PCOcare: PCOS Detection and Prediction using Machine Learning Algorithms

Article *in* Bioscience Biotechnology Research Communications · December 2020

DOI: 10.21786/bbrc/13.14/56

2 authors:

Shreyas Vedpathak
MIT World Peace University
3 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Vaidehi Sunil Thakre
MIT World Peace University, Pune, India
4 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 14 (2020) Pp-240-244

BBRC
Bioscience Biotechnology Research Communications

PCOcare: PCOS Detection and Prediction using Machine Learning Algorithms

Vaidehi Thakre[1*], Shreyas Vedpathak[2], Kalpana Thakre[3] and Shilpa Sonawani[4]

[1,2] *Student (Third Year B.Tech. CSE), MIT – World Peace University Pune, India,*

[3]*Professor, Sinhgad College of Engineering Pune, India,*

[4]*Assistant Professor, MIT – World Peace University Pune, India*

ABSTRACT

Polycystic Ovary Syndrome (PCOS) is a medical condition which causes hormonal disorder in women in their childbearing years. The hormonal imbalance leads to a delayed or even absent menstrual cycle. Women with PCOS majorly suffer from excessive weight gain, facial hair growth, acne, hair loss, skin darkening and irregular periods leading to infertility in rare cases. The existing methodologies and treatments are insufficient for early-stage detection and prediction. To deal with this problem, we propose a system which can help in early detection and prediction of PCOS treatment from an optimal and minimal set of parameters. To detect whether a woman is suffering from PCOS, 5 different machine learning classifiers like Random Forest, SVM, Logistic Regression, Gaussian Naïve Bayes, K Neighbours have been used. Out of the 41 features from the dataset, top 30 features were identified using CHI SQUARE method and used in the feature vector. We also compared the results of each classifier and it has been observed that the accuracy of the Random Forest Classifier is the highest and the most reliable. The dataset used for training and testing is available on KAGGLE and owned by Prasoon Kottarathil.

KEY WORDS: MACHINE LEARNING, POLYCYSTIC OVARY SYNDROME, SUPPORT VECTOR MACHINE, LOGISTIC REGRESSION, RANDOM FOREST CLASSIFIER, GAUSSIAN NAIVE BAYES, CHI-SQUARE.

INTRODUCTION

In the past few decades, technology has revolutionized our universe and affected our lives, making them easier day by day. Emerging technologies are reshaping mankind in a lot of ways. These days, machine learning, a field of study that gives computers to learn without being explicitly programmed, is playing a key role in the healthcare sector. Machine learning can deal with obscenely huge datasets, convert analysed data into clinical insights and help in the diagnosis of various ailments. Polycystic Ovary Syndrome (PCOS) is a medical condition which causes hormonal disorder in women in their childbearing years. PCOS occurs as a result of hormonal imbalances. In this disorder, the ovaries develop small collections of fluids called follicles (cysts) and fail to release eggs, which is why women suffering from PCOS tend to have complications in conceiving [Zhang, 2018]. A lot of women have PCOS, but do not get diagnosed with it at an earlier stage. In a study, 69 to 70 percent of women did not have a pre-existing diagnosis [Dewailly, 2013].

While the actual causes of PCOS remain a mystery, studies say that it is generally inherited. It is a very unpredictable condition as the cure is uncertain since there is no observable trend for this medical condition. The time and cost of taking innumerable medical tests and scanning is a burden for the patients and the doctors too. Hence, early diagnosis and treatments are important as long-

ARTICLE INFORMATION

*Corresponding Author: vaidehithakre21@gmail.com
Received 18th Oct 2020 Accepted after revision 29th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31)

Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.14/56

term health risks like type-2 diabetes, cardiovascular diseases can be avoided by simple changes in lifestyle. Common symptoms include irregular periods, excessive androgen levels (male hormones), polycystic ovaries. Hence, parameters such as Follicle-Stimulating Hormone (FSH), Luteinizing Hormone (LH), Human Chorionic Gonadotropin (HCG), number of follicles, Thyroid Stimulating Hormone (TSH), Age, cycle length, cycle regularity, etc. are taken into account to formulate the feature vector for our machine learning models. Early detection can help make necessary lifestyle changes beforehand and hence reduce risks of the condition as women with PCOS are three times more likely to undergo miscarriages in early stages of pregnancy, suffer from infertility and in rare cases, gynaecological cancer.

Literature Survey: From 1 in 10 women suffering from PCOS worldwide to currently 3-4 in 10 women, PCOS is now exponentially increasing among women due to an unhealthy lifestyle. The literature says that 1 in every 5 women in India suffers from PCOS. PCOS symptoms differ in every patient. The major diagnosis includes scanning for follicles, their number and sizes using Ultrasound imaging. In the existing literature, several various techniques have been used to analyse and detect PCOS. We need to refer to the categories of PCOS standards to gain complete understanding of what PCOS is. Even though it is called Polycystic Ovary Syndrome, it is not essentially described by ovarian cysts. It is defined by examining at least two of three diagnostic criteria. These criteria which are used for diagnosis have been evaluated multiple times separately by the National Institutes of Health (NIH, in 1990), by the European Society of Human Reproduction and Embryology (ESHRE) and the American Society for Reproductive Medicine (ASRM, in 2003) (popularly known as the Rotterdam criteria). In 2012, the 2003 Rotterdam criteria were endorsed by NIH for PCOS. Table 1 illustrates the criteria used for diagnosis of PCOS which have been set as a standard by NIH.

For an accurate PCOS diagnosis, disorders that have specific signs and symptoms that match with those of PCOS must be dismissed. Hyperprolactinemia, Cushing's syndrome and non-classic congenital adrenal hyperplasia are few examples. [Zhang, 2018] have used different machine learning algorithms like K-nearest neighbour (KNN), decision tree and SVM with different kernel functions to predict PCOS from the identification of new genes. [P. Mehrotra, 2011] have used machine learning algorithms like Bayes and Logistic Regression (LR) to develop an automated system that will act as an assisted tool for the doctor for saving considerable time in examining the patients and hence reducing the delay in diagnosing the risk of PCOS by using metabolic and clinical factors in a feature vector. [Norman, 2007], have done a comprehensive study on the disorder and its three diagnostic criteria in depth giving us insights on not just PCOS but also abnormalities of insulin, gonadotropin and folliculogenesis. [Essah, 2006], have discussed how there exists an overlap between the metabolic syndrome and the polycystic ovary syndrome (PCOS).

That article discusses the existing data regarding the familiarity, characteristics, and treatment of the metabolic syndrome in women with PCOS. [Amsy Denny, 2011], have proposed a system for the early detection and prediction of PCOS from an optimal and minimal but promising clinical and metabolic parameter, which act as an early marker for this disease. [Dewailly, 2013] have illustrated in their literature how the diagnosis of PCOS depends on biological, clinical and morphological criteria. As ultrasonography has technologically advanced, the excess follicle has become the primary criterion of polycystic ovarian morphology (PCOM). Since 2003, most investigators have used a threshold of 12 follicles (measuring 2–9 mm in diameter) per whole ovary, but that now seems obsolete [A. Saravanan, 2018]. The fluctuations in the quantity of ovarian volume or area may also be considered as accurate markers of PCOS Morphology, yet their utility compared with excess follicle remains a puzzle.

METHODOLOGY

The formulation of a good machine learning model is an important aspect of project design. Having the correct patient data is very important because one cannot afford mistakes while devising healthcare services. We have used multiple machine learning models to check which model gives us the most accurate results. To support our claims and results obtained, use of plots and evaluation metrics has been made. A basic workflow diagram to explain the proposed system is given in Figure 1. The following sections will give a detailed insight into the system:

Figure 1: Architecture of the Proposed System

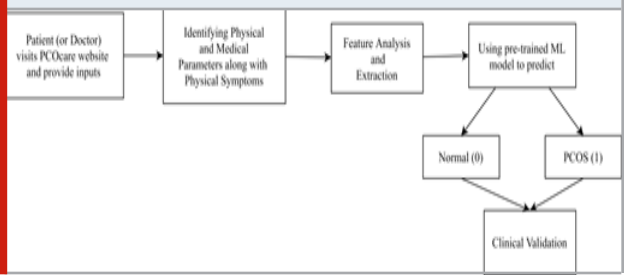
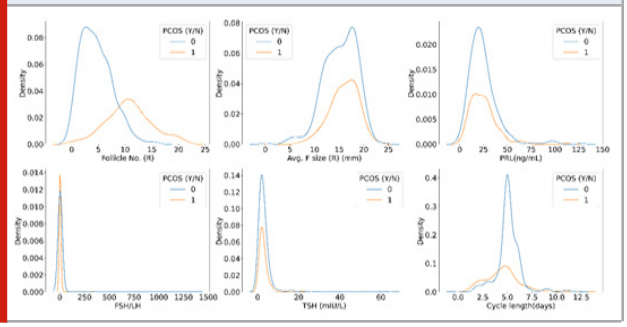


Figure 2: Kernel Density Estimation Plots



For the machine learning model implementation, Jupyter Notebook has been used. Our proposed web app PCOcare is intended to be developed in Python using Streamlit

Open Source Web App Framework. Data Pre-processing: The dataset found was cleaned. Hence, no data pre-processing was required. The dataset contains columns which have continuous as well as discrete observations. So, let us see if we can derive any useful insights from the columns which have continuous values.

These Kernel Density Estimation Plots demonstrate that patients who had PCOS have similar trends as the patients without PCOS. These distributions are not useful from the point of view of finding features that can help us differentiate between a patient who is diagnosed with PCOS and a patient who is not. So, in order to find important features, statistical help has been taken (Chi-Square Method). Feature Selection and Importance: Feature selection is performed to divide the set of features into a subset of significant features so that the classifier efficiency can be done. From 41 features present in the dataset, Top 30 features were selected using CHI Square method. A chi-square test is a test used in statistics to test the independence of two events. Essentially, what chi-square is doing is that it will calculate scores based on a formula:

$$(X_c)^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

Where, c= degrees of freedom, O= Observed Value and E= Expected Outcome. Once a score is calculated, we can form a conclusion about a particular feature by saying: The higher is the chi-square value for a feature, it is more dependent on the response and can be selected for model training.

Table 2. Chi-Square Score for Top 30 Features (Generated in Jupyter Notebook)

Rank	Feature	Score	Rank	Feature	Score
1	PRL(ng/mL)	9600.594	16	Vit D3 (ng/mL)	25.00828
2	No. of abortions	6899.359	17	Hair loss(Y/N)	23.56211
3	FSH(mIU/mL)	2572.754	18	Cycle length(days)	19.71094
4	II beta-HCG(mIU/mL)	1592.273	19	Height(Cm)	15.10558
5	I beta-HCG(mIU/mL)	1012.629	20	Skin darkening (Y/N)	8.910647
6	Follicle No. (L)	673.1438	21	Cycle(R/I)	8.230296
7	BP _Diastolic (mmHg)	564.5952	22	Follicle No. (R)	7.460844
8	TSH (mIU/L)	221.8157	23	FSH/LH	5.426396
9	LH(mIU/mL)	96.23587	24	Hip(inch)	5.219221
10	hair growth(Y/N)	85.66499	25	PRG(ng/mL)	4.779813
11	Weight gain(Y/N)	84.0381	26	Avg. F size (L) (mm)	3.352904
12	RBS(mg/dl)	65.01353	27	Avg. F size (R) (mm)	3.144839
13	Age (yrs)	50.85829	28	Pregnant(Y/N)	2.824165
14	Pimples(Y/N)	37.43732	29	Fast food (Y/N)	1.856357
15	Hb(g/dl)	27.7938	30	Blood Group	1.235629

Random Forest Classifier: Random Forest Algorithm is another example of Ensembling methods. It combines result from many decision trees to derive a conclusion. Used for solving problems based on both Regression and Classification.

Support Vector Classifier: Support Vector Machine algorithms are supervised machine learning algorithms which are used for regression, classification and outlier detection problems. In SVM, it basically plots the data as points in an n-dimensional space, where n is the number of features. The algorithm tries to find a hyperplane which can separate the plotted points into the required or identified number of classes. Logistic Regression: Logistic Regression is a classification algorithm. It is a supervised machine learning algorithm. It uses sigmoid function to perform its hypothesis. The outcome of the hypothesis is the estimated probability. It is in terms of binary i.e. will it happen or not basically 1 or 0 respectively.

Gaussian Naive Bayes: Naive Bayes are algorithms that use the Bayes' theorem to calculate the probability and decide to which class does the given data would belong. We have used Gaussian Naive Bayes for our hypothesis. It follows Gaussian Normal Distribution which means there will not be covariance between the features, and it supports continuous data. KNeighbours Classifier: The KNN algorithm assumes that similar data if plotted would exist nearby. We first load the data and choose how many classes we want the algorithm to classify the data into. The algorithm first calculates the distance of K number of neighbours using distance formula, then it takes the K nearest neighbours according to the distance we calculated. Among these classes, it counts the number of data items for each class. It then allocates the new data points to that class, where it has the greatest number of neighbours.

Comparison of Models: After implementing the machine learning algorithms, the following observations and

results were obtained - Table 3 and 4. These have been generated in Jupyter Notebook using Scikit Learn Library. Table 3 describes the K Fold cross-validation scores of each algorithm on training data and Table 4 describes the precision, recall and Fscore for each model on testing data. From the tables 3 and 4, Random forest Classifier is seen to perform better than the respective others. Hence, Random Forest Classifier is used for our final hypothesis which will predict results using test data. The given ROC curve is plotted using test data and is 89.0% accurate (Given by area under the curve). The train data accuracy of the same is 90.9% (Refer to Table 3).

Table 3. K Fold Cross Validation Scores

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean Accuracy
Random Forest Classifier	0.895349	0.930233	0.872093	0.918605	0.929412	0.909138
Logistic Regression	0.918605	0.918605	0.872093	0.872093	0.917647	0.899808
Linear SVM	0.895349	0.883721	0.872093	0.848837	0.905882	0.881176
Radial SVM	0.941860	0.883721	0.813953	0.883721	0.882353	0.881122
KNeighbors Classifier	0.883721	0.883721	0.837209	0.883721	0.894118	0.876498
Gaussian Naive Bayes	0.895349	0.906977	0.779070	0.848837	0.917647	0.869576

Table 4. Classification Report

	Precision (Class 1, Class 2)	Recall (Class 1, Class 2)	Fscore (Class 1, Class 2)
Linear SVM	(0.911, 0.850)	(0.911, 0.850)	(0.911, 0.850)
Radial SVM	(0.855, 0.906)	(0.955, 0.725)	(0.902, 0.805)
Logistic Regression	(0.888, 0.888)	(0.941, 0.800)	(0.914, 0.842)
Random Forest Classifier	(0.891, 0.941)	(0.970, 0.800)	(0.929, 0.864)
KNeighbors Classifier	(0.820, 0.866)	(0.941, 0.650)	(0.876, 0.742)
Gaussian Naive Bayes	(0.923, 0.813)	(0.882, 0.875)	(0.902, 0.843)

RESULTS

Figure 3: Confusion Matrix for Random Forest Classifier

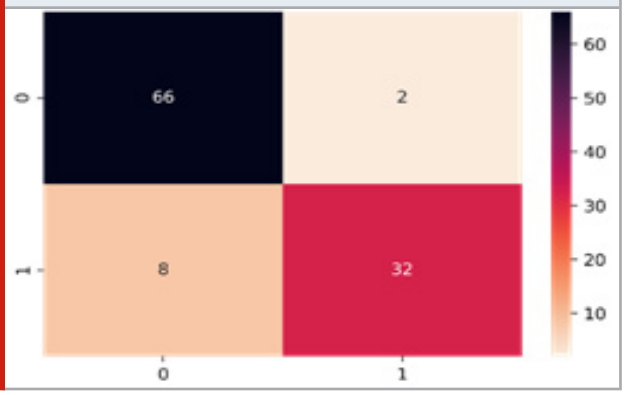
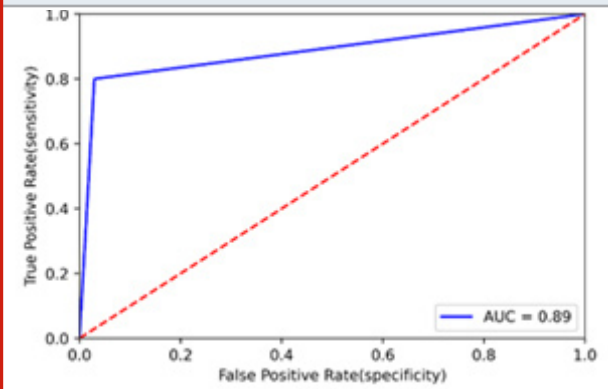


Figure 4: ROC Curve for Random Forest Classifier



CONCLUSION

Polycystic Ovary Syndrome (PCOS) is a medical condition which causes hormonal disorder in women in their childbearing years. The hormonal imbalance leads to a delayed or even absent menstrual cycle. Women with PCOS majorly suffer from excessive weight gain, facial hair growth, acne, hair loss, skin darkening and irregular periods leading to infertility in rare cases. Our proposed system helps in early detection and prediction of PCOS treatment from an optimal and minimal set of parameters which have been statistically analysed using the chi-square method. The Random Forest Classifier was found to be the most reliable and most accurate among 4 others with accuracy being 90.9%. The proposed system can be used by both patients and by doctors too, as a doctor can filter new patients with basic information and give priority to treat patients with PCOS first and then meet patients who do not have PCOS.

REFERENCES

A. Saravanan, S. Sathiamoorthy, “Detection of Polycystic Ovarian Syndrome: A Literature Survey,” Asian Journal of Engineering and Applied Technology, 2018.

Amsy Denny, Anita Raj, Ashi Ashok, Maneesh Ram C, Remya George, “i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques,” 2019 IEEE Region 10 Conference (TENCON 2019), 2019.

Dewailly, D., Lujan, M.E., Carmina, E., Cedars, M.I., Laven, J., Norman, R.J. and Escobar-Morreale, H.F., 2013. Definition and significance of polycystic ovarian morphology: a task force report from the Androgen Excess and Polycystic Ovary Syndrome Society. Human reproduction update, 20(3), pp.334-352.

Essah, P.A. and Nestler, J.E., 2006. The metabolic syndrome in polycystic ovary syndrome. Journal of endocrinological investigation, 29(3), pp.270-280.

Norman, R.J., Dewailly, D., Legro, R.S. and Hickey, T.E., 2007. Polycystic ovary syndrome. The Lancet, 370(9588), pp.685-697.

P. Mehrotra, J. Chatterjee, C. Chakraborty, B. Ghoshdastidar and S. Ghoshdastidar, ”Automated screening of Polycystic Ovary Syndrome using machine learning techniques,” 2011 Annual IEEE India Conference, Hyderabad, 2011, pp. 1-5.

Rotterdam EA-SPCWG. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. Fertil Steril. 2004;81(1):19-25.

Zhang, X.Z., Pang, Y.L., Wang, X. and Li, Y.H., 2018. Computational characterization and identification of human polycystic ovary syndrome genes. Scientific reports, 8(1), p.12949.