ISSN No:-2456-2165

Early Identification of PCOS using Machine Learning Techniques

D.P. Sangeetha[1] (Assistant Professor)
Department of ECE, Sona College of Technology,
Salem Tamilnadu.

P. Nithish Raj[2] (Final year Student)
Department of ECE, Sona College of Technology,
Salem Tamilnadu.

R. Shurthika[3] (Final year Student)
Department of ECE, Sona College of Technology,
Salem Tamilnadu.

Abstract:- PCOS stands for Polycystic Ovary Syndrome. It is a common hormonal disorder that primarily affects women of reproductive age. It can cause various symptoms, including irregular periods, excess androgen levels, and small fluid-filled sacs (cysts) in the ovaries. PCOS is not typically found in men, as it is associated with the female reproductive system. While it's not life-threatening, it can lead to serious health problems if left untreat- ed. PCOS can increase the risk of conditions such as type 2 dia- betes, heart disease, and infertility. Early diagnosis and proper management are important for managing its effects on women's health.

Keywords:- *PCOS, Diabetes, Heart Disease, Obesity, ML*

I. INTRODUCTION

Polycystic ovary syndrome is a reproductive disorder and hormonal that affects one in 10 women of childbearing age. It is a leading cause of infertility in women and is associated with an increased risk of type two diabetes and heart disease The National Institutes of Health reports that more than 50% of women with PCOS will develop diabetes before age 40. Additionally, PCOS is associated with increased rates of anxiety, depression, and suicide attempts, especially in af- fected and educated women what's alarming, PCOS most commonly affects adolescents, those in their teens, and those under the age of 20, often because of an unbalanced diet. Early detection is important to help manage the physi- cal, emotional and internal effects of PCOS, and reduce the risk of more serious related diseases PCOS is common in India, with one in 5 women 1 struggle with the disease, and 5 of reproductive age in general Affects 10% of women PCOS syndrome is caused by many factors, including ab- normal insulin signaling, increased oxidative stress, ovarian a uncontrollable, and a combination of genetic and envi- ronmental factors This condition results in excess androgen secretion despite the presence of dietary factors.

Research has shown that insulin resistance is a major factor in the prevalence of PCOS, and that women with PCOS have an increased risk of developing diabetes or pre-diabetes, and about 40% are expected to when they are 40. Additionally, PCOS is associated with obesity.

II. PROBLEM DEFINITION

PCOS is a disease in India where 1 in 5 women suffering from it. About 10 % of women of childbearing age have PCOS. The vast body of knowledge on PCOS pathogenesis considers it a complex condition including abnormal insulin signaling, increased oxidative stress, uncontrolled ovarian steroid hormones, and environmental genetic factors Pro- vides evidence meaning lice Most people don't care about obesity because This is one of the most common explana- tions for health. Also, they think it won't affect their health. It's just the external structure of their bodies. But the sad truth is that most diseases are linked to obesity. As marked by diabetes, heart disease, malignancy, arthritis, chronic kidney disease, stroke, hypertension, and epidemics of dead- ly diseases, sometimes the cause of death can be obesity in adults and children.

III. CONCEPTUAL DEFINITION

A. Machine Learning:

Machine learning is a form of artificial intelligence that fo- cuses on algorithms and mathematical models that enable computer systems to improve on a specific task by learning from data in an unstructured way. Here are some key points about how machine learning can be useful and its benefits in real-time projects:

➢ *Benefits of Machine Learning:*

- Data-Driven Intelligence: It can analyze vast amounts of data to uncover patterns, trends and insights not apparent through traditional methods.
- Automation: ML automates repetitive tasks, saving time and reducing the risk of human error.
- Personalization: It enables personalized recommenda- tions such as e-commerce or content recommendations, im- proving user experiences
- Prediction and forecasting: ML can predict future trends, demands or outcomes, helping in decision making.
- Natural Language Processing: It improves chatbots and language translation, communication and customer service.

➢ *Advantages of Real-Time Programs:*

- Efficiency: ML automates tasks, making processes more efficient and cost-effective
- Scalability: ML models can handle large datasets and adapt as the data grows
- Accuracy: ML can make predictions and decisions with high accuracy, reducing errors.
- Continuous learning: Models can learn and adapt to changing data, ensuring relevance over time.
- Competitive advantage: ML can give businesses a competitive edge through better insights and customer engage- ment.

➢ *Here are the Main Characteristics of Machine Learning in Pro- Projects.*

- *Data Dependence:*

Machine learning is heavily depend- ent on data, and the quality, quantity and relevance of data directly affects the success of the project.

- *Automation:*

ML automates tasks by learning patterns from data and reducing the need for manual intervention in decision-making processes.

- *Predictive Analytics:*

ML is used for predictive tasks such as generating future forecasts, detecting anomalies and making recommendations based on historical data.

- *Adaptability:*

ML models can adapt and improve over time as they receive more data and feedback, improving their performance.

- *Feature Engineering:*

The process of selecting and pre- paring relevant features from data is important to improve the performance of the model.

- *Model Training and Testing:*

ML models are trained on a subset of data, then generalize to new data to evaluate their performance.

➢ *Software:*

Software plays an important role in the field of machine learning by providing the tools and frameworks needed to build, train, and deploy machine learning models. Here is a description of the different types of software used in ma- chine learning:

- *Programming Languages:*

✓ *Python:*

Python is the most popular programming lan- guage for machine learning. Libraries make them a great choice for data scientists and machine learning engineers.

- *Machine Learning Frameworks:*

✓ Developed by Google, TensorFlow is an open, source machine learning framework known for its flexibility and scalability. It is widely used for deep learning tasks.
✓ Developed by Facebook's AI research lab, PyTorch is easily used for its dynamic computational graph and deep learning projects.

- *Scikit-Learn:*

This library provides simple and efficient tools for data mining and data analysis. It is an excellent choice for classical machine learning tasks.

- *Data Pro- Cessing and Analysis:*

✓ *Num Py:*

A library for numerical calculations in Python.

✓ *Pandas:*

A data manipulation and analysis library for han- dling structured data.

✓ *Jupiter Notebook:*

An interactive, web-based environ- ment for data analysis and machine learning proto-typing.

- *Data Visualization:*

✓ *Matplotlib:*

A versatile library for creating static, ani- matted or interactive plots and charts.

✓ *Seaborn:*

A high-level interface for creating attractive and informative statistical graphics.

- *Deployment and Production:*

✓ *Docker:*

A container technology that makes it easy to compile and deploy machine learning models in different environments.

✓ *Kubernetes:*

A container orchestration platform for man- aging and scaling machine learning deployments.

- *Cloud Service:*

✓ Amazon AWS, Google Cloud, Microsoft Azure: These cloud platforms provide machine learning services and re- sources to build, train and deploy models.

- *Version Control:*

A distributed version control system used to track changes to code, essential for collaboration in machine learning pro- jects.

These software tools and frameworks are used at different stages of the machine learning project lifecycle, from data pre-processing to model training and deployment.

➢ *Working:*

- *Data Collection and Pre-Processing:*

Machine learning starts with relevant data collection. This data is then cleaned and preprocessed to prepare it for analysis, which may in- clude tasks such as data cleaning, normalization, and feature engineering.

- *Model Training:*

In this step, machine learning algorithms are applied to the pre-processed data to build predictive models. These models learn patterns and relationships with- in the data, allowing them to make predictions or classifica- tions.

- *Evaluation and Deployment:*

After training, the models are evaluated to evaluate their performance using various metrics. If the model meets the desired criteria, it can be used to make predictions or make decisions in real-world applications. Continuous monitoring and retraining can be part of the process to ensure the accuracy of the model over time.

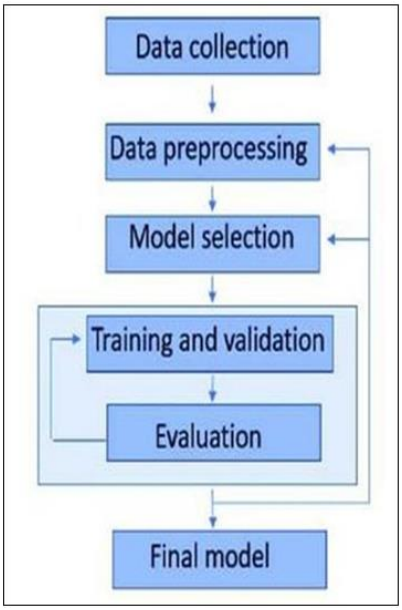

Fig 1 Working Flow of Machine Learning.

IV. EXISTING SYSTEM

PCOS, or polycystic ovary syndrome, is a common hormo- nal disorder in people with ovaries. It can cause a variety of symptoms and lead to long-term health concerns. Here's a brief overview:

➢ *Causes and Symptoms:*

- *Causes:*

The exact cause of PCOS is not fully known, but it involves hormonal imbalance, insulin resistance and ge- netic factors.

- *Symptoms:*

PCOS presents with a variety of symptoms including irregular periods, excess androgen hormones (which lead to acne and excessive hair growth) and multiple small cysts in the uterus. Other symptoms include weight gain, hair loss and fertility problems.

➢ *Diagnosis:*

- *Medical Diagnosis:*

To diagnose PCOS, a medical profes- sional usually does a combination of the following:

✓ Medical history and symptom assessment.
✓ Physical examination.
✓ Blood tests to check hormone levels (such as androgens, LH, FSH and insulin).
✓ Pelvic ultrasound to examine the ovaries.

➢ *Cost of PCOS Test:*

- *Cost:*

The cost of PCOS testing varies depending on your location, health care provider and insurance age. Blood tests and ultrasounds may be covered by insurance, but you should check with your healthcare provider and insurance company for specific costs.

If you suspect you have PCOS, it is important to consult a healthcare professional as they can provide a proper diagno- sis and guidance on managing the condition. Untreated PCOS can lead to infertility, increased risk of type 2 diabe- tes and cardiovascular disease, and problems related to hor- mone imbalances, including mood disorders.

V. PROPOSED SYSTEM

PCOS is a condition that mostly affects women and can lead to problems like obesity, heart disease, diabetes, and high blood pressure. Unlike these common issues, PCOS often goes undiagnosed. The researchers aim to find PCOS early by looking at things like insulin resistance, which is related to how the body handles sugar, and hormone imbalances, like having too much male hormone in the blood.

They used computer techniques to figure out which factors are important for diagnosing PCOS. After studying these factors, they found that they could predict PCOS by consid- ering features related to obesity, diabetes, heart disease, and high blood pressure. This study used both supervised and unsupervised learning methods to make these predictions. The results and how they did it are explained in the next part of the research.

ISSN No:-2456-2165

Table 1 A view of the Amalgamated Dataset

A view of the amalgamated dataset.

Age	BMI	Glucose	cp	ca	Chol	trestbps	fbs	target
21	28.1	89	1	0	210	150	1	1
30	25.6	116	1	0	210	150	0	0
36	25.9	95	0	0	183	138	0	0
35	32.3	116	0	0	183	138	0	0
35	43.4	93	0	0	183	138	1	1
35	35	136	0	0	183	138	0	0
38	34.1	117	2	0	215	152	0	0
37	40.2	133	2	0	215	152	1	1
38	32.5	109	2	0	215	152	1	1
37	48.8	137	2	0	215	152	1	1
37	29.5	144	2	0	215	152	0	0
38	34	124	2	0	215	152	1	1
38	39.5	106	2	0	215	152	0	0
37	21.9	114	2	0	215	152	0	0
37	39.1	130	2	0	215	152	1	1
38	41.8	151	2	0	215	152	1	1
37	25.2	119	2	0	215	152	0	0
36	42.9	151	0	0	183	138	1	1
35	24.5	90	0	0	183	138	0	0

VI. METHODOLOGY

A machine learning algorithm is a computer program or mathematical model that processes data and learns to make predictions, decisions, or tasks without explicit planning. These algorithms are designed to identify patterns, relation- ships, or trends in datasets and use that knowledge to make informed responses or predictions. They are fundamental components of machine learning and artificial intelligence. Machine learning algorithms can adapt and improve their performance as they are exposed to more data, making them valuable tools for data analysis and automation.

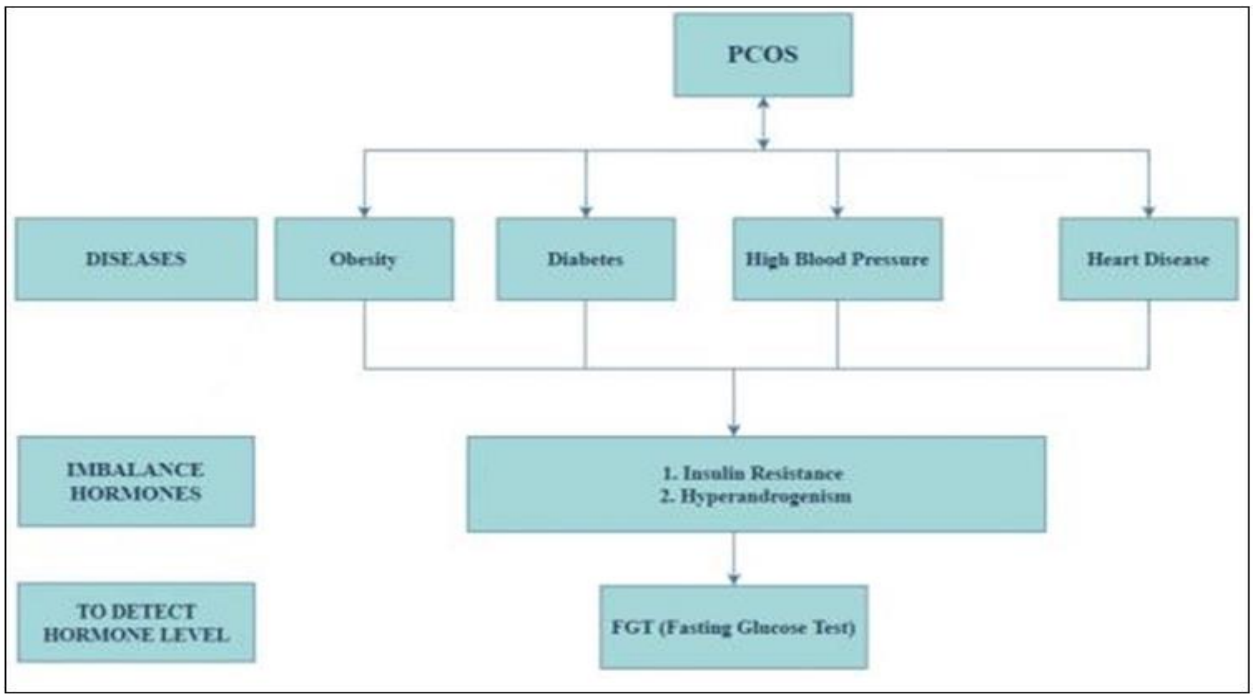

Fig 2 Workflow of the Proposed Methodology

➢ *Of Course, here is a Brief Description of Supervised, Unsu- Pervised and Reinforcement Learning Algorithms:*

- Supervised Learning:

✓ In supervised learning, the algorithm is trained on a la- beled dataset, where the input data is associated with an appropriate output or goal.
✓ The goal is to learn a mapping from input to output so that the algorithm can make accurate predictions on new, unseen data.
✓ Common applications include classification (assigning data to predefined categories) and regression (priority- ring numerical values).

- *Unsupervised Learning:*

✓ Unsupervised learning deals with unlabeled data and aims to find patterns, structures or relationships within the data. - It has no explicit output labels, and the algorithm examines the data to find inherent groups or associations.
✓ Common tasks include clustering (grouping similar data points) and dimensioning

- *Reinforcement Learning:*

Reinforcement learning involves learning by an agent in- teracting with the environment and receiving rewards or punishments based on its actions.

✓ The agent's goal is to learn a policy that maximizes its overall reward over time.
✓ It is commonly used in dynamic decision-making scenar- ios such as sports, robotics, and autonomous systems.
✓ Example algorithms: Q-learning, deep Q-networks and proximal policy optimization.

These three categories represent different approaches to predict the inherent structure of the data, and reinforcement learning that focuses on learning through interaction and feedback.

- *Decision Tree:*

✓ It is a tree-based model in which each internal node repre- sents a test of an attribute and each leaf node represents a class label.
✓ Decision tree algorithms iteratively build the tree by se- lecting the best-fit attribute for segmentation using various measures of impurity such as entropy, index, and log loss. Once the tree is created, it can be traversed to find the clas- sification of the given test data.
✓ We have chosen to use entropy as a criterion for the best split decision, the reason for choosing it is explained below.

- *Slope Increase:*

✓ A gradient boosting classifier is a supervised learning al- gorithm that uses a boosting technique to generate models that attempt to minimize subsequential sampling errors.
✓ Promotional technique works to support a strong learner by bringing together several weak learners. One of the most notable features is the loss function in gradient increments. of this data

- *Logistic Regression:*

✓ Logistic regression is a supervised machine learning algorithm used for classification problems.
✓ Logistic regression modeling was able to classify new individuals as more or less likely to have PCOS based on the presence of heart disease and diabetes.
✓ Overall, logistic regression is a useful tool for pre- dicting the likelihood of PCOS based on independent variables such as BMI, glucose levels, and blood pres- sure.

- *Support Vector Machine:*

✓ SVM learning algorithm divides the data into differ- ent classes by constructing a hyperplane.
✓ Once the hyperplane is generated, the SVM algo- rithm is used to predict the probability of new individ- uals having PCOS based on their independent variable values.
✓ The algorithm assigns each new observation to a class based on which side of the higher plane it falls on.

- *Random Forest:*

✓ It is an ensemble learning method that uses multiple decision trees and combines their repeated results to produce a prediction.
✓ In a random forest, each decision is generated from a subset of the training data and set of random features, which helps improve the accuracy and stability of the model and reduces overfitting.
✓ Decision Tree Finds its best split-ting feature and constructs its tree, for which various measures are ap- plied. Supplementary Conclusion Best split fire in trees

VII. RESULT AND DISCUSSION

➢ *Supervised Learning*

Table 2 Supervised Learning Results

Algorithm	Accuracy (IF)	Accuracy (AF)
Random Forest	98.5	98.8
Gradient Boosting	98.5	98.8
KNN	93.75	85.8
Logistic Regression	94.8	94.1
SVM	94.1	90.5
Decision Tree	98.9	96.4
Hybrid RFLR	98.1	96.4

➢ *Unsupervised Learning:*

Table 3 Unsupervised Learning Results

Algorithm	Silhouette Score	Davies Bouldin Score
k-means	68.6	37.1

From the obtained results, we can infer that the decision tree classification algorithm provides the best result with random forest and gradient boosting algorithm. This classification is validated with the help of an unsupervised learning algo- rithm called K-means algorithm. The K-means algorithm was applied to our PCOS dataset, and outcome measures were calculated using the Silhouette score and Davis Boldin score. The best value of K was verified using an elbow dia- gram and a Scutt-Der plot was created for the results of the algorithm to obtain clusters of our PCOS classes. The reason why decision tree algorithm provides better accuracy than random forest and gradient boosting is as follows. The most important factor is the size of the dataset. We have a rela- tively small dataset, so a de-scission tree can capture the underlying patterns and relationships in the data more accu- rately than a random forest or gradient boosting algorithm. The problem with our scheme is simple. The features have a clear hierarchical-complexity relationship with the target variable, so a decision tree is sufficient to provide accurate results, while not requiring the additional complexity and gradient scaling of random forests.

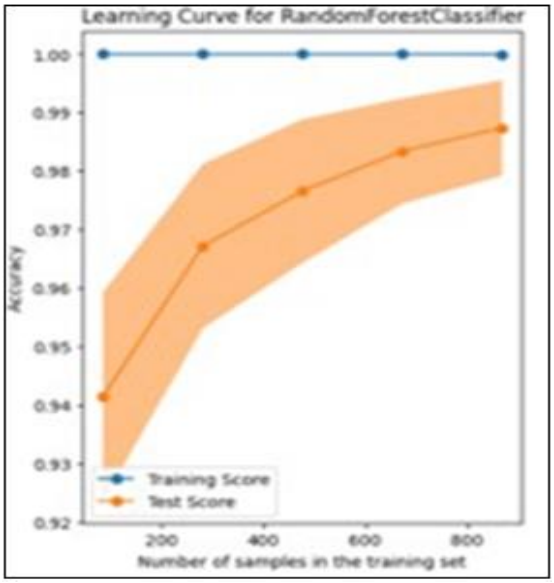

Fig 1 Learning Curve of Random Forest Algorithm

A learning curve graph is a visual representation of the vari- ous metrics of a machine learning model that demonstrate incremental learning data and thus are plotted against the training dataset.

➢ *Two main Types of Learning Curve:*

- *Optimization Learning Curve:*

These learning curves are calculated based on improved metrics such as loss or mean square error.

- *Performance Learning Curve:*

These learning curves are calculated in metrics that can be evaluated such as model accuracy or pre-selection.

So, let's plot the graph against the accuracy and training dataset

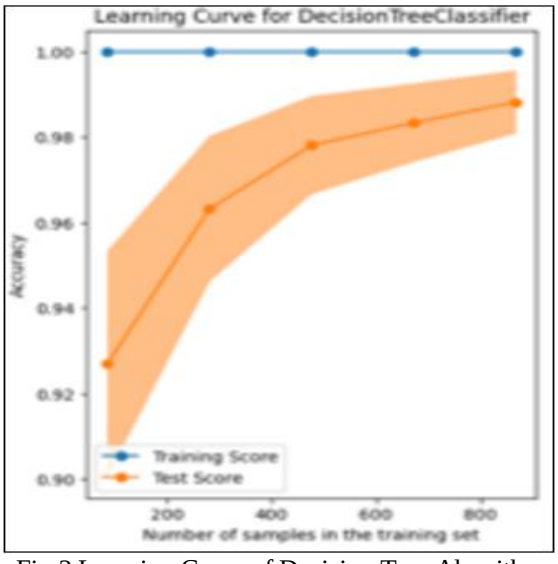

Fig 2 Learning Curve of Decision Tree Algorithm

ISSN No:-2456-2165

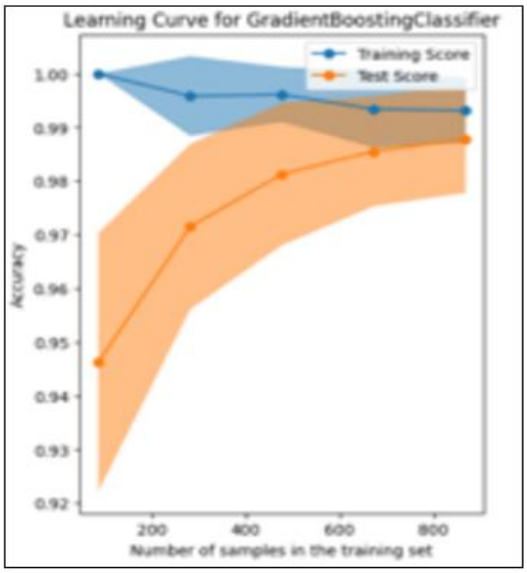

Fig 3 Learning Curve of Gradient Boosting Algo- rithm

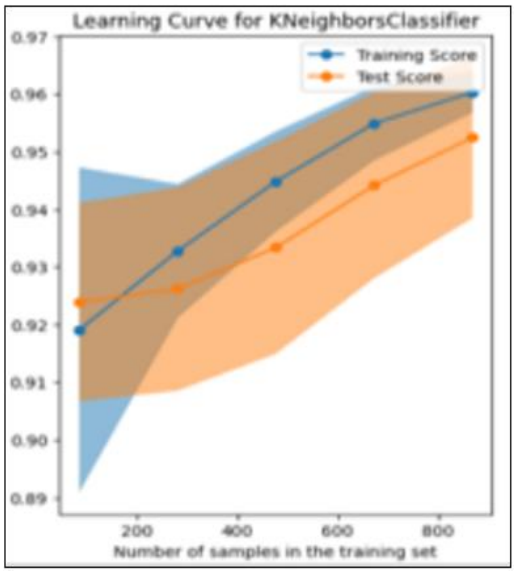

Fig 6 Learning Curve of KNN

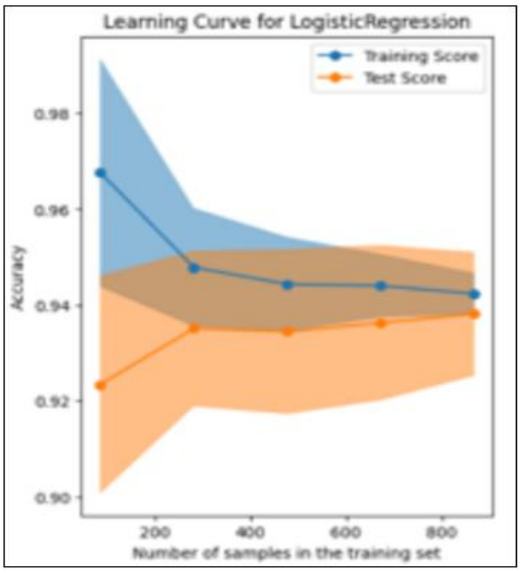

Fig 4 Learning Curve of Hybrid RFLR

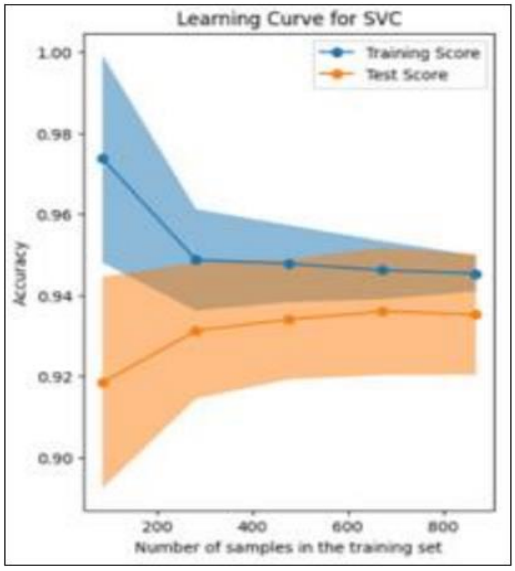

Fig 7 Learning Curve of

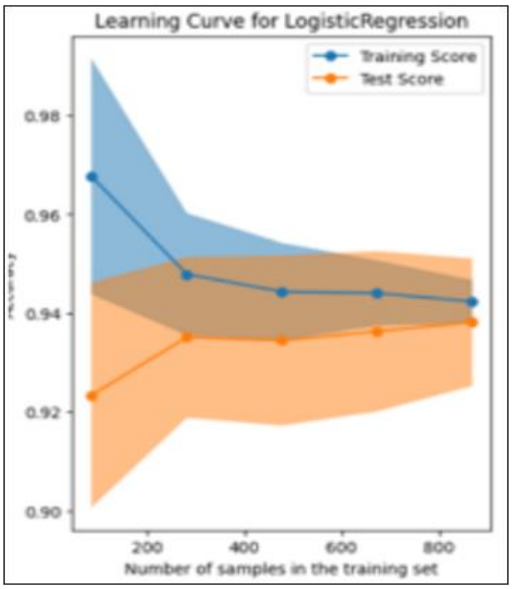

Fig 5 Learning Curve of Logistic Regression

- *Unsupervised Learning:*
  Snapshots of Unsupervised Learning Algorithms

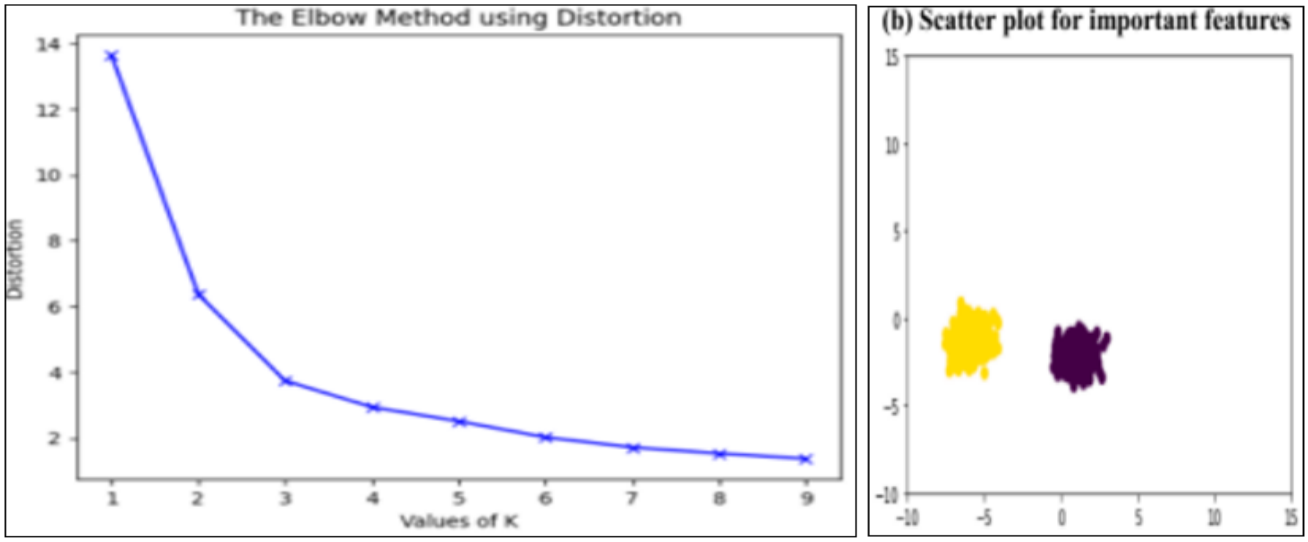

Fig 8 Elbow Plot for K-Means

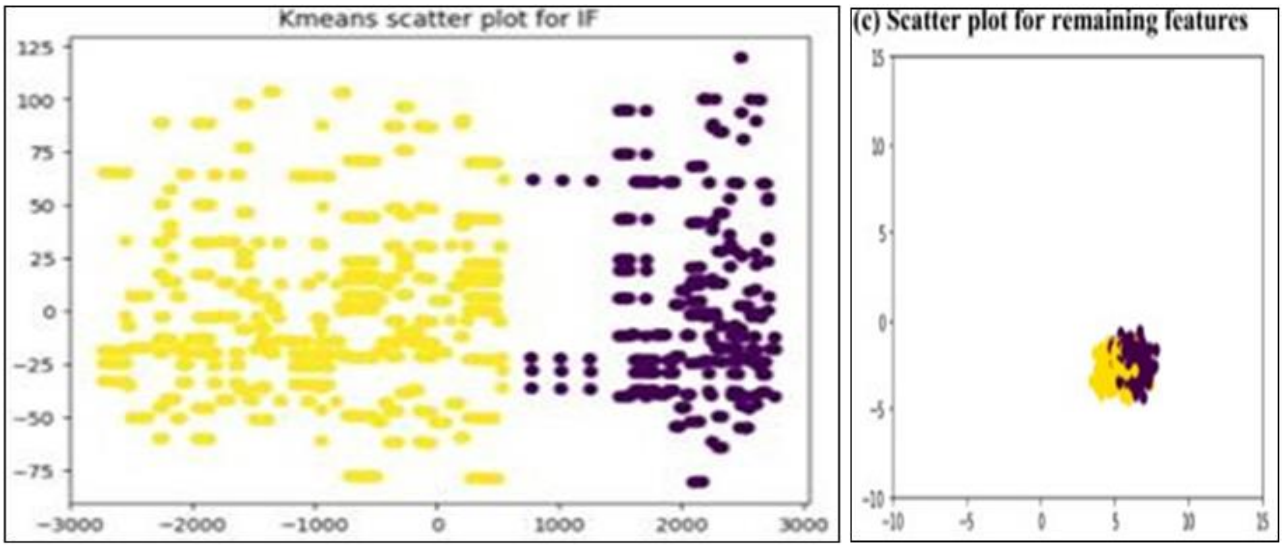

Fig 9 Scatter Plot of K-Means

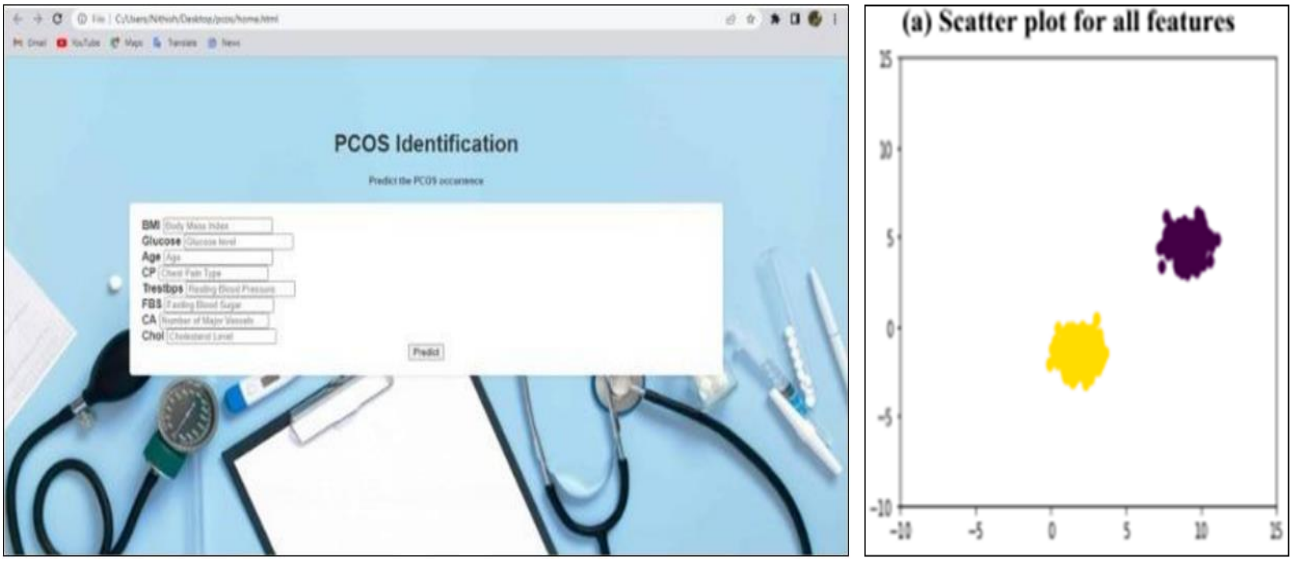

Fig 10 Final Output 1

ISSN No:-2456-2165

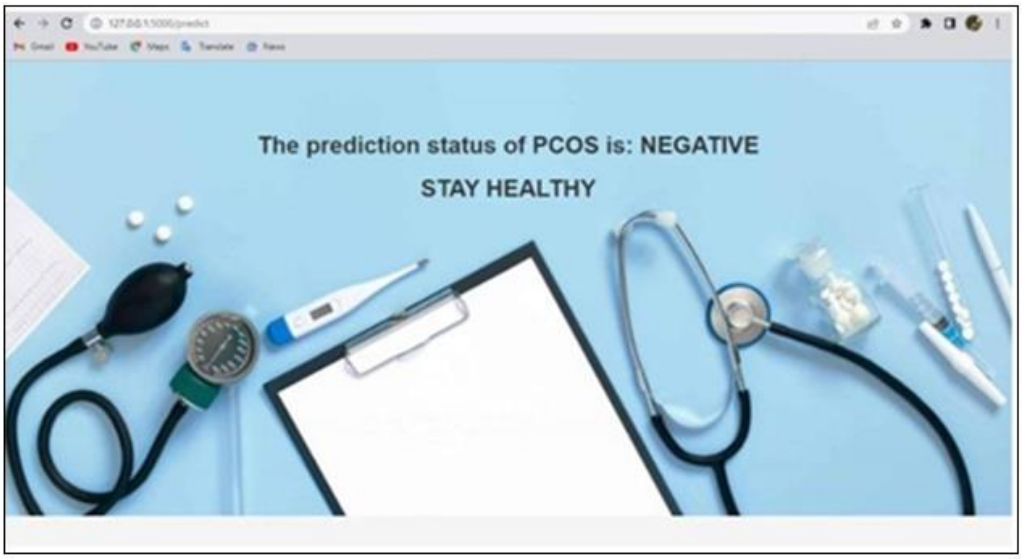

Fig 11 Final Output 2

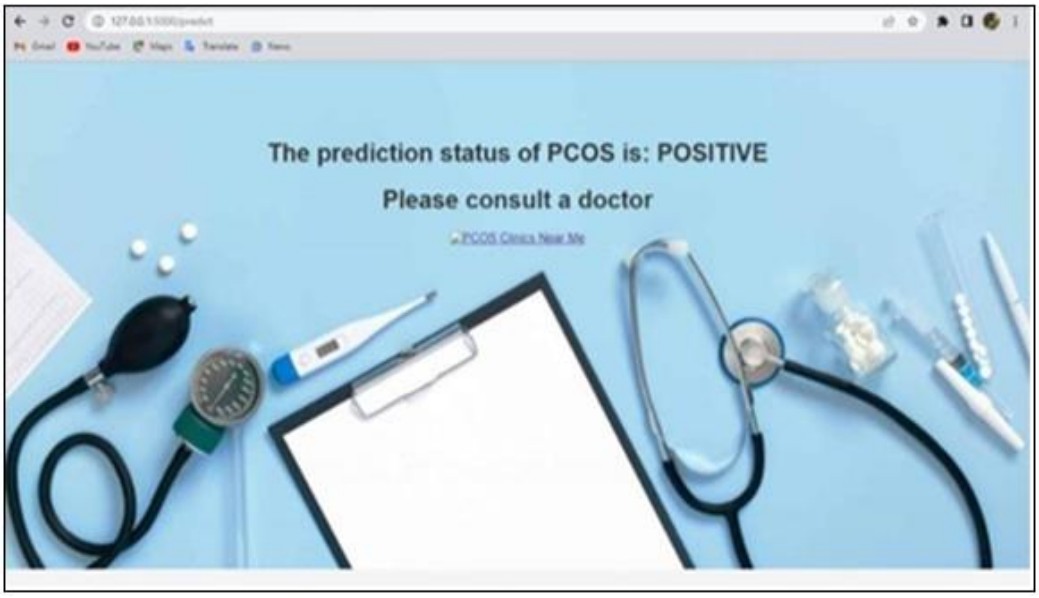

Fig 12 Final Output 3

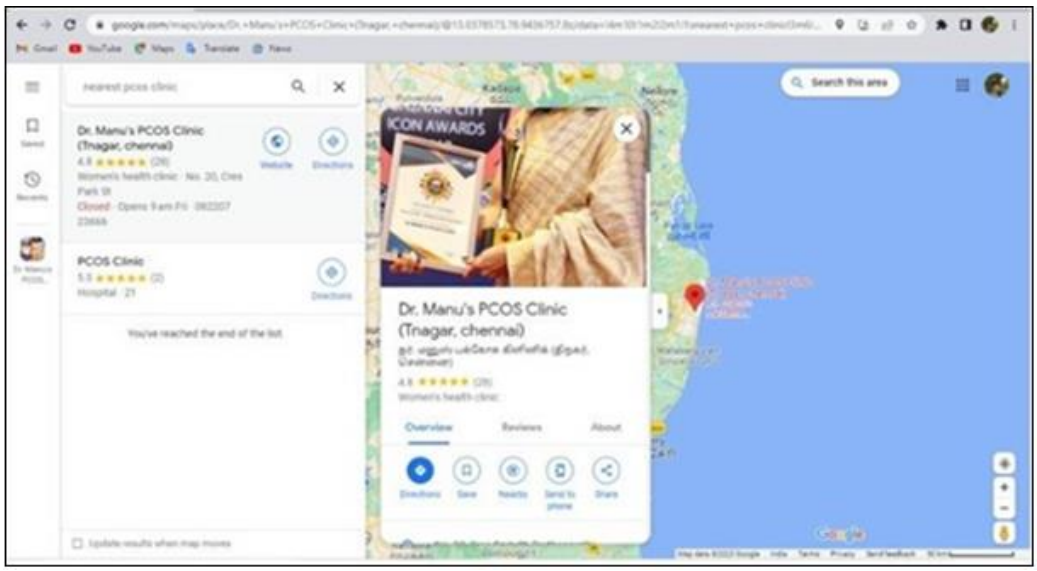

Fig 13 Final Output 4

Thus, PCOS is identified using machine learning algorithms and the output is obtained as “positive” and “negative”. If PCOS is diagnosed positive, a referral clinic for PCOS treatment will be shown.

Thus, PCOS is identified.

VIII. CONCLUSION

From the performance analysis of the above algorithms the best accuracy is reported to be decision tree algorithm. The learning curve graph and scatter plot is verified.

- Decision tree - 98.9 %
- Random Forest - 98.5 %
- Gradient Boosting - 98.5 %
- K-Nearest - 93.75 %
- Logistic regression 94.8 %
- Support Vector Machine - 94.1 %
- Hybrid RFLR – 98 %

Therefore, we conclude that the decision tree has the best metric among the supervised algorithms tested for the present dataset, followed by the gradient boosting algorithm.

FUTURE PLANS

We are planning to release the model as an awareness website accessible to all who need it.

According to the survey, cited based on data from the original paper, “PCOS is largely unknown to most people, despite affecting millions of women and with serious health consequences. It is shocking that studies reveal that about 50% of women living with PCOS go undiagnosed; we are planning to create a separate awareness website that includes details and possibilities related to PCOS. Treatments and health conditions.

This predictive model and with the added feature of predicting obesity at an early age. Regarding the model, we plan to test with additional supervised learning algorithms and check for better accuracy.

REFERENCES

[1]. Aggarwal, S., & Pandey, K. (2021). An Analysis of PCOS Disease Prediction Model Using Machine Learning Classification Algorithms. Recent Patent of Engineering, 15(6), 53–63. https://doi.org/10.2174/1872212115999201224130204

[2]. Aggarwal, S., & Pandey, K. (2022). Determining the representa- tive features of polycystic ovary syndrome via Design of Experiments. Multimedia Tools and Applications, 81, 29207–29227. https://doi.org/10.1007/s11042-022-12913-0

[3]. Ali, S. E., & Ali, F. E. (2020). A Study of Apelin-36 and GST Levels with Their K.Suriyakrishnaan, “Recommendation system for Agriculture using Machine learning and Deep learning” in the International Conference on Inventive Systems and Control (ICISC 2022),organized by Department of Electronics and Com- munication Engineering, JCT College of Engineering and Tech- nology, Coimbatore during 6th - 7th January 2022 (Scopus).

[4]. Sudha, V.,Ganesh Babu, T.R, Vikram, N., Raja, R. Compari- son of detection and classification of hard exudates using arti- ficial neural system vs. SVM radial basis function in diabetic retinopathy MCB Molecular and Cellular Biomechanics , 2021, 18(3), pp. 139–145

[5]. Relationship to Lipid and Other Biochemical Parameters in the Prediction of Heart Diseases in PCOS Women Patients. Baghdad Science Journal, 17(3), 924–930. https://doi.org/10.21123/bsj.2020.17.3(Suppl.).0924.

[6]. Anagnostis, P., Tarlatzis, B. C., & Kauffman, R. P. (2018). Poly- cystic ovarian syndrome (PCOS): Long-term metabolic consequences. Metabolism: Clini- cal and Experimental, 86, 33–43. https://doi.org/10.1016/j.metabol.2017.09.016

[7]. Bloice, M. D., & Holzinger, A. (2016). A tutorial on machine learning and data science tools with python. In Machine Learning for Health Informatics. https://doi.org/10.1007/978-3-319-50478-0_22

[8]. Causes of Sleep Apnea. (2021). WebMD.https://www.webmd.com/sleep-disorders/sleepapnea/obstructive- sleep-apnea-causes.

[9]. Centers for Disease Control and Prevention. (2020). PCOS (Poly- cystic Ovary Syndrome) and Diabetes. (n.d.).https://www.cdc.gov/diabetes/basics/pcos.html.

[10]. Accessed February, 2022.

[11]. Chen, W., & Pang, Y. (2021). Metabolic Syndrome and PCOS: Pathogenesis and the Role of Metabolites. Metabolites, 11(12). https://doi.org/10.3390/metabo11120869

[12]. Condorelli, R. A., Calogero, A. E., Mauro, M. D., & La, S. (2017). PCOS and diabetes mellitus : From insulin resistance to altered beta-pancreatic func- tion, a link in evolution. Gynecological Endocrinology, 33(9), 665–667. https://doi.org/10.1080/09513590.2017.1342240

[13]. Doroszewska, K., Milewicz, T., Mrozi´nska, S., Janeczko, J., Rokicki, R., Janeczko, M., et al. (2019). Blood pressure in postmenopausal women with a history of polycystic ovary syndrome. Przegląd Menopauzalny=. Menopause Review, 18(2), 94–98. https://doi.org/10.5114/pm.2019.84039

[14]. El Hayek, S., Bitar, L., Hamdar, L. H., Mirza, F. G., & Daoud, G. (2016). Poly Cystic Ovarian Syndrome: An updated overview. Frontiers in Physiology, 7(APR), 1–15. https://doi.org/10.3389/fphys. 2016.00124
[15]. Escobar-Morreale, H. F. (2018). Polycystic ovary syndrome: Defi- nition, aetiology, diagnosis and treatment. Nature Reviews Endocrinology, 14(5), 270–284. https://doi.org/10.1038/nrendo.2018.24
[16]. Osisanwo, F. Y., O Awodele, J. E. A., et al. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology, 48(3), 128–138. https://doi.org/10.14445/22312803/ijctt-v48p126
[17]. Wang, F. F., Wu, Y. Y. H. Z., et al. (2018). Pharmacologic therapy to induce weight loss in women who have obesity/overweight with polycystic ovary syn- drome : A systematic review and network. Obesity Reviews, 19(10), 1424–1445. https://doi.org/10.1111/obr.12720
[18]. Fauser, B. C. J. M., Van Rijn, B. B., Bekker, M. N., & De Wilde, M. A. (2019). Associations of preconception Body Mass Index in women with PCOS and BMI and blood pressure of their offspring Associations of preconception Body Mass Index in women with PCOS and BMI and blood pressure of their offspring. Gynecologi- cal Endocrinology, 35 (8), 673–678. https://doi.org/10.1080/09513590. 2018.1563885.
[19]. Glueck, C. J., & Goldenberg, N. (2019). CHARACTERISTICS OF OBESITY IN POLYCYSTIC OVARY. Metabolism, 92, 108–120. https://doi.org/10.1016/j. metabol.2018.1100%02