

ITA5007 DATA MINING AND BUSINESS INTELLIGENCE

Register Number: 22MCA0223

Name: Kamran Ansari

Register Number: 22MCA0233

Name: Bhooshan Birje

Program: Master of Computer Applications

School: SITE

EVALUATION OF MACHINE LEARNING TECHNIQUES FOR PCOS PREDICTION

1. ABSTRACT :

Polycystic Ovary Syndrome (PCOS) is a hormonal imbalance brought on by the ovaries producing too many male hormones, which affects a woman's reproductive system during her childbearing age. PCOS has been identified as one of the major factors of infertility in women. It also raises the likelihood of developing other illnesses including diabetes, high blood pressure, sleep difficulties, depression, and anxiety, among others. The goal of this research was to address this issue by implementing machine learning approaches for early PCOS identification based on several medical markers. The model used a dataset of 541 women from 10 hospitals in Kerala. This paper examined trends and correlation within the medical parameters and determined the attributes which are most important in the diagnosis of PCOS. 14 important features were found most relevant to the target using correlation matrix. Further Chi-square test and independent samples t-test were used to verify the association of these features with the target for categorical and numerical features respectively. Later, PCA was performed on the dataset, which reduced the dimensions to 8 to get 95% variance explained. We experimented with various Classification models and compared them using accuracy and sensitivity. Hyperparameter tuning was performed on the model to improve their performances. Out of all models, SVM gave the best results with 91% accuracy and 81% sensitivity. Furthermore, Number of follicles in the ovaries, skin darkening and age came out to be the most influential features for the prediction of PCOS.

Keywords: PCOS, dataset, feature selection, machine learning, classification models, hyperparameter tuning, feature importance

2. INTRODUCTION:

Polycystic Ovary Syndrome (PCOS) is one of the most prevalent endocrine diseases in women of child-bearing age. PCOS causes the ovaries to develop small collection of fluids called follicles which make them unable to release eggs, which in turn make a woman hard to conceive. Along with this, PCOS comes with many symptoms including irregular menstrual cycles, skin infections, weight gains, high blood pressure, diabetes. This further causes women to have anxiety and some fall into depression. Thus, early detection of PCOS is necessary so that further complications can be reduced. 12% to 21% of the women worldwide are affected by PCOS but most of them are undiagnosed due to lack of awareness. Also, diagnosing PCOS is difficult due to vagueness of the symptoms. Sometimes, a PCOS patient may not show any symptoms. Due to this, it is necessary to take advice of different physicians. But here the problem lies in lack of coordination between the doctors, which may take longer to deduce the results. Also, the medical tests needed for diagnosis are expensive and are not available everywhere. This is where Machine Learning comes to the rescue. Machine Learning is playing a key role in the health sector. ML can be used to process huge amount of patient data and produce clinical insights which can help in diagnosis of diseases. Therefore, this project aims to develop a machine learning diagnostic model that can predict the probability of PCOS of a patient, taking into account their medical parameters. It also reduces the possibility of human error and helps in early diagnosis. In addition, it makes diagnosis timely, cost-effective and easily available at home.

3. LITERATURE REVIEW

a. An Efficient Decision Tree establishment and performance analysis on Polycystic Ovary Syndrome

Publication Year : 2020

Author : A. S. Prapty and T. T. Shitu

Journal Name : 2020 IEEE

Summary : This research work had a dataset containing 542 data points 10 rows and 31 columns. This was the only research work that didn't have any kind of data processing before the models

were made and compared. The only indication of class imbalance handling was splitting the dataset into 70:30 ratio.

b. Diagnosis of Polycystic Ovary Syndrome using Machine Learning Algorithms

Publication Year : 2020

Author : S. Bharati, P. Poddar and M. R. Hossain Mondal

Journal Name : 2020 IEEE

Summary : In this research work, the significant features or columns were identified and ranked according to top-value. The significant features identified were; ratio of Follicle-stimulating hormone and Luteinizing hormone. Their results demonstrated that hybrid random forest and logistic regression on applying cross validation with 40 folds to the best 10 features demonstrated testing accuracy of 91.01% and recall of 90%

c. Automated Detection of Polycystic Ovary Syndrome using Machine Learning

Publication Year : 2021

Author : Y. A. Abu Adla, D. G. Raydan, M. Z J. Charaf, R. A. Saad, J. masreddine and M. O. Diab

Journal Name : 2021 IEEE

Summary : This research work had a dataset of 541 data points/rows and 39 features/columns. To avoid biased data splitting, K-Fold Cross validation (KVC) technique implemented filtered out the features that were considered insignificant by both the doctor and the statistical test. They employed the use of filters and wrappers for features selection, performed ANOVA test for feature significance and consulted a domain expert to make sure that the results obtained were aligned with the medical field standards. Final model selected is SVM with an accuracy of 91.6%, precision 93.655% and Recall 80.6%

d. Comparative Analysis of machine learning Algorithms for Prediction of PCOS

Publication Year : 2021

Author : P. Chauhan, P. Patil, N. Rane, P. Raundale and H. Kanakia

Journal Name : 2021 IEEE

Summary : -- A custom dataset that was normally distributed over various attributes of around 267 women of different groups was used. Feature/column rows were selected based upon visualizations generated using Tableau and Google sheets. The best performance was given by Decision Tree Classifier with an accuracy of 81%. It also gave a precision of 70% and specificity of 94%.

e. Machine Learning Approach: Detecting Polycystic Ovary Syndrome & its impact on Bangladesh Women

Publication Year : 2021

Author : N. Nabi, S. Islam, S. A. Khushbu and A. K. M. Masum

Journal Name : 2021 IEEE

Summary : Custom dataset was used for conducting research on PCOS detection in Bangladeshi women. They made a questionnaire in Bangla and English and circulated to collect the dataset. They used a train-test split of 80-20 and achieved 99.09% using SVM technique. XGB Classifiers and Gaussian Naïve Bayes both gave 98.18% accuracy. Logistic Regression and Decision Tree both gave 97.97% accuracy. Decision Tree and Gradient Boosting Classifiers and random Forest Classifier gave same accuracy of 97.27%

4. METHODOLOGY :

4.1. Dataset:

The "Polycystic Ovary Syndrome" dataset, which is accessible on the internet website Kaggle, was gathered from 10 hospitals in the state Kerala, India. It contains information on 541 women, 177 of whom have been diagnosed with PCOS and 364 of whom are healthy. The information is broken down into a variety of hormonal, biochemical, clinical, and physical factors,

including weight, blood type, age, blood pressure, height, endometrial thickness, levels of the hormone follicle-stimulating hormone (FSH), and cycle duration, among others. The dataset contains two biochemical parameters, nine metabolic parameters, sixteen physical parameters, three ultrasound imaging parameters, and nine hormonal parameters. This dataset is beneficial for machine learning professionals since it offers a variety of pertinent data on the subject of PCOS.

Target Characteristics	Bivariate	Number of Instances	541
Attribute Characteristics	Categorical, Integer, Real	Number of Attributes	42
Associated tasks	Classification	Missing Values ?	Yes

4.1. Dataset Description

4.2 Experimental Procedure

We followed a series of steps for our research work starting from data collection to finding the best model for our system to predict likelihood of PCOS.

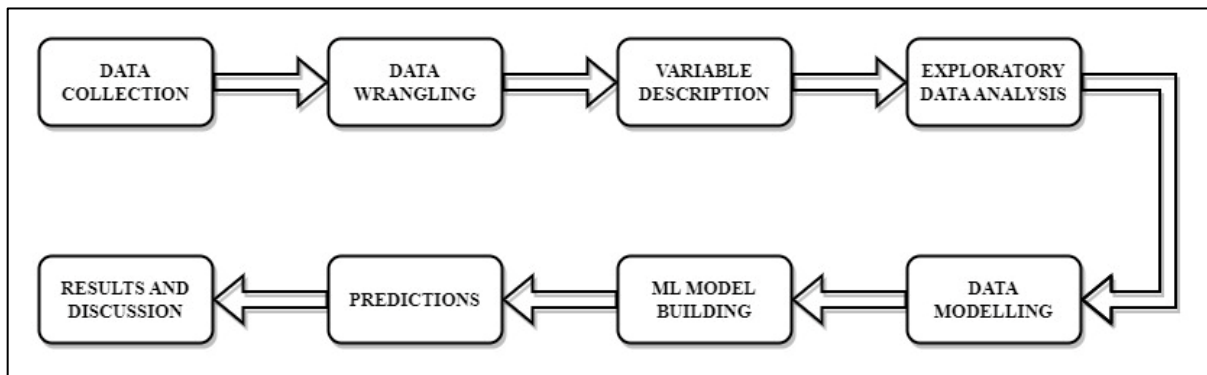


Figure 4.2.1. Flow of the research process

1. Data Wrangling -

- Dropped a column 'Unnamed:44' which had empty cells
- Also dropped all the duplicate columns

- c. Removed descriptive columns which that were not needed for processing such as “Serial No.” and “Patient file no.”
- d. Some columns had values encoded as objects. Standardized all of them to numeric values
- e. Filled the cells with null values with the median of the column values
- f. Removed leading and trailing spaces in column names.

2. Variable Description- This includes:

- a. The name of the variable and a brief description of what it represents
- b. The type of the variable (e.g., categorical, numerical, continuous, ordinal, etc.)
- c. The range of values that the variable can take on (e.g., for numerical variables, the minimum and maximum values; for categorical variables, the possible categories)

3. Feature Selection

Correlation Matrix is used for feature selection. After plotting correlation matrix top 11 features with highest positive correlation and top 3 features with highest negative correlation are considered for further processing

Correlation matrix shows features (Table 1) that have correlation with the Target value.

1. Follicle No. (R)	2. Follicle No. (L)
3. Skin darkening (Y/N)	4. hair growth(Y/N)
5. Weight gain(Y/N)	6. Cycle(R/I)
7. Fast food (Y/N)	8. Pimples(Y/N)
9. AMH (ng/mL)	10. Weight (Kg)
11. BMI	12. Cycle length(days)
13. Age (years)	14. Marriage Status (years)

Table 4.2.3. List of Features selected for processing

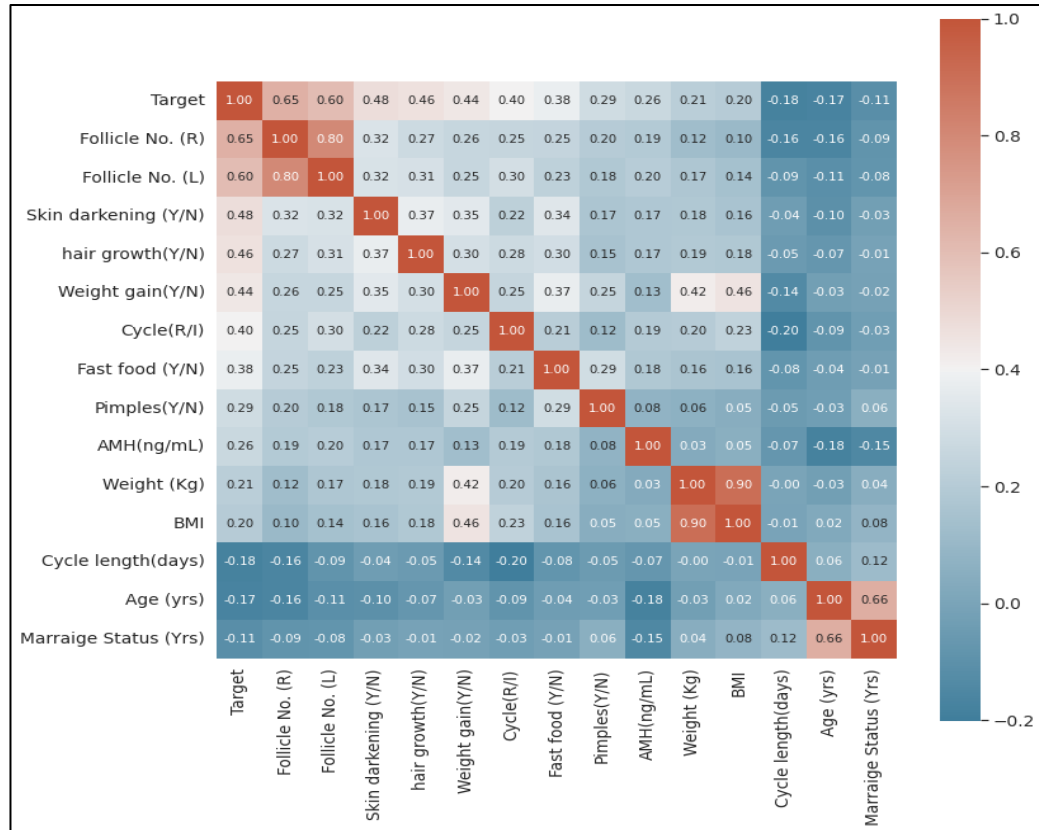


Figure 4.3.3. Correlation Matrix

Further, *Chi Square Test of Independence of attributes* was used to test if the categorical features considered have association with the target or not.

Null Hypothesis: Variables are independent of each other.

Alternative Hypothesis: Variables are not independent of each other.

All P values come out to be less than 0.05. Therefore, we reject the Null Hypothesis.

This confirms association of Target Variables with Selected Categorical Variables.

Similarly, *Independent Samples t-test* was employed to test the influence of numerical variables namely on the target. Here, numerical variables were divided into 2 groups, variables with positive patients and variables with negative patients. Then, t-test was used to check if mean of both the groups are equal or not. If both the means(averages) are equal that means numerical variables do not affect the target variables.

P values come out to be less than 0.05. Therefore, we reject the Null Hypothesis.

Null Hypothesis: Means of both groups are not significantly different

Alternative Hypothesis: Means of both groups are significantly different.

All P values come out to be less than 0.05. Therefore, we reject the Null Hypothesis.

This confirms effect of numerical variables on Target Variable.

4. Exploratory Data Analysis- Here, we explored all the features selected using various graphical components. Association of variables with the target and also with each other.

Inferences:

- i. Menstrual Cycle length seemed to be shorter and also the length increased as the patient age was greater. While the negative target patients had same cycle length throughout the age groups.
- ii. BMI of the positive patients increased drastically as the age increased while the negative patients had stable BMI.
- iii. Also, patients with PCOS had higher irregularity in the menstruation as compared to negative patients.
- iv. Number of follicles in both left and right ovaries of the PCOS patients were significantly higher than that of negative ones.
- v. Surprisingly, a greater number of younger patients had PCOS than the elder ones. But this could be because of the unawareness and readiness to treat the disorder.
- vi. Weight of patients with PCOS was significantly higher than that of the non-PCOS patients.
- vii. PCOS patients had slightly lower Hb levels.

5. Data Modelling-

- a. **Splitting Data-** Data was split into 2 groups for training and testing the models. 70% of the data was considered for training while remaining 30% for testing. Assigning 14 independent variables to X and target variable to Y.

Training – 378 samples

Testing – 163 samples

b. **Data Standardization-** Standardization is a way to transform data so that it has a mean of 0 and a standard deviation of 1. This can be useful for some machine learning algorithms that assume that the input variables are centered around 0 and have a standard deviation of 1. Standardization is done using Standard Scaler function.

6. Model Building-

a. List of Models employed in this system:

- i. XGBRF
- ii. Logistic Regression
- iii. K-Nearest Neighbors
- iv. SVM
- v. Decision Tree
- vi. Random Forest
- vii. Cat Boost Classifier

We built above models using ‘sklearn’ library in python. Metrics were calculated for all the models for further evaluation. Models used default hyperparameters for building.

b. **Hyperparameter tuning-** We have used Grid Search CV in our system. Range of values for each parameter were given. Scoring parameter for the CV was set to ‘accuracy’. Grid Search returned with the best parameter for all the models which helps the model achieve highest accuracy.

After finding best parameters. Again, models were trained and tested but with their best hyperparameters

5. RESULTS AND DISCUSSIONS

We compared accuracies between Untuned and Tuned models.

Models	Accuracy	Precision	Sensitivity	Specificity	F1-Score	AUC
SVM	0.91	0.81	0.81	0.96	0.86	0.89
Cat Boost	0.91	0.78	0.78	0.97	0.85	0.88

Logistic	0.90	0.80	0.80	0.95	0.84	0.88
XGBRF	0.89	0.76	0.76	0.95	0.82	0.86
Random Forest	0.88	0.70	0.70	0.97	0.80	0.84
KNN	0.86	0.65	0.65	0.96	0.75	0.81
Decision Tree	0.83	0.69	0.69	0.91	0.73	0.80

Table 5.1. Table of models and their metric values sorted by first highest **Accuracy** and **Sensitivity**

SVM performs better than all the other models for our system with an accuracy of 91% and a sensitivity of 81%

Accuracy and Sensitivity are used as important metrics for this study. This is because, we hope to achieve highest possible accuracy for our prediction model. Higher the accuracy, greater the chances of accurate predictions. Moreover, in medical inspections, it is intended to miss as few positive cases as possible and higher the sensitivity values, lower the chances of positive cases missed.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

Where, TP, TN, FP, FN stand for True Positive, True Negative, False Positive, and False Negative respectively.

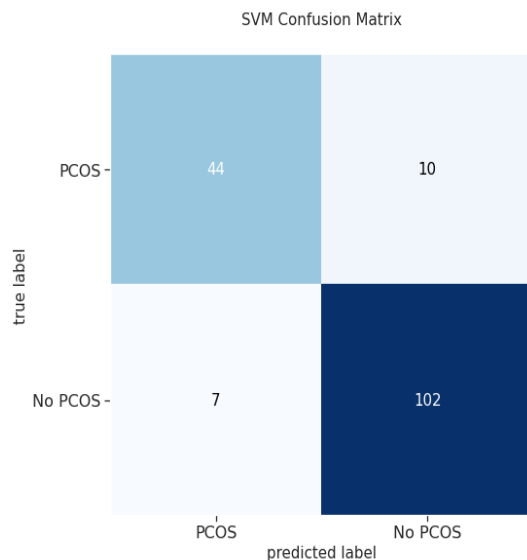


Figure 5.2 Confusion Matrix

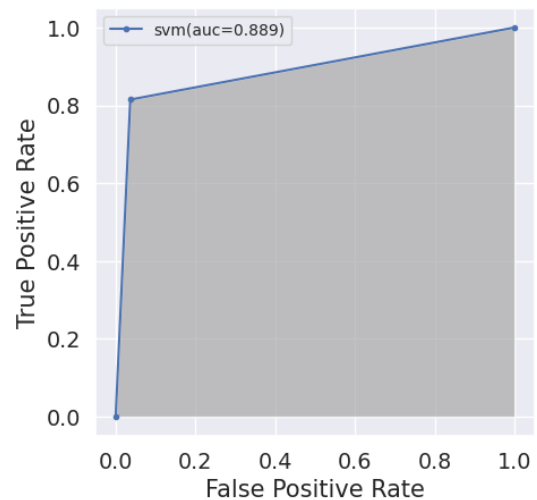


Figure 5.3. ROC-AUC

The Area Under the Curve for SVM is 0.889 which is greatest amongst all the models.

Confusion matrix gives the following inferences:

- 1) **True Positive:** SVM predicted 44 PCOS patients who actually had it.
- 2) **False Positive:** SVM predicted 10 patients to have PCOS, but they did not.
- 3) **False Negative:** 7 patients were predicted not to have PCOS by SVM, but they actually had it.
- 4) **True Negative:** SVM predicted 102 patients to be unaffected by PCOS who were in fact unaffected.

5.1. Feature Importance- For a given model, feature importance refers to the method of calculating a score that represents the importance of each input feature. If a feature has a high score, it has a greater impact on the model that is being used to predict a certain variable.

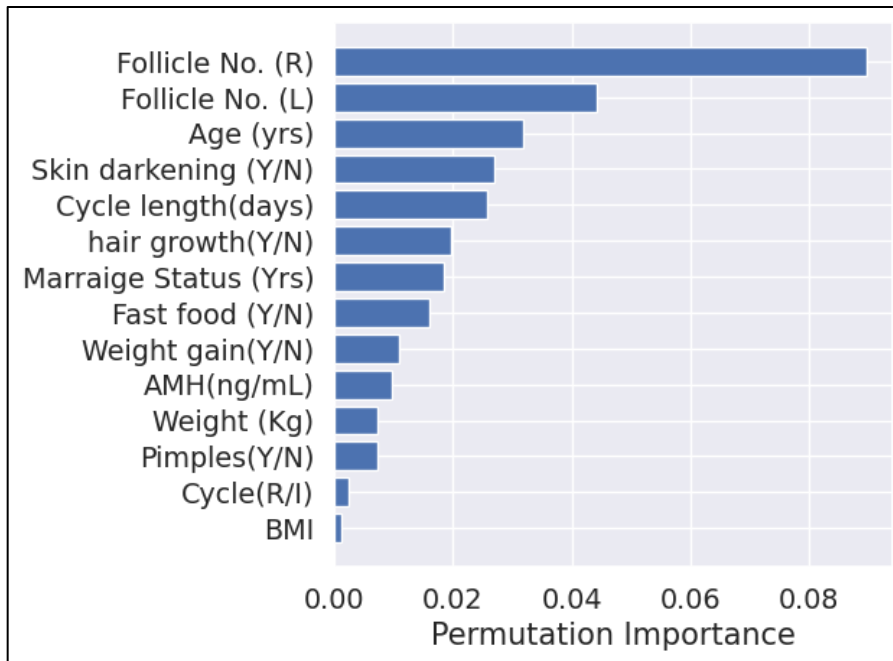


Figure 5.1.1. Importance of features in the model

Plotting Permutation importance graph, we get the top 4 features which SVM found most influential in prediction of PCOS are

1. Follicle No. (R)

2. Follicle No. (L)
3. Age (yrs.)
4. Skin darkening (Y/N)

From the tuned SVM we can still get false results in that case to avoid misleading results, we can take the help of feature importance from which we will see the 4 most important features and compare them with the ideal parameters and if they are violating it, we can tell the patient that he/she is at the risk of having PCOS.

6. CONCLUSION :

Today, millions of patients are affected by PCOS but the tragedy is that most of them are unaware of it. Patients suffer significant health consequences as a result. Our main goal was to develop a system that would allow a patient to determine their risk of PCOS by entering a few basic health information. In this study, we explored various medical parameters, how they are related to each other and how they vary for a person with PCOS. Then we explored various machine learning models. Tuned their hyperparameters to get accurate results. Out of all, SVM performed better than others with an accuracy of 0.91 and Sensitivity of 0.81. Further, Number of follicles in the ovaries, Age and darkening of skin came out to be the most important features. The dataset used for investigation had a smaller number of samples and it was concentrated on a single geographical location. For improving the diagnosis model large number of datasets have to be collected and awareness about PCOS should be increased.

7. REFERENCES :

Dataset Link –

<https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos>

A. S. Prapty and T. T. Shitu, “An Efficient Decision Tree establishment and performance analysis on Polycystic Ovary Syndrome”

S. Bharati, P. Poddar and M. R. Hossain Mondal, “Diagnosis of Polycystic Ovary Syndrome using Machine Learning Algorithms”

Y. A. Abu Adla, D. G. Raydan, M. Z J. Charaf, R. A. Saad, J. masreddine and M. O. Diab, “Automated Detection of Polycystic Ovary Syndrome using Machine Learning”

P. Chauhan, P. Patil, N. Rane, P. Raundale and H. Kanakia, “Comparative Analysis of machine learning Algorithms for Prediction of PCOS”

M. S. Khan Inan, R. E. Ulfath, F. I. Alam, F. K. Bappe and R. hasan, “Improved Sampling and Feature Selection to Support Extreme Gradient Boosting For PCOS Diagnosis”

N. Nabi, S. Islam, S. A. Khushbu and A. K. M. Masum, “Machine Learning Approach: Detecting Polycystic Ovary Syndrome & its impact on Bangladesh Women”