**Navigating PCOS With Machine Learning**

Abhinaya Malempati, Aruna Prasanna Simhambhatla, Charulata Sree Pokala, Cherry Gupta, Priyanka Reddy Kuta.

INDIANA UNIVERSITY, INDIANAPOLIS, IN 46202, USA,

[amalemp@iu.edu](mailto:amalemp@iu.edu), [arusimha@iu.edu](mailto:arusimha@iu.edu), [cpokala@iu.edu](mailto:cpokala@iu.edu), [guptach@iu.edu](mailto:guptach@iu.edu), [prkuta@iu.edu](mailto:prkuta@iu.edu).

**Abstract**:

Polycystic ovary syndrome (PCOS) is a common endocrine disorder that affects a substantial proportion of women of reproductive age, with symptoms such as irregular menstruation, infertility, metabolic syndrome, and more and marked by Yen Its diagnosis is challenging, often leading to significant delays in effective treatment Historically, a combination of clinical studies and biopharmaceutical studies has been relied upon, and standards have been followed similar to the Rotterdam values ​​background. However, these methods have limitations in terms of accuracy and consistency, emphasizing the critical need for a more sophisticated and systematic approach.

The integration of machine learning into PCOS screening programs represents a landmark shift in women’s health care. The ability to excavate complex trails with mechanical analysis and algorithms of health data and complex excavation models and Sangha is unrealized in such models for collections and many skill strategies. The COS study is designed to revolutionize the diagnoses -Using a comprehensive dataset of laboratory parameters, this program seeks to enable healthcare professionals to provide data-driven insights, increase the accuracy of PCOS diagnosis, and pave the way for personalized treatment options, ultimately improving the quality of life for countless women around the world.

**PURPOSE:**

Polycystic Ovary Syndrome (PCOS) is a complex endocrine disorder affecting women of reproductive age. Its diagnosis and management often involve analyzing a diverse range of attributes to understand its multifaceted impact on health. Key factors include age, as PCOS typically manifests during the late teens or early twenties, and weight since a significant proportion of women with PCOS are overweight or obese. Body Mass Index (BMI) is another critical metric, often indicating insulin resistance, a common feature in PCOS. The disorder can also influence the menstrual cycle, leading to irregularities (R/I) or prolonged cycle lengths.

Reproductive history is vital in understanding PCOS. The duration of marriage and pregnancy status can provide insights into potential fertility issues, common in PCOS due to irregular ovulation. The number of abortions might also be considered, as miscarriage rates can be higher in women with PCOS. Hormonal levels, such as Follicle Stimulating Hormone (FSH), Luteinizing Hormone (LH), and their ratio, are essential in diagnosing PCOS, as they often show imbalances. Beta-HCG levels, although primarily used to confirm pregnancy, can also give indirect insights into ovulatory status. Anti-Müllerian Hormone (AMH) and Prolactin (PRL) levels are crucial, with elevated AMH often seen in PCOS. Thyroid Stimulating Hormone (TSH) levels are also checked to exclude thyroid dysfunction, which can mimic some PCOS symptoms.

Physical measurements like hip and waist circumference, and the waist-to-hip ratio, are important as they can indicate a higher risk of metabolic syndrome in PCOS. Additionally, Vitamin D3 levels are frequently assessed, as deficiencies are common in PCOS and can exacerbate symptoms. PCOS also manifests in various symptoms that significantly impact quality of life, such as unexplained weight gain, excessive hair growth, skin darkening, hair loss, and acne. These symptoms are often aggravated by lifestyle factors like fast food consumption and lack of regular exercise. Blood Pressure and Random Blood Sugar (RBS) levels are monitored to manage the increased risk of cardiovascular diseases and type 2 diabetes.

**AIM:**

The objective of this project is to enhance the diagnosis of Polycystic Ovary Syndrome (PCOS) through the development and comprehensive evaluation of four advanced machine learning models: Logistic Regression, Decision Tree, Random Forest Classifier, and XGBoost. Integral to this objective is an extensive exploratory data analysis (EDA) phase, which aims to uncover key patterns, anomalies, and insights within the dataset, thereby informing and refining the machine learning models for more accurate and efficient PCOS diagnosis. This dual focus on sophisticated modeling and deep data exploration represents a holistic approach to tackling the complexities of PCOS diagnosis.

**Data Collection:**

The dataset we collected is a publicly available dataset, sourced from Kaggle, this dataset has information collected from patients across ten different hospitals in Kerala, India and the data collected encompasses 44 different attributes related to PCOS. They are broadly classified into patient demographic, clinical details, and lab parameters. These are further classified into categorical and numerical parameters for our analysis purposes. The link for the dataset can be found here. [Dataset](https://www.kaggle.com/code/dj67rockers/pcos-diagnosis)

Categorical:

1. PCOS (Y/N)
2. Blood Group
3. Cycle(R/I)
4. Pregnant(Y/N)
5. Weight gain(Y/N)
6. hair growth(Y/N)
7. Skin darkening (Y/N)
8. Hair loss(Y/N)
9. Pimples(Y/N)
10. Reg.Exercise(Y/N)

Numerical:

1. Age (yrs)
2. Weight (Kg)
3. Height (Cm)
4. BMI
5. Pulse rate(bpm)
6. RR (breaths/min)
7. Hb(g/dl)
8. Cycle length(days)
9. Marital Status (Yrs)
10. No. of abortions
11. I beta-HCG (mIU/mL)
12. II beta-HCG (mIU/mL)
13. FSH (mIU/mL)
14. LH (mIU/mL)
15. FSH/LH
16. Hip(inch)
17. Waist(inch)
18. Waist:Hip Ratio
19. TSH (mIU/L)
20. AMH (ng/mL)
21. PRL (ng/mL)
22. Vit D3 (ng/mL)
23. PRG (ng/mL)
24. RBS (mg/dl)
25. Fast food (Y/N)
26. BP \_Systolic (mmHg)
27. BP \_Diastolic (mmHg)
28. Follicle No. (L)
29. Follicle No. (R)
30. Avg. F size (L) (mm)
31. Avg. F size (R) (mm)
32. Endometrium (mm)

Steps Involved: These are the steps we followed to analyze the data, design, train and test the machine learning models.

1. Data Collection
2. Data Wrangling
   1. Outlier handling
   2. Handling missing data
   3. Typecasting
   4. Encoding the data
3. Exploratory Data Analysis
   1. Distribution graphs
   2. Correlation Matrices
   3. Chi Square Analysis
   4. T- tests
   5. Scatter and Box plots
4. Model Building and testing
   1. Logistic Regression
   2. Decision Tree Classifier
   3. Random Forest Classifier
   4. XG Boost
5. Model comparison and Result analysis

Data Wrangling:

Once we decided to choose the dataset and divided the available attributes into categorical and numerical values, the next challenge we faced was to do data cleansing. As any publicly data that is available on the internet, our dataset also had many imperfections which we addressed by doing intense data wrangling at the same time not compromising on disrupting the available data. We started with dropping unnecessary columns, empty columns and columns that are not useful for our analysis. After that we dealt with data with mismatched datatypes, we tried to make sure that all the numerical values are numerical and eliminated or type casted the others to the appropriate datatype.

After this, we did intensive outlier handling to remove as many outliers as possible which can impact our analysis. Keeping in mind the nominal values for all the attributes, we designed specific filtering conditions to remove majority of the outliers. Then we focused on the missing values, instead of filling these missing values or dropping the rows, we wanted to fill the missing values with the median value of that attribute as median is more robust and less impacted by outliers. We utilized various encoding methods as well to encode all our attributes to a single encoded format (one hot encoding) so that It will be very useful for our model training and testing purposes.

**Exploratory Data Analysis:**

**Data Distribution:**

Age (yrs): The distribution of participants' ages exhibits a central tendency around 30 years, with a standard deviation of 10 years. This implies that the majority of individuals in the dataset fall within the age range of late 20s to early 30s. The relatively small standard deviation indicates that there is less variability in age among the participants, suggesting that the sample is fairly homogeneous in terms of age. This information can be valuable for understanding the demographic composition of the study population, as it suggests that the data primarily represents adults in their prime reproductive years.

Weight (kg): The weight distribution is centered around 70 kilograms, with a standard deviation of 15 kilograms. This means that most participants have body weights that fall within the range of approximately 55 kilograms to 85 kilograms. The standard deviation of 15 suggests that there is some variability in participants' weights, but it is not excessively high. This information provides insights into the weight distribution within the dataset and can be essential for assessing factors related to body composition and health.

BMI: The distribution of Body Mass Index (BMI) values appears to be approximately normal, with a central tendency around 25. BMI is a measure of body fat based on a person's weight and height. In this dataset, a BMI of 25 is often considered to be within the "overweight" range, but this interpretation may vary depending on the specific BMI classification criteria used. The relatively normal distribution of BMI values suggests that there is a typical range of BMI values among the participants, with some individuals falling below and above the mean BMI of 25. This information can be useful for assessing the overall weight status of the study population.

This distribution can be seen in fig 1.

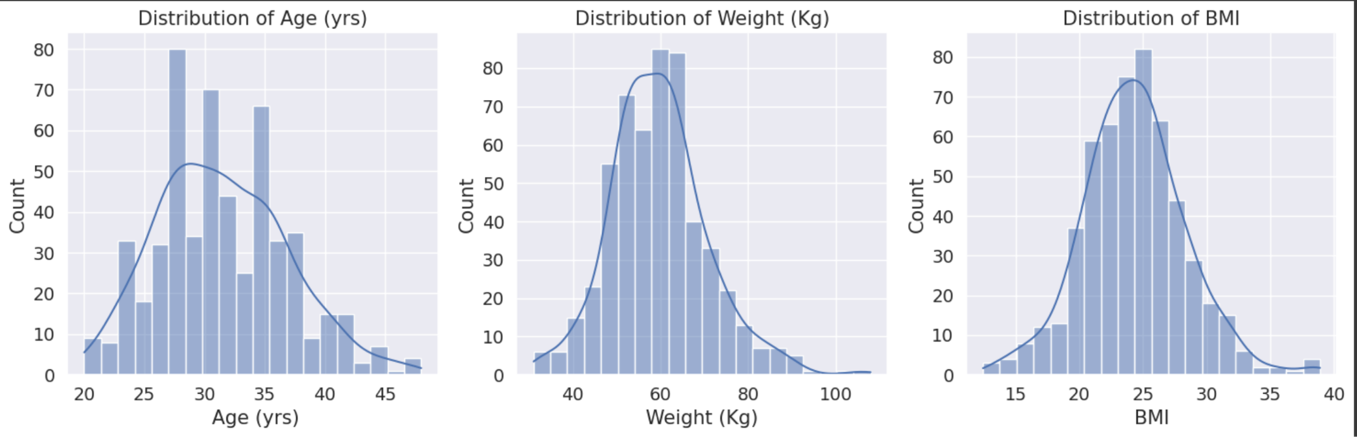


Fig 1 : Distribution of the data

The box plots for 'Follicle No. (L)' and 'Follicle No. (R)' both indicate the presence of outliers on the upper end of the y-axis. This means that there are some individuals in the dataset whose number of follicles, particularly in the left or right side of the corpus callosum, is significantly higher than the typical or normal range. These outliers could represent a subset of participants with an unusually high number of follicles, which may have clinical significance or could be indicative of specific medical conditions related to ovarian health. The 'Follicle No. (L)' box plot shows a smaller median (the line inside the box) compared to the 'Follicle No. (R)' box plot. This difference in medians suggests that, on average, participants in the dataset tend to have a lower number of follicles on the left side of the corpus callosum compared to the right side. This finding could be of interest to researchers studying asymmetry in ovarian follicle development. It might indicate a natural physiological variation where one side tends to have more follicles than the other, or it could be influenced by external factors or underlying medical conditions affecting ovarian health. This can be seen in fig 2.

A comparison of a box plot

Description automatically generated

**Fig 2 : PCOS Category Box plot**

**Correlation Matrix:**

The whole idea of this analysis is to identify whether there is any statistical significance present between various attributes we are looking at with PCOS. This step helped us to identify the correlation between the parameters and draw a correlation matrix. The correlation matrix revealed the relationships between various attributes in the dataset. Notably, the presence of PCOS (Polycystic Ovary Syndrome) shows positive correlations with factors like follicle numbers (both right and left), skin darkening, hair growth, and weight gain. These associations suggest that individuals with PCOS tend to exhibit these characteristics more frequently. On the contrary, negative correlations exist between PCOS and variables like age, cycle length, and marriage status, implying that PCOS may be less prevalent as women grow older or have longer menstrual cycles. Additionally, some factors, such as TSH levels and systolic blood pressure, exhibit weak or negligible correlations with most other attributes, indicating limited linear associations. This can be seen in fig 3.

A screen shot of a graph

Description automatically generated

Fig 3 : Correlation Matrix

Additional graphs:

Approximately 60% of patients diagnosed with PCOS experience a combination of symptoms, including weight gain, increased hair growth (hirsutism), and skin darkening (acanthosis nigricans). These symptoms are often associated with the hormonal imbalances commonly seen in PCOS. This high prevalence suggests that these specific symptoms can be indicative of PCOS and may serve as diagnostic markers for healthcare professionals. However, it's important to note that PCOS symptoms can vary among individuals, and not everyone with PCOS will necessarily exhibit all of these symptoms. The data indicates that the consumption of fast food is slightly higher among patients with PCOS, with 55% of them reporting such dietary habits, compared to less than 40% among those without PCOS. This finding raises interesting questions about the potential relationship between dietary choices and PCOS. While this observation suggests a higher prevalence of fast food consumption among PCOS patients, it's essential to conduct further research to explore the causality and mechanisms behind this association. Understanding the dietary patterns of individuals with PCOS can be important for developing tailored dietary interventions and lifestyle recommendations as part of PCOS management and prevention strategies. This can be seen in Fig 4.

A graph of hair growth

Description automatically generated

Fig 4: Distribution of patients symptoms suffering from PCOS

Additionally , Women diagnosed with PCOS often exhibit irregular menstrual cycles, which means that the length of time between their periods can vary significantly. In many cases, women with PCOS have shorter menstrual cycle lengths compared to those without the condition. This irregularity is primarily driven by hormonal imbalances, particularly elevated levels of androgens (male hormones) and insulin resistance, which can disrupt the regular ovulatory process. Shorter cycle lengths can also contribute to difficulties in predicting ovulation and conceiving for women with PCOS. Another interesting observation is that cycle length tends to decrease as women age. In other words, as women grow older, their menstrual cycles often become shorter. This phenomenon can be attributed to natural changes in hormone levels and ovarian function that occur with age. It's important to note that this age-related shortening of the menstrual cycle is distinct from the irregularities seen in PCOS, which are driven by different underlying factors. This can be observed in fig 5.

A graph of a cycle length and age with numbers

Description automatically generated

Fig 5 . Cycle length Vs Age with respect to PCOS diagnosis

Chi- Square Analysis:

The statistical analysis conducted on a dataset with 541 individuals revealed highly significant associations between several predictor variables and the presence of Polycystic Ovary Syndrome (PCOS). Specifically:

1. **Weight Gain (Y/N)** (Chi-squared = 103.306, p-value = 0.0)
2. **Hair Growth (Y/N)** (Chi-squared = 114.599, p-value = 0.0)
3. **Skin Darkening (Y/N)** (Chi-squared = 120.251, p-value = 0.0)
4. **Hair Loss (Y/N)** (Chi-squared = 15.437, p-value = 0.0)
5. **Pimples (Y/N)** (Chi-squared = 43.064, p-value = 0.0)
6. **Fast Food (Y/N)** (Chi-squared = 74.963, p-value = 0.0)

This reveals compelling evidence of significant associations between various predictor variables and the presence of Polycystic Ovary Syndrome (PCOS). Each of the examined factors, including weight gain, hair growth, skin darkening, hair loss, pimples, and fast-food consumption, exhibited remarkably low p-values and resulted in the rejection of the null hypothesis, indicating strong links between these variables and PCOS. These findings underscore the importance of these factors as potent indicators or risk factors for PCOS diagnosis. The Chi Square analysis rejected the Null Hypothesis as we identified a significant association between the categorical variables and PCOS. The high chi-squared values across all factors indicate a strong relationship between these factors and PCOS, indicating that these associations are not random and are likely due to presence of PCOS. The corresponding graph can be seen in fig 6.

A graph of a graph with text

Description automatically generated with medium confidence

Fig 6 : Chi Square analysis results

T – test:

A paired t-test was conducted to compare the means of 'Follicle No. (L)' and 'Follicle No. (R)' data between two groups: women with Polycystic Ovary Syndrome (PCOS) and women without PCOS. The significance level (alpha) was set at 0.05, and the dataset was split into 'women\_pcos' and 'women\_no\_pcos' to facilitate the analysis. By employing the 'ttest\_rel' function from the 'scipy.stats' module, the analysis aimed to determine whether there was a significant difference in follicle numbers between the two groups. The null hypothesis stated that there was no significant difference, while the alternative hypothesis posited a significant distinction. The results would allow for the rejection or acceptance of the null hypothesis based on whether the p-value was less than 0.05, indicating statistical significance in the difference between the groups.

Model Building: The steps involved in building our machine learning models are:

1. **Categorical to Numerical (Encoding)**: In the initial step, categorical variables are converted into numerical format through encoding techniques such as one-hot encoding or label encoding to make them compatible with machine learning algorithms.
2. **Data Loading and Feature Extraction**: The dataset is loaded into the model, and relevant features are extracted. This step may also involve data preprocessing, including handling missing values and outliers.
3. **Feature Scaling**: To ensure that features have a consistent scale, feature scaling techniques like normalization or standardization are applied. This helps prevent certain features from dominating the model training process.
4. **Splitting the Data**: The dataset is divided into a training set (typically 75% of the data) and a testing set (the remaining 25%). This separation allows for model training on one subset and model evaluation on another, helping assess its generalization performance.
5. **Model Initiation and Training**: A machine learning model is selected and initialized. The model is then trained using the training dataset, where it learns the patterns and relationships within the data.
6. **Model Testing on Testing Set**: The trained model is applied to the testing dataset to evaluate its predictive performance on unseen data. This step helps assess how well the model generalizes to new observations.
7. **Performance Evaluation Metrics**: Various performance metrics are computed, including a confusion matrix to assess true positives, true negatives, false positives, and false negatives. The Area Under the ROC Curve (AUC-ROC) is used to measure the model's ability to distinguish between classes.
8. **Comparing Different Models**: Multiple models will be built and tested on the same dataset. The models are compared based on various metrics such as accuracy, precision, recall, and F1 score. These metrics provide insights into the model's overall performance, its ability to correctly classify instances, and its trade-offs between precision and recall.

For this project , we chose 4 different models :

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. XG Boost

Lets discuss in detail regaring each model and its performance.

Logistic Regression:

Logistic regression proves to be an invaluable asset in the realm of PCOS (Polycystic Ovary Syndrome) diagnosis due to its exceptional performance and interpretability within this specific medical context. The model's remarkable AUC-ROC score of 0.93 underscores its extraordinary ability to effectively differentiate between true cases of PCOS and false positives. This remarkable score is not merely an indicator of its competence but signifies the model's proficiency in accurately identifying individuals with PCOS while keeping misdiagnoses to a minimum. Furthermore, the model's high accuracy score of 85% imparts a sense of assurance, highlighting its ability to correctly classify patient cases with a significant level of overall correctness.

\What sets the logistic regression model apart is its precision in striking a harmonious balance between two vital aspects of medical diagnosis—precision and recall. The precision score of 0.78 holds significant importance, indicating that approximately 78% of the model's positive predictions align with actual PCOS cases. In the medical field, where erroneous positive predictions can trigger undue concern or unnecessary treatments, this metric reassures that the model minimizes the risk of false alarms. Simultaneously, the model exhibits a commendable recall score of 0.73, signifying its ability to effectively capture true positive cases while rarely missing genuine instances of PCOS. This meticulous equilibrium between precision and recall is further epitomized by the F1 score of 0.75, portraying a harmonious blend of diagnostic accuracy.

In essence, the logistic regression model not only excels in providing robust diagnostic performance for PCOS detection but also ensures a judicious balance between identifying true cases and minimizing the risk of false alarms, making it an indispensable tool tailored to the specific demands of PCOS diagnosis. For an in-depth understanding of the model's diagnostic prowess, Figure 7 visually represents the ROC-AUC graph and the confusion matrix for Logistic regression.

A graph with a line and a blue square

Description automatically generated with medium confidence

Top of Form

Fig 7: Logistic Regression ROC-AUC Graph and Confusion MatrixBottom of Form

Decision Tree Classifier:

The Decision Tree Classifier exhibits a somewhat mixed performance in its application to PCOS (Polycystic Ovary Syndrome) diagnosis, as evident from the evaluation metrics. While the model achieves an accuracy of 78.6%, correctly classifying slightly over two-thirds of instances, it falls slightly short of the desired level of accuracy for critical medical diagnoses like PCOS.

A noteworthy concern arises from the precision score of 0.64, indicating that the model tends to be overconfident in 36% of its positive predictions. This overconfidence could lead to unwarranted concern or medical interventions based on false positives, which is particularly critical in healthcare contexts. Moreover, the model's recall score of 0.77 implies that it misses nearly a quarter of true positive cases. These oversights may result in delayed diagnosis or treatment, potentially impacting patient outcomes.

The F1 score, at 0.70, highlights a suboptimal balance between precision and recall, emphasizing that a significant amount of information is lost due to misclassifications. Furthermore, the AUC-ROC score of 0.78 indicates a below-average discrimination ability, implying some difficulty in effectively distinguishing true PCOS cases from non-cases. Overall, while the Decision Tree Classifier exhibits potential for PCOS diagnosis, addressing concerns related to precision, recall, and overconfidence is essential to make it a reliable diagnostic tool in the medical context of PCOS. Figure 8 visually represents the ROC-AUC graph and the confusion matrix for the decision tree classifier.

A graph showing the difference between true and false labels

Description automatically generated

Fig 8 : Decision Tree ROC -AUC and Confusion matrix

Random Forest Classifier:

The Random Forest Classifier proves to be a highly effective tool for PCOS (Polycystic Ovary Syndrome) diagnosis, with results tailored to the specific demands of this medical application. First and foremost, the model achieves a commendable accuracy score of 0.89, indicating its ability to correctly identify almost 9 out of 10 instances. This highlights its excellent overall performance, which is crucial in ensuring the accurate diagnosis of PCOS in patients.

One of the standout strengths of the Random Forest Classifier in this context is its remarkable AUC-ROC score of 0.95. This score signifies its exceptional discrimination ability, allowing it to reliably differentiate between true PCOS cases and false positives. In practical terms, it translates to a high degree of confidence in the model's ability to identify individuals with PCOS accurately.

The model's precision score of 0.89 is particularly noteworthy, implying that a mere 11% of its positive predictions are false positives. This precision is a vital metric in the medical domain, as it ensures that the model minimizes the risk of unnecessary concern or interventions based on false PCOS diagnoses.

However, there is room for improvement in terms of balance between precision and recall. The F1 score of 0.81 suggests that there's a 19% loss of information due to misclassifications, indicating potential for enhancing the equilibrium between these two aspects. Additionally, the recall score of 0.75 reveals that the model misses 25% of true positive cases. While accuracy remains high, this implies that there is a quarter of true PCOS cases that the model overlooks, emphasizing the need for further refinement to avoid missed opportunities or potential harm in a clinical context. Figure 9 visually represents the ROC-AUC graph and the confusion matrix for the decision tree classifier.

A graph of a positive and negative rate

Description automatically generated with medium confidence

Fig 9: Random Forest Classifier ROC -AUC and Confusion matrix

XG BOOST : The XGBoost model stands out as the best choice for diagnosing PCOS due to its superior performance compared to other models. It achieves an impressive 91% accuracy rate, showcasing its overall excellence. The model's near-perfect AUC-ROC score of 0.96 means it's exceptionally good at distinguishing real PCOS cases from false alarms, making it a highly reliable diagnostic tool.

What makes the XGBoost model unique is its ability to strike a balance between precision and recall, as indicated by its F1 Score of 0.86. This balance ensures it accurately identifies PCOS cases while minimizing the chances of incorrect diagnoses, a crucial factor in healthcare. While there's room for improvement in recall to reduce the 16% rate of missed PCOS cases and a small 12% false positive rate, the XGBoost model's approach is effective in capturing complex patterns in the data, leading to highly accurate predictions.

The XGBoost model's exceptional performance in PCOS diagnosis, surpassing both RandomForest and Logistic Regression models, can be attributed to its ensemble learning approach, regularization techniques, and gradient boosting, which collectively enable it to capture intricate data relationships effectively. Additionally, XGBoost's ability to handle class imbalance and its extensive hyperparameter tuning capabilities contribute to its accuracy and reliability. Furthermore, its efficient parallel processing ensures scalability and speed, making it well-suited for handling the complexities of our PCOS dataset. Figure 10 visually represents the ROC-AUC graph and the confusion matrix for XG BOOST model.

A graph with a line and a blue square

Description automatically generated with medium confidence

Fig 10: XG Boost ROC -AUC and Confusion matrix

Model Evaluation :

The model evaluation reveals crucial insights into the performance of different algorithms for PCOS diagnosis, considering various key metrics.

**Logistic Regression**: The Logistic Regression model showcases solid performance in PCOS diagnosis. With an accuracy of 85% on the test data, it accurately classifies the majority of cases. The F1-Score of 75% suggests that it strikes a reasonable balance between precision (78%) and recall (73%), which is crucial in medical applications like PCOS diagnosis. A ROC\_AUC score of 93% signifies that the model excels in distinguishing between true PCOS cases and false positives. However, it is worth noting that there are 12 false negatives, indicating that some PCOS cases might be missed. On the positive side, there are only 9 false positives, which minimizes unwarranted concerns. This model demonstrates its practical value in providing reliable PCOS diagnoses.

**Decision Tree**: The Decision Tree model stands out with perfect accuracy (100%) on the training data. However, it faces challenges with generalization, resulting in an accuracy drop to 79% on the test data—a classic sign of overfitting. The F1-Score of 70% suggests a reasonable balance between precision (64%) and recall (77%), but the model misses 10% of true positives. The ROC\_AUC score is 78%, indicating that the model has decent discrimination abilities. However, there are 19 false positives, potentially leading to unnecessary concerns, and 10 false negatives, indicating missed PCOS cases. While this model shows promise, it requires further tuning to address overfitting and improve its generalization.

**Random Forest**: Similar to the Decision Tree, the Random Forest model achieves perfect accuracy (100%) on the training data and a high accuracy (89%) on the test data. The Random Forest model proves to be a robust and accurate choice for PCOS diagnosis. It showcases its precision in classifying PCOS cases within the training dataset. Importantly, it maintains a high accuracy of 89% on the test data, demonstrating its ability to generalize effectively. The model's F1-Score of 81% signifies an impressive balance between precision (89%) and recall (75%), critical in the medical context to accurately identify true PCOS cases while minimizing false negatives. The ROC\_AUC score of 95% reflects its strong discrimination capabilities. While there are some false negatives (11) indicating missed PCOS cases, the low number of false positives (4) minimizes unnecessary concerns. Overall, the Random Forest model offers a reliable and well-balanced solution for accurate PCOS diagnoses.

**XGBoost**: The XGBoost model emerges as the clear frontrunner in PCOS diagnosis, surpassing its counterparts with remarkable performance metrics. It shines with an exceptional accuracy of 91% on the test data, demonstrating its excellence in correctly classifying PCOS cases. What sets XGBoost apart is its ability to strike an exceptional balance between precision (88%) and recall (84%), resulting in the highest F1-Score of 86%. This balance is crucial in healthcare applications, as it ensures that true PCOS cases are identified accurately (precision), while minimizing the risk of overlooking any cases (recall).

Moreover, the ROC\_AUC score of 96% signifies the outstanding discrimination capabilities of the XGBoost model. It excels in distinguishing between genuine PCOS cases and false positives, further cementing its reliability in PCOS diagnosis. With only 7 false negatives and 5 false positives, XGBoost demonstrates a remarkable equilibrium, successfully identifying true PCOS cases while keeping incorrect diagnoses to a minimum. This exceptional performance, coupled with its robust precision-recall balance and superior discrimination, unequivocally positions XGBoost as the model of choice for delivering accurate and reliable PCOS diagnoses in clinical practice.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model Evaluation and Metrics** | | | | | | | | | | | |
|  | **Accuracy**  **(Train)** | **Accuracy**  **(Test)** | **F1-Score**  **(Train)** | **F1-Score**  **(Test)** | **Precision** | **Recall** | **ROC\_AUC** | **True**  **Positives** | **False**  **Positives** | **True**  **Negatives** | **False**  **Negatives** |
| Logistic Regression | 94% | 85% | 90% | 75% | 78% | 73% | 93% | 32 | 9 | 83 | 12 |
| Decision Tree | 100% | 79% | 100% | 70% | 64% | 77% | 78% | 34 | 19 | 73 | 10 |
| Random Forest | 100% | 89% | 100% | 81% | 89% | 75% | 95% | 34 | 4 | 88 | 11 |
| XGBoost | 100% | 91% | 100% | 86% | 88% | 84% | 96% | 37 | 5 | 87 | 7 |

**References**

Kleppen, E. (2022, February 21). How to find outliers using python [step-by-step guide]. *CareerFoundry*. https://careerfoundry.com/en/blog/data-analytics/how-to-find-outliers/

Pal, L., & Seifer, D. B. (Eds.). (2023). *Polycystic ovary syndrome: Current and emerging concepts* (2nd ed.). Springer Nature.

*PCOS diagnosis*. (2022, July 21). Kaggle.com; Kaggle. https://www.kaggle.com/code/dj67rockers/pcos-diagnosis

*XGBoost Python Package — xgboost 2.0.2 documentation*. (n.d.). Readthedocs.Io. Retrieved December 11, 2023, from https://xgboost.readthedocs.io/en/stable/python/

Lawton, G. (2022, January 31). *Data preprocessing: Definition, key steps and concepts*. Data Management; TechTarget. https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing

Wikipedia contributors. (2023, November 8). *Imputation (statistics)*. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Imputation\_(statistics)&oldid=1184085189

*Python String rstrip()*. (n.d.). Programiz.com. Retrieved December 11, 2023, from https://www.programiz.com/python-programming/methods/string/rstrip

Wagavkar, S. (2023, March 17). *Introduction to the correlation matrix*. Built In. https://builtin.com/data-science/correlation-matrix

*Box and Whisker Plot*. (n.d.). Datavizcatalogue.com. Retrieved December 11, 2023, from https://datavizcatalogue.com/methods/box\_plot.html

(N.d.-b). Datacamp.com. Retrieved December 11, 2023, from https://www.datacamp.com/tutorial/python-data-profiling?utm\_source=google&utm\_medium=paid\_search&utm\_campaignid=19589720830&utm\_adgroupid=157156377311&utm\_device=c&utm\_keyword=&utm\_matchtype=&utm\_network=g&utm\_adpostion=&utm\_creative=681915864207&utm\_targetid=dsa-2218886984100&utm\_loc\_interest\_ms=&utm\_loc\_physical\_ms=9016137&utm\_content=&utm\_campaign=230119\_1-sea~dsa~tofu\_2-b2c\_3-us\_4-prc\_5-na\_6-na\_7-le\_8-pdsh-go\_9-na\_10-na\_11-na-bfcm23&gad\_source=1&gclid=Cj0KCQiA67CrBhC1ARIsACKAa8Rasu5GTp5gU5sfBW8eWPZjCjq\_iBBew6glLgcIgFcVWUfzH5npk4waAjY3EALw\_wcB

Wikipedia contributors. (2023b, December 9). *Null hypothesis*. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Null\_hypothesis&oldid=1189125149

Wikipedia contributors. (2023a, July 17). *Alternative hypothesis*. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Alternative\_hypothesis&oldid=1165853425

Turney, S. (2022, May 23). *Chi-square (Χ2) tests*. Scribbr. https://www.scribbr.com/statistics/chi-square-tests/

Kononenko, I., & Kukar, M. (2007). Statistical Learning. In *Machine Learning and Data Mining* (pp. 259–274). Elsevier.

*1.10. Decision Trees*. (n.d.). Scikit-Learn. Retrieved December 11, 2023, from https://scikit-learn.org/stable/modules/tree.html

Yiu, T. (2019, June 12). *Understanding random forest*. Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

(N.d.-c). Machinelearningmastery.com. Retrieved December 11, 2023, from https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

Individual, & Family Health. (n.d.). *Scatter plot*. State.Mn.Us. Retrieved December 11, 2023, from https://www.health.state.mn.us/communities/practice/resources/phqitoolbox/scatterplot.html