Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 10, Issue 3, July 2019: 386-405

An Optimal Data Preparation and Feature Extraction Methodology for Classification Algorithms

Rakshitha kiran P[a] , Dr. Naveen N C[b]

[a]Research Scolar at ISE Dept, Assistant Professor, Department of MCA , Dayananda Sagar College of Engineering, Bengaluru India-560078

[b] Professor and Head, Department of CSE, JSS Academcy of Technical Education, Bengaluru,India -560060

Abstract: Data preprocessing is a most important stage in data mining which is often neglected. This stage involves transforming the raw data into readable format. The real world data tends to be noisy, incomplete, inconsistent and lacks certain behavioral patterns. It is very important to preprocess such data before using for analysis. This paper summarizes various data preprocessing methodology for PCOS (Polycystic Ovary syndrome) datasets. PCOS is a common hormonal problem faced by ladies in the age group of 19-35's. Initially the PCOS dataset is preprocessed by preprocessing methods like Multiple Imputation, Discretization method which converts data into into discrete values, Standard scaler, Min-Max scalar methods are used for feature scaling, RobustScaler() to used remove the outliers. After the preprocessing stage feature extraction procedure is carried out where the most relevant features are extracted. Then the data sets are classified using various classification techniques like K Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Network (ANN) are used. The classification model is evaluated for accuracy, precision, recall and F1-score performance metrics. This paper compares the model performance with and without data preprocessing stage and also with feature extraction. And the results have proved that the preprocessing method with feature extraction technique has significantly improved the model performance.

Keywords: *Discretization, Min-Max scalar, RobustScaler(), KNN, LR, Random forest, SVM, ANN.*

1. Introduction

The Healthcare data can be of different forms: structured tables, images, audio files or video format. It contains huge numbers of rows and columns, so it's very important to preprocess the data before analyzing them. In this stage the data gets transformed into a state which the machine can easily understand. A dataset is a collection of data objects called as record / points / pattern[1,3]. These data objects have a number of features which show characteristics of an object .These features are variables, characteristics, fields, attributes, or dimensions. Pre-processing the datasets before using, gives better and accurate results. In this paper preprocessing of PCOS (Polycystic Ovary Syndrome) datasets are carried out and the best features are extracted using Random forest method. The PCOS

datasets had contained few erroneous, ambiguous, missing values. All these problems are addressed during preprocessing. There are different steps involved in this pre-processing process[2]. Not all steps are applicable to every problem. Some of the methods are as mentioned below[4]:

- **Data Quality Assessment**: The quality of your data tends to go down over time, even with the best data collection method. Hence it's important to clean the data before using it for analysis. Data Quality Assessment identifies those records that are inaccurate and rectifies the data.
- **Feature Aggregation**: This method takes the various local features from an instance of a dataset and creates a single global feature vector for the same feature[17,18].
- **Feature Sampling**: Sampling is a method of selecting a subset of the dataset to be used for analysis. It is very tedious, time-consuming and expensive to work with a complete dataset, so the portion of the dataset is selected that has approximately the same properties as the original dataset, which means the sample is representative. To obtain an accurate sample, you must choose sample size and sampling strategy.
- **Dimensionality Reduction**: In this process the number of random variables is reduced, by obtaining a set of principal variables. There are 2 stages: feature selection and feature extraction.
  - Feature selection: Here the subset of the original set of variables or features are obtained to get a smaller subset that can be used to model the problem. This can be done using Filter method, Wrapper method and Embedded method.
  - Feature extraction: here data is reduced from a higher dimensional space to a lower dimension space.
- **Feature Encoding**: Since Machine learning models can only work with numerical values it is needed to convert the categorical values of the appropriate features into numerical ones. This process is called *feature encoding*. In Data frame analytics this process happens automatically. The input data is pre-processed with the following encoding techniques:
- one-hot encoding: In this method vectors are allotted based on category i.e if similar features are present then its represented as ‘1’ else it's represented as ‘0’.
- target-mean encoding: Here categorical values are replaced with the mean value of the target variable.
- frequency encoding: This method checks the number of times a given categorical value is present in relation with a similar feature.

This paper summarizes the importance of data pre-processing and feature extraction and selection process. In the section 2 various data preprocessing techniques with the python implementation is explained. In the section 3 the description of PCOS datasets has been made. The section 4 shows the research methodology where data pre-processing and feature selection using Random forest has been elaborated. The section 5 and 6 shows the Evaluation matrix and Results of the pre-processing and feature selection stage. Finally section 8 summarizes the conclusion of the paper.

2. Data Preprocessing techniques with Python

There are 3 stages in this process that is data cleaning, data transformation and data reduction.

Data Cleaning: The goal of data cleaning is to provide simple, complete, and clear sets of examples for machine learning. The data taken from the real-world problem is seldom clean and complete, specially the healthcare field. The data cleansing/cleaning step deals with the treatment of missing values, errors and inconsistencies in the dataset. To deal with this issue various methods or techniques are evolved since past. Choosing the appropriate technique depends on various factors[5,13].

- **Missing Value Treatment** Missing value (MV) is defined as the data value that is not present in the cell of the particular column [6,15]. There can be multiple reasons for missing values in healthcare context, few of them are as follows: Missing value due to human omission or not applicable instance or the patient condition is unrelated for a particular variable or not recorded electronically by the sensor or patient not present on the ventilator due to a medical decision or due to some electricity failure or database synchronization and so on. Working on missing value may lead to undesired or biased result and could result to misleading conclusions [9,10].

- **Discarding the Missing Values** The most usual approach to discard the MVs but this approach is not so practical[9] because if the train data have a large number of missing values then the produced result must biased. If the dataset has a small number of missing values, then we must assure that analysis on the remaining part will not produce the inference bias. The deletion of the MVs can be done in following ways:

- **Listwise deletion (Deleting rows)**: In this case, a full case analysis is performed and all observed cases with more than one missing value are removed. But this approach is helpful in small number of missing cases. It works well when dataset has the MCAR missing pattern and which is rare [13].

- **Pairwise deletion:** This method attempts to minimize the error in list wise deletion. In this the attribute with MVs is deleted if it is not use as a case of another attribute while analyzing the data. It strengthens the analysis power but create other complications like producing standard error[14].

- **Dropping attribute completely**: This is very rare and, in my opinion, sometimes you can drop the attribute completely if the missing observations are more than 60% and the attribute looks insignificant in the analysis. Sometime attributes with missing values should be kept due to high relevance.

Data Transformation: In this stage the data is transformed into a format. The data which is independent is normalized within a particular range. This transformation stage helps in reducing the computation time and speeding up the calculations[36,37]. There are various ways to perform data transformations, few of them are listed below:

- **Rescaling Data**: This method gives values between 0 and 1. It is used in normalizing the data. In python MinMaxScaler class function is used for this purpose. This method helps in calculating

distance measure in various classification and regression algorithms like KNN, neural networks and so on.

- **Standardizing Data:** This process helps in finding the z-score for input datasets in order to format the data. In this process the input data set is standardized to the range of features. The z-score is obtained by subtracting the input feature value and the mean and then dividing by the standard deviation for each value of each feature.

$$Z=\frac{value-mean}{standard\ deviation}$$

The output of standardization is that, all the features will have a mean equal to zero, a standard deviation equal to one, and thus, the same scale. In python, StandardScaler class is a function used to perform standardization.

- **Normalizing Data**: This technique is used when the data distribution is unknown or when the distribution is not Gaussian (a bell curve)[38] or when data has varying scales . Here the observations are rescaled to a length of 1 and does not make assumptions about the distribution of data. In python Normalizer() class is used to perform normalization. This function is obtained from skleran library and it can transform inputs to unit norms which can be used for classification.

- **One Hot Encoding:** It is very difficult to work with categorical data in Machine learning. These categorical data must be converted into numerical data. To perform this we use "One Hot Encoding function". This function represents categorical variables as binary vectors. Here initially categorical values are mapped into integer values. Each integer value is represented as a binary vector i.e., all zero values are marked as 1 except the index of the integer. OneHotEncoder() is defined in sklerarn library in Python and it derives the categories based on the unique values in each feature.

- **Label Encoding**: In this methodology the training data will be given with a label in textual format to make it readable. These word labels are then converted into numbers for the machine-readable form i.e. the categorical value will be replaced with a numeric value between 0 and the number of classes minus 1. The Machine learning algorithms will work on these labels. Label encoding is a very important preprocessing step for the structured dataset in supervised learning [39] In python LabelEncoder() from scikit-learn library is the function used to perform label encoding**.**

- **Smoothing:** This technique is used to remove noise from the dataset using some algorithms. Smoothing highlights important features present in the dataset and predicts the patterns. This technique eliminates or reduces any variance or any other noise form.

- **Robust Data Scaling:** is used to remove the outliers**.** This methodology takes out the median value and scales the data based on the InterQuartile Range (IQR). This function scles the features that are robust to outliers. The robust scaler transforms are available in the scikit-learn Python machine learning library via the RobustScaler() class.

Data reduction: The dimension of data is one of the main criteria in data analysis [5, 6]. Data reduction is the process of reducing the dimension of the data in order to store the data easily and

also to increase efficiency of the Data. This technique has been used in various data analysis field and also to overcome the curse of dimensionality. Dimension reduction (DR) is direction proportional to the curse of dimensionality. The main reason for the "curse of dimensionality" is c intractable problems caused due to high computing time in data analysis. It is observed that as the number of variables increases the computing cost also increases exponentially. So to avoid this problem, DR methods have been used [7]. Some of the popular DR techniques are Principal component analysis (PCA) and factor analysis (FA). These techniques reduce the number of variables in the dataset and increase the computing time exponentially to avoid the problem of "curse of dimensionality".[11,12].

- **PCA** This algorithm performs the linear dimensionality reduction to transform a set of correlated variables (p) into fewer k (k <p) uncorrelated variables. These uncorrelated variables are called as principal components. PCA is mainly used to find correlation between features, if the correlation value is above the threshold the function will merge those features and produce that data for those fewer linearly uncorrelated features. PCA algorithm works until correlation is reduced to some extent and is done by indentifying the maximum variance in the original high-dimensional data and projecting them onto a smaller dimensional space. [16]**.**

- **Factor Analysis (FA)** Like PCA, Factor Analysis is used to reduce the dimensionality of the data, as well as to find hidden variables in datasets that are not directly measured in one variable, but produced from other variables.[17]. These hidden variables are called factors.

3.Dataset Description

The dataset consist of Polycystic ovary syndrome is a disorder observed in ladies involving irregular prolonged or infrequent menstrual periods[7,8]. This also involves excess male hormone (androgen) levels in ladies suffering from PCOS resulting in male features like excess body hair including in chest, back and face[21]. The ovaries in PCOS patients will have numerous small collections of fluid called follicles. These ovaries fail to release eggs regularly. The dataset is obtained from Kaggle dataset repository [40] and it contains all physical and clinical parameters to determine PCOS and infertility-related issues. The dataset consists of PCOS data of the ladies belonging to age group from 20-45.

BMI is calculated from height and weight[20]. If BMI ratio is below 18 then the patient is underweight. In the datasets there are 18 underweight patients out of which 10 patients are suffering with PCOS. If the BMI ratio value is in between 18.5 to 23 its considered as normal and the dataset consist of 113 normal weight patients, out of which 48 have PCOS. If the BMI ratio is between 23 to 25 its Overweight—At Risk, there are about 180 patients in this range. If the BMI ratio is between 25- 30 its Overweight—Moderately Obese, there are 40 patients within this range. And finally if the BMI ratio is above 30 then the condition is called as overweight, there are 15 overweight patients details.

The graphical representation for the datasets is as shown below:

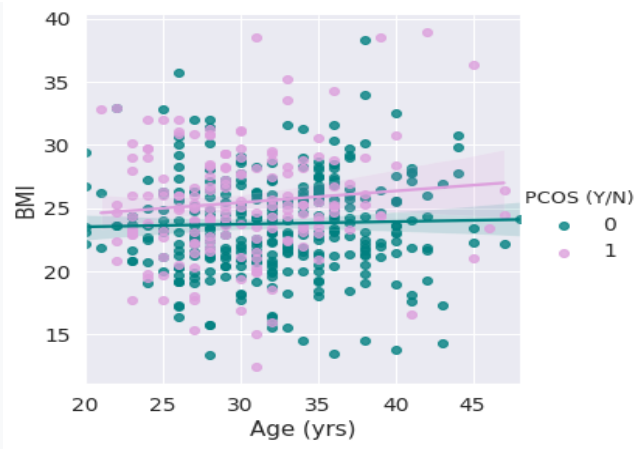

Figure 1: Graphical representation of BMI vs. Age for PCOS datasets

Body mass index (BMI) is showing consistency for normal cases, Whereas for PCOS the BMI increases with age.

Blood group: According to the survery by Rahul , Pratik Kumar [20] on Polycystic ovary syndrome, blood group & diet: A Correlative study in south Indian females, it is observed that O+ females are more prone to get PCOS. After O+ patients B+ individuals in India are more prone to get PCOS. In this study it is observed that along with blood group O+ other contributing factors for PCOS are mixed diet & alcohol intake. This paper also discussed about how the early screening of 'O+' &'B+' females of reproductive age-group could help in the timely diagnosis of PCOS, better management &also prevention of complications. The blood.

Pulse rate(bpm): According to the study by Yildirir, Aylin & Aybar, [22], PCOS patients has adverse cardiovascular risk when compared to non PCOS patients. The HRV(Heart Rate variability) parametersin PCOS patients had significantly higher LF (Lower frequency), LF/HF (Lower frequency/ High frequency), and significantly lower HF when compared to controls. The researchers have concluded saying "Autonomic innervations of the heart can be affected in PCOS with increased sympathetic and decreased parasympathetic components of HRV, As a result, sympathetic to parasympathetic ratio may increase in PCOS".

Hb(g/dl): According to research by Lerchbaum E, Schwetz V, Giuliani A, Obermayer-Pietsch B. [22] it is observed that the PCOS patients with showed a significant increase in HbA1C(Glycated hemoglobin) levels (5.799Â±1.022; 4.96Â±0.625, p=0.001) when compared to the normal non PCOS patients. The normal range for the hemoglobin A1c level is between 4% and 5.6% if Hemoglobin A1c levels between 5.7% and 6.4% means there is higher chance of getting diabetes.

Cycle length(days)/Cycle(R/I): Menstrual cycles varies from one body to another. From the recent research [23] it is observed that few ladies suffering from PCOS have regular periods but their androgen level is very high and also the production of insulin in their bodies is very high. This can interrupt their monthly cycle of ovulation and menstruation .

The average menstrual cycle is of 28 days for an egg from the ovary to release. The cycle lasts between 21 and 35 days is considered to be normal [24]. But in case of 'irregular' period cycle a minimum of eight or fewer menstrual cycles per year and these menstrual cycles is shorter than 21

days and longer than 35 days. In case of young women the menstrual cycle should have it within three years of starting periods, not longer than 45 days. For teenagers, 'irregular periods' means that the periods did not begin until the age of 15 or more than a year after periods began, and menstrual cycles are longer than 90 days. Cycles are also shorter than 21 days or longer than 45 days. Regular period's cause's excess thickening of the lining of the uterus (womb).It can also lead to abnormal cells building up inside the womb. For a healthy body it is important to have at least four cycles a year.

The graph below is obtained for the datasets used in the study here. It shows that the length of the menstrual phase is overall consistent over different age's groups for normal patients. Whereas the PCOS patients have their cycle length increased with age.

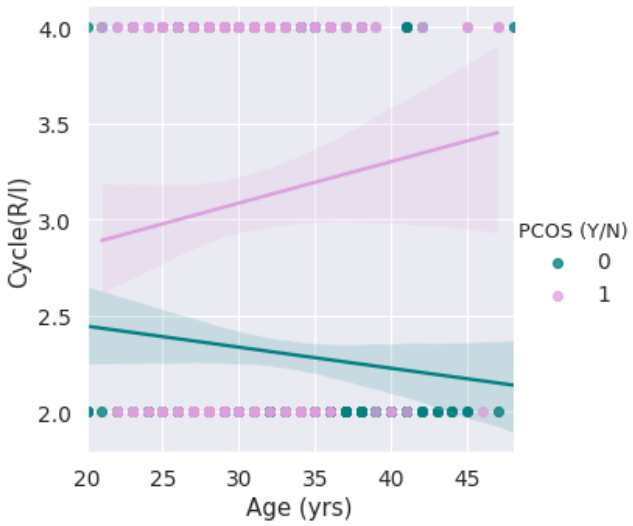

Figure 2: Graphical representation of cycle(R/I) vs. Age(yrs) for PCOS datasets

FSH(mIU/mL)& LH(mIU/mL The follicle-stimulating hormone (FSH) and Luteinizing hormone (LH) level are the primary factors to eliminate ovarian failure in ladies [25]. From the latest research it is observed that ladies with PCOS have FSH levels within the reference range or sometimes low. The LH levels in PCOS ladies are elevated in Preadolescent stage [27]. According to research the LH-to-FSH ratio is mostly higher than 3.

It is observed that PCOS patients have LH and FSH within the range of 5-20 mlU/ml and LH level is mostly twice or thrice times more that of the FSH level i.e., 18 mlU/ml and a FSH level of about 6 mlU/ml[FSH:LH=1:3]. This change in the LH to FSH ratio disrupts the ovulation and this condition is important factor for diagnosing PCOS.

AMH(ng/mL):AMH (Anti-Mullerian Hormone) is a laboratory test to detect woman's ovarian reserve or egg count[27]. This hormone is produced by cells from the small follicles inside ovaries and is used as a marker test of oocyte (immature egg) quantity. If the AMH level is over 3.0 ng/ml then the patient has greater changes of having PCOS. It is observed that more than 97% of women has AMH >10 ng/mL was suffering from PCOS. The table 1 shows AMH blood level classification for PCOS patients.

Table 1: AMH blood level interpretation

Interpretation	**AMH Blood Level**
High (often an indicator of PCOS)	Over 3.0 ng/ml
Normal	Over 1.0 ng/ml
Low Normal Range	0.7 – 0.9 ng/ml
Low	0.3 – 0.6 ng/ml
Very Low	Less than 0.3 ng/ml

TSH (mIU/L): TSH is Thyroid hormone disorder which is closely related to the elevated risk of infertility, unavoidable miscarriages, preterm delivery, and other metabolic dysfunctions and is observed in most of the PCOS patients [28]. The patients with PCOS have high risks of subclinical hypothyroidism and thyroid autoimmunity than normal women [29]. Also the research between PCOS and TSH shows that thyroid autoimmunity and metabolic parameters is very high in PCOS patients.

Endometrium (mm): Endometrium is the mucous membrane lining in the uterus. This membrane thickens during every menstrual cycle and it mainly happens to prepare the uterus for the possible implantation of an embryo. According to the research by Shah B, Parnell L, Milla S, Kessler M, David R[30], the mean thickness of the endometrium will much higher in the PCOS group (11.1mm). Endometrial receptivity is one of the main limiting factors for the formation of pregnancy in a huge count of gynecological disorders which includes PCOS.

PRL(ng/mL): Prolactin is a hormone which is released by the pituitary gland and it measures the amount of prolactin in the blood. PCOS patients have PRL level exceeds 85.2 ng/mL and there is a high possibility of prolactinoma warranting pituitary imaging. According to research, the Pituitary MRI is present in young PCOS patients with moderate PRL elevation [31].

Follicle No. (L) Follicle No. (R): According to the study by Mr. S. Jonard, Mr. Y. Robert [32] on 214 ladies having PCOS and compared with other 112 non PCOS ladies is it observed that the biological and clinical factors for PCOS can be obtained during the early follicular phase. Also the mean FNPO (follicle number per ovary) of follicles is 2–5 mm in size . It size is very high in polycystic ovaries than in normal non PCOS patients.

BP_Systolic(mmHg) & BP_Diastolic(mmHg): The metabolic syndrome is a condition where blood pressure (BP) is elevated [33]. It is calculated that over 30% of PCOS ladies have Blood Pressure greater 130/85 mmHg. Also the adolescent ladies with PCOS are three times more likely to be at higher risk of hypertension when correlated to others [34].

Table 2: PCOS Dataset description

Sl. No	Attributes	Description
1	Patient File No.	Patient ID
2	PCOS (Y/N)	Categorical data Yes-1 No-0
3	Age (yrs)	Age of the patient
4	Weight (Kg)	Weight in kilograms
5	Height(Cm)	Height in centimeters
6	BMI	Body Mass Index (kg/m^2)
7	Blood Group	A+, A-, B+, B-, O+, O-, AB+, AB-
8	Pulse rate(bpm)	Number of times heart beats in one minute.
9	RR (breaths/min)	normal respiration rate
10	Hb(g/dl)	Hemoglobin, or Hb, is usually expressed in grams per deciliter (g/dL) of blood.
11	Cycle(R/I)	No of menstrual cycle in a year
12	Cycle length(days)	No of days menstruation last in a month
13	Marraige Status (Yrs)	No of years of been married
14	Pregnant(Y/N)	Categorical data Yes-1 No-0
15	No. of abortions	No of times patient had abortion

16	FSH(mIU/mL)	***follicle-stimulating hormone level in patient body***
17	LH(mIU/mL)	level of luteinizing hormone (LH) in your blood.
18	FSH/LH	Ratio of follicle-stimulating hormone by luteinizing hormone
19	Hip(inch)	Hip circumference
20	Waist(inch)	Waist circumference
21	Waist:Hip Ratio	Ratio of waist/hip circumference.
22	TSH (mIU/L)	thyroid stimulating hormone
23	AMH(ng/mL)	Anti-Mullerian hormone or AMH is a hormone produced by the granulosa cells in your ovarian follicles
24	PRL(ng/mL)	Prolactin (PRL) measures of a hormone called prolactin in blood.
25	Vit D3 (ng/mL)	Vitamin D3 count in body.
26	PRG(ng/mL)	Progesterone count (nanogram per milliliter)
27	RBS(mg/dl)	A random blood sugar test is the testing of the blood sugar level (**milligrams per deciliter)**
28	Weight gain(Y/N)	Categorical data Yes-1 No-0
29	hair growth(Y/N)	Categorical data Yes-1 No-0
30	Skin darkening (Y/N)	Categorical data Yes-1 No-0
31	Hair loss(Y/N)	Categorical data Yes-1 No-0
32	Pimples(Y/N)	Categorical data Yes-1 No-0

33	Fast food (Y/N)	Categorical data Yes-1 No-0
34	Reg.Exercise(Y/N)	Categorical data Yes-1 No-0
35	BP _Systolic (mmHg)	Blood pressure
36	BP _Diastolic (mmHg)	Blood pressure
37	Follicle No. (L)	Count of Follicle (number)
38	Follicle No. (R)	Count of Follicle (number)
39	Avg. F size (L) (mm)	Left Size of the follicle in mm
40	Avg. F size (R) (mm)	Right Size of the follicle in mm
41	Endometrium (mm)	endometrial thickness in millimetre

The table above shows the brief description about the datasets used in this study. The dataset contains 542 instances of patient data where 365 patients are non PCOS and the rest 177 are PCOS patients.

4.Research methodology

This stage demonstrates the overall flow of the model. The model has two stages, first one is Data preprocessing stage and second is Feature section using Random forest method.

Data pre-processing stage

The PCOS dataset is first preprocessed using preprocessing technique. Data is cleaned initially by eliminated missing values and replacing the values with its mean value. Once the data is cleaned i.e. elimination of missing, NaN , null values data transformation is done. In this stage transforms the data to a specific format hence MinMaxScaler class function is applied because features vary in magnitude and nature (categorical and non-categorical) and also to the obtain values between 0 and 1.

Feature selection using Random forest

The next stage is to perform feature selection which is done by Random Forest algorithm [35]. RF will select features randomly with replacement method. The grouping of every subset is done in a separate subspace which is also called as random subspace. Calculating the feature or variable importance is an important step in random forest model. There are two stages in RF methodology: first, select top N from most important features by applying feature selection mechanism. Second,

calculate the Feature Importance with Random Forest. In Random forest methodology, there is randomness which is assigned to each process and node is chosen randomly. In python initially Random forest model is trained using " RandomForestClassifier()" function and once the model is built features are extracted by using function "feature_importances_ ". Both the function is imported from sklearn python library.

This preprocessed data then applied to KNN, LR (Logistic Regression), Naïve Bayes, ANN (Artificial Neural Network), SVM (support vector machine) and Random Forest classifier to predict the PCOS and non- PCOS patients. All experimental work has been done on Jupyter Notebook v. IDE. The Python 3.0 is the programming language is used to for the analysis and for building prediction models. For this experiment libraries used are numpy, pandas, matplotlib libraries of python 3.0 for dataset representation, processing and visualization and scikit-learn for building the machine learning models for predictions.

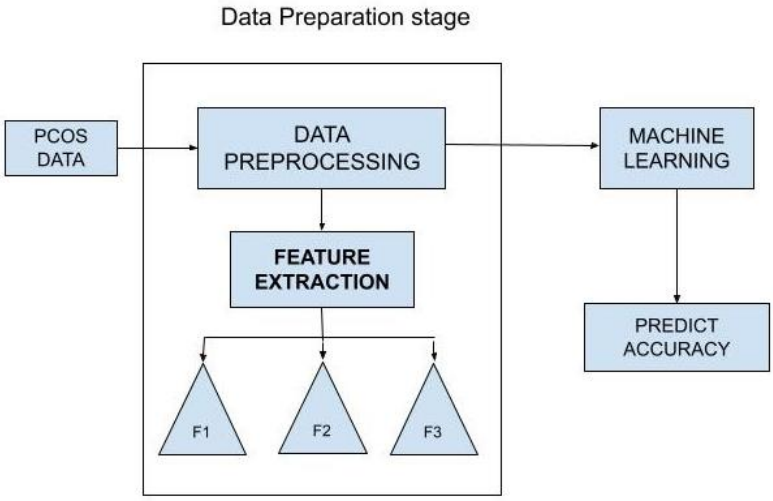

Figure 3: Research Methodology

Out of 40 features 15 most important features are removed. These features are extracted based on the research done by the doctors.

5. Evaluation Matrix

To evaluate the model performance Accuracy, precision recall and f1-score are used. Accuracy measure is always preferred choice in classification problem if target variable in dataset is approximately balanced. To calculate performance measures Accuracy, precision, recall and f1-score certain matrix is needed i.e. True Positive (TF), True Negative (TN), False Positive (FP) and False Negative (FN). The formulation of these matrices is given in table below:

Table 3: Model evaluation metrics

Accuracy	(TP+TN)/(TP+TN+FP+FN)
precision	TP/(FP+TP)

Recall	FP/(TP+FN)
F1-score	2*[precision *Recall/ precision +Recall]

6. Results and Discussion

The PCOS dataset, when used without preprocessing as an input to the classifiers KNN, LR (Logistic Regression), ANN (Artificial Neural Network), Random Forest and SVM (support vector machine) models exhibit the accuracies 75%, 72.5%, 65%, 79.41%,80.2% respectively. Here, it can be observed that Logistic regression model gives the best response without applying any preprocessing methods while Random forest is on second top performer. But ANN and SVM are not performing well. The reason why they are giving the worst result can be missing values and noise and both models require scaled values.

After preprocessing the dataset the accuracies increased by good value, for classifiers KNN, LR (Logistic Regression), ANN (Artificial Neural Network), Random Forest and SVM (support vector machine) are 84%, 84%, 86%, 82.6% and 88.6% respectively. This improves in the accuracies and other performance measures are because of elimination of missing values, null values and noisy data. Table below shows the performance measures with and without preprocessing stage.

Table 4: With and without preprocessing

Classification Algorithms	Preprocessing	Accuracy	precision	recall	f1-score
KNN	Without Preprocessing	75	0.78	0.78	0.74
	With Preprocessing	84	0.86	0.84	0.84
LR	Witho	72.5	0.73	0.73	0.72

	ut Preprocessing				
	With Preprocessing	84	0.86	0.84	0.84
Random forest	Without Preprocessing	65	0.66	0.64	0.68
	With Preprocessing	86	0.88	0.86	0.86
ANN	Without Preprocessing	79.41	0.79	0.79	0.79
	With Preprocessing	82.6	0.82	0.81	0.82
SVM	Without Preprocessing	80.2	0.78	0.81	0.81
	With Preprocessin	88.6	0.89	0.89	0.86

	g				

After the preprocessing stage feature extraction is carried out. Out of the 40 attributes, 15 features were extracted using random forest algorithm. These 15 features were processed with classification algorithms KNN, LR (Logistic Regression), ANN (Artificial Neural Network), Random Forest and SVM (support vector machine). The results of the feature extraction are quite surprising. The SVM classifiers showed very good performance when compared with other classification models.

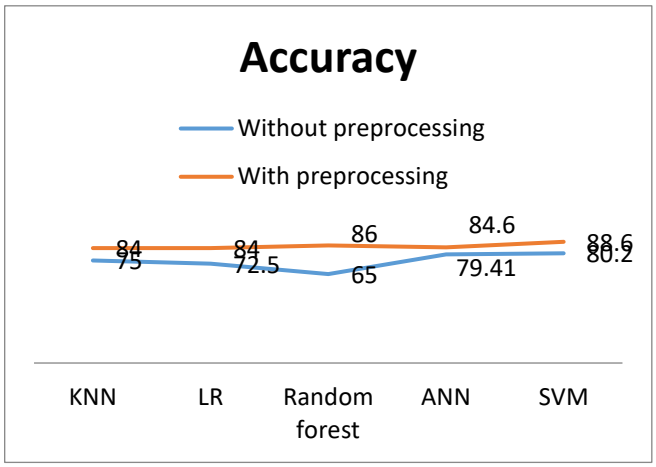

Figure 4: Graphical Representation Accuracies of with and without preprocessing stage

The figure 4 shows the graphical representation of the accuracy with and without preprocessing stage. There is the clear improvement in the accuracy after preprocessing stage. SVM classifier shows very high accuracy followed by random forest ANN, LR and finally KNN classifiers.

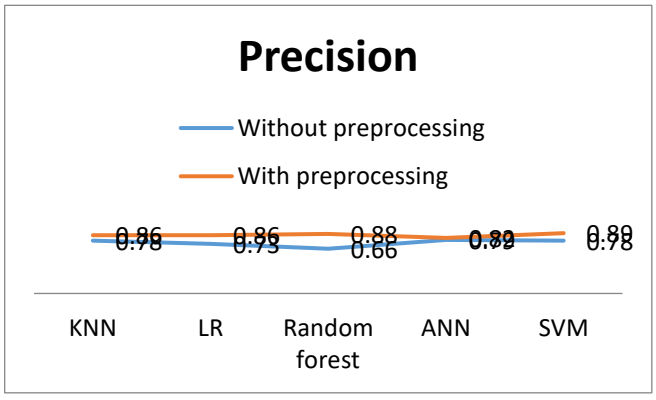

Figure 5: Graphical Representation of Precision of with and without preprocessing stage

The figure 5 shows the graphical representation of the precision with and without preprocessing stage. The precision graph shows improvement with data preprocessing stage when compared to without preprocessing stage. SVM classifier shows very high precision followed by random ANN, LR, KNN and finally Random forest classifiers.

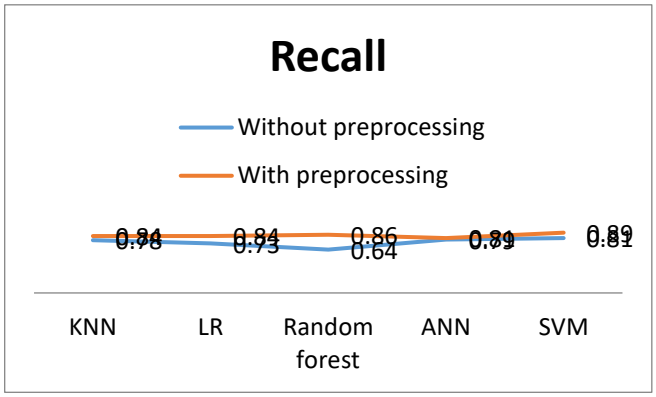

Figure 6: Graphical Representation of Recall of with and without preprocessing stage

The figure 6 shows the graphical representation of the Recall with and without preprocessing stage. The graph shows improvement with data preprocessing stage when compared to without preprocessing stage. SVM classifier shows very high precision followed by random ANN, LR, KNN and finally Random forest classifiers.

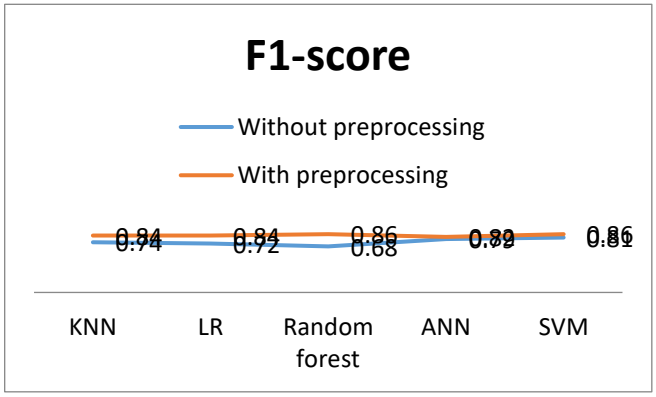

Figure 7: Graphical Representation of F1-score of with and without preprocessing stage

The figure 7 shows the graphical representation of the f1-score with and without preprocessing stage. The f1-score graph shows improvement with data preprocessing stage when compared to without preprocessing stage. SVM classifier shows very high precision followed by random ANN, LR, KNN and finally Random forest classifiers.

Table 5: Performance measures after Feature extraction process

Classification Algorithms	Accuracy	precision	recall	f1-score
KNN	80.90	0.80	0.81	0.80
Logistic	84.42	0.83	0.71	0.76

Regression				
Random forest	85.41	0.72	0.65	0.68
ANN	81.43	0.51	0.71	0.60
SVM	90.62	0.7	0.7	0.63

9.Conclusion

Data Pre-processing is a very important stage for solving any Machine Learning problem. Using raw data directly will not fetch proper results. Raw data consists of erroneous data, missing value, ambiguous values, improper format and so on. Hence it's very important to pre-process the data. This paper summarizes the various pre-processing techniques available using python. Feature extraction and selection pulls out the best features needed for the Machine Learning Model. Here feature selection which is done by Random Forest algorithm. In Random forest methodology initially Random forest model is trained using " RandomForestClassifier()" function and once the model is built features are extracted by using function "feature_importances_". The results have proved that the preprocessing method with feature extraction technique has significantly improved the model performance.

References (APA)

1. J. Han, M. Kamber, Data Mining Concepts and Techniques, 2nd edition, Morgan Kaufmann, 2006.
2. R.A. Johnson, D. W. Wichern, Applied Multivariate Statistical Analysis, 3rd edition, Prentice Hall, 1992.
3. J. H. Friedman, "On Bias, Variance, 0/1-loss, and the Curse of Dimensionality," Data Mining and Knowledge Discovery vol. 1, pp. 55–77, 1997.
4. M. A. Tanner, Tools for Statistical Inference, Springer, 1996.
5. Y. Youk, S. Kim, Y. Joo, "Intelligent Data Reduction Algorithm for Sensor Network based Fault Diagnostic System," International Journal of Fuzzy Logic and Intelligent Systems, vol. 9, no. 4, pp. 301-308, 2009.
6. J. Keum, H. Lee, M. Hagiwara, "A Novel Speech/Music Discrimination Using Feature Dimensionality Reduction," International Journal of Fuzzy Logic and Intelligent Systems, vol. 10, no. 1, pp. 7-11, 2010.
7. S. Ben-David and S. Shalev-Shwartz, Understanding Machine Learning: From Theory to Algorithms. 2014.

8. S. Batra and S. Sachdeva, “Organizing standardized electronic healthcare records data for mining,” Heal. Policy Technol., 2016.
9. R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S.K. Khatri, “Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India,” Int. J. Diabetes Dev. Ctries., vol. 36, no. 4, pp. 469–476, 2016.
10. F. Cismondi, A.S. Fialho, S.M. Vieira, S.R. Reti, J.M. C. Sousa, and S.N. Finkelstein, “Missing data in medical databases: Impute, delete or classify?,” Artif. Intell. Med., 2013.
11. H. Wang and S. Wang, “Mining incomplete survey data through classification,” Knowl. Inf. Syst., Vol. 24, No. 2, pp. 221–233, 2010.
12. A.K. Jain, “Data clustering: 50 years beyond K-means,” Pattern Recognit. Lett., Vol. 31, No. 8, pp. 651–666, 2010.
13. L. Beretta and A. Santaniello, “Nearest neighbor imputation algorithms: A critical evaluation,” BMC Med. Inform. Decis. Mak., Vol. 16, Suppl 3, 2016.
14. X.H. Cao, I. Stojkovic, and Z. Obradovic, “A robust data scaling algorithm to improve classification accuracies in biomedical data,” BMC Bioinformatics, Vol. 17, No. 1, pp. 1–10, 2016.

A. Idri, H. Benhar, J.L. Fernández-Alemán, and I. Kadi, “A systematic map of medical data preprocessing in knowledge discovery,” Comput. Methods Programs Biomed., Vol. 162, pp. 69–85, 2018.

15. P.R. Peres-Neto, D.A. Jackson, and K.M. Somers, “How many principal components? stopping rules for determining the number of non-trivial axes revisited,” Comput. Stat. Data Anal., Vol. 49, No. 4, pp. 974–997, 2005.
16. N. Poolsawad, L. Moore, C. Kambhampati, and J.G.F. Cleland, “Issues in the Mining of Heart Failure Datasets,” Int. J. Autom. Comput., vol. 11, no. 2, pp. 162–179, Apr. 2014.
17. R.K.A and L. George H. John b, “Wrappers for feature subset selection,” Artif. Intell., Vol. 97, No. 1–2, pp. 273–324, 1997.

A. Hapfelmeier and K. Ulm, “A new variable selection approach using Random Forests,” Comput. Stat. Data Anal., Vol. 60, No. 1, pp. 50–69, 2013.

18. Pal, Rahul & Chatterjee, Pratik & Chatterjee, Poulomi & Na, Vinodini & Mithra, Prasanna & Banerjee, Sourjya & Suman, VB & Pai, Sheila. (2014). Polycystic ovary syndrome, blood group & diet: A Correlative study in south Indian females. International Journal of Medical Research & Health Sciences. 3. 604. 10.5958/2319-5886.2014.00404.4.
19. Yildirir, Aylin & Aybar, Funda & Kabakci, Giray & Yarali, Hakan & Oto, Ali. (2006). Heart Rate Variability in Young Women with Polycystic Ovary Syndrome. Annals of noninvasive electrocardiology : the official journal of the International Society for Holter and Noninvasive Electrocardiology, Inc. 11. 306-12. 10.1111/j.1542-474X.2006.00122.x.
20. Lerchbaum E, Schwetz V, Giuliani A, Obermayer-Pietsch B. Assessment of glucose metabolism in polycystic ovary syndrome: HbA1c or fasting glucose compared with the oral glucose tolerance test as a screening method. Hum Reprod. 2013 Sep;28(9):2537-44. doi: 10.1093/humrep/det255. Epub 2013 Jun 11. PMID: 23756702.
21. Harris, H. R., Titus, L. J., Cramer, D. W., & Terry, K. L. (2017). Long and irregular menstrual cycles, polycystic ovary syndrome, and ovarian cancer risk in a population-based case-control study. *International journal of cancer*, *140*(2), 285–291. https://doi.org/10.1002/ijc.30441

22. Lim AJR, Huang Z, Chua SE, Kramer MS, Yong EL (2016) Sleep Duration, Exercise, Shift Work and Polycystic Ovarian Syndrome-Related Outcomes in a Healthy Population: A Cross-Sectional Study. PLOS ONE 11(11): e0167048. https://doi.org/10.1371/journal.pone.0167048
23. Saadia Z. (2020). Follicle Stimulating Hormone (LH: FSH) Ratio in Polycystic Ovary Syndrome (PCOS) - Obese vs. Non- Obese Women. *Medical archives (Sarajevo, Bosnia and Herzegovina)*, *74*(4), 289–293. https://doi.org/10.5455/medarh.2020.74.289-293.
24. https://www.contemporaryobgyn.net/view/hormone-levels-and-pcos
25. Tal R, Seifer DB, Khanimov M, Malter HE, Grazi RV, Leader B. Characterization of women with elevated antimüllerian hormone levels (AMH): correlation of AMH with polycystic ovarian syndrome phenotypes and assisted reproductive technology outcomes. Am J Obstet Gynecol. 2014 Jul;211(1):59.e1-8. doi: 10.1016/j.ajog.2014.02.026. Epub 2014 Mar 2. PMID: 24593938.
26. Cai, J., Zhang, Y., Wang, Y., Li, S., Wang, L., Zheng, J., Jiang, Y., Dong, Y., Zhou, H., Hu, Y., Ma, J., Liu, W., & Tao, T. (2019). High Thyroid Stimulating Hormone Level Is Associated With Hyperandrogenism in Euthyroid Polycystic Ovary Syndrome (PCOS) Women, Independent of Age, BMI, and Thyroid Autoimmunity: A Cross-Sectional Analysis. *Frontiers in endocrinology*, *10*, 222. https://doi.org/10.3389/fendo.2019.00222
27. Dayan CM, Daniels GH. Chronic autoimmune thyroiditis. *N Engl J Med*. (1996) 335:99–107. 10.1056/NEJM199607113350206
28. Shah B, Parnell L, Milla S, Kessler M, David R. Endometrial thickness, uterine, and ovarian ultrasonographic features in adolescents with polycystic ovarian syndrome. J Pediatr Adolesc Gynecol. 2010 Jun;23(3):146-52. doi: 10.1016/j.jpag.2009.07.002. Epub 2009 Sep 3. PMID: 19733099.
29. Kyritsi EM, Dimitriadis GK, Angelousi A, Mehta H, Shad A, Mytilinaiou M, Kaltsas G, Randeva HS. The value of prolactin in predicting prolactinoma in hyperprolactinaemic polycystic ovarian syndrome. Eur J Clin Invest. 2018 Jul;48(7):e12961. doi: 10.1111/eci.12961. Epub 2018 Jun 13. PMID: 29845629.
30. S. Jonard, Y. Robert, C. Cortet-Rudelli, P. Pigny, C. Decanter, D. Dewailly, Ultrasound examination of polycystic ovaries: is it worth counting the follicles?, *Human Reproduction*, Volume 18, Issue 3, March 2003, Pages 598–603, https://doi.org/10.1093/humrep/deg115
31. Mellembakken, J. R., Mahmoudan, A., Mørkrid, L., Sundström-Poromaa, I., Morin-Papunen, L., Tapanainen, J. S., Piltonen, T. T., Hirschberg, A. L., Stener-Victorin, E., Vanky, E., Ravn, P., Jensen, R. C., Andersen, M. S., & Glintborg, D. (2021). Higher blood pressure in normal weight women with PCOS compared to controls, *Endocrine Connections*, *10*(2), 154-163. Retrieved Jul 16, 2021, from https://ec.bioscientifica.com/view/journals/ec/10/2/EC-20-0527.xml
32. Glintborg D, Rubin KH, Nybo M, Abrahamsen B & Andersen M Cardiovascular disease in a nationwide population of Danish women with polycystic ovary syndrome. *Cardiovascular Diabetology* 2018 17 37. (https://doi.org/10.1186/s12933-018-0680-5)
33. Chen, RC., Dewi, C., Huang, SW. *et al.* Selecting critical features for data classification based on machine learning methods. *J Big Data* **7,** 52 (2020). https://doi.org/10.1186/s40537-020-00327-4
34. Vijayarani, S., & Tamilarasi, A. (2010). Data Transformation Technique for Protecting Private Information in Privacy Preserving Data Mining.

35. Km. Swati, Dr. Sanjay Kumar ,A Comparative Study of Various Data Transformation Techniques in Data Mining, International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume No.4 Issue No.3, pp : 146-148 01 March. 2015.
36. Muhammad Ali, Peshawa & Faraj, Rezhna. (2014). Data Normalization and Standardization: A Technical Report. 10.13140/RG.2.2.28948.04489.
37. Potdar, Kedar & Pardawala, Taher & Pai, Chinmay. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. International Journal of Computer Applications. 175. 7-9. 10.5120/ijca2017915495.
38. https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos.