samples for harder-to-learn examples, i.e., those that are harder to classify.

This helps improve recall and overall model performance on imbalanced datasets by making the classifier more sensitive to the minority class.

Since medical datasets often suffer from class imbalance (fewer PCOS cases), we apply ADASYN to generate more synthetic PCOS cases and ENN to remove noisy non-PCOS examples. This combination balances the dataset and prepares it for accurate classification.

We then use the balanced dataset to train a Support Vector Machine (SVM) model. We apply GridSearchCV, a technique that automatically tests different combinations of SVM settings (like kernel type, regularization strength, and gamma) to find the best configuration — with a specific focus on maximizing recall, which is critical in healthcare where missing real cases is dangerous.

Once the best SVM model is found, we evaluate its performance using several metrics: accuracy, precision, recall, F1-score, ROC curve, and a precision-recall curve. These help us understand how well the model is identifying true PCOS cases versus making mistakes. Finally, we assess which features influenced the model most using permutation-based feature importance.

Results and Interpretation

The tuned Support Vector Machine (SVM) model has demonstrated exceptional performance in classifying PCOS cases. With an overall test accuracy of 97.8%, the model is highly reliable. Most notably, it achieved a recall of 1.0 (100%) for the positive class (PCOS cases), meaning it successfully identified all true PCOS patients without missing any. This is particularly important in medical diagnostics where failing to detect a condition could delay treatment.

The precision for PCOS detection was also very high at 96.1%, indicating that most of the predicted positive cases were indeed correct. The F1-score of 98% shows a strong balance between precision and recall. The confusion matrix supports this, showing zero false negatives (none of the actual PCOS cases were misclassified) and only two false positives.

The ROC Curve further confirms the model's effectiveness, with an AUC of 1.00, suggesting perfect separation between the classes. The Precision-Recall Curve remains flat at high precision across all recall levels, reinforcing the model’s reliability for imbalanced datasets. Finally, feature importance analysis highlighted that variables like follicle count, cycle length, and average follicle size had the strongest influence on the predictions, aligning well with clinical indicators of PCOS.

In summary, the tuned SVM not only performs superbly in accuracy metrics but also excels at correctly identifying all PCOS patients — making it an excellent candidate for diagnostic support.

EXB:

The Explainable Boosting Machine (EBM) is a transparent, interpretable machine learning model designed to provide high performance while remaining understandable to humans. It's based on a collection of generalized additive models with interactions, where the model learns separate, readable shape functions for each feature. This makes EBM ideal for healthcare tasks like PCOS prediction, where both accuracy and interpretability are essential.

In this case, after tuning hyperparameters such as the number of interactions, learning rate, histogram bins, and maximum leaf nodes, the best EBM model achieved an excellent test accuracy of 95.6%. Both the precision and recall for the positive class (PCOS patients) were 96%, indicating that the model is not only correctly identifying most of the PCOS cases (high recall), but also making very few false positives (high precision). The F1-score — a harmonic mean of precision and recall — was also balanced at 96%, reflecting overall robust performance.

This shows that EBM, with the right tuning, can match the accuracy of complex models like SVM or neural networks, while also offering the interpretability needed to understand which clinical features (e.g., follicle count, cycle length) influence predictions most — a crucial factor for decision support in medical environments.

Conclusions:

Model Comparison Table (Performance Metrics)

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	87.1	94.0	64.0	76.0
Neural Network (Tuned)	86.0	83.0	72.0	77.0
SVM (with ADASYN+ENN)	97.8	96.1	100.0	98.0
Explainable Boosting Machine (EBM)	95.6	96.0	96.0	96.0

We started with Random Forest because it’s a strong, widely used baseline model. It handles non-linear relationships, is resistant to overfitting, and provides feature importance scores — which is useful in identifying key clinical variables. It works well even with limited preprocessing and performs decently on imbalanced data.

A neural network was chosen to test a more flexible, nonlinear model that can model complex patterns in the data. Since PCOS diagnosis may depend on intricate combinations of features (like hormone levels and physical indicators), deep learning can offer better performance than tree-based models — especially after hyperparameter tuning.

SVMs are excellent for binary classification tasks, particularly when combined with class balancing techniques like ADASYN and ENN. These methods helped correct the imbalance in PCOS cases, and SVM's margin-maximizing behavior helps make sharp, reliable decisions. We chose it to see how well it could perform after balancing and feature selection.

Finally, we chose EBM because it offers a transparent, interpretable approach that matches black-box models in performance. In healthcare, trust is critical — EBM gives clinicians direct insight into how each feature (e.g., follicle count, cycle length) contributes to the prediction.

Among the four models, the Support Vector Machine (SVM) clearly stood out:

It achieved the highest accuracy (97.8%) and perfect recall (100%) — meaning it detected all true PCOS cases.

It had a very high precision (96.1%), indicating that almost all predicted PCOS cases were correct.

The F1-score of 98% shows a near-perfect balance between precision and recall.

However, the Explainable Boosting Machine (EBM) also deserves recognition. It matched SVM closely in all metrics (96% recall, precision, and F1), but offered something SVM cannot: interpretability. In clinical settings where understanding *why* a patient is flagged matters, EBM is a powerful option.

Use SVM when the highest detection rate (recall) is critical and decisions are automated.

Use EBM when interpretability and trust are equally important, such as in clinical decision support systems.