Detecting PCOS using Machine Learning

Namrata Tanwani

[1](*Department Depart of Computer Engineering, G. H. Raisoni College of Engineering and Management, Pune, India,tanwaninamrata11@gmail.com)*

Abstract*—*Polycystic Ovary Syndrome or PCOS is an endocrine disorder that occurs in women of reproductive age. The condition once detected cannot be cured but treatment can help relieve its affects. The exact cause of PCOS is still unknown but there are certain factors that portray the risk of getting PCOS. The factors that result in this syndrome are: obesity, insulin resistance, blood pressure, depression, inflammation. While the symptoms include: hirsutism, Oligo-ovulation, acne, heavy bleeding, skin darkening. Using the causes and symptoms, a model is prepared in order to accept them as features and outputs the presence or absence of this condition. The machine learning models used for supervised classification are K-Nearest Neighbor and Logistic Regression .The reason behind to build multiple models is to find out the best one for the given dataset, in the known scope of knowledge

Keywords—component; formatting; style; styling; insert (key words)

1. INTRODUCTION

Polycystic Ovary Syndrome is a disease that occurs in women during their child-bearing years[1]. The reproductive organs of women called ovaries that produce progesterone and estrogen-hormones that regulate the menstrual cycle, are affected. Ovaries also produce little amounts of androgen also called as male hormones. The basic features of PCOS are:

- Cysts in ovaries.
- High levels of hormone: androgen.
- Irregular Periods
- Excessive body hair growth

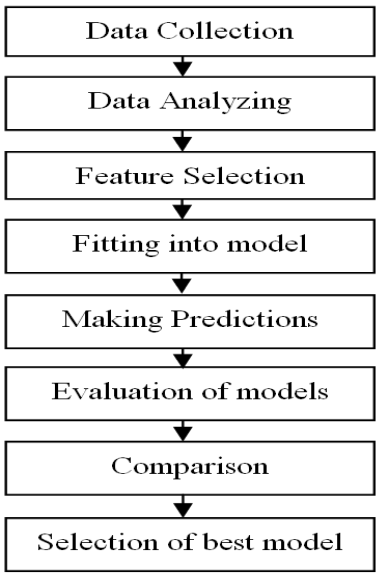
Since the condition is a syndrome, it has a collection of symptoms that suggest its presence. These symptoms play a vital role in detection of this condition. Along with these symptoms, causes which may lead to the risks of the syndrome can also be considered. It is necessary to detect PCOS as early as possible since it holds the risk of infertility, diabetes, endometrial cancer and cardiovascular disease at later stage of the condition[2].

Here, a couple of machine learning models are built in order to determine the presence of PCOS. Since the dataset does classify if the condition is present or not, supervised machine learning algorithms are used called: K-Nearest Neighbor (K-NN) and Logistic Regression. The former is a distance based technique and the latter is probability based, this is why these two techniques, that are poles apart, are used and their accuracy is compared[3]. Logistic Regression is more accurate with accuracy 92% while K-NN's is 90.74%.

The organization of paper is as follows: section 2 describes the methodology; section 3 and 4 discuss the results obtained by the method and the conclusions made respectively along with section 5 with references.

2. Methodology

In order to build an appropriate machine learning model, the data belonging a dataset needs to go through a sequence of steps. It is necessary, as to become a filtered and noise-less input to the algorithm. The results of the algorithms are compared and evaluated,



leading to the motive, that is, detecting PCOS using machine learning model with the highest accuracy. Following is the block diagram of the Methodology:

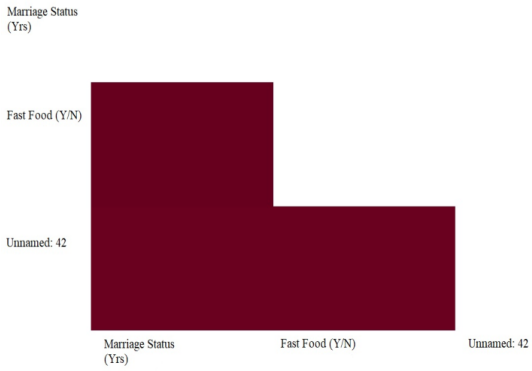
Figure 1: Block diagram of the methodology.

A. *Data Collection*

The first and crucial step is to collect data. Various platforms are available for this purpose. It contains samples from ten different hospitals across Kerala, India (downloaded from Kaggle). The identities of the patients have not been disclosed.

B. Data Analyzing

To move further, it is required to first understand what the dataset holds. Getting to know about samples, their attributes, and inconsistencies such as: negative



values, empty records, unwanted strings, is performed in this step.

Figure 1.2: Heat map to localize missing values.

The shape of the dataset is (540, 43) which means that it contains 540 samples and 43 attributes. In figure 1.2 we can see that there are only few records that have missing values. Hence, these records can be dropped. In addition to that, the data type of every value in the dataset should be either float or integer. This is required so that data can be processed by the algorithms.

C. Feature Selection

In order to improve the performance of the model and reduce the computational cost, only selected attributes of the samples act as features. Filter method is used to find the weights of the features in order to determine which of them have high correlation with the target.

Table: 1 displays the features and their correlation to the target, arranged in descending order. The more is the weight of the feature the more is its influence on the target, independently.

Features	Weights
Follicle No. (R)	0.650608
Follicle No. (L)	0.601035
Skin darkening (Y/N)	0.479679
hair growth(Y/N)	0.464623
Weight gain(Y/N)	0.441753
Cycle(R/I)	0.399746
Fast food (Y/N)	0.380246
Pimples(Y/N)	0.28672
AMH(ng/mL)	0.260287
Weight (Kg)	0.206051
BMI	0.195577
Hair loss(Y/N)	0.175055
Hip(inch)	0.156196
Waist(inch)	0.155068
Avg. F size (L) (mm)	0.126586

Table 1 Feature Weighing

The features in Table 1 are the top fifteen, amongst the forty three features. From the above table, we can conclude that the parameter **Follicle No. (R)** and **Follicle No. (L)** have the highest weight, that determine the number of follicles in right and left ovaries respectively. Features like **Skin darkening**, **hair growth**, **Weight gain**, **Fast food**, **Pimples**, **Hairloss** contain the values 0 or 1, 0 denoting the absence of that particular feature and 1 denoting its presence. **AMH** or Anti-Müllerian hormone is used as an indicator of egg count. Its unit is nanograms per milliliter. Following, **BMI** refers to Body Mass Index which is the ratio of patient's weight to their height. Along with Hip and Waist sizes in inches, there is also Average Follicle Size of the left ovaries, that is measured in millimeters.

D. Fitting into Model

With the data cleaned and selected, it is now ready to be processed by the models. The two models used for supervised machine learning are K-NN and Logistic Regression.

i) K-Nearest Neighbor

KNN is an Instance-Based Learning that compares new instances with instances stored in memory at the time of training of dataset. A new instance is classified by measuring its distances with the instances retrieved from memory, defined in terms of standard Euclidean Geometry, that is, distance between points in n-dimensional space. [4].

The accuracy of this model depends upon two factors:

- The value of 'K'
- The number of selected features.

Figure 1.3 Accuracy vs Value of K graph

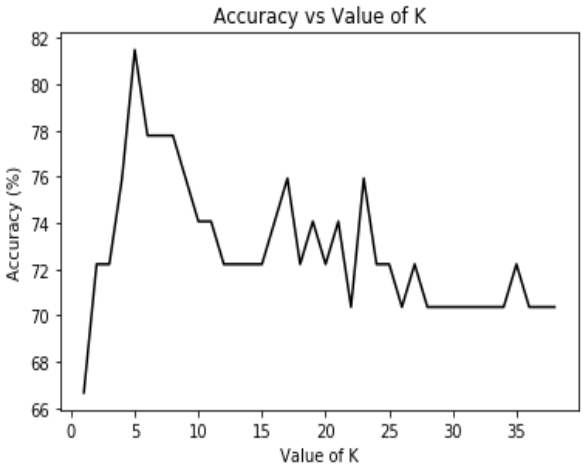
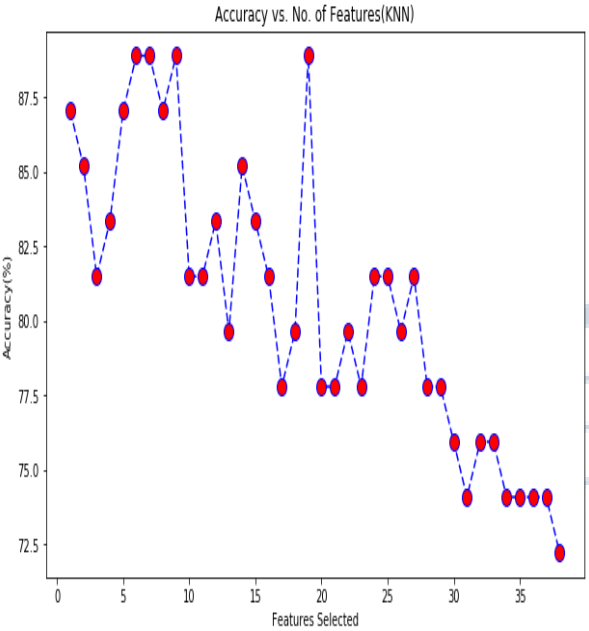


Figure 1.4 Accuracy vs No. of Features(KNN) graph

From figure 1.3, it is concluded that highest accuracy occurs when the value of K is 5. It means that when classification needs to be made, the numbers of neighbors whose votes are considered are the five closest ones.

From Figure 1.4, we have determined the numbers of features that need to be selected in order to give maximum accuracy. Therefore, the numbers of features included are 9.

Following image shows the nine features along with weights.



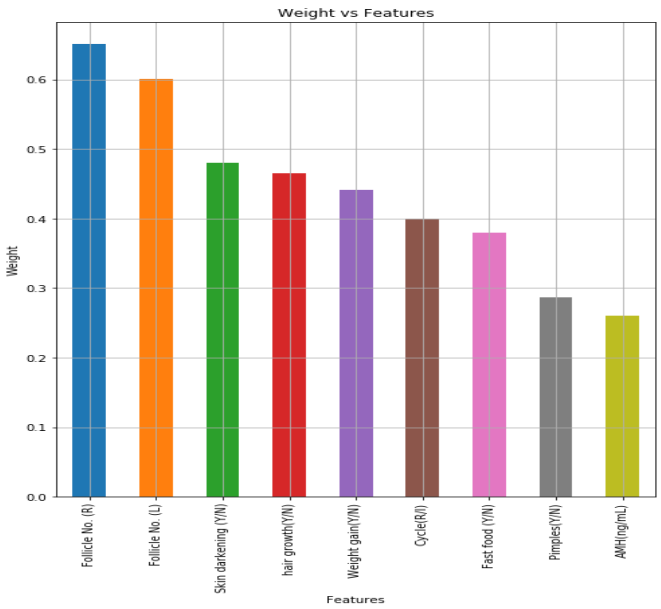
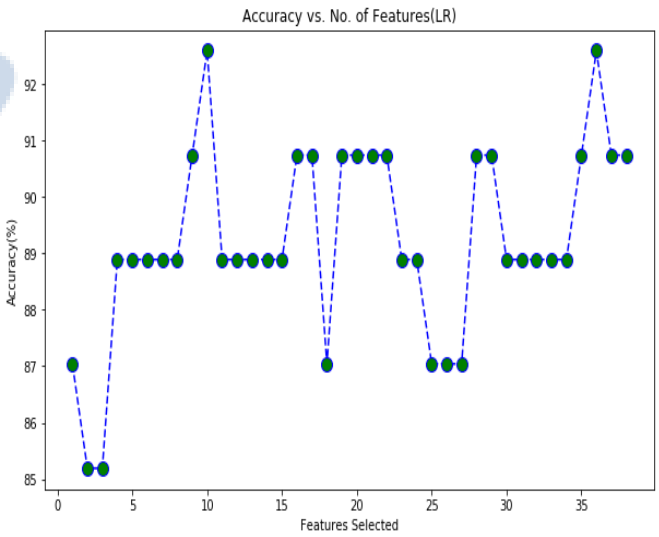


Figure 1.5: Selected features

ii) Logistic Regression



Logistic Regression is an extension of Simple Linear Regression that rounds off the result of input variables on the result variable as probability. Probability is what determines the relationship between the dependent and independent variables[5].

Here, The accuracy of Logistic Regression depends upon the number of features selected.

Figure 1.6 Accuracy vs. No.of Features(LR)

From figure 1.6, we can conclude that the highest accuracy occurs when the numbers of selected features are either 10 or 36. So, naturally, for the sake of computational cost, numbers of features that are selected are 10.

Following are the top 10 features:

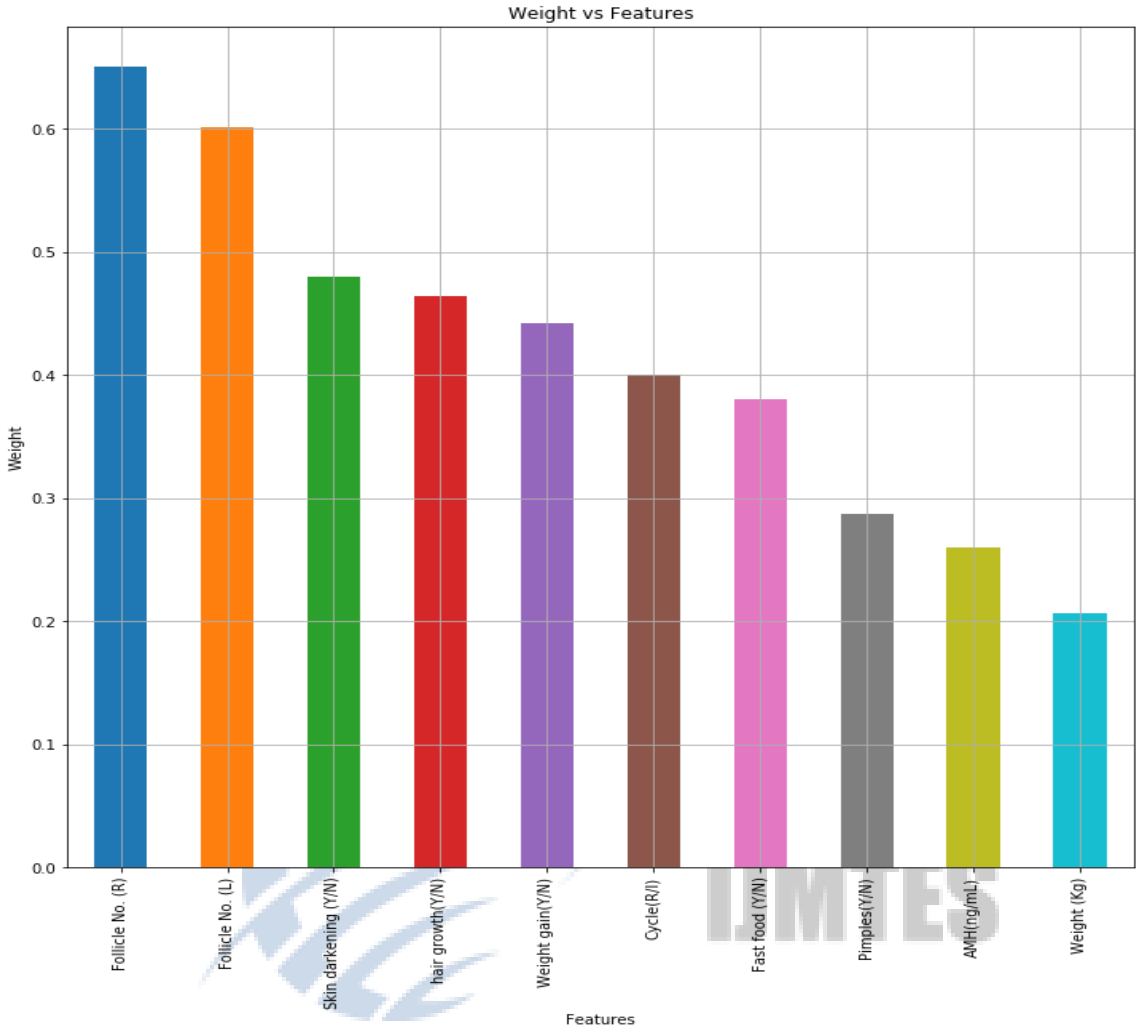


Figure 1.7 Weights vs Features

D) Making Predictions

Using the prepared models, predictions are made with testing set. In figure 1.8, Prediction using KNN model, and in figure 1.9, Prediction using Logistic Regression, the first scatter plot shows the sample numbers and actual output values,0: absence of PCOS or 1: presence of PCOS while the second scatter plot shows the sample numbers and their predicted output values.

In figure 1.8, only five predicted values differ from actual values.

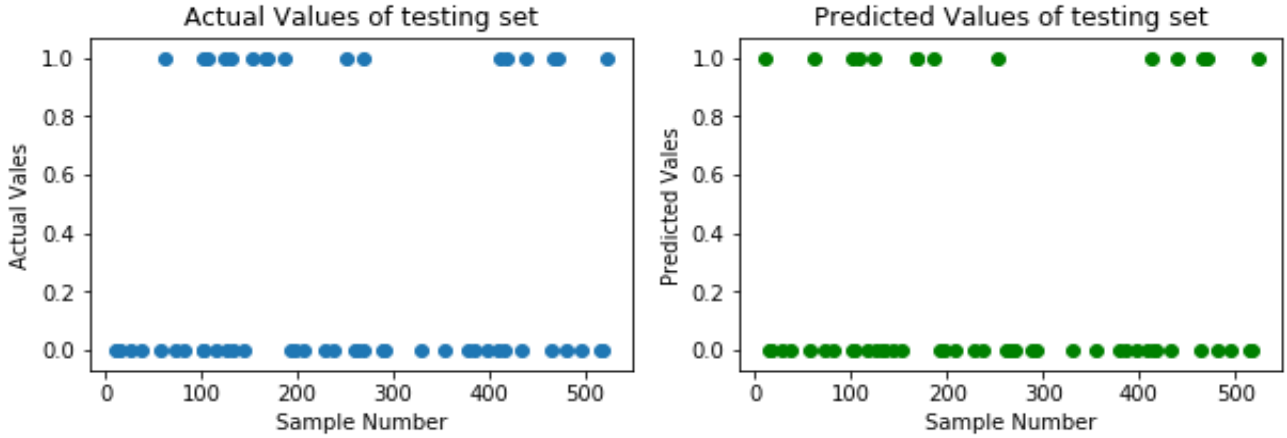
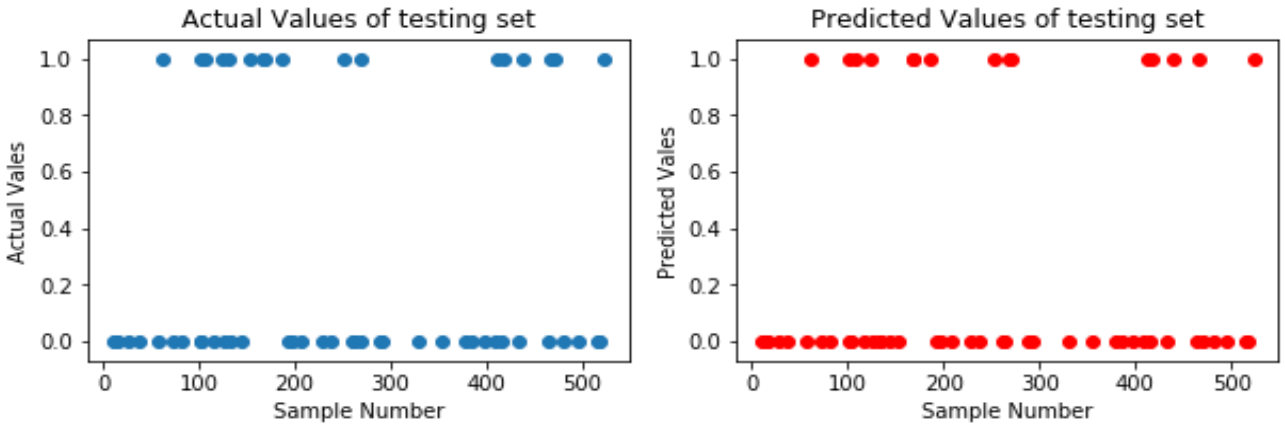


Figure 1.8: Prediction using KNN

In figure 1.9, only four predicted values differ from



actual values.

Figure 1.9: Prediction using Logistic Regression

E) Evaluation

The classification report includes: Precision, Recall and Support but before going through the classification reports of each model we need to understand the following terms:

- True Positive (TP)

The positives that are identified as positives. In other words, the number of patients with PCOS that are tested positive.

- False Positive (FP)

The negatives that are identifies as positives. That means the number of healthy (without PCOS) patients that are tested positive.

- True Negative (TN)

The negatives that were identified negatives. It means that healthy patients are tested negative.

- False Negative (FN)

The positives that were identified as positive. In this case, the patients with PCOS are tested negative.

	precision	recall	f1-score	support
0.0	0.90	0.97	0.94	37
1.0	0.93	0.76	0.84	17
micro avg	0.91	0.91	0.91	54
macro avg	0.91	0.87	0.89	54
weighted avg	0.91	0.91	0.90	54

Figure 2.0: Classification report of KNN model

Figure 2.1: Classification report of Logistic Regression model

	precision	recall	f1-score	support
0.0	0.92	0.97	0.95	37
1.0	0.93	0.82	0.87	17
micro avg	0.93	0.93	0.93	54
macro avg	0.93	0.90	0.91	54
weighted avg	0.93	0.93	0.92	54

Precision: It is the ability of the model not to classify a sample positive that is actually negative. For each class (0 or 1), it is defined as true positives divided by the sum of true positives and false positives. Therefore,

Precision = TP / (TP+FP)

Recall: It is the ability of the model to find all the positives. It is also called sensitivity and is defined, for each class, as the ratio of true positives and the sum of true positives and false negatives.

Recall = TP / (TP+FN)

F1 score: It is defined as the harmonic mean of precision and recall, a measure of test's accuracy where 1 is its best score and 0 is its worst. It is helpful for comparing two classifiers.

F1 score = 2*((Precision*Recall) / (Precision+Recall))

Support: It is the number of actual occurrences of the class in the specified dataset.

F) Comparison and Selection of Model

To be able to set a baseline for comparison of our models, we have selected different numbers of features for different models. Each of them has given respective results, hence, giving a level of comparison.

3. Result

A total of 538 samples of patients from ten different hospitals from Kerala, India made the dataset. There were total of 39 parameters out of which only 9 parameters, figure 1.5, with the highest weights were considered for KNN and 10 parameters, figure 1.7, were considered for Logistic Regression. A comparison was made between the two different classifiers: liner and non linear. The liner classifier is KNN while the non linear classifier is the model of Logistic Regression. The F1 score helps determine the best model between the two. The F1 score for KNN is 0.90 and for that of Logistic Regression is 0.92, hence, model of Logistic Regression is selected to determine the absence or presence of PCOS.

4. Conclusion

Therefore, using the methodology and techniques of machine learning, we have successfully determined the best supervised classification model to detect Polycystic Ovary Syndrome.

2. ACKNOWLEDGMENTS

The dataset used here is downloaded from Kaggle from the following link: https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos

REFERENCES

[1] Mehrotra, P., Chatterjee, J., Chakraborty, C., Ghoshdastidar, B., & Ghoshdastidar, S. (2011). Automated screening of Polycystic Ovary Syndrome using machine learning.

[2] Revised 2003 consensus on diagnostic criteria and longterm health risks related to polycystic ovary syndrome (PCOS) . Human Reproduction, Volume 19, Issue 1, January 2004, Pages 41-47.

[3] Supervised Learning-Classification Using K-Nearest Neighbors (KNN). (2019). Python® Machine Learning, 205–220. doi:10.1002/9781119557500.ch9

[4] Bichler, Martin and Kiss, Christine, "A Comparison of Logistic Regression, k-Nearest Neighbor, and Decision Tree Induction for Campaign Management" (2004). AMCIS 2004 Proceedings. 230.

[5] Murat KORKMAZ,Selami GÜNEY,Şule Yüksel YİĞİTERTHE, "IMPORTANCE OF LOGISTIC REGRESSION IMPLEMENTATIONS IN THE TURKISH LIVESTOCK SECTOR AND LOGISTIC REGRESSION IMPLEMENTATIONS/FIELDS".