Advanced Machine Learning for Diagnosis of Polycystic Ovarian Syndrome (PCOS)

Dr. U D Prasan[1], Y Sri Vidya[2], B L Prasanna[3], B Srija[4], Ch Ravi Teja[5], S Sai Siva[6]

[1]*Department of CSE & Aditya Institute of Technology And Management*
[2]*Department of CSE & Aditya Institute of Technology And Management*
[3]*Department of CSE & Aditya Institute of Technology And Management*
[4]*Department of CSE & Aditya Institute of Technology And Management*
[5]*Department of CSE & Aditya Institute of Technology And Management*
[6]*Department of CSE & Aditya Institute of Technology And Management*

---***---

Abstract- Polycystic Ovary Syndrome (PCOS) is one of the most common type of endocrine disorder in reproductive age women (15-49 years). An estimated one in five (20%) Indian women suffer from PCOS. The main focus of this paper is to detect PCOS patients based on clinical and metabolic parameters. In this paper we are using Advanced Machine Learning approaches like Random Forest, AdaBoost, XGBoost, Bagging.The results are analysed and performance of the algorithms is validated on the basis of accuracy, precision, recall, F1 Score, AUC_ROC. The performancemetrics indicate the highest i.e. 97% accuracy of XGBoostalgorithms in the diagnosis of PCOSon giving data.

***Key Words*:** Syndrome, endocrine, metabolic, AdaBoost, XGBoost, Bagging

1.INTRODUCTION

Polycystic Ovary Syndrome(PCOS) is a common health issue in women caused by an imbalance of hormones related to reproductive system.This may leads to infertility and anovulation.Women with PCOS majorly suffer from weight gain, ovarian cysts, Irregular periods, acne, facial hair, depression, anxiety, Skin pigmentation and heavy periods.The cause of PCOS has not well understood, but may involve a combination of genetic and environmental factors.Diagnosis is often challenging since there are many nonspecific signs and symptoms. If not monitored in time, the condition can have serious health impacts like type 2 diabetes, cardiovascular disease, hormonal dysfunction, endometrial and infertility. There is no cure for PCOS. Treatment helps in management of the disease and symptoms associated with it such as hirustism, acne, hormonal imbalance, infertility and obesity. Lifestyle modifications such as weight loss, suitable diet intake, regularexercise can help in the management of the disease. This paper focuses on the prediction of PCOS.

The rest of the paper is organized as follows. Section II describes the Proposed Algorithm, Section III describes Design Methodology, Section IV describes Experiments and Results, Section V describes Conclusion and References are given in Section V.

2. Proposed Algorithm

i. **XGBoost**: XgBoost stands for Extreme Gradient Boosting. XgBoost is an open-source implementation of Gradient Boosted decision trees. XgBoost dominates structured or tabular datasets on classification and regression predictive modelling problems.
In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It works on regression, classification, ranking and user-defined prediction problems.

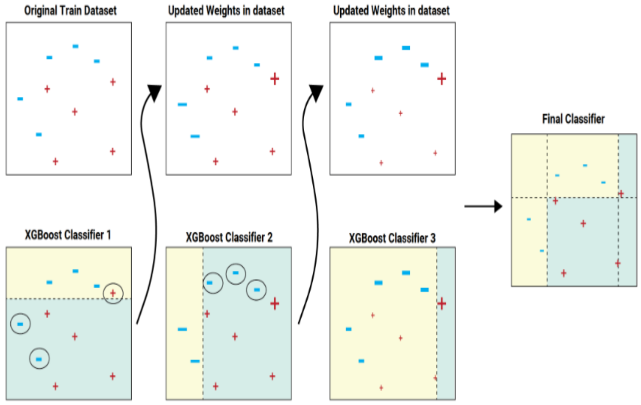


Fig.1: XGBoost

ii. **AdaBoost:** Adaptive boosting or AdaBoost is a Boosting technique used as an Ensemble Method in Machine Learning.Usually, decision trees are used for modelling. Multiple sequential models are created, each correcting the errors from the last model. AdaBoost assigns weights to the observations which are incorrectly predicted and the subsequent model works to predict these values correctly.

Below are the steps for performing the AdaBoost algorithm:

1. Initially, all observations in the dataset are given equal weights.
2. A model is built on a subset of data.
3. Using this model, predictions are made on the whole dataset.
4. Errors are calculated by comparing the predictions and actual values.
5. While creating the next model, higher weights are given to the data points which were predicted incorrectly.
6. Weights can be determined using the error value. For instance, higher the error more is the weight assigned to the observation.
7. This process is repeated until the error function does not change, or the maximum limit of the number of estimators is reached.

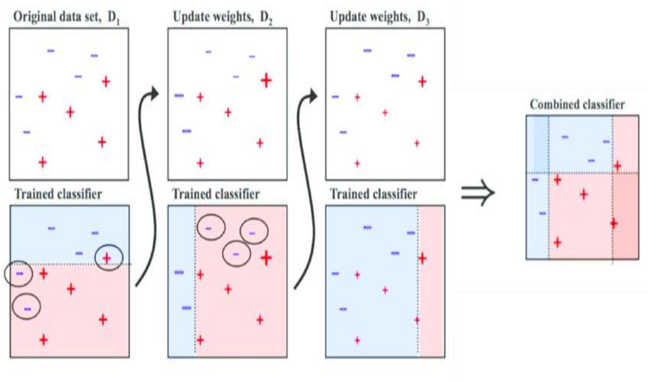


Fig.2: AdaBoost

iii. **Bagging:** Bagging also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset.

Working of Bagging algorithm is as follows:

1. Multiple subsets are created from the original dataset, selecting observations with replacement.
2. A base model (weak model) is created on each of these subsets.
3. The models run in parallel and are independent of each other.
4. The final predictions are determined by combining the predictions from all the models.

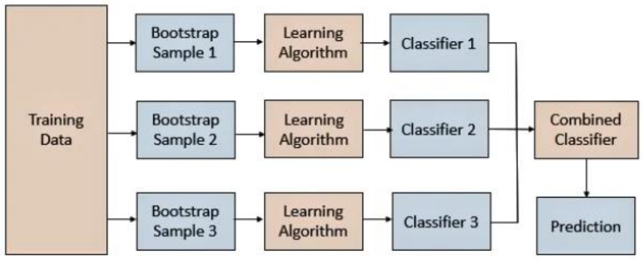


Fig.3: Bagging

iv. **Random Forest:** Random Forest is another ensemble machine learning algorithm that follows the bagging technique. It is an extension of the bagging estimator algorithm. The base estimators in random forest are decision trees. Unlike bagging meta estimator, random forest randomly selects a set of features which are used to decide the best split at each node of the decision tree.

Looking at it step-by-step, this is what a random forest model does:

1. Random subsets are created from the original dataset (bootstrapping).
2. At each node in the decision tree, only a random set of features are considered to decide the best split.
3. A decision tree model is fitted on each of the subsets.
4. The final prediction is calculated by averaging the predictions from all decision trees.

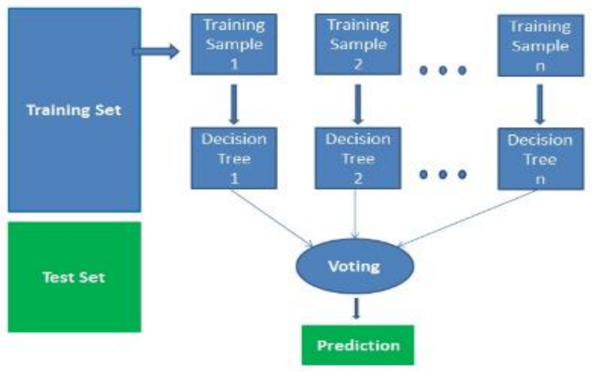


Fig.4: Random Forest

3. Design Methodology

The test set for this evaluation experiment Detecting PCOS is collected from kaggle. This section briefly explains to carried out for diagnosis of PCOS using Advanced Machine Learning algorithm. For this experiment, a dataset with 43 attributes of 541 women are collected from the Kaggle repository. Out of these 541 instances, 364 are for normal and the remaining 177 are for PCOS affected patients. For the experiments, Python programing language is used as a machine learning tool. For this, Anaconda distribution package, Scikit-learn library, Spyder, etc. are used for the deployment of Python.

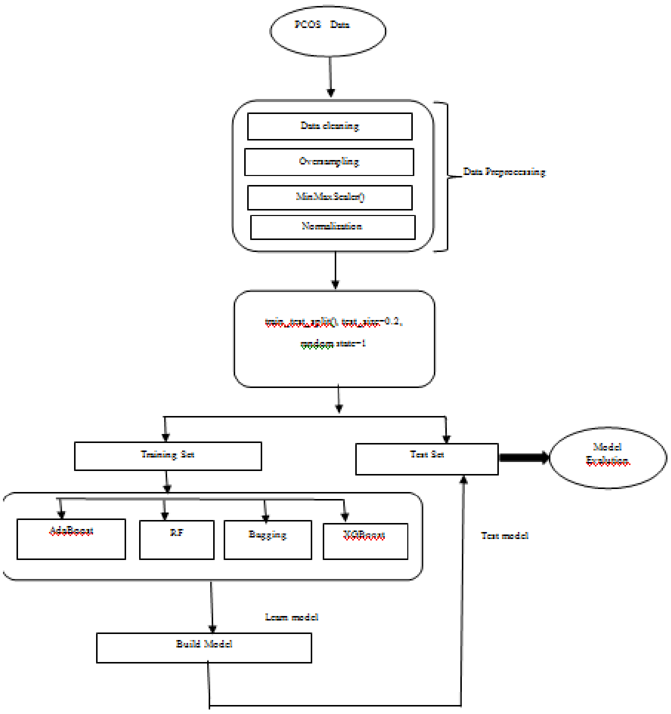


Fig.5: Work flow

Fig. 5 illustrates the work flow process. The first stage of the experiment is data preprocessing. At this stage, Data cleaning, encoding non-numeric columns into numeric, replacing missing values etc. are applied to the data. After that,Oversampling,MinMaxScaler and Normalization are applied. After data preprocessing, train_test_split() function is used to split the data into train data and test data. Next, a number of classification algorithms including AdaBoost, random forest, Bagging and XGBoost are used. We build the model using algorithm which gives best accuracy to test the data for model evaluation.

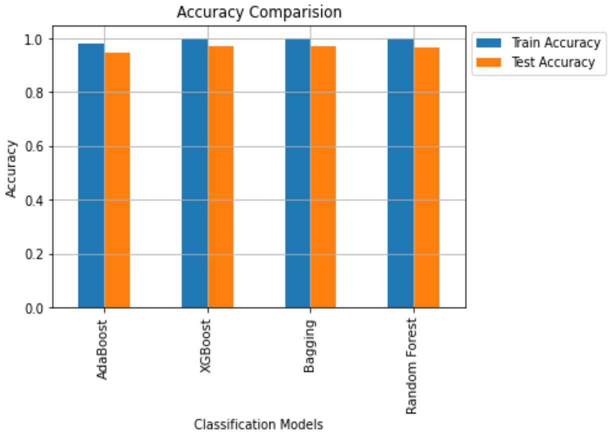
4. Experiments and Results

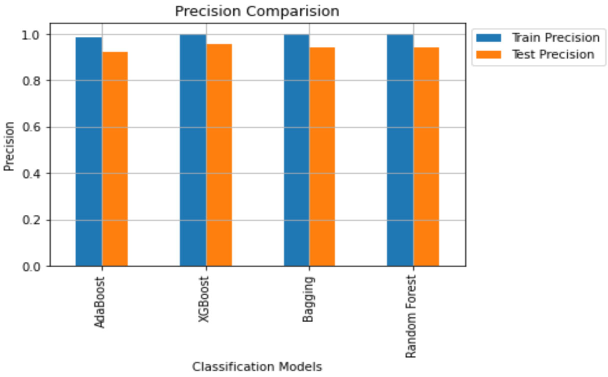
From Experimentation using algorithms AdaBoost, Bagging, Random Forest and XGBoost, we can observe the performance metrices Accuracy, Precision, F1_Score, Recall value and AUC_ROC.

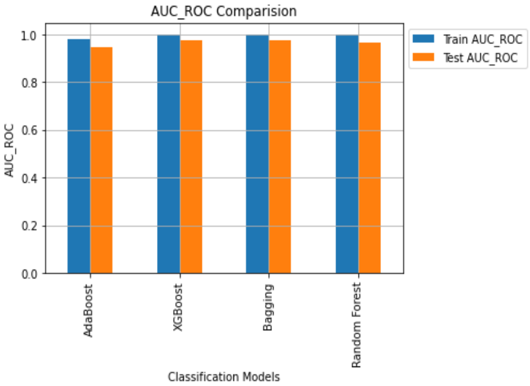
Fig.6 illustrates the values of performance metrices of above mentioned algorithms. The accuracy obtained by AdaBoost, Bagging, XGBoost, Random forest are 94.52%, 96.57%, 97.26%, 95.89% respectively. The best accuracy is obtained by using XGBoost method.

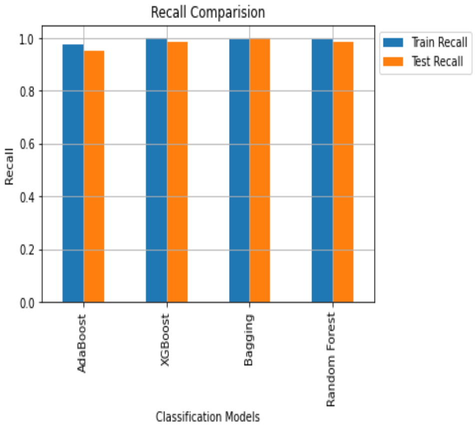
Classifier	Accuracy (%)	Precision (%)	F1_score (%)	Recall (%)	AUC_ ROC (%)
AdaBoost	94.52	92.42	93.84	95.31	94.60
Bagging	96.57	94.02	96.18	98.43	96.77
Random forest	95.89	95.31	95.31	95.31	95.82
XGBoost	97.26%	95.45%	96.29%	98.43	97.38

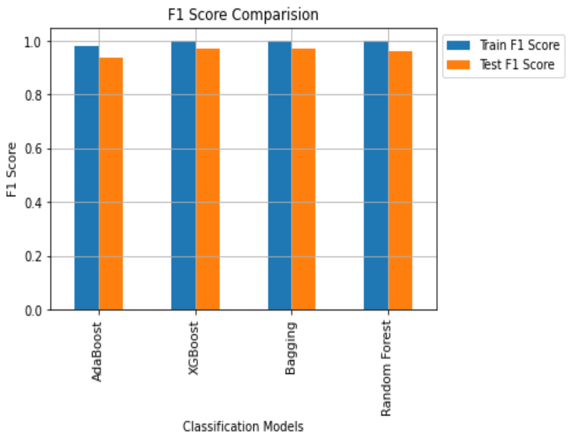
Fig.6: Performance metrics for various algorithms

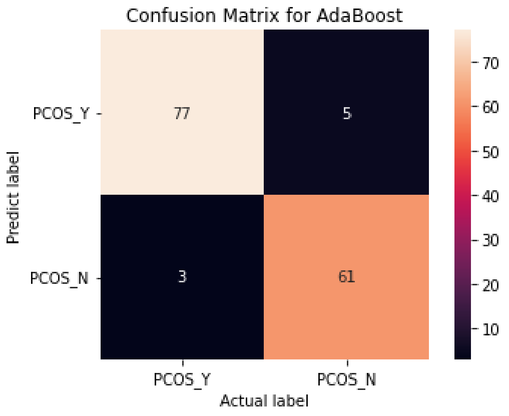


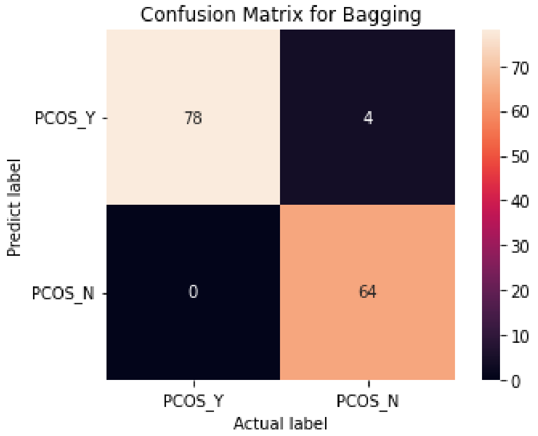


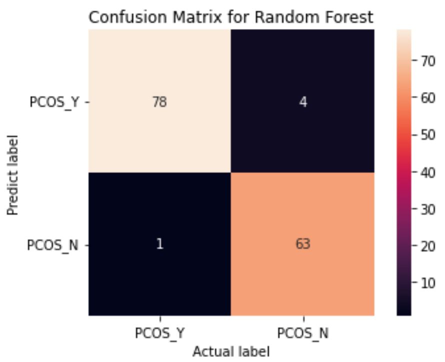












5. CONCLUSION

This paper presents data-driven diagnosis of PCOS disease in women using advanced machine learning algorithms. It is shown in the paper thatXGBoost has the best testing accuracy of 97.26%, precision value of 95.45%, F1_score of 96.29%, recall value of 98% and AUC_ROC of 97.38%. As a future work, the results obtained from this paper can be validated with a number of different datasets.

REFERENCES

[1] Bharati, Subrato, PrajoyPodder, and M. Rubaiyat Hossain Mondal. "Diagnosis of polycystic ovary syndrome using machine learning algorithms." *2020 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 2020

[2] Denny, Amsy, et al. "i-hope: Detection and prediction system for polycystic ovary syndrome (pcos) using machine learning techniques." *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019.

[3] Y. A. Abu Adla, D. G. Raydan, M. -Z. J. Charaf, R. A. Saad, J. Nasreddine and M. O. Diab, "Automated Detection of Polycystic Ovary Syndrome Using Machine Learning Techniques," 2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME), 2021, pp. 208-212, doi: 10.1109/ICABME53305.2021.9604905.

[4] Perry, Daniela Silvia, Jeremy Gunawardena, and Nicolas Orsi. "Identification of non-invasive cytokine biomarkers for polycystic ovary syndrome using supervised machine learning." *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. 2018.