Enhancement of Accuracy on a Medical Dataset by the Usage of Different Data Preprocessing Techniques

Manda Arpitha, Ramalakshmi.K, Venkatesan.R

***Abstract*: As the data is increasing exponentially, the real-time data has many inconsistencies. These factors of having inconsistent, incomplete and irrelevant data in the dataset would show its impacts on the knowledge development process. So, the quality of the data determines the success rate of the prediction by the machine learning model. By having a lot of missing and irrelevant values in the dataset would make the training and testing phase more troublesome. It is known that data preparation for analysis would take quite a long time. The primary concentration in this paper would be to clarify the stream of proficient advances that ought to be done amid the procedure of Data Mining.**

***Index Terms*: Data Mining, Data Preprocessing Machine Learning, Training, Testing.**

I. INTRODUCTION

Data Mining is an important technique that is used in industries to perform analytics on data. The analysis is an easy way of measuring or concluding knowledge. As the amount of data is being increased exponentially, handling this sort of data or deriving knowledge from it would be difficult [1]. The accuracy cannot be predicted without exploring the incomplete dataset. However, performing analysis over the dataset with missing values do not dive you into conclusions of knowledge.

II. ABOUT THE DATASET

The process of mining data would be explained in this paper with the help of the dataset named the “chronic kidney disease” dataset. The data set was gathered from the Kaggle[4]. The dataset was obtained in the un-preprocessed state. The following dataset has 400 instances and 25 attributes. There are 11 numerical values and 14 nominal values. Nearly 24 tests have been and performed and noted to conclude the presence of classification attribute which is either chronic kidney disorder or no chronic kidney disorder. The Tests conducted were represented in medical abbreviations, which are described in Table 1.1. We can divide the entire process of data mining into a few simple steps and explain
the same. The steps are

1. Understanding the dataset.
2. Dealing with the inconsistent dataset.
3. Finding the accuracy using the predefined models.

Manuscript published on 30 March 2019.
*Correspondence Author(s)

Manda Arpithais, B. Tech, Department of Computer Sciences, Karunya Institute of Technology, Coimbatore (Tamil Nadu), India.
Dr. K. Ramalakshmi, B. Tech, Madurai Kamaraj University, Madurai (Tamil Nadu), India.
R. Venkatesan. B. Tech, Madurai Kamaraj University, Madurai (Tamil Nadu), India.

III. UNDERSTANDING THE DATASET

The data set has un-countable null values for some of the attributes. The representation of the null values has been shown in the Fig-1. The figure was constructed using a visualization tool called “Tableau”. The construction of the Figure-1 involved the adding of classification and RBC(Red Blood Cells) From table-1 to the column level of the graph and the number of instances to the row. As a conclusion, Fig-1 shows us that there are more than 150 instances that possess the null values just in the RBC column. These null values would contribute backlogs in the process of the analysis. The dataset contains medical test abbreviations that are elaborated in the Table – 1. The Table -1 represents all the test levels predicted and the target class is also included.

Table – 1 Description of the dataset

S.no	abbreviations	Generic test name
1	Age	age
2	Bp	blood pressure
3	Sg	specific gravity
4	Al	albumin
5	Su	sugar
6	RBC	red blood cells
8	PC	pus cell
9	Pcc	pus cell clumps
10	Ba	bacteria
11	Bgr	blood glucose random
12	Bu	blood urea
13	Sc	serum creatinine
14	Sod	sodium
15	Pot	potassium
16	Hemo	hemoglobin
17	Pcv	packed cell volume

18	Wc	white blood cell count
19	Rc	red blood cell count
20	Htn	hypertension
21	Dm	diabetes mellitus
22	Cad	coronary artery disease
23	Appet	appetite
24	Pe	pedal edema
25	Ane	anemia
26	Class	classification

The missing values as shown in Fig - 1 have to be handled to improve the consistency of knowledge.

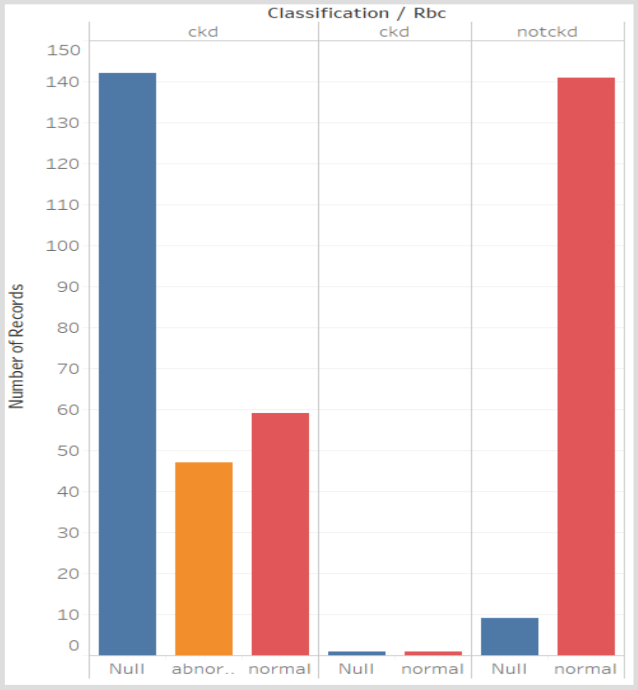


Fig.1 Representation of Null values

IV. DATA PREPROCESSING

Real-world data is commonly incomplete, inconsistent, and lacks in the certain behaviors or trends, and is probably going to contain several errors. Data Preprocessing [2] is one among the steps involved before mining data that involves reworking data into a lucid format. In this article, we focus on getting rid of the missing values using different methods using some python libraries.

The python libraries that are used are discussed below Pandas [3] is a python library that is used for data manipulation and analysis. Pandas is an opensource framework. There are different methods to for handling missing data using pandas data frame. We import the data in the .csv format to pandas which would be stored in the format of a Data Frame. Handling data using data frame would make handling missing values reliably easy. It would make analysis and method prediction easy by storing values of columns, mutable size and the type of data stored in each columns. Arithmetic Operations can also be performed over the data that has been stored in a data frame. The different methods that can be used to fill the missing values in the dataset are:

Generic methods:

Filling with random value: The values of the dataset can be filled randomly. A nominal value can replaced by creating a new category for missing values and a numerical value can be filled with ‘0’s or another considerable random value.

Forward filling: When this method of forward filling is chosen. The data frame fills or replaces the missing values with the value that is present above it

Backward filling: When this method of backward filling is chosen. The data frame fills or replaces the missing values with the value that is present below it

Filling data using Imputers:

The filling of missing values with a forward fill or a backward fill is cannot be a considerable remedy for filling the missing medical data. It might cause error and invalid predictions. So, the commonly used method for filling the missing values of the medical Data is by using Imputers. There are about two different methods or strategies that can be implemented by the Imputers. They are

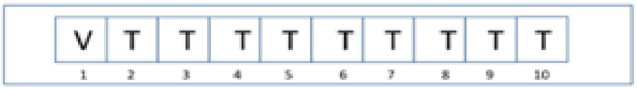
The rows with missing values have to be dropped.

Impute the data based on the datatype: A generic dataset would contain numeric and nominal data. It is preferred that numerical data is imputed by using mean and the nominal data is imputed by the mode.

V. TESTING AND TRAINING

Separating data into training and testing sets is a vital part of evaluating data processing models. Typically, after you separate a data set into a training set and testing set, most of the data is employed for training, and a smaller portion of the data is used for testing. After a model has been processed by using the training set, we check the model by making predictions against the test set. Because the data in the testing set already contains familiar values for the attribute that you need to predict. In this paper, we will discuss the two different sampling techniques available. We will also look at the accuracy that can be obtained by each resampling technique and some classification models. The two resampling techniques used are *Percentage Split*: This method can also be called as Holdout or Fixed Re-Sampling method [9]. Here we divide the entire dataset into two parts asymmetrically, such that, there is a greater percentage of data that can be trained. After training a dataset with a Classification model, the remaining lesser part of the split data has to be tested. Therefore, after the training and testing phase, the accuracy of the dataset with the model can be predicted.

K-folds with the Cross Validation (Where K = 10): The value of K in reference to this paper is considered 10. So, the second Re-Sampling technique that we have here is 10-fold Cross-Validation Approach.



V	T	T	T	T	T	T	T	T	T
1	2	3	4	5	6	7	8	9	10

Fig-2 Representation of 10-Cross Folds

As shown in the Fig-2, The entire dataset is divided into a 10 folds are 10 divisions. We will have a total of 10 iterations over the dataset

Where the first 9 folds are used for Testing and the remaining or the last fold is used for Training in the first iteration. In the second Iteration fold-2 to fold-10 are considered for testing and the first fold is used for training. The same process continues until each fold undergoes the testing and training phase. After the training and testing all the folds, the accuracy of the dataset with a model is predicted.

The following figures show the rate of accuracy with three classification models. The classification models that are used are Naive Bayes, 1R and Decision Table. The accuracy is Predicted by using both the Re-Sampling techniques which are Percentage Split and 10-fold Cross-Validation.

Classification model	Percentage Split	10-folds Cross-Validation
Naive Bayes	96.94%	94.5%
1R	93.38%	92%
Decision Table	96%	97%

Table 2 Accuracy after Pre-Processing

The accuracy of the dataset has been shown in the table-2as shown above. To improve the accuracy of the dataset with the model we will further continue with Data Cleaning and sampling techniques. The data cleaning has been performed by using a data science platform called the “Rapid Miner”.

VI. DATA CLEANING

Data cleansing is the method of detection and correcting (or removing) corrupt or inaccurate records from a dataset and refers to distinguishing incomplete, incorrect, inaccurate or unsuitable components of the data. So we replace, modify, or delete the dirty data that is present in the dataset. The entire process of the Data Cleaning Passes through four stages. They are

1. Identifying the Target column.
2.Removal of low-quality columns.
3.Replacing Missing Values
4.Principle Component Analysis
5. Normalization

Let us see how the data is modified in each stage of the data cleaning process.

1. Identifying the Target column: The Target column which is “Classification” has to be selected. It has two nominal values which are “Chronic Kidney Disorder” or “Not Chronic Kidney Disorder”. The dataset has to train and learn based on this target column.

2. Removal of low-quality columns: The columns of “High Stability” and “Many Values” are to be removed. These values have to be removed as they don't contribute to the analysis.

3. Replacing Missing values: The data columns with missing values will not contribute to the analysis of machine learning. This phase of filling missing values has already been done using the imputes.

4. Principle Component Analysis: The Principle Component analysis [5] is a mathematical technique that helps is reducing the data dimensionality. This method PCA achieves dimension reduction by making new, artificial variables known as artificial variables.

5. Normalization: Data normalization [6] is that the method of rescaling one or additional attributes to the vary of 0 to 1. The interpretation would be easy when a standard normal distribution is used.

VII. DATA SAMPLING

Data sampling [7] is a statistical analysis technique that can be used to select a subset of data. Data sampling enhances the execution speed and produces easily visual sable data when handling large datasets.

Classification model	Percentage Split	10-folds Cross-Validation
Naive Bayes	100%	90%
1R	92.8%	95.2%
Decision Table	92%	95%

Table 3 Accuracy after Cleaning and Sampling

The accuracy of the three classification models after the Data cleaning and Data sampling with both the Re-Sampling techniques has been shown in the Table 3. Therefore it has been predicted that the accuracy is higher for the classification model after performing Data cleaning and sampling.

VIII. CONCLUSIONS

The Chronic kidney disease data set has been preprocessed and cleaned. The preprocessing technique of imputation is used which the best fit for preprocessing medical data. The nominal values of the Dataset are replaced with the mode and the numerical values often dataset are replaced with the mean of the values. Normalization and Principal Component Analysis has been performed to get optimized output. We, therefore, see that a 100% of accuracy can be obtained by using the above-mentioned techniques.

REFERENCES

1. Frawley and G. Piatetsky -Shapiro, “Knowledge Discovery in Databases: An Overview”. *The MIT Press, Menlo Park, C.A 1996*
2. Hackernoon -Mohit Sharma (2018, July, 25). “What Steps should one take while doing Data Preprocessing?” Available: https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa
3. “DataFrame in pandas”-Tom, Available :https://pandas.pydata.org/pandas-docs/version/0.23.4/generated/pandas.DataFrame.html
4. Chronic Kidney Disorder. dataset https://www.kaggle.com/mansoordaku/ckdisease
5. MiodragLovric - “Principal Component Analysis” *International Encyclopedia of Statistical Science.*

6. S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas - “Data Preprocessing for Supervised Leaning ”*International Journal Of Computer Science Volume 1 Number 1 2006 ISSN 1306-4428J.*
7. Margaret Rouse – “Data Sampling” – Available : https://searchbusinessanalytics.techtarget.com/definition/data-sampling
8. DursunDelen, Glenn Walker, Amit Kadam – “Processing data using imputers” Available : http://lijiancheng0614.github.io/scikit-learn/modules/generated/sklearn.preprocessing.Imputer.html
9. Re-Sampling - “Percentage Split or Holdouts” Published by https://gerardnico.com/data_mining/validation_set
10. Jason Brownlee, “A Gentle Introduction to k-fold Cross-Validation” –Available : https://machinelearningmastery.com/k-fold-cross-validation

AUTHORS PROFILE

Manda Arpithais a B.tech, Final year in the field of Computer Sciences at Karunya Institute of Technology, Coimbatore. Her Interest is in the field of Machine Learning and Data Mining.

Dr. K. Ramalakshmi has completed her B.E degree from Madurai Kamaraj University and M.E degree from Anna University and Phd in Information and communication Engineering from Anna University area of research include data mining, database, wireless sensor network, artificial intelligence

R. Venkatesan. has completed his B.E degree from Madurai Kamaraj University and M.E degree from Anna University and Phd in Information and communication Engineering from Anna University area of interest IOT, could computing, datamining, machine leaning