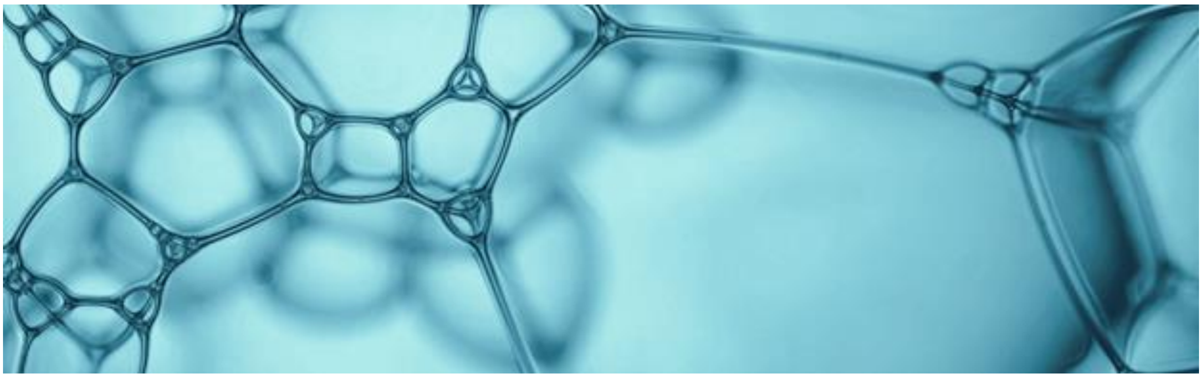


PROJECT REPORT

ON

Prediction of Polycystic Ovarian Syndrome (PCOS) using Machine Learning Techniques

Msc.Big Data Analytics, St Joseph's University Bangalore

Submitted by:

222BDA06 – Jaini Maria John

222BDA17 – Lajiya Aleena Saji

222BDA28 – Keethana P P

222BDA30 – Gladys Nathan

Submitted to:

JAYATI KAUSHIK

Assistant Professor

Department of Advanced Computing

St. Joseph's University

AIM

Polycystic Ovary Syndrome (PCOS) is a health problem that affects women of childbearing age. Women with PCOS have a hormonal imbalance and metabolism problems that may affect their overall health. The aim of this project is to build a predictive model that can identify the women who are at risk of getting PCOS based on their medical reports, lifestyle and other relevant factors. We want to identify trends and risk variables that can be used to develop an accurate predictive model by analysing a large dataset of women with and without PCOS. Our goal is to equip healthcare professionals with a tool that will enable them to recognise and treat women who are at risk of getting this debilitating ailment early on. In order to achieve that, we will combine statistical analysis with machine learning methods to create a solid model that can precisely predict the likelihood of PCOS in specific patients, thereby enabling earlier diagnosis and better treatment options. We want to improve the lives of the millions of women suffering by PCOS around the world through our effort.

For hypothesis purpose we defined null and alternative hypothesis. The alternative hypothesis(H1) would be there is a significant association, while the null hypothesis(H0) would be that there is no significant relationship between the independent factors and PCOS.

DOMAIN – Health sector

PROBLEM STATEMENT

Study the demographic affected by PCOS. What are the likely causes? What are some good predictive features to predict PCOS?

INTRODUCTION

Polycystic ovary syndrome (PCOS) is a health problem that affects almost 10% of women population of reproductive age. Some of the symptoms include ovarian cysts, increased testosterone production, and irregular menstrual cycles. PCOS can have a severe influence on a woman's health, fertility, and quality of life. A large number of women in the world today is significantly impacted by preterm abortions, infertility, and other problems. It has been highlighted that PCOS, a condition that affects a lot of women in their reproductive age, is one of the main factors contributing to infertility. More than five million women of reproductive age suffer from PCOS worldwide. It is an endocrine condition marked by changes in female hormone levels and aberrant male hormone production. This condition causes ovarian malfunction, which increases the risk of pregnancy and gravidity. Some of the symptoms of PCOS include obesity, irregular menstruation, an overproduction of the male hormone, acne, etc. PCOS is extremely difficult to diagnose because of the vast spectrum of symptoms and the occurrence of other linked gynaecological issues. The time and money-consuming various clinical tests and ovary scanning procedures are now a hardship for PCOS patients.

PCOS can be treated to a certain extent with regulated medications and lifestyle changes. Due to the wide range of symptoms associated with this illness, doctors are compelled to request numerous clinical test findings and pointless radiological imaging treatments. As PCOS directly causes ovarian dysfunction with an increased risk of miscarriage, infertility, or even gynaecological cancer and mental distress for the patients due to time and financial waste, early detection and diagnosis of the condition with minimal tests and imaging procedures is of the utmost importance.

LITERATURE REVIEW

Among the in-numerous problems that exist around us, the problems that are related to the reproductive health of women was selected as an area of our interest, due to its importance in this contemporary society. A detailed survey of studies on PCOS and systems to support its diagnosis was carried out. Literature says that about 5-10% of Indian women in reproductive age are affected by the multifaceted endocrine disorder called Polycystic Ovary Syndrome (PCOS). The symptoms for PCOS might be varying from patient to patient. Some of them are irregularity in menstrual periods, acne, overweight, increased tendency for infertility, intense hair fall, balding of front head, increased facial hair growth. When there are more than 12 follicles per unit area and they are evident on a radiological scan, PCOS is traditionally suspected. According to some scientists, the cut-off point should be raised from 12 to 20 follicles, and ultrasound testing should be replaced with other biomarkers like serum Anti-Mullerian Hormone (AMH) or other biomarkers. The accurate application of a few well-established, standardised diagnostic techniques is all that is necessary for the diagnosis of PCOS. Delays in PCOS diagnosis have a significant impact on patients' wellbeing. It is recommended that treatment be symptom-focused, long-term, dynamic, and tailored to the patient's unique requirements, expectations, and changing circumstances. The relationship between PCOS and the infertility rate among women in this group is also taken into account, and PCOS-positive individuals were substantially more likely to utilise fertility hormone therapy. Strategies to improve PCOS diagnosis and the factors affecting fertility are crucial given the prevalence of PCOS and the financial and medical burden of infertility. This is due to the fact that, regardless of BMI, infertility is said to be 15 times more common among women who report having PCOS. Obesity and PCOS are correlated in opposite directions.

AIM OF THE WORK

The aim of the work for is to develop a reliable and accurate method for identifying individuals who may have or be at risk for developing PCOS. This include developing predictive models of accurate and efficient method for identifying individuals who may have Polycystic Ovary Syndrome (PCOS). Various algorithms such as logistic regression, random forest, ada boost can be used to analyse the clinical and/or genetic data and make predictions about the likelihood of an individual having PCOS. Using some hypothesis-testing techniques to validate the project.

Using machine learning techniques other relevant clinical and/or genetic data, it is possible to develop a highly accurate and efficient method for diagnosing PCOS. This method can significantly improve the speed and accuracy of PCOS diagnosis, leading to better clinical outcomes for affected individuals. It is possible to create predictive models that can be used for PCOS management, diagnosis, and therapy. ML systems can find patterns and connections in vast quantities of patient data that may be challenging for humans to notice. Nearly 12–21% of women of reproductive age have PCOS, and among them, 70% are still waiting for a diagnosis. Thus, this project aims to detect the PCOS in early stage depending on symptoms and help in curing them.

METHODS AND MATERIALS

For the development of an appropriate machine learning model based diagnostic aid for PCOS, a comparison of performance of various existing algorithms in our data set need to be presented. Preparation of the model is the most crucial step that provides the outline of the research. The following are the methods and materials used

- We have used Pandas and NumPy library for basic operations with dataset
- We have used Matplotlib and Seaborn for plotting graphs for Exploratory Data Analysis
- We have used math library to perform some mathematical operations.
- We have used sklearn library to create a predictive model. From sklearn library we imported Logistic Regression and created the model for predicting the outcome of our model. We also imported the accuracy score to know how accurate the model is.
- From sklearn library we used train_test_split which is used to split the data into training data and testing data is compared with the predicted output and is checked whether the model is working correctly or not.
- We installed Scipy and stats models to do the hypothesis testing.

Software used - python jupyter notebook

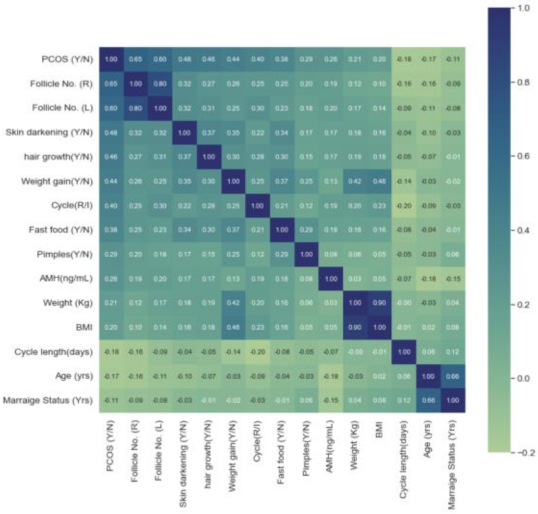
STUDY DESIGN

PCOS (polycystic ovarian syndrome) is a hormonal disorder found in women of reproductive age that are characterized by symptoms like irregular menstrual cycles, high levels of male hormones, and ovarian cysts. PCOS can even lead to various complications including infertility, type 2 diabetes, and even cardiovascular diseases. So, it is necessary to detect and treat them. Following are the steps that we implemented for the prediction of PCOS.

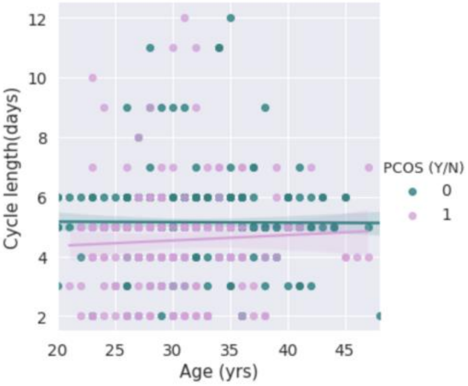
- Study population – study the demographic affected by PCOS, analysing the symptoms of women with PCOS and without PCOS.
- Feature selection – identifying the most relevant features that can be used to distinguish between individuals with and without PCOS.
- Data collection and Pre-processing – the first step involves collecting dataset, which is typically done from Kaggle. Our dataset contains total 44 parameters. Then the dataset is pre-processed for data analysis.
- Exploratory data analysis – creating visualisation such as boxplot, heatmap, scatterplot to explore the distribution of the data and examining the correlation between the different features. EDA let us know about the different features and how they affect our present population.
- Model development – developing machine learning model for predicting PCOS in different women. Since the feature which we are predicting is categorical dependent variable we use logistic regression. We also use ada boost classifier to boost the performance of the model.
- Model Evaluation – the performance of the models is evaluated using the accuracy.
- Interpretation and analysis – the most important features that cause PCOS are identified and interpreted. This step helps in realising the key features, importance of diagnosing PCOS in women and provide guide to future research.
- Reporting – the findings of this study are reported in the project report including the methods, results, conclusion, future development.

EXPLORATORY DATA ANALYSIS

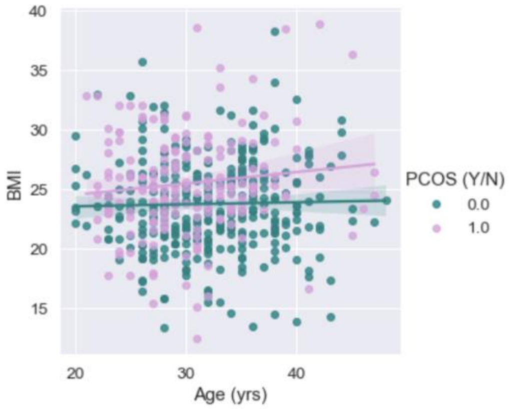
Exploratory data analysis (EDA) is an essential step in understanding and interpreting data in PCOS diagnosis. EDA involves visualizing and summarizing the main features and patterns of the data to identify potential outliers, missing values, or trends that may affect the validity of the results. We had plotted boxplot, heatmap, scatterplot for some of the important features.



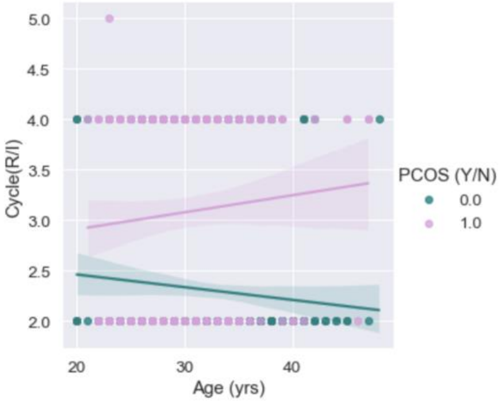
By examining the correlation between the independent variables and dependent variable we found that PCOS is highly correlated to follicle numbers, skin darkening, hair growth i.e., an increase in number of Follicle No. (L) and Follicle No. (R) affects the PCOS.



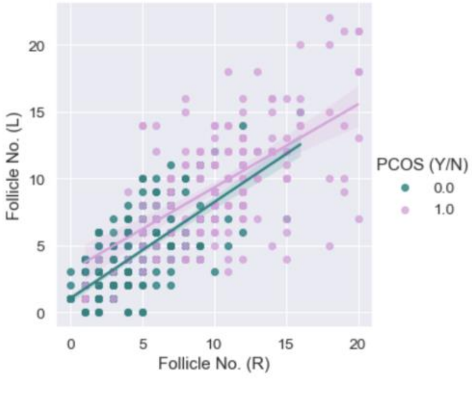
From the above plot, length of the menstrual phase is overall consistent over different ages for normal cases. Whereas in the case of PCOD the length increased with age.



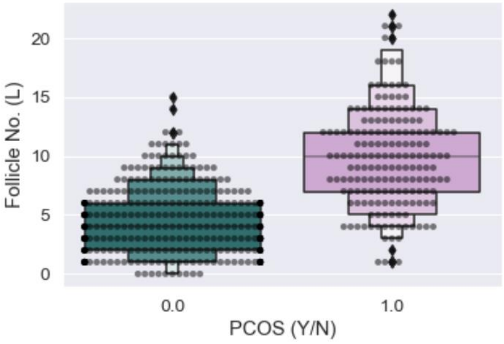
Body mass index (BMI) is showing consistency for normal cases, whereas for PCOS the BMI increases with age.

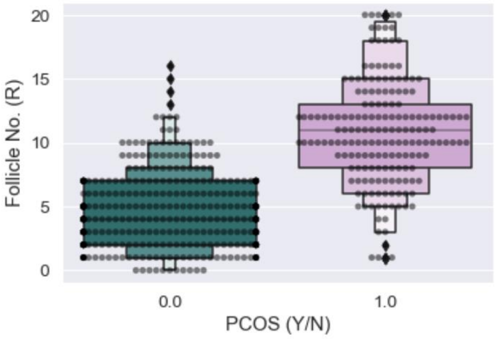


The mensural cycle becomes more regular for normal cases with age. Whereas, for PCOS the irregularity increases with age.



The distribution of follicles in both ovaries Left and Right are not equal for women with PCOS in comparison with the "Normal" patient.





The number of follicles in women with PCOS is higher, as expected. And are unequal as well. Overall, the visualization validated the important features in our dataset.

RESULTS

After EDA we did model testing and hypothesis testing. Since our data contains large number of features and in order to acquire more accuracy, we did random forest classifier model. We got ang accuracy of 89.57%. While plotting confusion matrix for this model we were about to interpret that 63% of correctly predicted data as well as 27% of wrongly predicted data is there in the model from which we were about to reach an accuracy of 89.57%. Since we have more than one categorical variable our data is more suitable for logistic regression, so we gave it a try. We predicted the model using this classifier, but our accuracy was reduced to 84.66%. The confusion matrix provides us with the information that the model correctly predicts 60% of data and wrongly predicts 25% of data.

For doing hypothesis, as we had more than one categorical variable in our dataset, we chose chi-square test for the features PCOS[Y/N] and hair growth[Y/N]. In the output we got chi-square statistic as 114.599, the p value as $9.634e^{-27}$ and degrees of freedom as 1. So, the result suggests us there is significant association between the two categorical variables. From the expected frequency table, it shows that observed frequencies in each cell differ from the expected frequencies which contributes to the large chi-square statistic and small p value. While examining the correlation between the dependent and independent variable we can see that follicle no is highly correlated, so we did t test over follicle no and PCOS [yes/no], and we got t statistic as 34.048 and p value as $1.5605351809577597e^{-136}$. The large T statistic value and small p value it appears that follicle number is a significant predictor of PCOS.

However, it is important to consider our data size, so we did ANOVA also. From analysis of variance, we took our null hypothesis as there is no significant difference in follicle no between individual with PCOS and without PCOS and alternative hypothesis as there is significant difference in follicle no between individual with PCOS and without PCOS. We got p value as 0 which is less than 0.05 and f value as 390.83593 which shows we could reject our null hypothesis it indicates there is a significant difference with follicle no between the individual with PCOS and without PCOS. We did it for both follicle number left and right.

CODE: https://github.com/JainiMariaJohn/ADS-project

CONCLUSION

Polycystic ovary syndrome (PCOS) is a complex condition with multiple potential causes. However, based on current study, there are several factors that are believed to contribute to the development of PCOS. Based on our study we concluded that follicle no, skin darkening, hair growth, weight gain, cycles are some of the major symptoms of PCOS. PCOS tends to run in families, suggesting that there is a genetic component to the disorder, Insulin resistance occurs when the body's cells become resistant to the hormone insulin, which can lead to high blood sugar levels. Women with PCOS are more likely to have insulin resistance, which can cause the ovaries to produce more androgens (male hormones). Women with PCOS have higher than normal levels of androgens, which can interfere with the development and release of eggs from the ovaries. Some studies suggest that chronic low-grade inflammation may play a role in the development of PCOS. Exposure to certain environmental toxins, such as bisphenol A (BPA), may contribute to the development of PCOS. Obesity, poor diet, and lack of exercise may also contribute to the development of PCOS. The exact cause of PCOS is unknown, but researchers believe that a combination of genetic, environmental, and lifestyle factors may play a role. Treatment options for PCOS vary depending on the individual's symptoms and may include lifestyle changes, medications, or surgical interventions. Overall, PCOS is a complex and challenging condition that requires careful management and support from healthcare professionals.

FUTURE WORK

Machine learning (ML) methods have a great deal of potential for use in PCOS prediction. To increase the precision and applicability of PCOS prediction models, we can concentrate on the following topics in the future:

Data collection and standardisation: The accuracy and completeness of the data used to develop and test machine learning algorithms can have a big impact on how well those algorithms work. Future research should concentrate on gathering sizable and varied datasets that include many PCOS characteristics, including clinical traits, biomarkers, and imaging evidence.

Model selection and assessment: The performance and scalability of the models can be influenced by the hyperparameters of the ML algorithms that are chosen. Future research should concentrate on evaluating and choosing the best machine learning (ML) methods, as well as findings in deep learning models for PCOS prediction.

Clinical translation and validation: Improving clinical judgement and patient outcomes is the ultimate goal of PCOS prediction models. Future research should concentrate on testing the models' functionality and generalizability in actual environments, such as primary care offices or fertility clinics. The models can also be translated into clinical practise for analysing their effects on patient outcomes like time to diagnosis, therapy response, or quality of life.

KEY LEARNINGS

Some of the key learnings from our project include importance of interdisciplinary collaboration, data quality and standardization, feature selection and engineering, model selection and evaluation, potential for personalized diagnosis and treatment.

PCOS is a hormonal disorder that calls for knowledge from endocrinology, gynaecology, genetics, and data science, among other fields. Working together with professionals in these domains can help to improve clinical practise and assure the validity and generalizability of the findings. The validity and generalizability of PCOS findings might be strongly impacted by the quality and completeness of the data used in the study. The comparability and stability of the results can be improved, as well as the precision of PCOS diagnosis and therapy, by ensuring data quality and standardisation. The choice of ML algorithms and their hyperparameters can affect the performance and scalability of PCOS prediction models. PCOS is a diverse condition with a range of clinical manifestations and treatment results. It is possible to increase the precision and efficacy of PCOS therapy and boost patient outcomes by using ML algorithms to forecast individualised diagnosis and treatment.

REFERNCES

- https://www.sciencedirect.com/science/article/abs/pii/S1472648310616446
- https://ieeexplore.ieee.org/document/8929674
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6952118/
- https://www.researchgate.net/profile/Namrata-Tanwani/publication/348415641_Detecting_PCOS_using_Machine_Learning/links/5ffda7f892851c13fe0717fd/Detecting-PCOS-using-Machine-Learning.pdf