# Project Scope Statement

## POLYCYSTIC OVARY SYNDROME DATA(PCOS)

Tianqi Zhou, Zixuan Li, Yifei Wu
GEORGETOWN UNIVERSITY |

# Table of Contents

# 1 Project Outlines

| *Project #* | *Project Description* | *Date Submitted* | *Project Priority* |
|---|---|---|---|
| 1 | Georgetown University is committed to cultivating an environment centered on personalized care tailored to the specific needs of patients. In pursuit of this goal, they are embarking on the development of an application designed to swiftly detect and diagnose PCOS (Polycystic Ovary Syndrome) as soon as a patient's laboratory results are input into the system. To support this initiative, our project group designed a machine learning application that leverages existing patient data to enhance the speed and accuracy of PCOS diagnosis, uncovering intricate details that may prove challenging for human assessment. | | Priority 01 |

## *1.1 Step 1. Project Deliverables*

Please list *all project deliverables* listed in the Project Charter and, if necessary, elaborate on them. *Do not list dates*. Add more rows as necessary.

| **Deliverable ID#** | **Description** |
|---|---|
| 1 | Problem Identification |
| 2 | Method evaluation |
| 3 | Solution by Hyperparameter Optimization |
| 4 | Deployment Pipeline/Platform |
| 5 | Project Demo In class |

## *1.2 Step 2. List of Project Tasks*

Please list ***all project tasks*** to be completed, based on the "Deliverables" specified in the Project Charter. *Do not list dates*. Add more rows as necessary. Optional: You may substitute a work breakdown structure (WBS) or mind map in lieu of Step 2. Please attach WBS or mind-map to the document.

| **Task ID#** | **Task to be completed** | **Delivery Date** | **For Deliverable #** |
|---|---|---|---|
| 1 | Submit Project Charter | 09/07/2023 | 1 |
| 2 | Method evaluation - research on techniques, quality assurance, project approach. | 09/28/2023 | 2 |
| 3 | Solution by Hyperparameter Optimization - actual effective techniques used. | 10/12/2023 | 3 |
| 4 | Deployment Pipeline/Platform. | 10/26/2023 | 4 |
| 5 | Project Demo In class - Final report of the project in addition to the cloud deployed link. | 11/16/2023 | 5 |

## *1.3 Step 3. Out of Scope*

| | |
|---|---|
| This project **will NOT accomplish or include** the following: | This project will not include any analysis of other diseases of the patient and other background details. |

## *1.4 Step 4. Project Assumptions*

Please list any project factors that will be considered to be true, real, or certain. Assumptions generally involve a certain degree of risk.

| # | **Assumption** |
|---|---|
| 1 | Filling in missing data with the mean value of the feature won't our model accuracy |
| 2 | The distribution of variables won't affect the model accuracy |
| 3 | Resampling won't cause generalization issues |

## *1.5 Step 5. Project Constraints*

| | |
|---|---|
| Project Start Date | 09/07/2023 |
| Launch/Go-Live Date | 10/26/2023 |
| Project End Date | 11/16/2023 |
| List any hard deadline(s) | 09/07/2023<br>09/28/2023<br>10/12/2023<br>10/26/2023<br>11/16/2023 |
| List other dates/descriptions of key milestones | None |
| Budget constraints<br>Enter information about project budget limitations | N/A |
| Quality or performance constraints<br>Enter any other requirements for the functionality, performance, or quality of the project | N/A |
| Equipment/personnel Constraints<br>Enter any constraints regarding equipment or people that will impact the project | N/A |
| Regulatory constraints<br>Enter any legal, policy or other regulatory constraints | N/A |

## *1.6 Step 6. Updated Estimates*

| | | | |
|---|---|---|---|
| Estimate T&C hours required to complete project | Enter total # of T&C hours | If charge-back project, list total estimated T&C cost | Enter N/A if not applicable. |

# 2 Introduction

Polycystic Ovary Syndrome (PCOS) is a hormonal disorder that impacts women of reproductive age, potentially causing symptoms like hormonal imbalances, irregular menstrual cycles, and emotional disturbances, as recognized by the World Health Organization (WHO). It affects approximately 8–13% of women in their reproductive years, with as many as 70% of cases going unnoticed. Diagnosing PCOS can be challenging since its symptoms can be attributed to various factors, such as heavy menstrual bleeding.

Georgetown University is deeply committed to fostering an environment focused on personalized patient care, tailored to individual needs. In pursuit of this mission, the university is embarking on the development of an application aimed at promptly identifying and diagnosing PCOS upon the entry of a patient's laboratory results into the system. The ultimate goal is to increase the rate of diagnosis.

To support this endeavor, our project group designed an application based on machine-learning approaches that harness existing patient data to enhance the speed and precision of PCOS diagnosis. The data used is from Kaggle called PCOS Diagnosis. The dataset contains 45 variables including the patient's biometric measurements, blood test, hormone data, and other symptoms. All of them are numerical and binary data. Model selection will be performed to select the most effective models for this dataset. Models for comparison include Logistic Regression, Ridge Regression, Lasso Regression, Decision Tree Classifier, Random Forest Classifier, Bagging Classifier, Gradient Boosting Classifier, and XGB Classifier. This innovative approach seeks to better evaluate the patient's situation based on the likelihood of PCOS, ultimately improving the overall management and treatment of PCOS.

# 3 Analysis of the Dataset and Trained Model

## *3.1 Exploratory Analysis and Visualization*

The dataset has 541 samples with 1 missing column in marriage status and 1 missing value in AMH. The mean of the feature was used to fill in the missing value. Our project group first did the exploratory analysis to obtain the distribution of our data and the relationship between the target and feature variables.

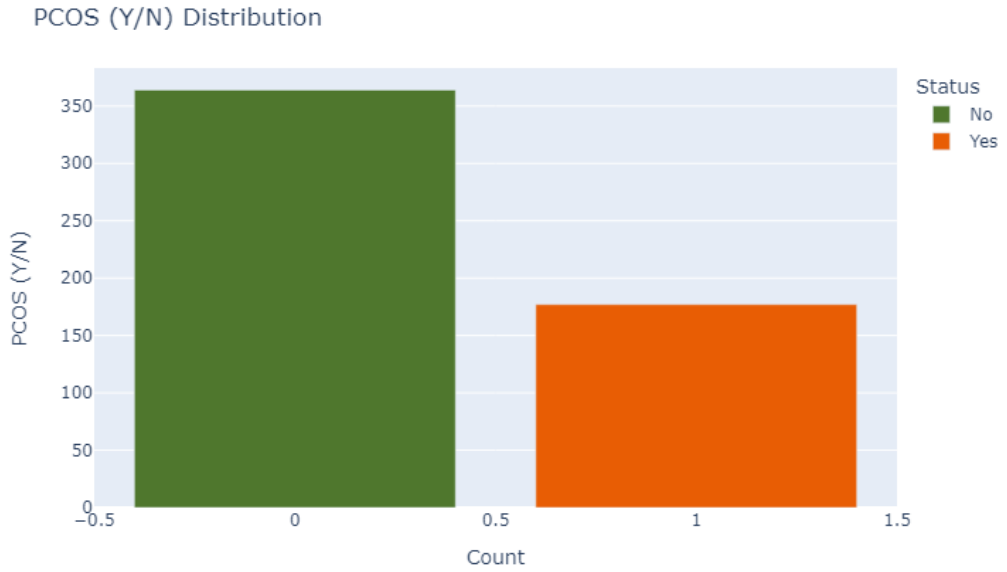


*Figure 1 Distribution of the dependent variable (PCOS)*

Figure 1 shows the distribution of the dependent variable, PCOS. Our dataset contains 177 positive samples and 364 negative samples. Due to the imbalanced distribution, we considered using a resampling method called SMOTE to increase the number of positive samples.

*Table 1 Summary of the Numerical Variables*

| param. | Age (yrs) | Weight (Kg) | Height (Cm) | BMI | Pulse rate (bpm) | RR (brths /min) | Hb(g/dl) | Cycle length (days) | Marriage Status (Yrs) | No. of absorptions |
|---|---|---|---|---|---|---|---|---|---|---|
| mean | 31.43 | 59.64 | 156.48 | 24.32 | 73.25 | 19.24 | 11.16 | 4.94 | 7.68 | 0.29 |
| std | 5.41 | 11.03 | 6.03 | 4.05 | 4.43 | 1.69 | 0.87 | 1.49 | 4.8 | 0.69 |
| min | 20 | 31 | 137 | 12.42 | 13 | 16 | 8.5 | 0 | 0 | 0 |
| 25% | 28 | 52 | 152 | 21.71 | 72 | 18 | 10.5 | 4 | 4 | 0 |
| 50% | 31 | 59 | 156 | 24.24 | 72 | 18 | 11 | 5 | 7 | 0 |
| 75% | 35 | 65 | 160 | 26.64 | 74 | 20 | 11.7 | 5 | 10 | 0 |
| max | 48 | 108 | 180 | 38.9 | 82 | 28 | 14.8 | 12 | 30 | 5 |

| param. | I beta-HCG(mIU/mL) | II beta-HCG(mIU/mL) | FSH(mIU/mL) | LH(mIU/mL) | FSH/LH | Hip (inch) | Waist (inch) | Waist: Hip Ratio | TSH (mIU/L) | AMH(ng/mL) |
|---|---|---|---|---|---|---|---|---|---|---|
| mean | 664.55 | 5.62 | 14.6 | 6.47 | 6.9 | 37.99 | 33.84 | 0.89 | 2.98 | 5.62 |
| std | 3348.92 | 5.88 | 217.02 | 86.67 | 60.69 | 3.97 | 3.6 | 0.05 | 3.76 | 5.88 |
| min | 1.3 | 0.1 | 0.21 | 0.02 | 0 | 26 | 24 | 0.76 | 0.04 | 0.1 |
| 25% | 1.99 | 2.01 | 3.3 | 1.02 | 1.42 | 36 | 32 | 0.86 | 1.48 | 2.01 |
| 50% | 20 | 3.7 | 4.85 | 2.3 | 2.17 | 38 | 34 | 0.89 | 2.26 | 3.7 |
| 75% | 297.21 | 6.9 | 6.41 | 3.68 | 3.96 | 40 | 36 | 0.93 | 3.57 | 6.9 |
| max | 32460.97 | 66 | 5052 | 2018 | 1372.83 | 48 | 47 | 0.98 | 65 | 66 |

| param. | PRL(ng/mL) | Vit D3 (ng/mL) | PRG(ng/mL) | RBS(mg/dl) | BP_S(mmHg) | BP_D(mmHg) | Follicle No. (L) | Follicle No. (R) | Avg. F size (L) (mm) | Avg. F size (R) (mm) | Endometrium (mm) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mean | 24.32 | 49.92 | 0.61 | 99.84 | 114.66 | 76.93 | 6.13 | 6.64 | 15.02 | 15.45 | 8.48 |
| std | 14.97 | 346.21 | 3.81 | 18.56 | 7.38 | 5.57 | 4.23 | 4.44 | 3.57 | 3.32 | 2.17 |
| min | 0.4 | 0 | 0.05 | 60 | 12 | 8 | 0 | 0 | 0 | 0 | 0 |
| 25% | 14.52 | 20.8 | 0.25 | 92 | 110 | 70 | 3 | 3 | 13 | 13 | 7 |
| 50% | 21.92 | 25.9 | 0.32 | 100 | 110 | 80 | 5 | 6 | 15 | 16 | 8.5 |
| 75% | 29.89 | 34.5 | 0.45 | 107 | 120 | 80 | 9 | 10 | 18 | 18 | 9.8 |
| max | 128.24 | 6014.66 | 85 | 350 | 140 | 100 | 22 | 20 | 24 | 24 | 18 |

Table 1 presents the statistics of the numerical variables in our dataset.

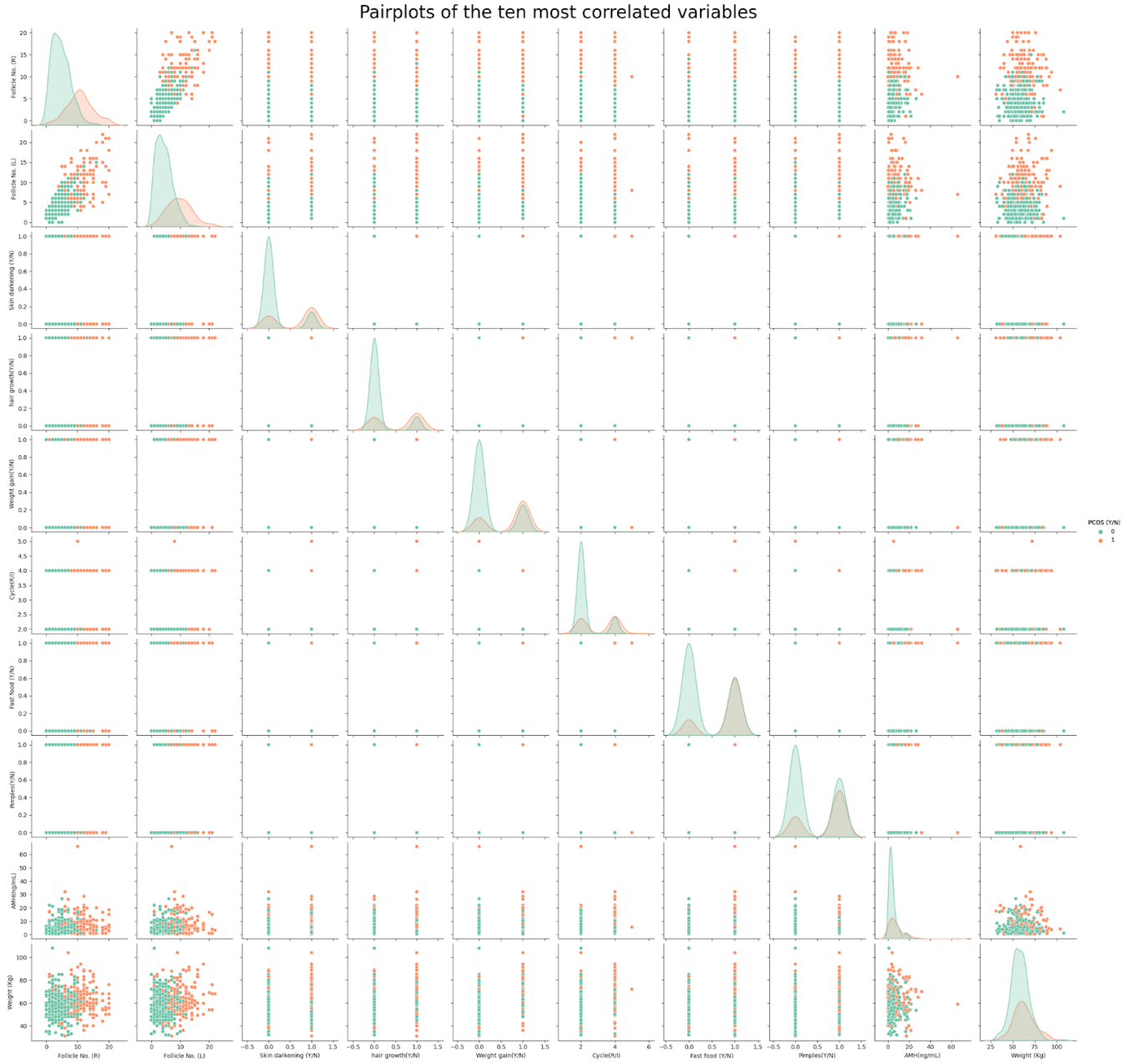


*Figure 2 Pairplot of ten most correlated variables*

Since our dataset contains 42 different features, we selected the ten most correlated variables based on the correlation coefficient and visualized their distributions. Different colors in Figure 2 indicate distinct diagnostic results. The results indicate that only weight is normally distributed, while both Follicle No. (right and left) are right-skewed and AMH gathers near 0. However, speaking of different diagnostic groups, Follicle numbers and AMH become closer to normal distributions. Moreover, from Figure 2 we can primarily assume that phenomena like skin darkening, hair growth, weight gain, and lifestyles of eating fast food have important influences on the occurrence and diagnosis of PCOS. Highly correlated variables, notably BMI, FSH/LH, and Waist(inch), were excluded from the dataset to ensure independence. The resultant clean dataset comprises 38 variables, with the target being the diagnosis of PCOS.

### *3.2 Baseline Model*

In this study, logistic regression served as our baseline model to classify the provided dataset. Logistic regression is a widely used method for binary classification due to its simplicity, interpretability, and efficiency. The model's coefficients and intercept were computed, providing essential insights into the influence of predictor variables on the target. From Table 2, based on the large positive values of hair growth, skin darkening, and fast food, we can indicate that these features have significant positive influences on the probability of being diagnosed with PCOS. We also evaluated the model's performance, resulting in an $R^2$ score of 0.576 and an accuracy of 0.894, affirming it is somehow competent in predicting PCOS diagnosis accurately. To enhance predictive performance, we introduced ensemble techniques, with logistic regression acting as a secondary level within the stacking framework. This strategic integration aims to synergize the strengths of various models and improve the overall predictive accuracy, paving the way for a more reliable diagnostic model for PCOS based on the provided dataset.

$$Logit(P(PCOS)) = -0.046 - 0.036 \times Age + 0.079 \times Weight - 0.041 \times Height - 0.125 \times BloodGroup + \ldots + 0.068 \times Endometrium$$

*Table 2 Logistic Regression Coefficients*

| Features | coeff | Age (yrs) | Weight (Kg) | Height (Cm) | Blood Group | Pulse rate (bpm) | RR (breaths/min) | Hb(g/dl) | Cycle (R/I) | Cycle length (days) | Marriage Status (Yrs) | Pregnant (Y/N) | No. of abortions |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Est. Coeffs | -0.046 | -0.036 | 0.079 | -0.041 | -0.125 | 0.115 | -0.241 | 0.007 | 0.529 | -0.191 | -0.009 | -0.436 | -0.753 |
| **Features** | **I beta-HCG** | **II beta-HCG** | **FSH (mIU/mL)** | **LH (mIU/mL)** | **Hip (inch)** | **Waist:Hip Ratio** | **TSH (mIU/L)** | **AMH (ng/mL)** | **PRL (ng/mL)** | **Vit D3 (ng/mL)** | **PRG (ng/mL)** | **RBS (mg/dl)** | **Weight gain (Y/N)** |
| Est. Coeffs | -0.000077 | 0.00002 | -0.069 | 0.014 | -0.011 | -0.087 | 0.005 | 0.038 | -0.001 | 0.0004 | -0.419 | 0.001 | 0.563 |
| **Features** | **Hair growth (Y/N)** | **Skin darkening (Y/N)** | **Hair loss (Y/N)** | **Pimples (Y/N)** | **Fast food (Y/N)** | **Reg.Exercise (Y/N)** | **BP_Systolic (mmHg)** | **BP_Diastolic (mmHg)** | **Follicle No. (L)** | **Follicle No. (R)** | **Avg. F size (L) (mm)** | **Avg. F size (R) (mm)** | **Endometrium (mm)** |
| Est. Coeffs | 0.987 | 1.136 | -0.177 | 0.421 | 1.004 | -0.267 | -0.018 | -0.041 | 0.243 | 0.374 | 0.010 | 0.018 | 0.067 |

# 4 Model Selection

Obtaining a baseline provided us necessary insight for model selection. Following the preliminary regression analysis, we have decided to evaluate LogisticRegression, KNeighborsClassifier, DecisionTreeClassifier, SVC, GaussianNB, RandomForestClassifier, BaggingClassifier, GradientBoostingClassifier and XGboostclassifier as candidate models. Figure 3 shows the model results, and we can see that LogisticRegression, RandomForest, Bagging, GradientBoosting, and XGboost perform the best (have relatively highest accuracy). Because of this, we will select them as the candidate models for level 0 in the stacking classifier. We will use LogisticRegression as the level 1 combiner or metamodel to aggregate the results of the level 0 models.

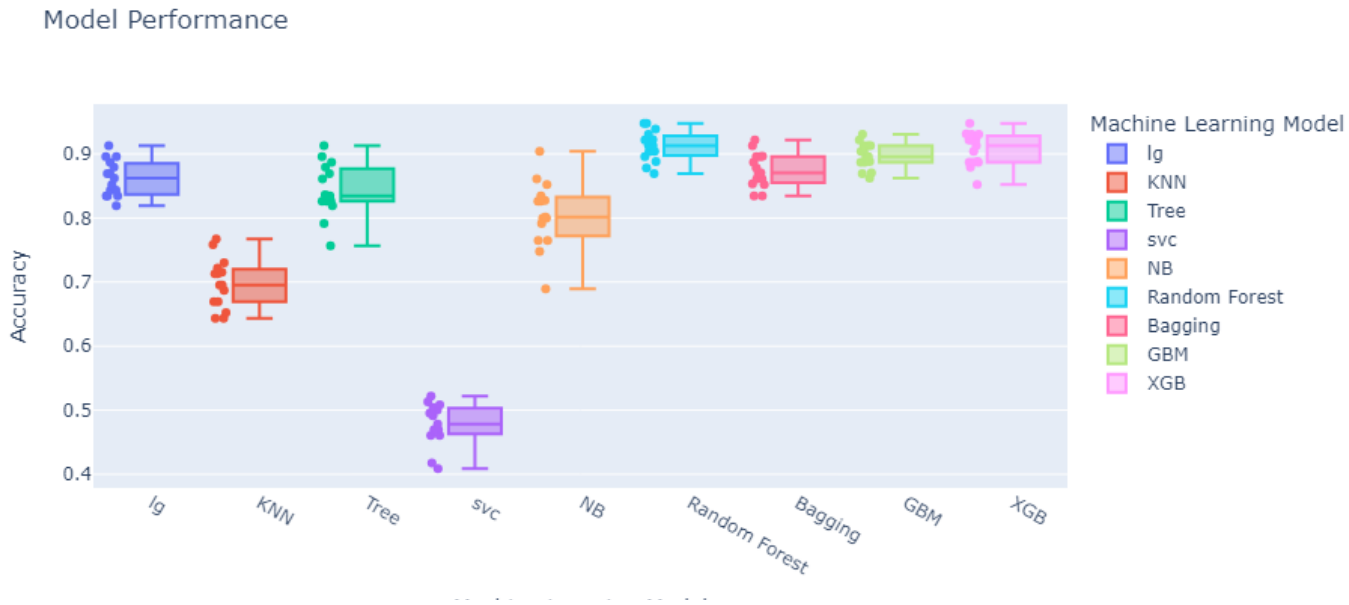


*Figure 3 Machine Learning Model Result*

In addition to the selected models, we have also opted for cross-validation of the data set. For the current iteration of the model, we will use five-fold validation and repeat three times. Figure 4 shows the performance of the standalone and stacked models. The performance of the stacked model works better than most of the models with the highest mean accuracy. The performance of the stacked model can also be improved by tuning the hyperparameters to have a smaller accuracy range. For now we will focus on the PCOS diagnostic classification by using a stacked model.

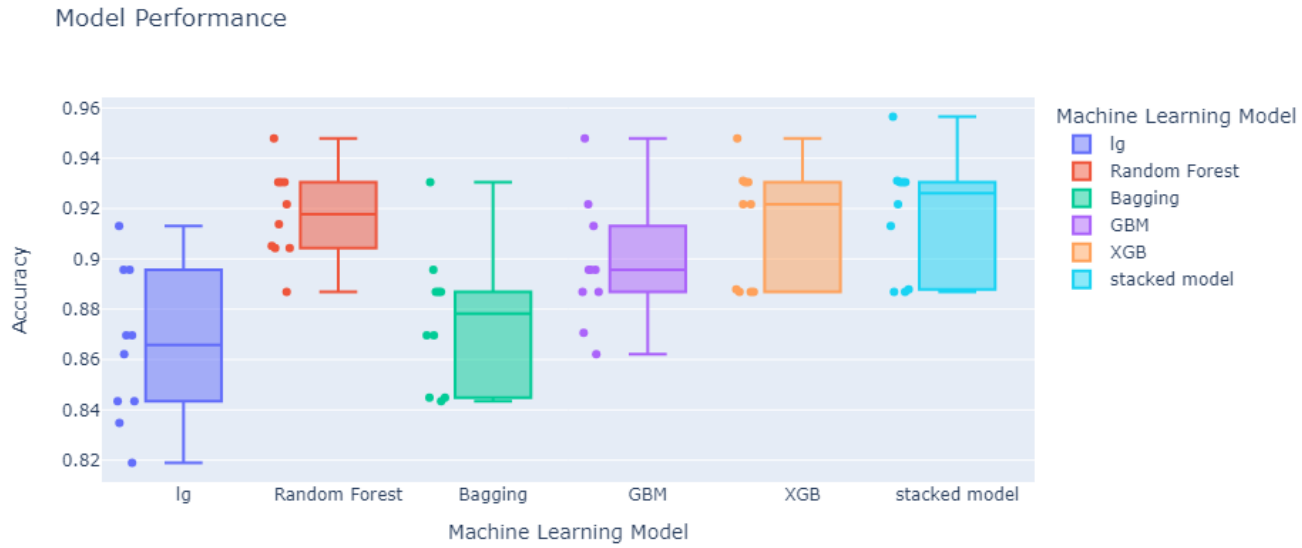


*Figure 4 Candidate model and stacked model performance*

## *4.1 Model Performance Evaluation*

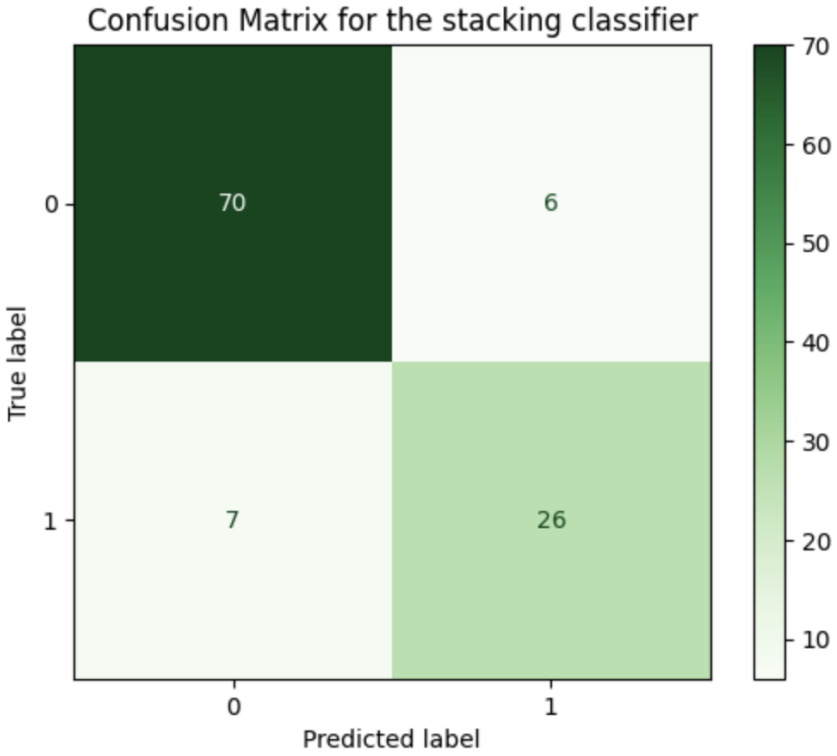


*Figure 5 Confusion matrix of Stacking Classifier*

The above Figure 5 illustrates the confusion matrix of the stacked classifier. The high count of true negatives (70) underscores the model's proficiency in accurately identifying individuals without PCOS, showcasing its specificity to capture true negative cases. Additionally, the low count of false positives (6) indicates the model's ability to distinguish non-PCOS cases, minimizing misclassifications of individuals without the condition. However, the false negatives (7) represent missed opportunities for accurate PCOS diagnosis, while the true positives (26) reflect the correct identification of individuals with PCOS, emphasizing the model's sensitivity.