Predicting Presence of Polycystic Ovary Syndrome

Lao, Kevin Christian, Ambita
kelao@up.edu.ph
University of the Philippines Visayas College Of Arts And Sciences, *Division Of Physical Sciences And Mathematics*

1. INTRODUCTION

Polycystic Ovary Syndrome or PCOS is common in women of childbearing age with an irregular menstrual cycle and androgen excess. The occurrence of PCOS in women of reproductive ages today is between five to 20 percent. It is a serious disease that forms cysts in the ovary due to excess androgen (male hormone). [1]

PCOS has been prevalent in endometrial cancer patients who belong to the premenopausal age group (30-39 years old), and in the Philippines, this prevalence is apparent currently. In a study of Ortega, G., & Aguilar, A., they gathered data from 487 Filipino endometrial cancer patients, and 61 (12.56%) have PCOS. 34 out of 61 belong to premenopausal and menopausal age groups. According to them, Filipino women with endometrial cancer are more prone to PCOS and are more likely related to obesity, nulligravida (women who never got pregnant), nulliparous (had pregnant but never had a child), and records of abnormal blood glucose. [2] Some common symptoms of PCOS include:

- Infrequent or absent menstrual periods
- Excess body hair (hirsutism)
- Acne
- Weight gain or difficulty losing weight
- Balding or thinning hair on the scalp
- Skin tags
- Darkened skin on the neck, groin, or underarms [3]

For this study, an available PCOS dataset from Kaggle repository is used to train the machine learning models. Since the dataset does classify the presence of PCOS, the classification accuracy is the important metric to consider when evaluating the model performance. It is important to accurately classify whether or not a person has PCOS in order to provide the appropriate treatment and management. The paper is arranged as follows: section 2 includes the methodology, sections 3 and 4 are the results and discussion, and conclusion. Section 5 is the reference.

2. METHODOLOGY

In our study of PCOS, we aimed to understand the factors that contribute to the development of PCOS and to identify potential predictors of the condition. To accomplish these goals, we used the dataset provided by Kottarathil. It has 42 attributes of 541 women. We then applied various statistical techniques, including univariate feature selection and machine learning algorithms, to identify the most important predictors of PCOS. We also performed correlation analyses and summary statistics to better understand the relationships between the different variables in our dataset. The dataset needs to go through several steps to preprocess and clean the data.

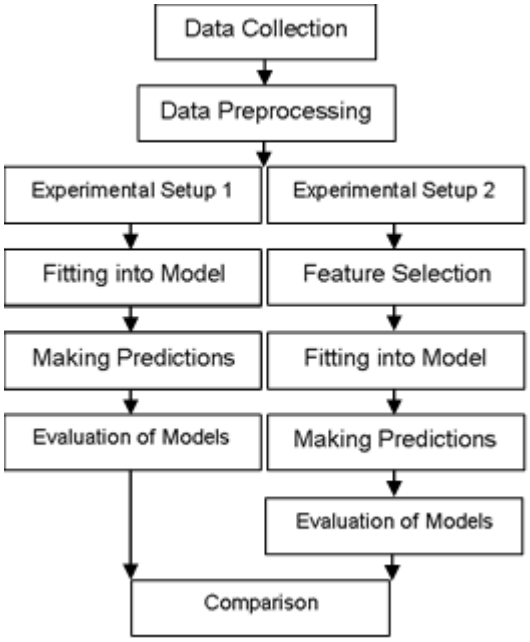


Figure 1. Block Diagram of the Methodology.

a. Data Collection

The first step is to collect the dataset. Multiple online repositories provide free and available datasets. In this study, the PCOS dataset provided by Kottarathil from Kaggle will be used. The data was gathered across 10 different hospitals in Kerala, India. Identities of the female subjects remained undisclosed.

b. Data Preprocessing

It is important to preprocess and clean the data to evaluate the performance without noise. Categorical data were converted to ordinal values. The dataset was also standardized.

Experimental setups were varied into two setups: (1) setup where all features are selected, and (2) setup where features selection was implemented.

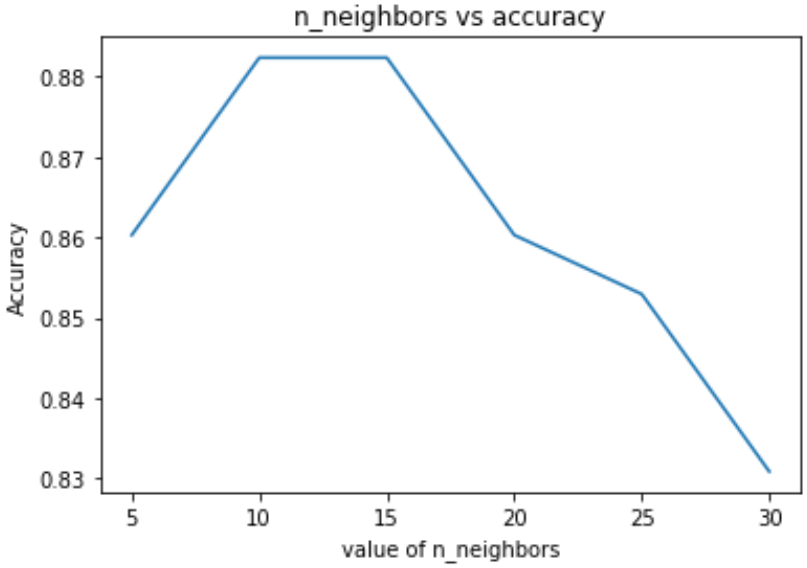
c. Experimental Setup 1

a. Fitting into Model and Making Predictions

With the data cleaned and selected, it is now ready to be processed by the models. The models used were KNN, SVM, logistic regression, RF, and XGboost. Using the prepared models, predictions were made with the standardized testing set. The k-fold crossvalidation approach was used to determine the performance of each model across different splits of data.

1. **K-Nearest Neighbor**

K-nearest neighbors (KNN) is a simple and effective technique for classification and regression. It works by storing all available cases and classifying new cases based on a similarity measure (e.g., distance functions). Classification is done by a majority vote to its neighbors. [4]



As shown in figure 2, it is concluded that k=10 and k=15 got the highest accuracy value of ~88.23%, with ~96% precision and ~64% recall. All features were included. Therefore, k=10 and k=15 were used on the KNN classifier for this setup.

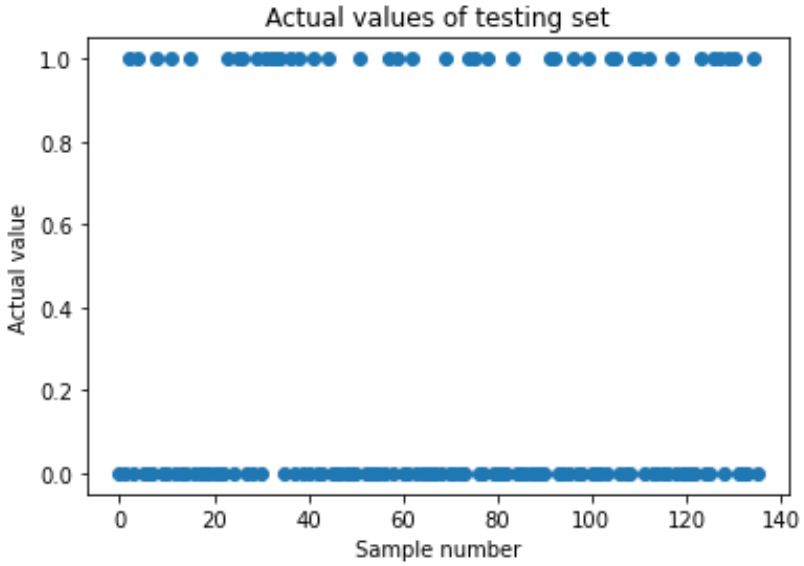


Figure 3. Actual values of testing set for KNN

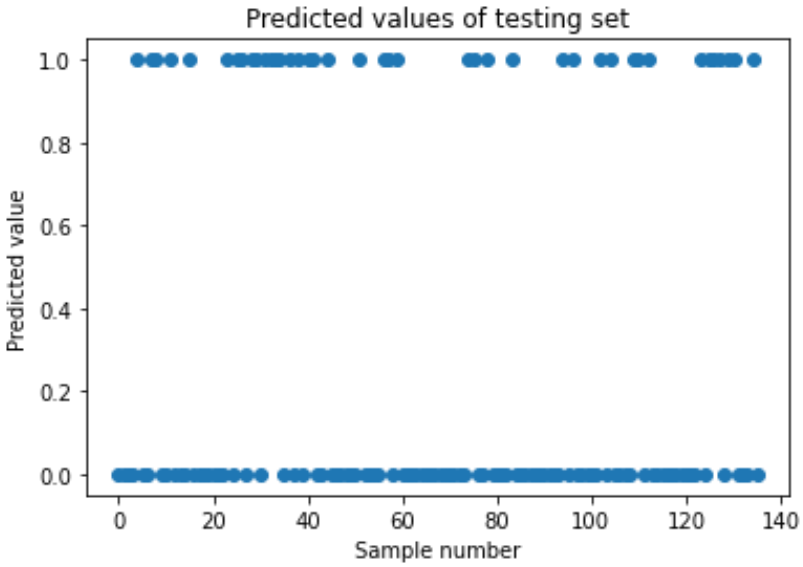


Figure 4. Predicted values of testing set for KNN

Figure 3 and Figure 4 shows the scatter plot of actual and predicted values of testing set for KNN with all features. Many a predicted values differ from the actual values, with an accuracy score of ~88%, precision ~96%, and ~64% recall. Which means, KNN only classified ~64% of the samples correctly.

The mean validation score of KNN is about 86%. This suggest that the model is performing consistently in both training and testing set.

2. **Support Vector Machine**

SVM algorithm finds the hyperplane in a high-dimensional space that maximally separates the different classes.

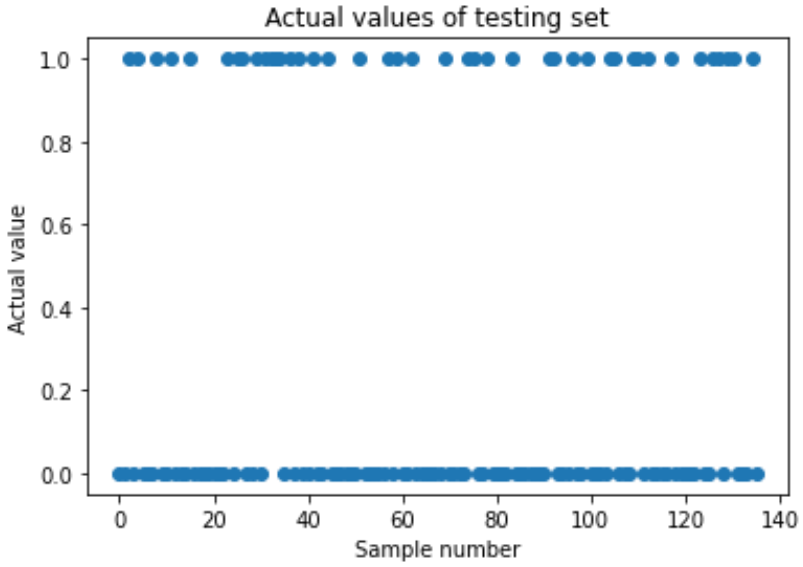


Figure 5. Actual values of testing set for SVM

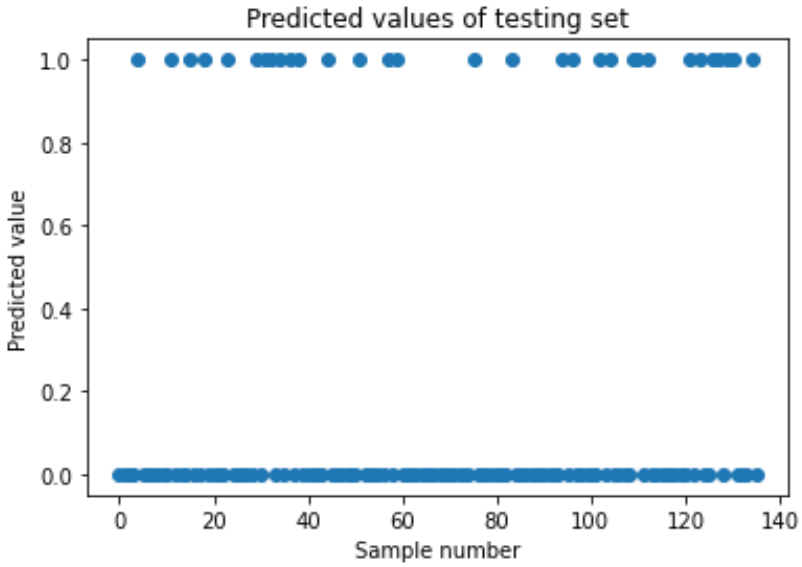


Figure 6. Predicted values of testing set for SVM

Figure 5 and Figure 6 shows the scatter plot of actual and predicted values of testing set for SVM with all features. Many a predicted values differ from the actual values, with an accuracy score of ~86%, precision ~87%, and ~64% recall. Which means, SVM only classified ~64% of the samples correctly.

The mean validation score of SVM is about 90%. This strongly suggest that the model is performing consistently in both training and testing set for predicting presence of PCOS.

3. **Logistic Regression**

Logistic regression is a statistical model that is used to predict a binary outcome from a set of independent variables. In this study, the scaled training set was fitted into the model and calculated the accuracy, precision and recall.

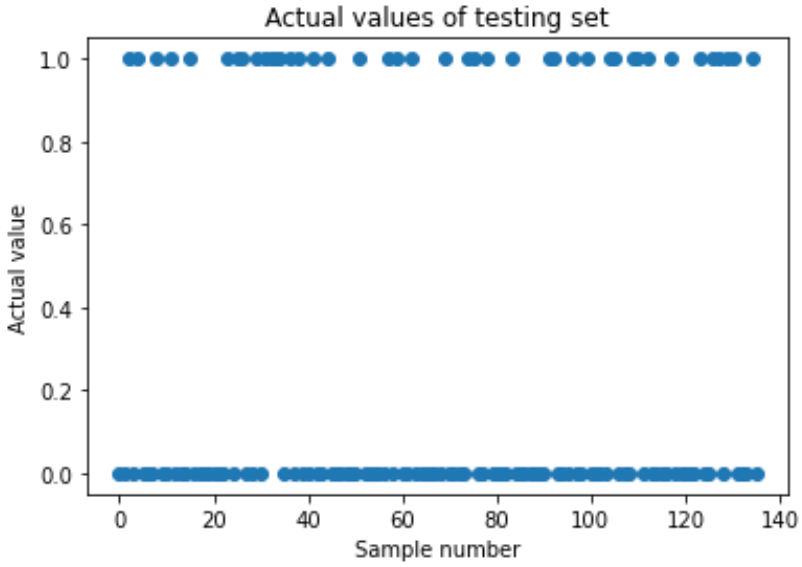


Figure 7. Actual values of testing set for LR

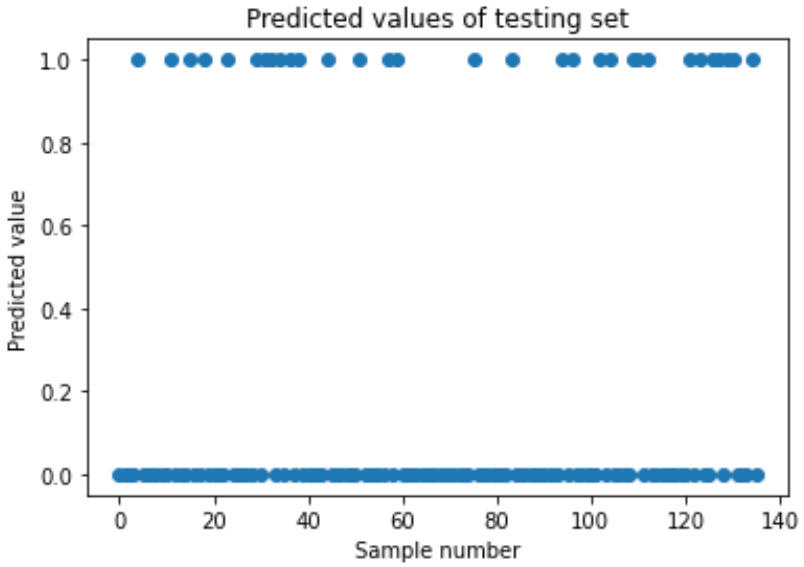


Figure 8. Predicted values of testing set for LR

Figure 7 and Figure 8 shows the scatter plot of actual and predicted values of testing set for LR with all features. Many a predicted values differ from the actual values, with an accuracy score of ~84%, precision ~78%, and ~69% recall. Which means, LR only classified ~69% of the samples correctly.

The mean validation score of LR is about 88%. This suggest that the model is performing consistently in both training and testing set.

4. **Random Forest**

The concept behind random forests is that a large number of decision trees, each trained on a random subset of the training data, can work together to make more accurate predictions than any individual decision tree.

Figure 9 and Figure 10 shows the scatter plot of actual and predicted values of testing set for RF with all features. Many a predicted values differ from the actual values, with an accuracy score of ~91%, precision ~91%, and ~78% recall. Which means, LR only classified ~78% of the samples correctly. This is the highest accuracy among the other models for this setup.

The mean validation score of RF is about 91%. This strongly suggests that the model is performing consistently in both training and testing set.

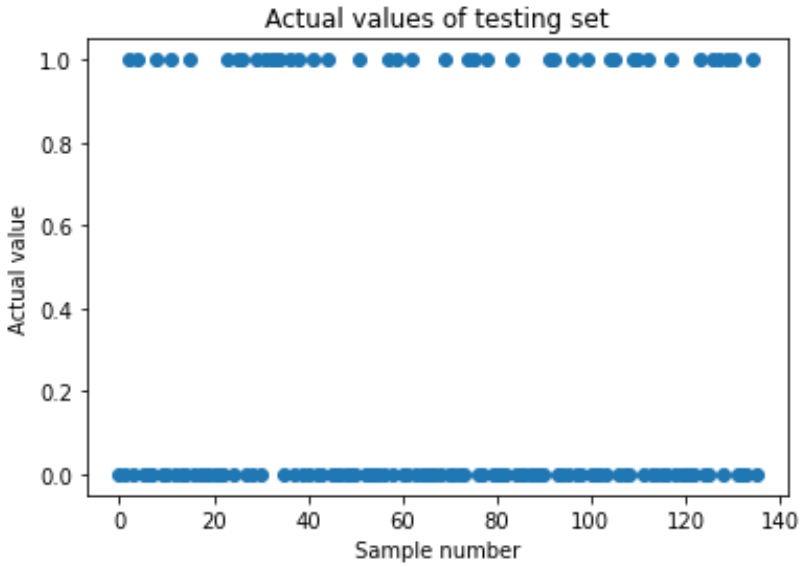


Figure 9. Actual values of testing set for RF

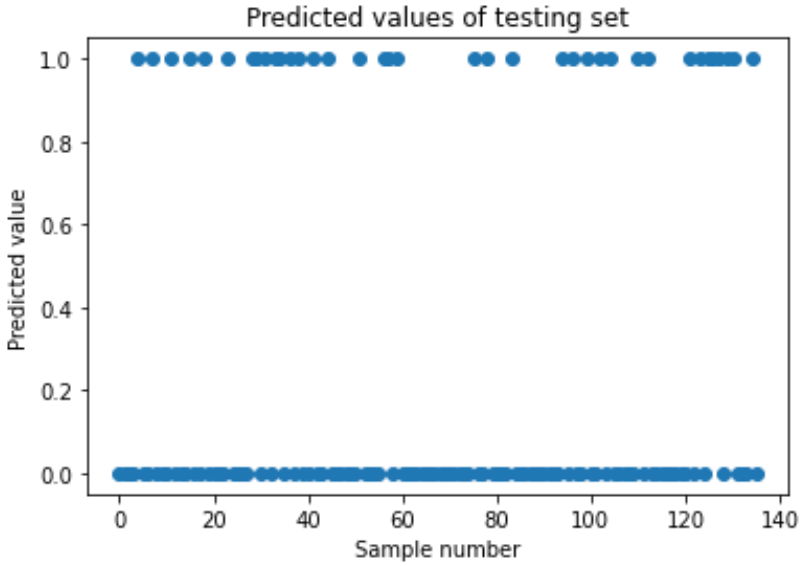


Figure 10. Predicted values of testing set for RF

5. **Extreme Gradient Boosting**

XGBoost works by building a model in the form of an ensemble of weak learners (e.g., decision trees), and iteratively improving the model by adding new weak learners that correct the mistakes of the previous ones.

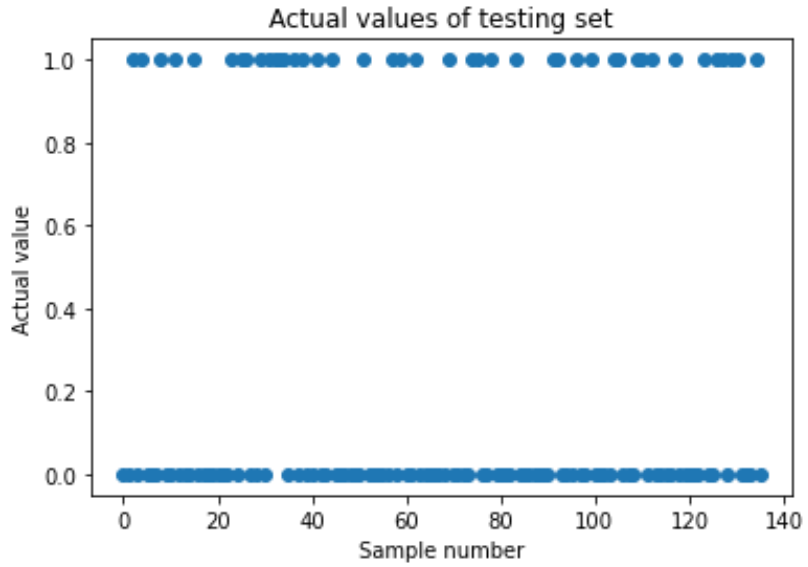


Figure 11. Actual values of testing set for XGB

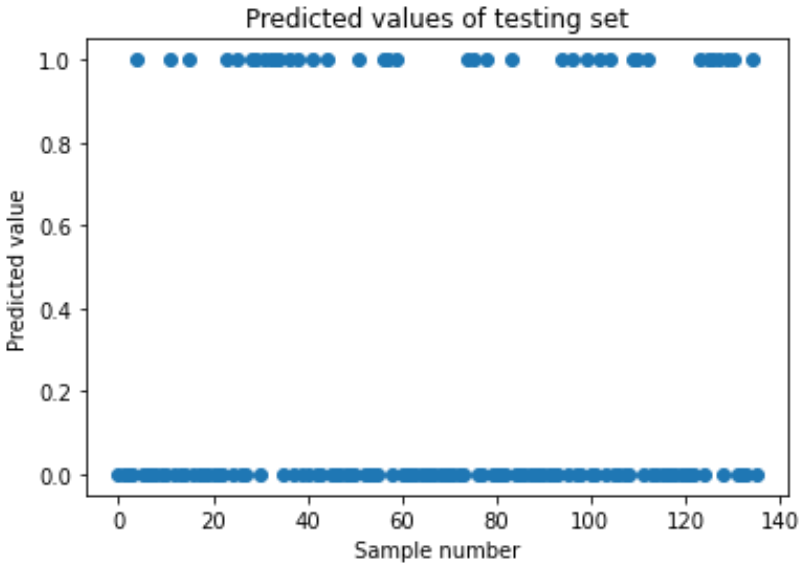


Figure 12. Predicted values of testing set for XGB

Figure 11 and Figure 12 shows the scatter plot of actual and predicted values of testing set for RF with all features. Few a predicted values differ from the actual values, with an accuracy score of ~89%, precision ~85%, and ~80% recall.

The mean validation score of RF is about 89%. This suggests that the model is performing consistently in both training and testing set.

b. **Evaluation of Models**

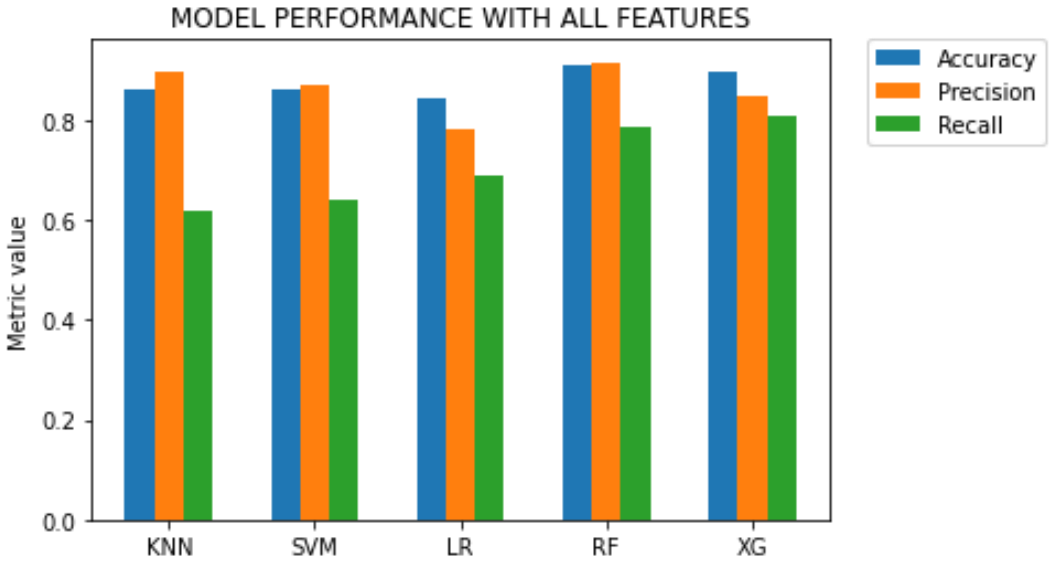


Figure 13. Evaluation metrics for all models

Figure 13 is a visual comparison of accuracy, precision, and recall of the different models. RF has the highest accuracy of 0.91 and XG has the second highest accuracy of 0.89. LR has the lowest accuracy for this setup with 0.84. In terms of precision, RF also has the highest score of 0.91. Second highest is KNN with 0.89 and LR having the lowest precision of 0.78. in terms of recall, XG has the highest with 0.8 and followed by RF with 0.78. Table 1 shows the numerical comparison of the evaluation scores of each model.

	Model	Accuracy	Precision	Recall
0	KNN	0.860294	0.896552	0.619048
1	SVM	0.860294	0.870968	0.642857
2	LR	0.845588	0.783784	0.690476
3	RF	0.911765	0.916667	0.785714
4	XG	0.897059	0.850000	0.809524

Table 1. Evaluation metrics of models with all features

d. Experimental Setup 2.

a. **Feature Selection**

imageIn this study, we used the SelectKBest method for feature selection. This method selects a specified number of the highest scoring features based on statistical tests. In this case, we selected the top 10 features based on their pvalues, which indicated the probability that the relationship between the feature and the target variable was due to chance. There were no significant difference if we selected k number of features other than 10. This allowed us to narrow down the number of features in our dataset and improve the interpretability of the models. The 10 selected features include Weight (Kg),

BMI, Cycle(R/I), weight gain(Y/N), hair growth(Y/N), skin darkening (Y/N), pimples(Y/N), fast food (Y/N), follicle No. (L), and follicle No. (R), with the top three are BMI, follicle No. (L), and follicle No. (R). The top three were identified using the chi-squared approach.

b. Fitting into Model and Making Predictions

With the data cleaned and selected, it is now ready to be processed by the models. The models used were KNN, SVM, logistic regression, RF, and XGboost. Using the prepared models, predictions were made with the standardized testing set. The k-fold crossvalidation approach was used to determine the performance of each model across different splits of data.

1. K-Nearest Neighbor

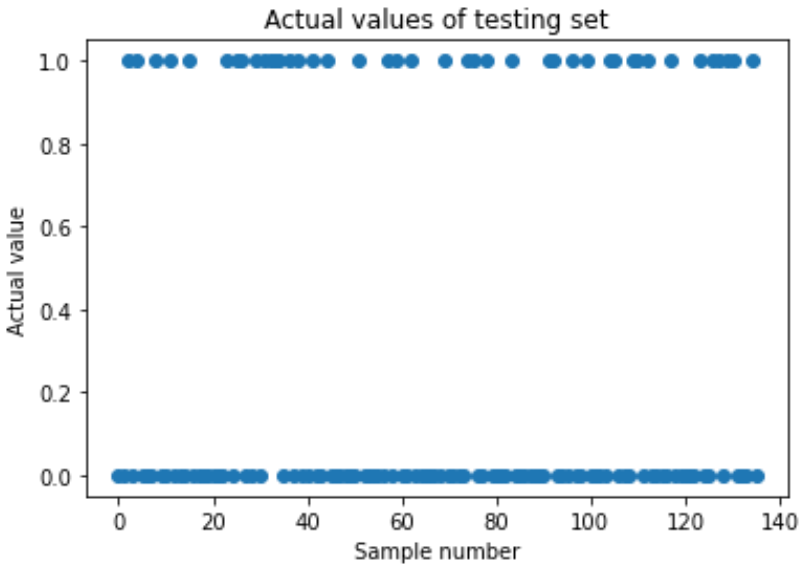


Figure 14. Actual values of testing set for KNN

Figure 14 and Figure 15 shows the scatter plot of actual and predicted values of testing set for RF with all features. Many a predicted values differ from the actual values, with an accuracy score of ~88%, precision ~96%, and ~64% recall.

The mean validation score of KNN is about 87%. This suggest that the model is performing consistently in both training and testing set.

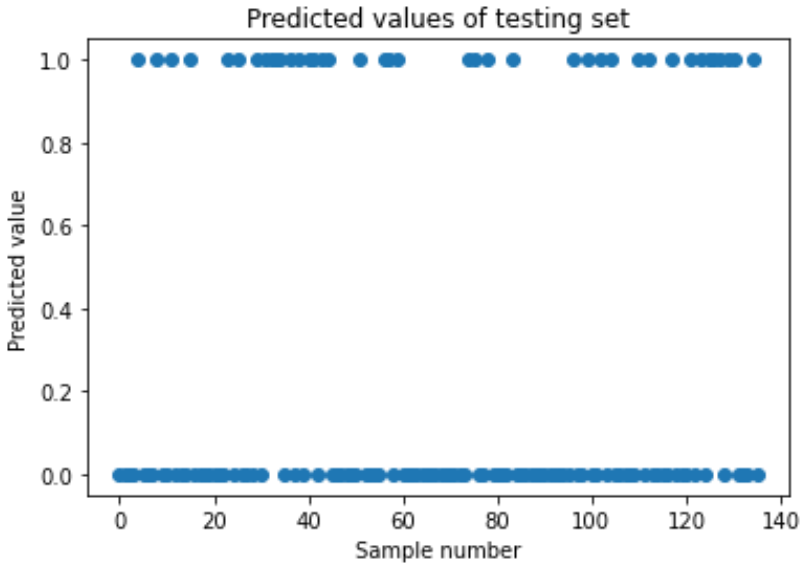


Figure 15. Predicted values of testing set for KNN

2. **SVM**

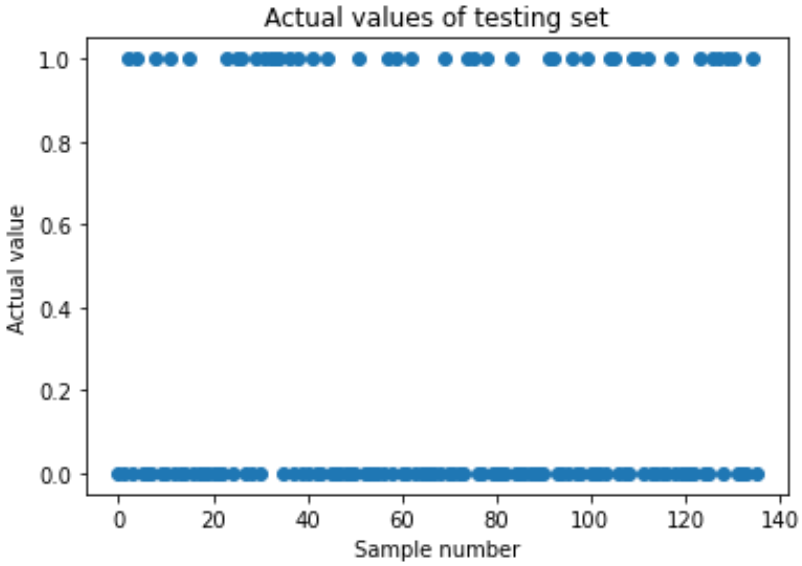


Figure 16. Actual values of testing set for SVM

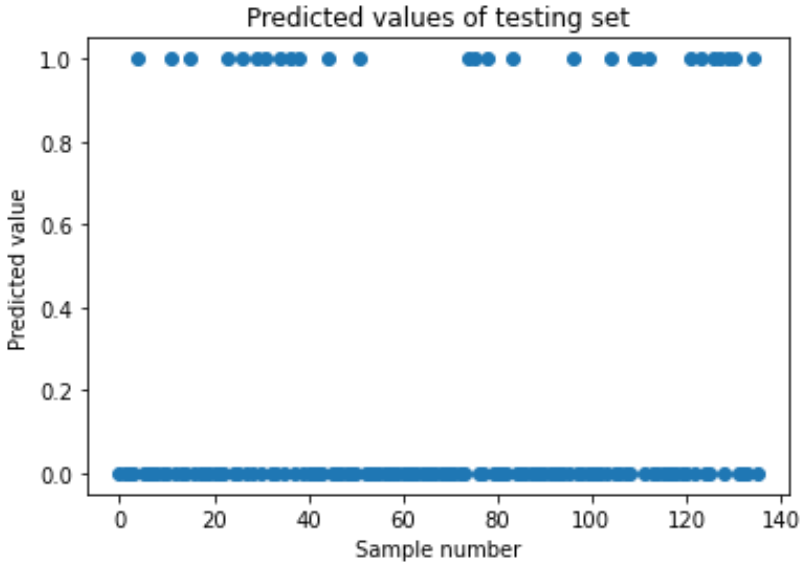


Figure 17. Predicted values of testing set for SVM

Figure 16 and Figure 17 shows the scatter plot of actual and predicted values of testing set for SVM with all features. Only a few predicted values differ from the actual values, with an accuracy score of ~91%, precision ~91%, and ~80% recall. Which means, SVM classified ~80% of the samples correctly.

The mean validation score of SVM is about 90%. This strongly suggests that the model is performing consistently in both training and testing set.

3. **Logistic Regression**

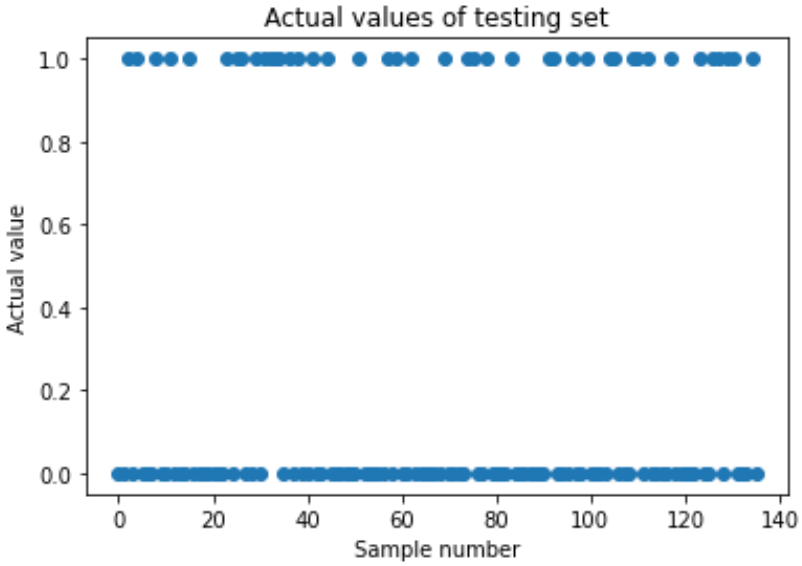


Figure 18. Actual values of testing set for LR

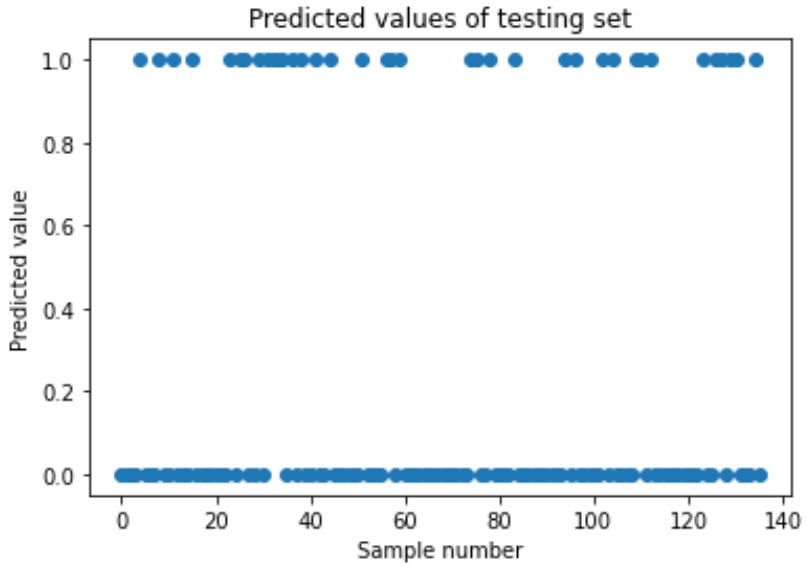


Figure 19. Predicted values of testing set for LR

Figure 18 and Figure 19 shows the scatter plot of actual and predicted values of testing set for SVM with all features. Only a few predicted values differ from the actual values, with an accuracy score of ~90%, precision ~87%, and ~80% recall. Which means, SVM classified ~80% of the samples correctly.

The mean validation score of SVM is about 92%. This strongly suggests that the model is performing consistently in both training and testing set.

4. **Random Forest**

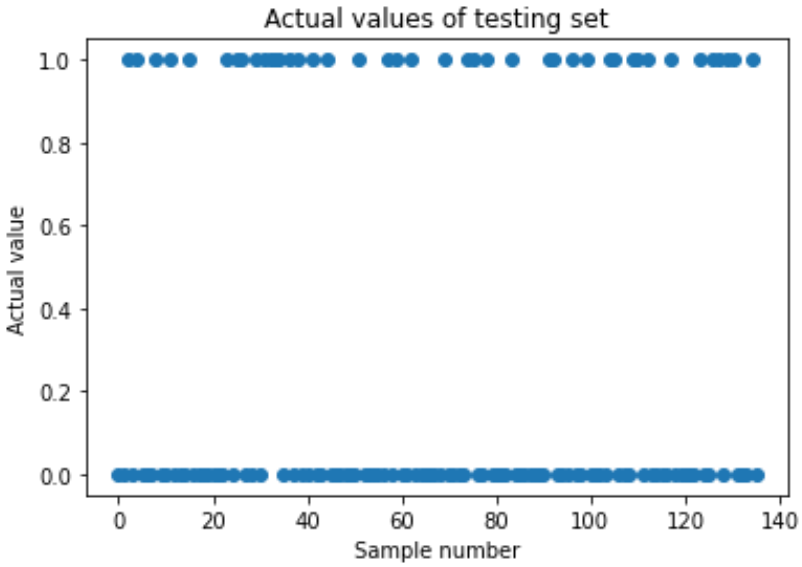


Figure 20. Actual values of testing set for RF

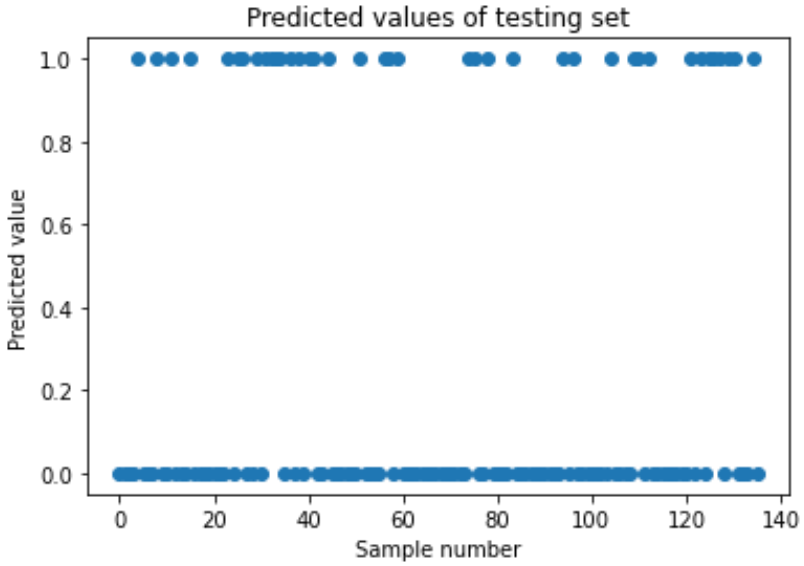


Figure 21. Predicted values of testing set for RF

Figure 20 and Figure 21 shows the scatter plot of actual and predicted values of testing set for SVM with all features. Only a few predicted values differ from the actual values, with an accuracy score of ~88%, precision ~82%, and ~80% recall. Which means, RF classified ~80% of the samples correctly.

The mean validation score of RF is about 90%. This suggest that the model is performing consistently in both training and testing set.

5. **Extreme Gradient Boosting**

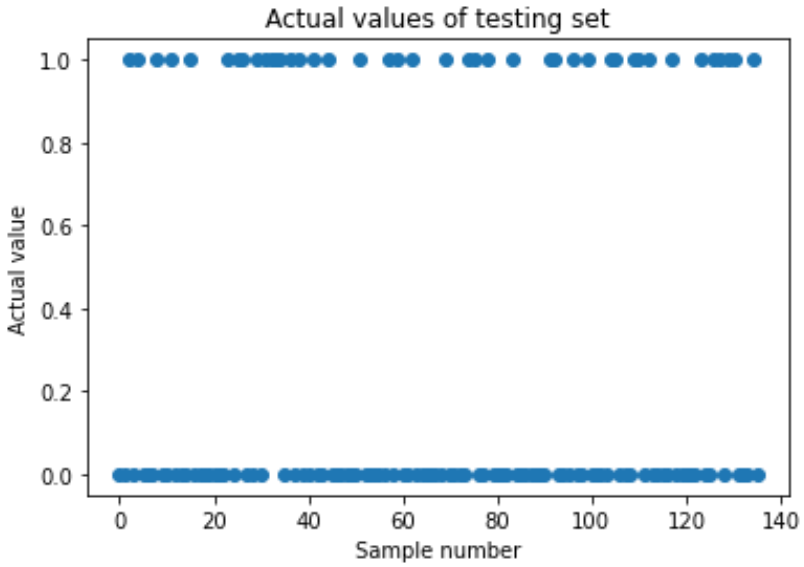


Figure 22. Actual values of testing set for RF

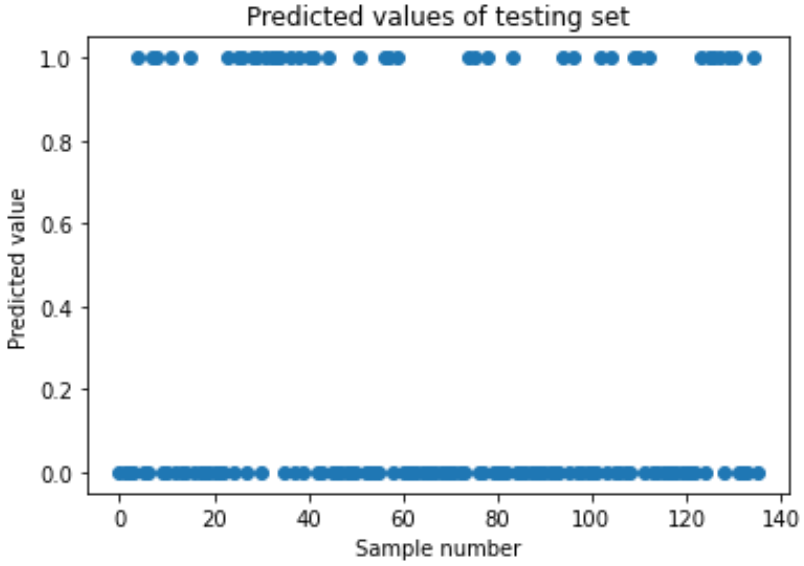


Figure 23. Predict values of testing set for RF

Figure 22 and Figure 23 shows the scatter plot of actual and predicted values of testing set for SVM with all features. Many a predicted values differ from the actual values, with an accuracy score of ~88%, precision ~84%, and ~78% recall. Which means, RF classified ~80% of the samples correctly.

The mean validation score of RF is about 89%. This suggest that the model is performing consistently in both training and testing set.

c. Evaluation of Models

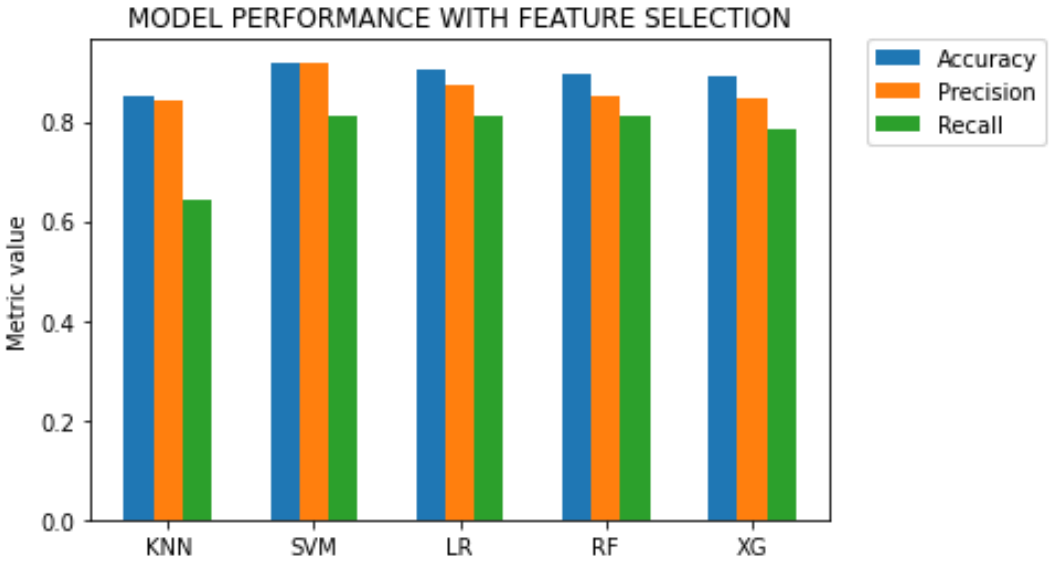


Figure 24. Evaluation metrics for all models (with feature selection)

	Model	Accuracy	Precision	Recall
0	KNN	0.852941	0.843750	0.642857
1	SVM	0.919118	0.918919	0.809524
2	LR	0.904412	0.871795	0.809524
3	RF	0.897059	0.850000	0.809524
4	XG	0.889706	0.846154	0.785714

Table 2. Evaluation metrics for all models (with feature selection)

Figure 24 is a visual comparison of accuracy, precision, and recall of the different models. SVM has the highest accuracy of 0.91 and LR has the second highest accuracy of 0.90. KNN has the lowest accuracy for this setup with 0.85. In terms of precision, SVM also has the highest score of 0.91. Second highest is LR with 0.87 and KNN having the lowest precision of 0.84. In terms of recall, SVM, LR, and RF has the highest with 0.8 and KNN as the lowest with 0.64. Table 1 shows the numerical comparison of the evaluation scores of each model.

e. **Correlation between selected features**

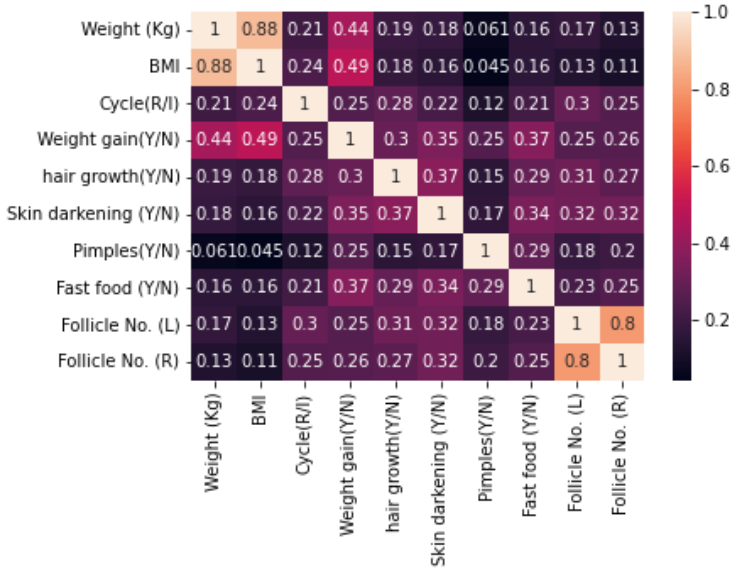


Figure 25. Heatmap of the correlation of different features selected

The feature "BMI" appears to have a strong positive correlation with the feature "Weight (Kg)" as indicated by the light shade of orange. On the other hand, the feature " follicle No. (L)" appears to have a strong positive correlation with the feature " follicle No. (R)" as indicated by the light shade of orange.

f. Comparative Analysis of Experimental Setup 1 and Experimental Setup 2

In table 1, random forest achieved the highest accuracy with ~0.91and precision ~0.91 among the tested models in setup 1. Random forest had the second highest recall, indicating that it was able to correctly identify a significant number of the positive cases.

It is followed by XGBoosting with ~0.89 accuracy, ~0.85 precision and ~0.80 recall. Possible implication of this is that all of the features included in the dataset may be useful for predicting PCOS. However, in setup 2, the performance of the SVM model improved significantly, with an accuracy and precision of ~0.92 and ~0.81 recall. This could be due to the application of feature selection, which likely helped the model to focus on the most relevant features and improve its performance.

We used F1 score to compare the performance of both models. We found that SVM from setup 2 has better performance than Random Forest from setup 1 with a F1 score of 0.86 while RF has 0.82. However, XGboosting has higher F1 score than random forest. It is possible that XGboosting performs better than random forest.

3. RESULTS

The study used 541 samples of data of patients from 10 different hospitals in India. A total of 42 features were used and out of which only 10 features were selected using univariate feature selection (SelectKBest method), with the strong positive correlation between weight and BMI, and Follicle No. (L) and Follicle No. (R) of females. A comparison was made between the two different classifiers from different setups: Random Forest from setup 1, where all features were selected, and SVM from setup 2 where 10 features were selected. The F1 score helped determine the best model between the two. The F1 score for RF is 0.825 and for that of SVM is 0.86, hence, model of Support Vector Machine is selected to determine the absence or presence of PCOS.

4. CONCLUSION

In this study, we propose a model of predicting the presence of PCOS in females. It was shown experimentally that the SVM can be used predicting the presence of PCOS. In future works, we are very interested in verifying the robustness of our model by testing it different feature selection methods. We can also try other machine learning algorithms and collect more or use different datasets to see if it improves model performance.

5. REFERENCES

[1] O. Negis, D. Brown, I. Galic, L. Zhaunova, Jain and T. Jain, "SUN-LB3 Relationship Between BMI and PCOS Symptoms Among Flo App Users in the United States.," 2020. [Online]. [Accessed 30 December 2022]

[2] G. Ortega and A. Aguilar, "Prevalence and Characteristics of Polycystic Ovary Syndrome (PCOS) in Filipino Women Diagnosed with Endometrial Cancer: A five-year Retrospective Study," *Philippine Journal of Reproductive Endocrinology and Infertility, ,* p. 13, 2016. [3] N. I. o. C. H. a. H. Development, "Polycystic Ovary Syndrome (PCOS)," 2020. [Online]. Available: https://www.nichd.nih.gov/health/topics/pcos. [Accessed 30 December 2022].

[4] S. Learn, "Nearest Neighbors: Scikit Learn Documentation," [Online]. Available: https://l.facebook.com/l.php?u=https%3A%2F%2Fscikit-learn.org%2Fstable%2Fmodules%2Fneighbors.html%3Ffbclid%3DIwAR3CWJLam9nwAGDsr8rQUClKOMD4bfIp8IvOi2twqXHk-x99lDqBL44Q4L8&h=AT3jA5o83tiCk4yUKQJzd4t_ARFQCrNmnwc2LxtD594-wu1iuJkxday5QB5kP6oo3Wyf7fdD7PoKH3x9U