

PCOS Disease Prediction Using Machine Learning Algorithms

M.Priyadharshini¹, A.Srimathi², C.Sanjay³, K.Ramprakash⁴

¹Assistant Professor, Computer Science and Engineering, SRM Valliammai Engineering College, Kattankulathur, Tamilnadu, India.

^{2, 3, 4} UG Computer Science and Engineering, SRM Valliammai Engineering College, Kattankulathur, Tamilnadu, India.

Emails: priyadharshinim.cse@srmvalliammai.ac.in¹, srimathianandssa@gmail.com², sanjaymangayarkarasi@gmail.com³, ramprakashkalyanasundaram@gmail.com⁴

Orcid ID: <https://orcid.org/0009-0004-3080-770X>

Abstract

Polycystic Ovary Syndrome (PCOS) is a prevalent endocrine disorder affecting reproductive-aged women, characterized by hormonal imbalances, irregular menstrual cycles, and the presence of cysts on the ovaries. One of the most worrying diseases today is Polycystic Ovarian Syndrome (PCOS) which is highly dangerous to an extent of affecting women's reproductive life to a huge extent. The dataset includes a diverse set of features such as age, body mass index (BMI), hormonal levels, menstrual irregularities, and lifestyle factors. We explore various machine learning algorithms, including linear regression, decision tree, and random forests to identify the most effective model for PCOS prediction. [1] The study focuses on the development of a robust and clinically applicable predictive model that can aid healthcare professionals in early identification of individuals at risk of PCOS. The results obtained from this research have the potential to significantly impact the field of women's health by offering a reliable and efficient tool for PCOS prediction. Early identification of individuals at risk can facilitate timely interventions, personalized treatment plans, and improved outcomes. Furthermore, the study contributes to the growing body of literature on the application of machine learning in healthcare, demonstrating its potential as a valuable tool for predictive modelling in complex endocrine disorders such as PCOS.

Keywords: PCOS, Lifestyle factors, Machine learning techniques, Endocrine disorders

1. Introduction

Polycystic Ovary Syndrome (PCOS) stands as one of the most prevalent endocrine disorders affecting women of reproductive age, with a spectrum of manifestations that include irregular menstrual cycles, hormonal imbalances, and the presence of ovarian cysts [1, 2] Since PCOS is a hard to diagnose widespread hormonal disorder, blood tests, symptoms, and other parameters with the help of a computer can form a new and easy method to diagnose it. [3]. This research focuses on employing three distinct machine learning algorithms—Linear Regression, Random Forest, and Decision Tree—to predict PCOS risk based on a comprehensive dataset

encompassing clinical and demographic parameters. Linear Regression, a foundational algorithm, is chosen for its simplicity and interpretability, making it an excellent baseline for comparison. Random Forest, known for its ensemble approach and ability to handle complex relationships, is selected to explore the benefits of a more sophisticated model. Decision Tree, with its intuitive decision-making process, is included to assess the impact of a tree-based algorithm on PCOS prediction. The dataset utilized in this study comprises a diverse array of features such as age, body mass index (BMI), hormonal levels, and

menstrual irregularities. The primary objectives of this research are to evaluate the predictive performance of each algorithm and to identify the most effective model for PCOS risk prediction. The early detection and timely intervention of PCOS are crucial for managing its symptoms and preventing associated complications. [4] However, there is a lack of efficient and accurate predictive models that can identify individuals at risk of developing PCOS based on a combination of lifestyle, and clinical factors. This project aims to develop a predictive model for PCOS that utilizes advanced machine learning algorithms to analyse diverse datasets, including lifestyle factors, and relevant health metrics. [5] The ultimate goal is to provide a reliable tool for early identification and intervention, contributing to improved health care outcomes for women susceptible to PCOS [6]. Our system can provide early detection, prognosis, and treatment suggestions for PCOS, which can improve the quality of life for women affected by this disorder. The data consists of information about the symptoms of the patients that includes BMI, Age, and any weight gain in recent years etc. Thus, the symptoms influencing the prediction of the disease, are collected. The process gets followed by data pre-processing. In here, the data is standardized for further use. [7] Next, strong training datasets are created and the model is trained to produce higher accuracy. Then, the model is tested with testing datasets. In this we used three different algorithms that includes decision tree, random forest and linear regression. Out of these three algorithm we found that decision tree gives much more accuracy compared to other two algorithms. [8]

2. Methodology

2.1 Data Set

First, the data is taken from certain sources. This data is used to determine the symptoms of the patient. The data is collected as raw data. The data is taken from careful observations of different patients reports for a long time. The values are observed and the details are noted. The details which are used for this prediction are parameters and symptoms that influence disease prediction. These data are carefully collected and then pre-processed. [9]

2.2 Data Preprocessing

The Raw data consists of many unwanted information. And so, the data is pre-processed to produce the required data set. This is done to handle missing values and to ensure data consistency. Then the data is standardized. The dataset is split into training, validation and test sets. This is done to evaluate the model properly. [10]

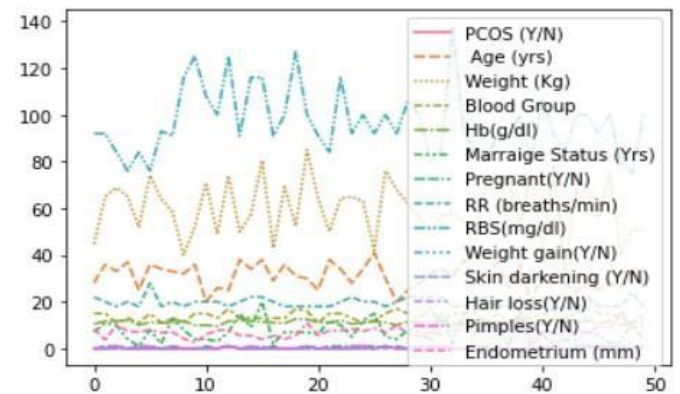


Figure 1 Graph That Shows the Fluctuation of Different Parameters

The above graph is the line plot of the factors like age, weight, blood group, RR rate, that influence the disease prediction. Blue coloured line represents the variation of RBS rate and RR rate of the patients. It shows highest amount of fluctuation. The brown coloured line has the medium fluctuation among all the other lines. The least fluctuation is shown by the purple line which represents the variation in the endometrium level. It is very consistent and has smaller fluctuations. So, it's not considered as one of the major factors which affect disease prediction. Then, the pink coloured line also shows similar fluctuations. It represents the variation in pimples affected years level of different data. The green coloured line shows smaller fluctuation which represents the hair loss level of different patients. The red coloured line also contains smaller fluctuations. It represents the skin darkening level of the patients. Thus, the priority of the factors to be considered is age, HB level, blood group, RR rate and RBS rate. [11] Figure 1 shows the fluctuation of different parameters.

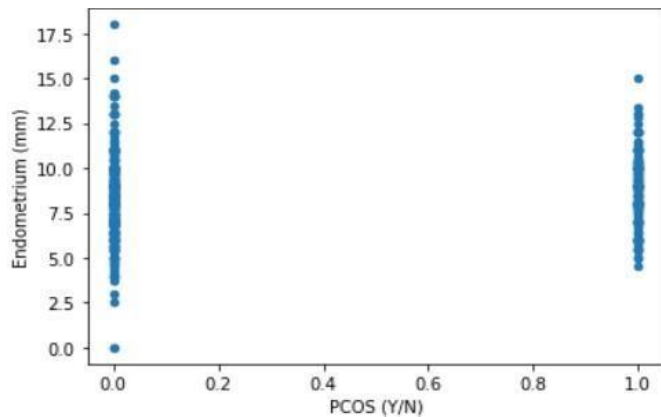


Figure 2 Scatter Plot Graph

The above graph of Figure 2 represents the Scatter plot of Endometrium thickness content with respect to PCOS . The scatter plot contains many dots plotted on it sequentially which can be seen as a single line. Each dot represents the endometrium thickness level of patient required for the disease prediction. [12]

2.3 Algorithm –Decision Trees

Although decision trees are shown in Figure 3. A supervised learning technique, they are primarily used to solve classification problems. However, they can also be used to solve regression problems. This classifier is tree-structured, with internal nodes standing in for dataset features, branches for decision rules, and leaf nodes for each outcome. The Decision Node and the Leaf Node are the two nodes that make up a decision tree. While leaf nodes are the result of decisions and do not contain any further branches, decision nodes are used to make any kind of decision and have multiple branches. The characteristics of the provided dataset are used to inform the decisions or the test. It is a graphical tool that shows all of the options for solving a problem or making a decision given certain parameters. It is named a decision tree because, like a tree, it begins with the root node and grows on subsequent branches to form a structure resembling a tree. A decision tree merely poses a question, and then divides the tree into subtrees according to the response (yes/no). [13]

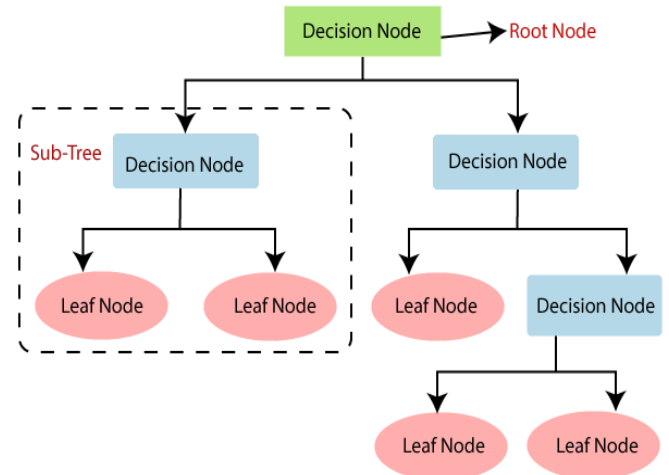


Figure 3 Decision Tree Representation

2.4 Working

The first step is to collect the datasets and preprocess it. Then features applicable are determined. Important Elements are found: The characteristics that had a major influence on disease prediction are determined using statistical techniques or domain expertise. Now Train-Test Split is performed. To assess how well the model generalizes, the dataset is divided into training and testing sets (i.e., 80% training, 20% testing). Then the Model is initialized. The model is selected and initialized using a suitable decision tree algorithm. The Model is mastered. Training dataset to train the decision tree model are used. Adjusted Hyper parameters (5.3): adjusted hyper parameters (such as minimum samples per leaf and tree depth) to improve model performance and prevent overfitting. Now, Decision Trees are visualized. To better understand the decision-making process, a visual representation of the decision tree is created. Machine learning libraries' visualization features or tools such as Graph viz are utilized. The main objective of our paper is to provide an early detection model for PCOS. The early detection of PCOS reduces the possibility of long-term complications. Several ML utilized to build the proposed stacking ensemble ML model. It combines diverse ML models (LR, DT, RF etc.) at the base learner level with RF at the meta-learner level is proposed to improve the performance of a single ML. [14] System architecture shown in

Figure 4.

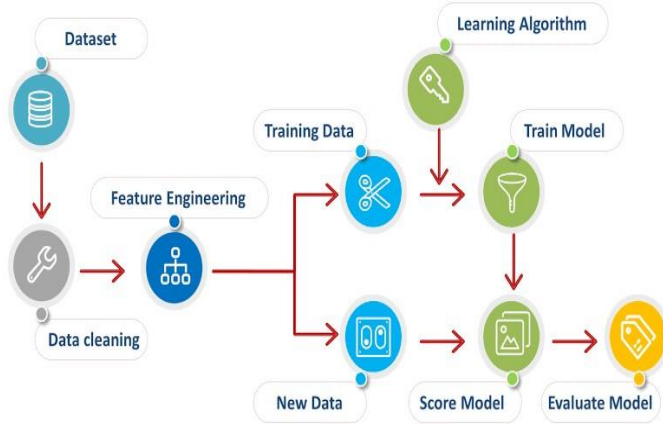


Figure 4 System Architecture

3. Results

The decision tree model has been created and trained.

```

In [26]: Result= regressor.predict(np.array([[28,44.6,15,10.48,7.0,0,22,92.0,0,0,0,8.5]]))

In [27]: if Result[0]==0:
          print('Output : ', 'PCOS DISEASE')
        else:
          print('Output : ', 'HEALTHY')

Output : PCOS DISEASE

In [28]: Result1= regressor.predict(np.array([[33,68.8,11,11.80,10.0,1,18,84.0,0,0,1,1,10.0]]))

In [29]: if Result1[0]==0:
          print('Output : ', 'PCOS DISEASE')
        else:
          print('Output : ', 'HEALTHY')

Output : HEALTHY
  
```

Figure 5 Output

The above image tells us about the prediction of the disease for the values that has been entered by the patients. [15]

Conclusion

In conclusion, the application of machine learning algorithms for predicting Polycystic Ovary Syndrome (PCOS) marks a significant leap forward in the field of healthcare. The research undertaken has demonstrated the effectiveness of various machine learning models in predicting and diagnosing PCOS, offering a promising avenue for early detection and intervention. In summary, the integration of machine learning algorithms for PCOS prediction represents a groundbreaking approach that holds great promise for the future of women's health.

In this we compared the outcomes with three different algorithms that includes decision tree, random forest, and linear regression. In this decision tree gives more accuracy compared to other two algorithms. The accuracy comparison is given below in Table 1. The final results of PCOS disease shown in Figure 5.

Table 1 Accuracy Comparison

ALGORITHM	ACCURACY
Linear Regression	0.34
Random Forest	0.88
Decision Tree	1.00

By leveraging the power of artificial intelligence, we can revolutionize early detection, intervention, and personalized treatment strategies, ultimately contributing to better outcomes. The future scope for PCOS prediction using machine learning algorithms is poised for significant advancements. Integration of cutting-edge technologies such as genetic sequencing, advanced imaging, and wearable devices holds immense potential for refining predictive models.

References

[1]. Kyrou, E. Karteris, T. Robbins, K. Chatha, F. Drenos, and H. S. Randeve, 617 “polycystic ovary syndrome (PCOS) and COVID-19: An overlooked 618 female

- patient population at potentially higher risk during the COVID-19 pandemic,”. 2020
- [2]. B. J. Sherman, N. L. Baker, K. T. Brady, J. E. Joseph, L. M. Nunn, and 621 A. McRae-Clark, “The effect of oxytocin, gender, and ovarian hormones 622 on stress reactivity in individuals with cocaine use disorder,” May 2020.
- [3]. X.-Z. Zhang, Y.-L. Pang, X. Wang, and Y.-H. Li, “Computational charac- 625 terization and identification of human polycystic ovary syndrome genes,” 626 *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. No. 2949. 627
- [4]. E. Khashchenko, E. Uvarova, M. Vysokikh, T. Ivanets, L. Krechetova, 628 N. Tarasova, I. Sukhanova, F. Mamedova, P. Borovikov, I. Balashov, and 629 G. Sukhikh, “The relevant hormonal levels and diagnostic features of 630 polycystic ovary syndrome in adolescents, Jun. 2020. 632
- [5]. M. Woźniak, R. Krajewski, S. Makuch, and S. Agrawal, *phytochemicals* 633 Jan 2021
- [6]. Dana Hdaib; Noor Almajali; Hiam Alquaran; 2022 “Detection of Polycystic Ovary Syndrome (PCOS) using Machine Learning Algorithm.
- [7]. Preeti Chauhan; Pooja Patil; Neha Rane; “Comparative Analysis of Machine Learning Algorithms for Prediction of PCOS” 2022.
- [8]. V Srinithi; R.Rekha; “Machine Learning for diagnosis of Polycystic ovary syndrome (PCOS/PCOD)” 2023.
- [9]. M Sneha; K. Sai Sravani; S. Nithya; Kaladevi R; “Analysis of Irregular Menstrual Cycle and its Consequences” 2023.
- [10]. Prajna K B; Balasubramanian V Iyer; Bhuvan C; “Implementation of various machine learning polycystic ovary syndrome” 2023
- [11]. P. Chitra; M. Sumathi; K. Srilatha; F.V. Jayasudha; S. Amudha; “Review of Artificial Intelligence based algorithm for prediction of Polycystic Ovary Syndrome (PCOS) from blood sample.” 2022
- [12]. Sakthipriya Dhinakaran; Chandrakumar Thangavel; “PCOS Perception analysis using machine learning algorithms.” 2022
- [13]. P. Sheruba; R. Madhumitha; Arivazhagi; “Prediction of Ovarian Syndrome using deep feature representation and learning approaches”. 2023
- [14]. Neeta Chavan; Sakshi karkera; Isha Chavan; “Comparative study of machine learning algorithms for prediction of polycystic ovary syndrome.” 2023
- [15]. Kajal Gupta; Rajesh prasad; “Machine learning solutions to polycystic ovary syndrome: A review”.2023