ORIGINAL RESEARCH PAPER

Volume - 9 | Issue - 12 | December - 2020 | PRINT ISSN No. 2277 - 8179 | DOI : 10.36106/ijsr

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

A MACHINE LEARNING APPROACH FOR SELECTION OF POLYCYSTIC OVARIAN SYNDROME (PCOS) ATTRIBUTES AND COMPARING DIFFERENT CLASSIFIER PERFORMANCE WITH THE HELP OF WEKA AND PYCARET

Biological Science

Dr. Ashok Munjal PhD; Department of Biochemistry and Genetics,Barkatullah University, Bhopal, MP, India.

Dr. Rekha Khandia* PhD; Department of Biochemistry and Genetics, Barkatullah University, Bhopal, MP, India.*Corresponding Author

Brijraj Gautam M.Sc.;Department of Biochemistry and Genetics, Barkatullah University, Bhopal, MP, India.

ABSTRACT

For any medical treatment there is a requirement of identification of features those are affecting the clinical condition the most. These are the parameters which decide the line of treatment and decide prognostic values. Process of diagnosis includes various aspects like physical examination through symptoms exhibited for a disease, person's previous medical history, and various type of medical tests. Diagnosis of a disease is often challenging since there are many nonspecific signs and symptoms and often are common with other ailments too. In present study we applied Advance Machine Learning approach to identify the major attributes those are involved in polycystic ovary syndrome (PCOS) disease progression as well as help medical professional to predict the disease with accuracy and minimal time. Present work encompasses the use of genetic algorithm a Machine learning approach for selection of major attributes (the sign and symptoms) for PCOS patients data which affect the disease condition most ,in present study various classifiers have been applied in our dataset and different accuracy parameters also have been used including Confusion matrix , Precision , F1 score and AUC (area under the curve) to select the best classifier which classify the diseased and non diseased patients with high accuracy.

KEYWORDS

PCOS, Genetic Algorithm, Machine Learning, Diagnosis, Accuracy

Introduction:

Polycystic ovarian syndrome (PCOS) is a common health issue in women caused by an imbalance of hormones related to reproductive system. This variance in hormones creates problems in the ovaries. The ovaries make the ova which released every month which is a sign of healthy menstrual cycle. In the condition of PCOS the egg either is not properly developed or it may not be released during ovulation.

The PCOS may lead to missed or irregular menstrual periods which further may result in infertility or formation of cysts (small fluid-filled sacs) within the ovaries (Stein and Leventhal., 1935).

Between 5% and 10% of girls and women between age of 15 to 44, or during the year where they became to have children, have PCOS. Most women have developed the symptoms of PCOS in their 20s and 30s, when they face problems on getting pregnant and consulted to doctor. But PCOS can happen at any age after puberty (Trivax et al., 2007).

Women of all races and ethnicities may acquire PCOS and the risk of PCOS goes higher in case of obese body or having family history of PCOS (Wendy M. Wolf et al., 2018). According World Health Organization (WHO) ,PCOS has affected 116 million women (3.4%) worldwide in 2012. Globally, prevalence estimates of PCOS are highly variable, starting from 2.2% to as high as 26%.

The prediction of PCOS is not a trivial one, because there are numerous symptoms associated with PCOS. The symptoms includes imbalance of hormones, hair loss , weight gain , irregular menstrual cycle(Stein and Leventhal.,1935) . Also it is not essential that all the symptoms will be appeared in case of PCOS. Very few women with PCOS exhibit the same set of symptoms, and the symptoms may vary at different stages of life. There exists no single test to diagnose PCOS. Therefore for diagnostic purpose, doctors are dependent on symptoms, blood tests, a physical examination; and some-times a pelvic ultrasound (Rotterdam Criteria., 2003). Though it sounds simple, but adding up various evidences might be time consuming, and sometimes a bit frustrating too. Major health organizations round the world disagree about how best to verify if an individual actually has polycystic ovary syndrome (Rotterdam Criteria., 2003).

Machine learning is an approach which might help in this regard. Machine learning came up with Technique for developing sophisticated, automatic, and objective algorithms for analysis of high-dimensional and multimodal biomedical data. It improves diagnosis and therapeutic monitoring of disease. Diagnosing the diseases using Machine learning allows healthcare experts to interchange the adapted care known as precision medicine (Paul Sajda et al., 2006).

In present study supervised machine learning approach has been used. In the Supervised machine learning, machine is trained with well labeled data and some of the data is already tagged with correct answer , in this the input variables consider as (x) and output variable as (Y) and different algorithms like Decision tree , Naïve Bayes etc, are used to learn the mapping function (depicted as f), from the input towards the output Y = f(X), The target is to approximate the mapping function (f) so well that when you have new input data (x) that you can predict the output variables (Y) for that data. This method named as Supervised learning since the correct answers for the training data are known, the algorithm iteratively makes predictions on the basis of training data . As the acceptable level of performance is reached learning will stop for result prediction of test data or new data (C.E. Brodley et al., 1999).

Along with supervised machine learning (Classification) model, in present study we used genetic algorithm for selection of attributes which play major role in disease and disease prediction (PCOS).Genetic Algorithm (GA) is a kind of search-based optimization technique that supports the principles of genetics and survival of the fittest. It is frequently accustomed to find optimal or near-optimal solutions to difficult problems which otherwise would take a lifelong to unravel. It is frequently used for solving of optimization problems, in research, and in machine learning(L. Haldurai et al., 2016).

Material and methods:

a. Data Retrieval:

There is enormous amount of data available in this digital world and its keep surge and we see a spurt on the healthcare data as the availability of EMR (electronic medical records) increases day by day, in present work we collected the data from well known and trusted website named Kaggle.com. This data is gathered from 10 different hospital across the Kerala, India, and uploaded by the user Prasoon Kottarathil (data scientist at Muriyad, Kerala, India)

In this dataset total no. of 541 patient data is being used. Some of the patients are diagnosed with PCOS while other did not. The PCOS diagnosis is carried out on the basis of large number of tests like medical test and physical examination. In present study total 40 attributes have been taken, some of them are categorical and some

numeric all are summarized in table no. 1.

Table 1: A complete list of input Attributes with the target attributes, input attributes are distinguish between categorical and Numerical

Target Attribute:
Diagnosis (women having a PCOS and women who don't have)
Key Attribute:
Patient File Number
Input Attributes:

Categorical: (1) Pregnant(Y/N) , (2) Weight gain(Y/N), (3) hair growth(Y/N), (4) Skin darkening (Y/N), (5) Hair loss(Y/N), (6) Blood Group, (7) Pimples(Y/N), (8) Fast food (Y/N) , (9) Reg. Exercise(Y/N)

Numeric: (1). Age (yrs), (2). Weight (Kg), (3). Height(Cm) (4) BMI, (5) Blood Group, (6) Pulse rate (bpm) (7) RR (breaths/min), (8)Hb(g/dl), (9) Cycle(R/I) (10) Cycle length(days), (11) Marriage Status (Yrs), (12) No. of abortion, (13) FSH(mIU/mL), (14) LH (mIU/ml), (15) FSH/LH (16) Hip(inch), (17) Waist(inch),(18) Waist:HiP(Ratio), (19) TSH (mIU/L) , (20) AMH (ng/ml), (21) PRL (ng/ml), (22)Vit D3 (ng/ml) , (23) PRG(ng/ml), (24) RBS(mg/dl), (25)BP _Systolic (mmHg) ,(26)BP _Diastolic (mmHg),(27)Follicle No. (L) (28)Follicle No. (R), (29)Avg. F size (L) (mm), (30)Avg. F size (R) (mm), (31) Endometrium (mm)

b. Dataset analysis:
a series of analysis has been conducted for this dataset. The analysis could be further classified into 3 sections In the first section through supervised machine learning model the classification has been performed. Three different classifiers including Decision tree, Random forest and Extra tree classifier have been selected for accurate classification of patient who has a PCOS. the performance of these classifiers has been analyzed on the basis of various statistical parameter, in second section work on Genetic algorithm is done which is for the selection of attributes and in the third section we again perform the classification models with the selected Attributes from the genetic algorithm, and analyze its result as well, after that we compare its result with the result of section one.

Section 1:
* Classifier and its evaluation (with the help of PyCaret):
Classification is supervised learning method where the algorithm is designed to learn by training data and apply this knowledge to test dataset. Classification methods are largely used in machine learning ,pattern recognition and Artificial intelligence Classification method have vast usage including Risk Analysis , credit card fraud detection targeted marketing , manufacturing and medical Diagnosis . In this section the task is done with the help of PyCaret . PyCaret is an open source, low-code machine learning library in Python that enables us to build machine learning models with minimal time and uses less line of code. In our dataset we work on three classifiers which are Decision tree, Random forest and Extra tree classifier, with the default parameter of PyCaret.

Decision tree:
Decision tree algorithm belongs to supervised learning algorithm this algorithm tries to solve the problem by using Tree representation .Each internal node of the tree corresponds to an attribute, and every leaf node corresponds to a category label. If the dataset has a N number of attributes then to decide which attribute to goes at base or in root or at different levels of the tree as internal nodes could be a complicate. for attribute selection researchers suggested using some criteria including Entropy, Information gain, Gini index, Gain Ratio, Reduction in Variance and Chi-Square These criterions will be calculate for every attribute. Then calculated values are sorted, and attributes are placed within the tree by following the score i.e. the attribute with a high value (in case of information gain) is preferred for the root and other for the leaf on basis of respective values.

Random Forest Classifier:
It is a type of ensemble tree-based learning algorithm. As the name depicted the Random Forest Classifier is a group of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to make a decision the ultimate class of the test object.

Random forest models reduce the problem of overfitting by introducing randomness by:

*building multiple trees (n estimators)
*drawing observations with replacement (i.e., a bootstrapped sample)
*a random subset of the features are selected at every node within splitting nodes on the best split.

Extra trees:
An "extra trees" classifier, otherwise known as an "Extremely randomized trees" classifier, is a other form of random forest. Unlike a random forest, at each step the entire sample is used and decision boundaries are picked at random, this is also a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result.

*Builds multiple trees with bootstrap = False by default, is telling it to sample observations without replacement
*nodes are split on basis of random splits among a random subset of the features selected at every node

Table 2: Differences between DT(decision tree), RF(random forest)and ET(extra tree) classifier on basic algorithmic features.

	Decision Tree	Random Forest	Extra Trees
Number of trees	1	Many	Many
Number of feature selected for split Et each decision node	All Features	Random subset of Features	Random subset of Features
Bootstrapping(Drawing Sampling without replacement)	Not applied	Yes	No
How split is made	Best Split	Best Split	Random Split

Analyzation of classifier performance:
Analyzing or testing of classifier performance is done with the help of various statistical method some of most common techniques are – confusion Matrix , AUC (area under the curve) - ROC , Precision and Recall value , F1 score , Kappa value with the help of these value we can select the best classifier for individual Dataset(S.Vijayarani et al., 2013).

Confusion matrix:
A Confusion matrix is basically N x N matrix and preferred for evaluating the performance of a classification model, where N represent the number of target classes. The matrix compares the prediction done by machine learning model with the actual target values. This gives us a holistic view for the performance of our classification model and what sorts of errors it is making. In the matrix there is value of True Positive, False Positive, True Negative and False Negative is given.

Recall

$$\text{Recall} = \frac{TP}{TP+FN}$$

Out of all the positive classes, what proportion we predicted correctly. It should be high as possible.

Precision:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Out of all the positive classes what we have predicted correctly, how many are actually positive.

Accuracy will be
Out of all the classes, what proportion we predicted correctly.

$$\text{Accuracy} = \frac{\text{Number of Object correctly Classified}}{\text{Total No. of object in the test set}}$$

F-Measure:

$$\text{F-Measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

It is not easy to comparing the two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score help us to measure Recall and Precision score at the same time.

Section 2:

Feature Subset Selection using Genetic algorithm (with the help of WEKA):

Feature Extraction can be defined as the process of detection and elimination of irrelevant, weakly relevant or redundant attributes or dimensions in a given data set. The feature selection process finds the minimal number of attributes in a manner to find probability distribution of data classes very close to original distribution obtained when all the attributes. Comparison can be consider as expensive operation . Generally, the computational cost of selected dataset set D is O(n x |D| x log(|D|)), where n – number of attributes, D – number of instances. The number of comparisons required for m attributes and n instances is m * n2. For the selected data set D, with n attributes, 2n subsets are possible. Search for an optimal subset would be highly expensive means it requires n number of operation to compute especially as the number of information classes increases. Due to the infeasibility of computationally expensive comparison operation selection of feature are done on the basis of heuristic methods. These heuristic methods try to explore possible reduced search space. Feature selection could be categorized into two techniques . First Technique is for Ranks the features and second one is feature subset selection. In the first technique, features are ranked on the basis of metric like information gain, chi-square etc. The features that fail to get the adequate score are eliminated. In the later one, the search is for optimal subset of features that could be similar to original subset of features. Evaluation of common features subset are done on basis of distance metrics (Euclidean, Hamming etc) or filter metrics (Entropy or Probabilistic distance) (Priyanka Khare et al.,2016). Greedy forward attribute selection, backward attribute selection, Genetic algorithms and simulated annealing are the approached that are commonly used.
Genetic Algorithm (GA): GA has been developed by John Holland, his students and colleagues at the University of Michigan, Genetic Algorithm supports the phenomenon of natural evolution . Initially genetic search had zero number of attributes, and an initial population with randomly generated rules. On the basis of phenomenon of evolution that is based on the theory of survival of the fittest, new population is generated in such a way that the fittest one rules the current population. Formation of new generation is done after cross over and mutation. The process of generation continues until it evolves a population P where every rule out P satisfies the fitness threshold, this is multistep process .

Steps Involved in Genetic Algorithm:

Initial Population

The process begins with a set of individuals which in our case every input attributes available in data and these individuals as a whole known as a Population. Each individual is a solution to the problem you want to solve (here algorithm assume that every attributes is responsible for PCOS).

Fitness Function

The fitness function determines how fit a specific one is (the strength of an individual to compete with other one).it means how well the each individuals determine the target attribute this fitness function provides a fitness score (responsible factor of each individual for developing PCOS in patient) for each individual, and this fitness score give us the idea which one has a capability to be selected, higher the score higher the probability will be.

Selection

The idea of selection phase is to pick the fittest individuals (attributes those are highly responsible for target (PCOS) attribute) and allow them to pass their genes to subsequent generation.

Two pairs preferred one (parents) are selected supported their fitness scores. Individuals with high fitness have more chance to be selected for further process..

Crossover

Crossover is one of most vital stage in a genetic algorithm. For each pair of oldsters to be mated, a crossover point is chosen randomly from within the genes (any random attributes from previous step).

Mutation

In certain new offspring formed, a number of their genes will be subjected to a mutation with lower random probability. This implies that a number of the bits within the bit string may be flipped. Which means one attributes in set of selected attributes is flipped we expect, in this case, that the fitness score of the offspring will be better than the ones of both its parents. In this stage we have a new population, a new generation, which will continue to go through natural selection, so we will go back to Step 2, natural selection, assign new scores to all the individuals of the new generation, and then continue the cycle, and create another generation and another generation, and so on.

We can rightly assume that, over time, over generations, the individuals in the population will improve, as the beneficial crossovers and mutations, together with the biased selection mechanism, will tend to strike roots and establish the population. What's so nice about computers is that we can set this process loose, and create an evolution of thousands and even millions of generation quite easily.

But every algorithm needs to stop and produce the required output which is final step.

Termination

The termination point has reached if the population has converged (does not produce offspring which are significantly different from the previous generation). that means we get final result, the genetic algorithm has come up with a set of solutions for our problem. Here the algorithm completed its task and give us a precise list of best attributes (set of attributes which are highly related to target (PCOS) attribute).

The application of genetic Algorithm search for the selection of Attributes in entire dataset, we took the help of well Known machine learning software called WEKA(Waikato Environment for Knowledge Analysis, developed at the University of Waikato, New Zealand) we work with its Default parameters :

CrossoverProb = 0.6
MaxGenerations = 20
MutationProb = 0.033
PopulationSize = 20

Section 3:

In this section we again perform the complete work as similar to the section one but this time the attributes are less (selected by Genetic algorithm) which is the result of previous section.

And analyzation of result for both section (1,3) is done, and confirm whether the performance of each selected classifier is decrease as no. of attributes are reduced or it remain same or it get increased .

Final results and Discussion:

Result of section 1:

PyCaret has automatically split the data into training set(70%) and test set (30%) and uses K fold(K = 10)cross validation ,Performance of selected classifier (Decision tree, Random forest, and Extra trees) in the different statistical parameter is given on table no. 3.

Table3: Rows of table represent the selected classifiers where the columns are statistical parameter here the accuracy of each classifier given in percentile and the rest are given the form of ratio (0-1):

S.N.	Classifier	Accuracy	AUC	Recall	Prec.	F1
1	Extra Trees	83	0.92	0.64	0.84	0.72
2	Random Forest	81	0.87	0.52	0.85	0.64
3	Decision Tree	77	0.74	0.63	0.68	0.65

Result of section 2.

After the performing of multistep process Genetic algorithm has select the nine attributes given on table no.4.

Table 4 : A complete list of input Attributes selected by GA(genetic

algorithm) with the target attributes input attributes are distinguish between categorical and Numerical attributes

Target Attribute:
Diagnosis (women having a PCOS and women who don't have)
Key Attribute:
Patient File Number
Reduced Input Attributes:

Categorical: (1) Weight gain(Y/N) , (2) hair growth(Y/N)
(3) Skin darkening (Y/N), (4) Fast food (Y/N)

Numeric: (1) Cycle length(days), (2)FSH/LH (Ratio)
(3) AMH(ng/mL),(4) Follicle No. (R) (5) Avg. F size (L) (mm)

Result of section 3:
Performance of selected classifier (Decision tree, Random forest, and Extra trees) with the selected attributes in the different statistical parameter is given on table no 5.

Table5: Rows of table represent the selected classifiers where the columns are statistical parameter here the accuracy of each classifier given in percentile and the rest are given the form of ratio (0-1):

S.N.	Classifier	Accuracy	AUC	Recall	Prec.	F1
1	Extra Trees	88	0.96	0.73	0.88	0.80
2	Random Forest	85	0.95	0.67	0.83	0.75
3	Decision Tree	85	0.88	0.69	0.82	0.75

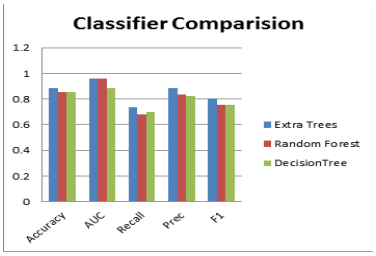

Figure 1
Histogram of all selected classifier performance in all selected parameters:

Confusion matrix for all the selected classifier:

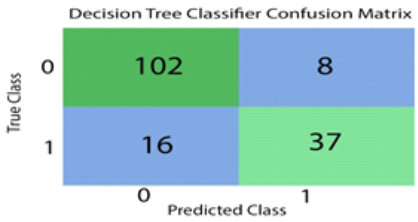

Figure 2.1
Confusion Matrix for classifier: Decision Tree, out of 163 correctly predicted instances 139 (102+37) and incorrectly 24(16+8).

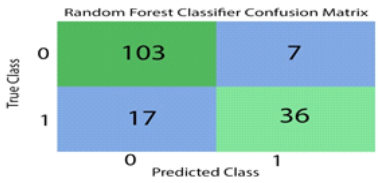

Figure 2.2
Confusion Matrix for classifier: Random Forest, out of 163 correctly predicted instances 139 (103+36) and incorrectly 24(17+7).

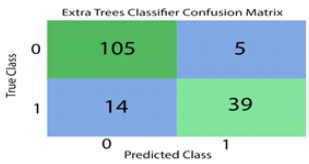

Figure 2.3
Confusion Matrix for classifier: Extra Trees out of 163 correctly predicted instances 144 (105+39) and incorrectly 19(14+5).

ROC (receiver operating characteristic) Curve for all Classifiers:

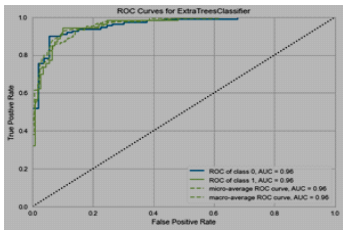

Figure 3.1
ROC(receiver operating characteristic curve) for Extra Tree Classifier.

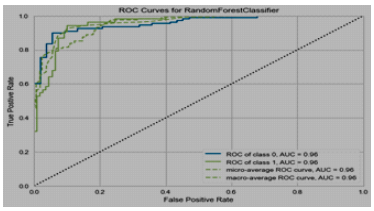

Figure 3.2
ROC(receiver operating characteristic curve) for Random Forest Classifier.

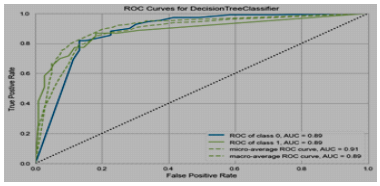

Figure 3.3
ROC(receiver operating characteristic curve) for Decision Tree Classifier:

Classification Report of Each Classifier:

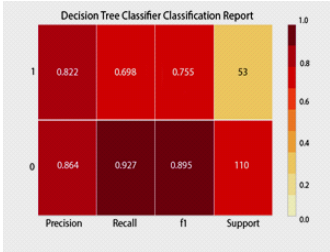

Figure 4.1
Classification report for the Decision Tree Classifier:

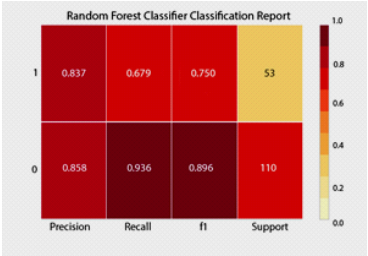

Figure 4.2
Classification Report for the Random Forest Classifier:

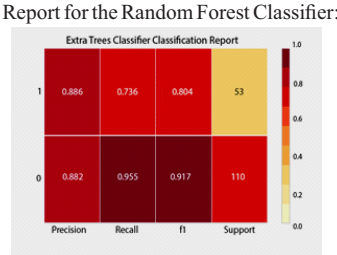

Figure 4.3 Classification Report for the Extra Trees Classifier:

Discussion:

Predictive accuracy has been widely used as the main criterion for comparing the predictive ability of classification systems, and AUC value is also on main focus which represent the degree or measure of separability it give us the idea about the capability of model in distinguishing between classes Higher the AUC better the model is (Karimollah Hajian-Tilaki,2013). After analyzing of all classifiers performance on the different Statistical parameter Extra Tree classifier has give the better result in compare to Decision tree and random forest , if we focus on the area under the ROC (Receiver Operating Characteristics) curve, or simply AUC, provides a better result in Extra Tree Classifier having the AUC value 0.92 and accuracy of 83% .

The section two the GA search With initial population of twenty instances generation continued until twentieth generation reached and having cross over probability of 0.6 and mutation probability of 0.033. For this Dataset the genetic search resulted in Nine attributes out of forty attributes, as given in table no 4 .these having higher fitness score or impact on result attribute (PCOS). To check the reliability of this result we again work on selected classifier as discussed in section 3 , and after analyzing the results of section 3 we found that not only the performance of classifiers maintained as earlier but it increases after the selection of Attributes if we focused on the result of Extra Tree classifier AUC value it get increased from 0.92 to 0.96 (figure no. 3.1) accuracy gets increased by 5% (83-88) and similarly the performance of other classifier (as on table number 3 and 5) also improved in selected statistical parameters which means after the removal of redundant Attributes the classifiers perform better and we get better accuracy . PCOS is counted as multifactorial disorder which is caused by varies abnormalities such as genetic, endocrine and environmental (V. De Leo et al., 2016). In the present study we have selected the attributes which related to physical abnormality and the major Hormonal imbalance in respect to PCOS but it doesn't include any genetic aspect of disorder which would also unearthed as the major attributes with the help of similar technique and support the study further.

Conclusion:

PCOS has a very high number of individuals who remain as undiagnosed when visiting to clinician, estimated to be as high as 75%(Wendy M. Wolf et al., 2018). This is likely because of variability of patient presentation and lack of provider knowledge. The benefit of pointing out more of these patients would be linkage to care, increased screening for comorbidities, and overall improvement in patient care. Giving a patient the diagnosis of PCOS makes the patient aware of possible fertility concerns, dysfunctional bleeding, endometrial cancer, obesity, diabetes, dyslipidemia, hypertension, and theoretical increased risk of cardiovascular disease(T.M.Barber et al.,2006) . Since PCOS could be genetic, it's going to bring awareness to members of the family and future children. Hence we worked on this segment and try to get some insight through the Machine Learning Approach.

From our Analysis with the help of Genetic Algorithm we are able to find the some major Attributes which would concluded as the major marker for diagnosis of patient , and this would provide the help for Medical Professional To Diagnose the disease with accuracy and minimal time , and further helpful and cost effective for the patient who went through large number of test for Diagnosis of disease , working on different classifier before and after applying the Genetic Algorithm and analyze the Statistical performance of each , we concluded that after the removing of redundant Attributes classifiers performance increases as Extra tree classifier from 83% Accuracy to 88% Accuracy , along with that if we compare the performance of each Classifier in our dataset we have concluded that the Extra tree classifier has give us the better result if we look at its confusion matrix , AUC (area under the curve) value, precision value and F1 score, this classifier give us 88% Accuracy as compare to Decision Tree and Random Forest classifier which having nearly 85% Accuracy. There is need of further study where the genetic attributes are also present with large number of dataset which gives more insight about PCOS Diagnosis Attributes selection and help the community larger.

References:

1. Stein I, Leventhal M. Amenorrhea associated with bilateral polycystic ovaries. Am J Obstet Gynecol 1935;29:181–191.
2. Trivax B, Azziz R. Diagnosis of polycystic ovary syndrome. Clin Obstet Gynecol 2007;50:168 –77.
3. Wendy M. Wolf, Rachel A. Wattick, Geographical Prevalence of Polycystic Ovary Syndrome as Determined by Region and Race/Ethnicity. Int J Environ Res Public Health. 2018 Nov; 15(11): 2589.
4. Rotterdam ESHRE/ASRM-Sponsored PCOS consensus workshop group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS). Hum Reprod 2004b;19:41–47.
5. P Sajda, Machine learning for detection and diagnosis of disease. Annu. Rev. Biomed. Eng., 2006;8: 537-565.
6. Brodley, C.E. and Friedl, M.A. Identifying Mislabeled Training Data. JAIR 1999; 11: 131-167.
7. L. Haldurai, T. Madhubala and R. Rajalakshmi, "A Study on Genetic Algorithm and its Applications", International Journal of Computer Sciences and Engineering Vol.-4(10), Oct 2016, E-ISSN: 2347-2693
8. Karimollah Hajian-Tilaki, Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med. 2013 Spring; 4(2): 627–635.
9. S. Vijayarani, S. Sudha, "Comparative Analysis of Classification Function Techniques for Heart Disease",International Journal of Innovative Research in Computer and Communication Engineering, vol. 1, Issue 3 , May 2013 .
10. R.Ade, Dhanashree, S. Medhekar, Mayur P. Bote, "Prediction using SVM and Naïve bayes", International Journal of Engineering Sciences and Research Technology, May 2013.
11. M.Anbarasi, N.CH.S.N.Iyengar, "Enhanced Prediction of Heart Disease With Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology, Vol. 2(10), 2010.
12. D. Goldberg," Genetic Algorithms in Search, Optimization, and Machine learning", Addison Wesley, 1989.
13. .Priyanka khare and Dr. Kavita Burse " Feature Selection Using Genetic Algorithm and Classification using Weka for Ovarian Cancer" / (IJCSIT) International Journal of Computer Science and Information Technologies,(1) , 2016, 194-196
11. Ahmed and Zeeshan, "Applying WEKA towards Machine Learning With Genetic Algorithm and Back-propagation Neural Networks" J Data Mining Genomics Proteomics 2014, 5:2 DOI: 10.4172/2153-0602.1000157
14. Pycaret.org/tutorial Python library for machine learning.
15. Dr. K. Meena and Dr. M. Manimekalai2 , S. Rethinavalli3 "A Novel Framework for Filtering the PCOS Attributes using Data Mining Techniques"
16. T. M. Barber, M. I. McCarthy, J. A. H. Wass and S. Franks, "Obesityand polycystic ovary syndrome", Clinical Endocrinology (2006) 65,137–145
17. V. De Leo, M. C. Musacchio, V. Cappelli, M. G. Massaro, G. Morgante & F. Petraglia, "Genetic, hormonal and metabolic aspects of PCOS: an update" Reproductive Biology and Endocrinology volume 14, Article number: 38 (2016).