International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Predicting the Presence of Poly Cystic Ovarian Syndrome using Classification Techniques

P.Gokila Bindha, R.R.Rajalaxmi, S.Poorani

Abstract— PCOS is an endocrine disorder which occurs due to hormone imbalance. PCOS may leads to infertility, diabetes mellitus and cardiovascular diseases. It may be identified by physical appearance, ultrasound scanning and clinical trials. The PCOS ovary can be identified as the follicles which are arranged peripherally and measuring 2-9mm of size. The dataset used in this paper consists of 119 samples with 17 features which represents the physical appearance and psychological characteristics such as stress, exercising methods, eating habits, etc. The classification algorithms can be applied on these data to predict the present of PCOS. The aim of the paper is to compare the accuracy of the classification model and find the algorithm which best suites for the dataset in predicting the occurrence of PCOS

Keywords: Accuracy, Classification, Metrics, PCOS

I. INTRODUCTION

The reproductive system of female comprises the uterus, ovaries and fallopian tubes. It carries out functions like production and transportation of gametes and sex hormones. It is compulsory to understand the women's reproductive system before studying PCOS because it is related to ovary[1]. The endocrinological disorder that pretends women is Poly Cystic Ovary Syndrome(PCOS). PCOS may leads to fertility related complications and a study says "PCOS is the main source for 75% of anovulatory infertility in women"[2].Women with PCOS has to undergo the medical treatment because it may leads to infertility or irregular menstrual cycles. PCOS has to be predicted before going to the medical treatments. Other than the clinical trials the psychological behaviors also have some impact on PCOS. Now a days the change in lifestyle and food habits make many hormonal changes in the adolescence girls and women. Other than the fertility related problem it also relates to the type2 diabetes, cardiovascular disease, endometrial cancer, anxiety and depression. This leads to the issues related to puppetry, ovulation and fertilization. The data mining technique can be applied in analyzing and predicting the presence PCOS. There are many classification algorithm available in data mining techniques. Some algorithms may be suited for some dataset while some may not. Based on the dataset and the accuracy of the result when applying the algorithm, the algorithm can be selected for the problem. This paper focuses on applying six different classification algorithms on the dataset and based on the accuracy it selects the Decision Tree classification algorithm suites best for this dataset. The paper runs in the following way: The Section 2 represents the data set and the Python language. The coding is developed using Python. In Section 3, the six data mining algorithms are described. The section 4 deals with experimental results and the metrics used to evaluate the algorithms. Conclusion, limitation of the paper and the future work is stated in Section 5.

Manuscript published on November 30, 2019.
* Correspondence Author
P.Gokila Bindha*, Department of Computer Technology - UG, Kongu Engineering College, Perunduai, Tamil Nadu, India. (Email: brindha@kongu.ac.in)
R.R.Rajalaxmi, Department of Computer Science and Engineering, Kongu Engineering College, Perunduai, Tamil Nadu, India. (Email: rrr@kongu.ac.in)
S.Poorani, Department of Computer Technology - UG, Kongu Engineering College, Perunduai, Tamil Nadu, India. (Email: brindha@kongu.ac.in)

II DATASET

The dataset [3] has been generated by taking the survey among 119 girls whose age ranges from 18 -22 years. The database consists of features related to regularity of menstrual cycles, weight gaining, excess growth in face or body, patches in skin, pimples, face depression and anxiety, history of diabetics and hyper tension, difficulty in maintaining the body weight, oily skin, hair loss, place where frequently eating, exercise regularly, newly admitted to hostel, personal problems, peer pressure, change in dietary habits and intake of fast food. Out of these attributes 14 are binary attributes and 3 are categorical attributes. The disease is the class variable which has two classes 'may be' and 'may be not'. The class variable is used to represent the presence of PCOS(maybe) or not(may be not).

Python is a powerful object Oriented language. It is one of the easiest language to execute the data mining algorithms. The pandas module is used to load the dataset and to store them as a dataframe. Since pandas has been built on the Numpy module it also helps to work with the n dimensional arrays. The SKlearn module having the methods related to different data mining classification algorithms. This module also provides the various methods which is used to measure the accuracy of the algorithm.

III MATERIALS AND METHODS

The preprocessing of the data is done so as to prepare the data for the classification algorithms. The coding is done using Jupyter Notebook.

The modules such as pandas, sklearn and classes needed for analysis is imported. The dataset in the excel format is loaded using pandas . The dataset is then converted into an array. The array is sliced such that it separates the feature variables(X) and the class variable(Y). Now, The dataset is divided into the training and testing set using the train_test_split() to which the X and Y
are passed as the arguments. The sklearn module provides the following methods for different classification algorithms.

TABLE I : Classification Algorithms In Sklearn Module

Method Name	Classification Algorithm
LogisticRegression()	Logistic Regression
SVC(kernel='linear')	Support Vector Machine
DecisionTreeClassifier()	Decision Tree
KNeighborsClassifier(n_neighbors=3)	K Nearest Neighbor
LinearDiscriminantAnalysis()	Linear Discriminant Analysis
GaussianNB()	Naive Bayes

Logistic Regression:

In statistical analysis Logistic Regression uses the logistic function to model the binary dependant variables. The main advance of using this method is it can be applied for the dense and sparse input. An example for the logistic regression equation can be given as

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)}) \quad (1)$$

where y represents the predicted class (0 or1), b0 represents the bias and b1 represents the coefficient of the input variable X.

Support Vector Machine:

Given a labeled database the SVM is used to identify the hyper plane which is used to divide each class. Suppose if the dataset consists of two classes the hyper plane divides the space into two parts having only one class in each part. There may be many hyper planes separating the two classes but the hyper plane whose margin is far away from the points (at the margin side)of both the classes is selected. The more distance from the points more accuracy in classification.

Decision Tree:

The leaf of the decision tree is considered as the class label, the root and intermediate nodes represent the features that may be used to make decision. All the features may not be needed to identify the class. As the decision tree follows the top down approach , at each step it is very important to select the feature as an intermediate node that best splits the given dataset. Gini Index, Gain Ratio and Information Gain metrics can be used to identify the splitting criterion . Prediction can be made by traversing the tree from the root and reaching a leaf node based on the decision made at each node.

K Nearest Neighbor:

The dataset may consists of n features, therefore each and every training tuple are stored in the n dimensional pattern space. When the testing tuple is given it searches the pattern space for the k nearest training tuples and these k nearest tuples are considered as the 'K nearest neighbors ' of the testing tuple. The class label of the testing tuple is assigned as the most common class among the k nearest neighbours.

Linear Discriminant Analysis:

Considering a set of features for each instances(training) with the known class, the classifier has to find the good predictor for the class of any given instance of the same distribution . LDA models the distribution of predictors separately in each of the response classes, and then it uses Bayes' theorem to estimate the probability The linear Discriminant analysis estimates the probability that a new set of inputs belongs to every class. The output class is the one that has the highest probability.

Naive Bayesian Classification:

Considering the dataset consisting of n attributes and associated with the class label, the class of the given tuple can be predicted by evaluating the posterior probability conditioned on that tuple. The class which is having the highest posterior probability that is conditioned the testing tuple is taken as the class label for that tuple. For example the the tuple X belongs to the class Ci, if the following condition gets satisfied.

$$P(Ci|X)>P(Cj|X) \text{ for } 1<=j<=m, j<>I \quad (2)$$

where, P(Ci|X) is the posterior hypothesis.

IV EXPERIMENTAL RESULT ANALYSIS

After implementing the algorithms, a model is generated for each algorithm by applying the training dataset. The SVM classifier methods can be applied with different kernals, here it has been taken as 'Linear' kernal. The number of neighbors considered is 3 and it has been passed as an argument to the K Nearest Neighbor method. Then the prediction of the class label for the testing data has been done. The accuracy of the model depends on the percentage of correct classification of the testing tuple. The accuracy can be measured using various metrics such as confusion matrix, true positive, true negative ,false positive, false negative, sensitivity, specificity and precision. The confusion matrix can be given as

TABLE I:CONFUSION MATRIX

Actual Class	Predicted Class	
	C1	***C2***
C1	TP	FN
C2	FP	TN

Where C1 and C2 are the class labels . TP represents the True Positive, the positive tuples are correctly labeled by the classifier as positive. TN represents the True Negative means the negative tuples that are correctly identified as the negative tuples by the classifier. FP is the False Positive which represents the negative tuples identified as positive by the classifier. FN is the False Negative means the positive tuples identified as negative by the classifier. In python the confusion matrix can be generated using confusion_matrix() by passing the class labels of the test tuples and the class labels predicted by the model(same set of tuples). Accuracy of the algorithm can be determined using Accuracy=(TP+TN)/(TP+TN+FP+FN). Classification Error can be calculated using Classification Error=(FP+FN)/(TP+TN+FP+FN).

The sensitivity is the true positive rate and the specificity is the true negative rate which can be given by Sensitivity=TP/P and Specificity=TN/N. Where P is the number of positive tuples and N is the number of negative tuples. Precision is the percentage of the tuples labeled as positive that are actually positive tuples. Percision=TP/FP+TP.

After applying the testing dataset on the models generated these metrics are used to measure the accuracy of the model . The following table shows the resulting values of these metrics.

Table II :Accuracy Table Of The Classifiers

Classification Algorithms	Confusion Matrix	Accuracy	Classification Error
Logistic Regression	$\begin{pmatrix} 24 & 0 \\ 6 & 0 \end{pmatrix}$	0.8	0.2
Support Vector Machine	$\begin{pmatrix} 24 & 0 \\ 6 & 0 \end{pmatrix}$	0.8	0.2
Decision Tree	$\begin{pmatrix} 24 & 0 \\ 1 & 5 \end{pmatrix}$	0.966	0.033
K Nearest Neighbors	$\begin{pmatrix} 24 & 0 \\ 2 & 4 \end{pmatrix}$	0.933	0.066
Linear Discriminant Analysis	$\begin{pmatrix} 23 & 1 \\ 6 & 0 \end{pmatrix}$	0.766	0.233
Naive Bayes	$\begin{pmatrix} 5 & 19 \\ 0 & 6 \end{pmatrix}$	0.366	0.633

From the Table II it is understood that the prediction has been exactly done the Logistic Regression and SVM ,but the learning of the records with/without PCOS is very low so the accuracy is only 80%.

TABLE III :Other Metrics Of The Classifiers

Classification Algorithms	Sensitivity	Specificity	False Positive Rate	Precision
Logistic Regression	0.0	1.0	0.0	0.64
SVM	1.0	1.0	0.0	0.64
Decision Tree	0.833	1.0	0.0	1.0
KNN	0.666	1.0	0.08	1.0
LDA	0.0	0.9583	0.0416	0.63
Naive Bayes	1.0	0.208	0.791	0.24

Table II shows the accuracy measures of the different classifiers . Amongst the all Decision Tree classifier has accuracy rate as 96.6% and K Nearest Neighbors has 93.3% accuracy. On the other hand Naive Bayes and Linear Discriminant Analysis has 76.6% and 36.6% of accuracy. The different metrics used for providing the support for accuracy of the classification algorithm is shown in Table III. From the above result the decision tree classifier and K Nearest Neighbors classifier can be used for future prediction of PCOS.

The limitation here is that the dataset contains only 119 observations which is a small dataset. The prediction will be more accurate when the size of the dataset is large.

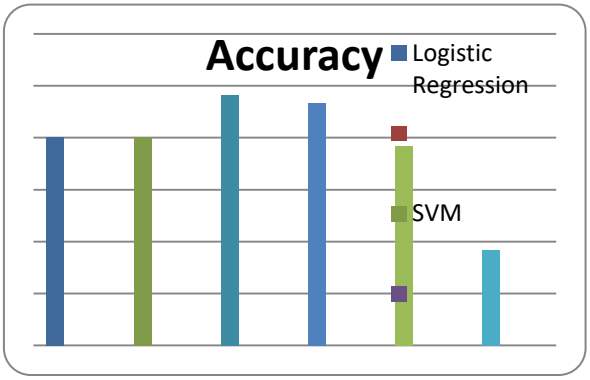

Fig. 1. Accuacy of the classifiers

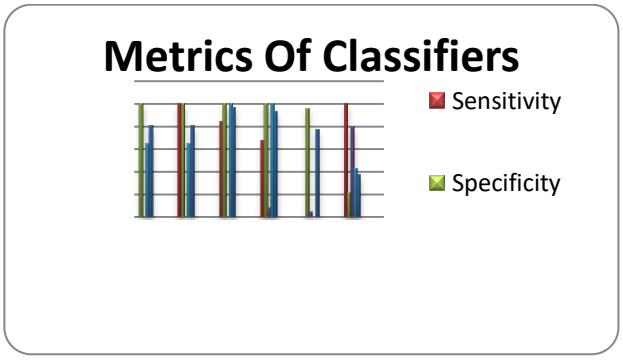

Fig. 2. Other Metrics of the classifiers

The Logistic Regression and SVM predict the absence of PCOS very accurately when compared with all other algorithms. As the splitting of data is 70:30 for training and testing, only few records having the presence of PCOS has been selected for testing so the accuracy of these two algorithms has been dropped down. At the same time when the number of neighbors of KNN is increased to 4 or more the accuracy falls down. The Naive Bayes identifies the presence of the PCOS to some extend but the fails in predicting the absence of PCOS.

V CONCLUSION AND FUTURE WORK

Thus prediction of PCOS can be done with the help of the Decision Tree classifier and K nearest classifiers, as it has more accuracy when compared to all other classifiers. Another limitation on the dataset is class labels are not equally distributed. In future the size of the dataset can be increased with the equal distribution of the classes and the different algorithms can be applied. The PCOS can be best identified with the ultra sound scan where more number of follicles are found in the ovaries. The data mining algorithms can also be applied on those scanned images to predict the presence of PCOS..

REFERENCES

1. S.Sheela, M.Sumathi,"Study and Theoretical Investigations on PCOS", IEEE International Conference on Computational Intelligence and Computing Reseach,2014
2. B.Vikas, B.S. Anuhya, K.SAnthosh Bhargav, Sipra
3. Sarangi, and Manaswini Chilla," Application of Apriori Algorithm for Prediction of Polycystic Ovarian Syndrome (PCOS)", Information Systems and Design and Intelligent Applications, pp 934-944, First Online :March 2018.

4. https://github.com/PCOS-Survey/PCOSData
5. https://www.shadygrovefertility.com/blog/diagnosing-infertility/pcos-one-size-doesnt-fit-all
6. Roy Homburg, "Pregnancy complications in PCOS", Best Practice & Reseach Clinical Endocrinology & Meabolism, Vol.20 No. 2., pp.281-292,2006
7. Vikas B. ,B.S.Anuhya , Manaswini Chilla and Sipra Sarangi "A Critical Study of Polycystic Ovarian Syndrome (PCOS) Classification Techniques", IJCEM International Journal of Computational Engineering & Management, Vol. 21Issue4,July2018
8. Ming-Yang Chang,Chien-Chou Shih,Ding-An Chiang and Chun-Chi Chen," Mining a Small Medical Data Set by Integrating the Decision Tree and t-Test",Jounal of Software,Vol.6 No.12,December 2011

AUTHORS PROFILE

Ms. P.Gokial Brindha completed M.Sc (Information Technology) from Anna University in the year 2007. She has 11 years of teaching expeiance. She published 2 research papers in International journals and presented 2 papers in the conferences. Presented seminars on the “R Tool”, ”Tableau – Visualization Tool” and “Data Analytics using R”. She is working as an Assistant Professor in the department of Computer Technology-UG, Kongu Engineering College affiliated to Anna University, Chennai,Tamilnadu. Perusing PhD under Anna University in the area of Data Mining and Machine Learning.

Dr. R.R.Rajalaxmi completed M.E. Computer Science and Engineering under Bharathiar University in the year 2000. She completed her PhD in Information and Communication Engineering under Anna University in the year 201. Currently she is working as the professor and Head of the Department, Department of Computer Science Engineering, Kongu Engineering College, Tamil Nadu. She is a life member of CSI and ISTE. Her area of interest is Data Mining, Data Analytics and Machine Learning. She guided 13 UG and 8 PG projects. She organized 9 sponsored seminars/workshops/training prorammes. She published 11 research papers in the international journals and 6 papers in the international conferences. She also completed 3 research projects sponsored by various funding agencies.

Ms. S. Poorani received M.Sc degree at Sri Vasavi College, Bharathiar University from the Department of Computer Science, India, in 2004. She has 12 years of teaching experience. She published 4 articles in International journals. She has also presented papers in National Conferences. Her area of interest includes data mining and big data. She is currently pursuing the Ph.D. degree working with Dr. P. Balasubramanie. Simultaneously she is also currently working as an Assistant Professor in the department of Computer Technology, Kongu Engineering College, affiliated to Anna University Tamilnadu, India.