International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Understanding Clinical Data using Exploratory Analysis

Owk Mrudula, A.Mary Sowjanya

Abstract*: In today's world the data plays an indispensable role. The proper understanding of data and its interpretation lays the foundation for the growth and also the success of company or an organization. As in domains such as business, finance and banking, health sector also produces huge amounts of data. This data needs to be properly analyzed and summarized before the data is modeled for a specific purpose. Generally, clinical data involves stakeholders like doctors, technicians, lab analysts, hospital managers, care providers and insurance agents. Exploratory Data Analysis plays an important role in providing the complete picture of the dataset along with identifying new insights and hidden patterns in the data. As such it becomes the most significant step before actually preprocessing the data. In our paper we have implemented EDA on Statlog heart disease dataset to identify the important variables, correlations between any variables, missing values, outliers and PCA. To verify, whether the process of EDA actually impacts the performance we have utilized machine learning algorithms like Naïve Bayes, Logistic regression, Decision Tree, Support Vector Machine, Random forest. Results indicate that the performance of the prediction model considerably increases after performing EDA regardless of the type of prediction algorithm used. Also the analysis of the dataset with graphical results helps the stakeholders to make better decisions regarding their patients and their treatments. Understanding any clinical data before modeling would prevent erroneous models later and exploratory analysis helps in achieving it.

Keywords: Data Analytics, EDA, Variable importance, Missing data, Outliers, Machine Learning, Clinical data.

I. INTRODUCTION

Almost every area has the applications of the data science, nowadays. The term "Exploratory Data Analysis" was coined by Turkey. EDA can be defined as the art and science of performing initial investigation on the data by means of statistical and visualization techniques that can bring out the important aspects in the data that can be used for further analysis [1].The well highlighted aspects of data science are the various statistical and machine learning techniques applied for solving a problem. For any data science application, activity starts with an Exploratory Data. In this paper, we are going to discuss in detail about concepts of EDA.

Manuscript published on January 30, 2020.
* Correspondence Author
Owk Mrudula*, Ph.D, College of Engineering (A), Department Computer Science, Andhra University.
A.Mary Sowjanya, Assistant Professor in College of Engineering (A), Andhra University.

Exploratory data analysis (EDA) is an essential step in any research analysis. Its primary aim is to examine the data for distribution, outliers and anomalies to direct specific testing of your hypothesis. Feature selection techniques are also required for EDA Analysis. According to John W.Tukey, a mathematician who first coined the term Exploratory Data Analysis. This is a very important step regarding the analytics process; it helps us to make senses of our data. Before performing a formal analysis, it is very
valuable to explore a dataset. No first models should be done without a proper EDA. EDA is a necessary and informative step in any data analysis. "Most EDA techniques are graphical in nature with a few quantitative techniques. The main role of EDA is to explore, and graphics gives the analysts unparalleled power to do so, while being ready to gain insight into the data. There are many ways to categorize the many EDA techniques" [6]. The objectives of EDA can be summarized as follows:

1. Maximize insight into the database/understand the database structure;
2. Visualize potential relationships (direction and magnitude) between exposure and outcome variables; 3. Detect outliers and anomalies (values that are significantly different from the other observations);
4. Develop parsimonious models (a predictive or explanatory model that performs with as few exposure variables as possible) or preliminary selection of appropriate models;
5. Extract and create clinically relevant variables.

II. LITERATURE REVIEW

Currently, huge voluminous information on healthcare field in different formats and data types is saved in datasets. This is due to the advancements in technology. It is practically infeasible for an individual physician to analyze the information through conventional methods. Description and targets of healthcare mining methods are presented. This work presents mining targets in light of duration, precision, decision support mechanism and obtaining useful insights. Potential aspects of information and its issues directly impact the quality of obtained insights. The basic limitation of data is that incorrect feed results in incorrect outcomes. Several classification methods have been implemented to diagnose methods for heart disease [2] [3]. A diagnosing system for heart disease generally contains two important parts: feature selection and classification. Feature selection enables to curtail unwanted features, which are not significant and thereby provides efficient classification/prediction. Chosen attributes is correspondingly utilized as input aimed at classification methods.

Thereby classification is performed effectively as per characteristic features chosen for it [3] [4] [2] [9]. A recent study states in the year 2010, considering all the reasons of deaths [5] in the world, the contribution of cardiovascular disease (CVDs) was greater comparatively. A study also finds that by the year 2030 about 76% of deaths in the world will be due to non-communicable diseases (NCDs) [6] of which cardiovascular related diseases are on the rise due to, alcohol consumption, use of tobacco, lack of exercise, diet, etc. [7]. Medical and health care professional's capability of accurately diagnosing heart disease is multiplied with an implementation of strategies, and techniques of data mining such as, classification, regression, neural networks (NN), decision trees (DT), genetic algorithms (GA), support vector machine (SVM) etc. However, a reduction of test data and number of class labels shows better performance using, Decision Tree (DT) and a comparable precision sometimes is obtained with Bayesian classification. In this test, 909 heart diseases patient records are considered where the dataset is divided into two sets equally, training dataset comprising of 455 records and testing dataset comprising of 454 records. In the learning process 13 regular attributes are used [8] with 2 class labels "Heart Disease" and "No Heart Disease". The prediction process using the algorithms DT and NBC based on Weighted Associate Classifier (WAC) attains accuracy of 81.51% maximum. The accuracy is improved to 99.2% with DT technique, compared to other techniques when data size is reduced and attributes reduced to 6 from 13 with genetic algorithm (GA).

III. PROPOSED SYSTEM

Today's healthcare organizations are submerged by a huge amount of data like hospital data, patient data, insurance data etc. Hospital database management systems store their daily basis of data in electronic form which may be in form of, structured database, text or images. Since EDA is a fundamental early step after data collection and preprocessing it helps to assess the quality of the data before model building. Since the data needs to be interpreted and used by different users like doctor's, technicians, insurance agents EDA techniques can either be graphical or nor non graphical. Non graphical methods are called descriptive statistics as they provide insights into the distribution and the characteristics of the variables in the data whereas graphical methods provide more qualitative and a complete picture of the data. The steps in Exploratory Data Analysis process are as follows

A.Data Exploration

1. Variable importance

We identify the input, output, the datatype and the category of all the variables in the dataset.

2. Univariate Analysis

It examines one variable in the data at a time. From the previous step if the variables are continuous characteristics like central tendency, measure of dispersion are studied and histogram, boxplots are used to depict the various statistical metrics. If the variables are categorical a frequency table containing the count, frequency of the data of each category is built and bar chat is used to depict it.

3. Bivariate Analysis

We look for association between two variables by using scatterplot, correlation, covariance and variance.

B.Handling Missing values

When a dataset has missing values it may lead to wrong prediction which becomes dangerous especially in health care datasets. As such missing values in the dataset need to be identified and treated accordingly. Generally health datasets have missing values at the time of data collection. These values can be deleted but it may considerably reduce the size of the dataset depending upon the number of the missing values. Hence we have used imputation methods to fill in the missing values using MICE.

C.Outliers Detection

Since outliers are observations that do not fit with the overall data they need to be detected as they can impact the results of data analysis. In health datasets errors in data entry, instrument measurements, experiments, data processing lead to outliers. We used visualization methods like quantile plot and boxplot to detect outliers. Once the outliers are detected they can either be deleted or imputed. Since the outlier values are due to error and very small in number we have detected the observations from the original data.

D.Feature Engineering

This is used to extract more information from already existing data. Variable transformation is done to change the scale or standardized the values of a variable for better understanding and implementation. Logarithmic, square slash cube root or binning methods can be used. Feature creation can be used to generate new variables which highlight the hidden relationship in a variable. In contrast, feature reduction reduces the amount of data by identifying less important variables through techniques like common factor analysis and Principal Component Analysis (PCA). The main aim of this work is to support EDA as an extension to data preprocessing for performing efficient data analysis and check for the quality of data without user interference.

IV. RESULTS AND DISCUSSION

We have implemented Exploratory Data Analysis on the Statlog dataset containing 303 observations and 14 variables [11]. The dataset was first preprocessed then data exploration, Handling of missing values and outliers and feature engineering was performed. Accuracy, Precision, Recall, F-Score have been calculated for the different types machine learning algorithm models using Naïve Bayes, SVM, Logistic Regression, Decision Tree classifiers, and Random Forest have been applied and compared with each other [10]. The performance metrics indicate that after performing Exploratory data analysis on the models gives higher accuracy on the given input data set. Performance metrics like Accuracy, Precision, Recall, and F-measure have been calculated to show the performance of the model before and after performing the Exploratory Data Analysis.

```
> str(Statlog)
Classes 'tbl_df', 'tbl' and 'data.frame':       270 obs. of  14 variables:
 $ age                                          : num  70 67 57 64 74 65 56 59 60 63 ...
 $ sex                                          : num  1 0 1 1 0 1 1 1 1 0 ...
 $ chest pain type                              : num  4 3 2 4 2 4 3 4 4 4 ...
 $ resting bp                                   : num  130 115 NA 128 120 120 130 110 140 150
 ...
 $ serum cholestral                             : num  322 564 261 263 269 177 256 NA 293 407
 ...
 $ fasting blood sugar                          : num  0 0 0 0 0 0 1 0 0 0 ...
 $ resting electrocardiographic results         : num  2 2 0 0 2 0 2 2 2 2 ...
 $ maximum heart rate achieved                  : num  109 160 141 105 121 140 142 142 170 15
4 ...
 $ exercise induced angina                      : num  0 0 0 1 1 0 1 1 0 0 ...
 $ oldpeak                                      : num  2.4 1.6 0.3 0.2 0.2 0.4 0.6 1.2 1.2 4
 ...
 $ slope of the peak exercise ST segment        : num  2 2 1 2 1 1 2 2 2 2 ...
 $ number of major vessels (0-3) colored by flourosopy: num  3 0 0 1 1 0 1 1 2 3 ...
 $ thal                                         : num  3 7 7 7 3 7 6 7 7 7 ...
 $ class                                        : num  2 1 2 1 1 1 2 2 2 2 ...
> |
```

Fig1: Structure and type of variables in the dataset.

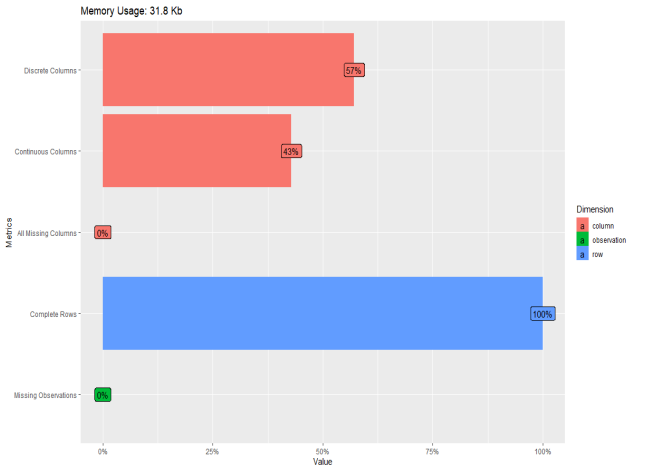

Fig2: Continuous and discrete columns for data exploration.

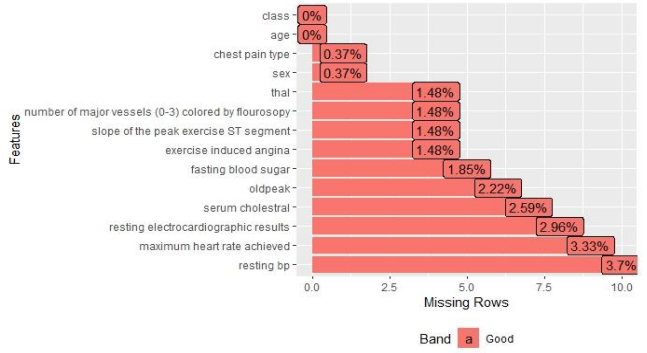

Fig3: Missing observations in variables for data exploration.

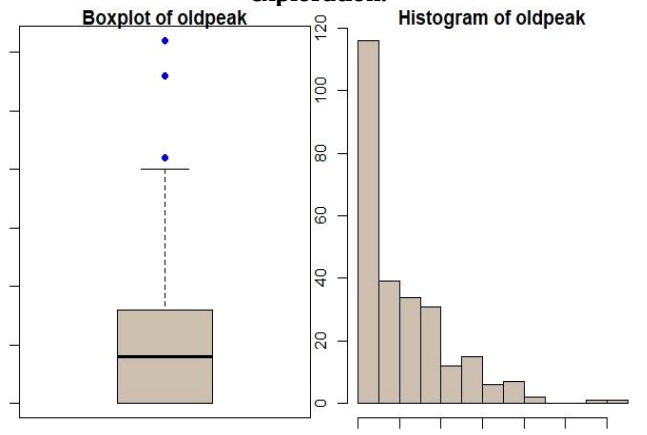

Fig4: Outliers identification and visualize using boxplot and histogram.

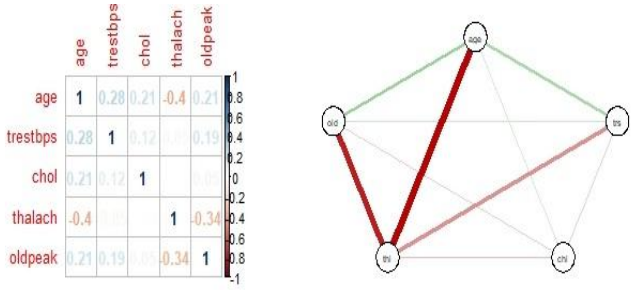

Fig5: Correlation matrix and network mapping for important variables.

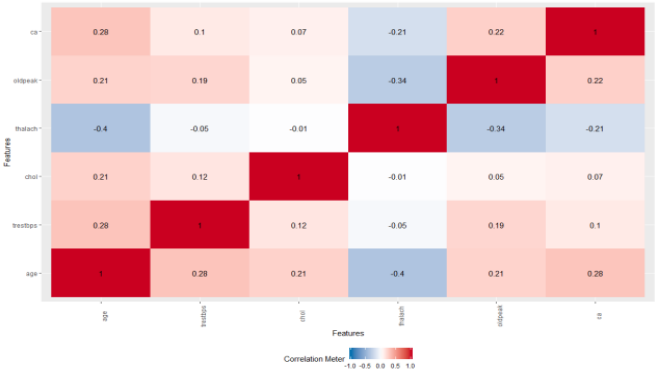

Fig6: Correlation matrix for the dataset.

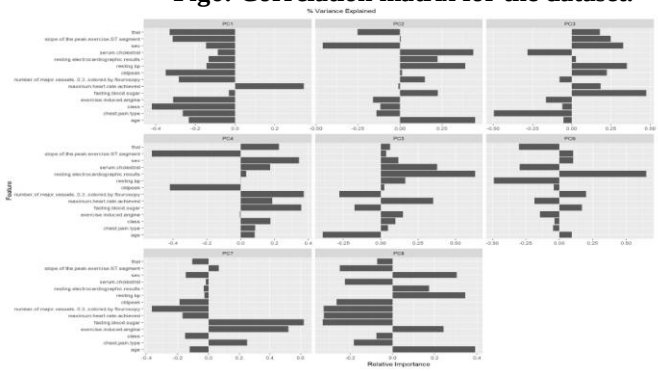

Fig7: PCA for the variables in the dataset.

Classifiers	Accuracy	
	Without EDA	With EDA
Naïve Bayes	0.830	0.853
Regression	0.859	0.892
Decision Tree	0.828	0.831
SVM	0.652	0.797
Random Forest	0854	0.909

Fig9: Results using Classification Models without and with performing EDA.

V.CONCLUSIONS

This paper highlights the use of EDA on clinical datasets for better understanding of the stakeholders. We first perform EDA on Statlog heart disease dataset and then build the model to predict whether a patient has heart disease or not. For this purpose different machine learning algorithms like Naïve Bayes, Logistic Regression, Decision Tree, Support vector machine, and Random forest have been used.

From the above visualization, it can be seen that the dataset has been thoroughly analyzed. The results also indicate an increase in accuracy after performing EDA, In future, in depth study of various EDA techniques with huge amounts of data can be used to provide deterministic results for the classifiers accuracy.

REFERENCES

1. Hands-On Exploratory Data Analysis with R: Become an expert in exploratory data analysis using R packages, Radhika Datar and Harish Garg.
2. Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Computational intelligence for heart diagnosis: A medical knowledge driven approach Expert Systems with Applications, 40(1), 96-104.
3. Shilskar S. & Ghatol A. (2013). Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. Expert Systems with Applications, 40(10), 4146-4153.
4. Sanz, J. A., Galar, M., Jurio, A., Brugos, A., Pagola, M., &Bustince, H. (2014). Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. Applied Soft Computing, 20, 103-111.
5. Mathers, C. (2008). The global burden of disease: 2004 update. World Health Organization.
6. World Health Organization, Public Health Agency of Canada, & Canada. Public Health Agency of Canada. (2005). preventing chronic diseases: a vital investment. World Health Organization.
7. Srinivas, K., Ranin, B. K., &Govrdhan, A. (2010). Applications of data mining techniques in healthcare and predication of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255.
8. Patil, B., Kumaraswamy, Y. S., Soni, J., Ansari, U., & Sharma, D. (2011). Predictive data mining for medical diagnosis of heart disease prediction, IJCSE, 17.
9. Khemphila, A., &Boonjing, V. (2011, August). Heart disease classification using neural network and feature selection. In 2011 21st International Conference on Systems Engineering (pp. 406-409). IEEE.
10. https://archive.ics.edu/ml/machine-learning-databases/heart-disease/
11. Statlog + (heart). (2017). Retrieved from http://archive.ics.uci.edu/ml/datassets/statlog+ (heart).

AUTHORS PROFILE

O.Mrudula has done her B.Tech in Information Technology from M.V.G.R College of Engineering and M.Tech in Computer Science from Andhra University. She is at present working on her Ph.D in College of Engineering (A), Computer Science Department from Andhra University. Research is in the area of Data Analytics.

Dr.A.M.Sowjanya has done her B.Tech and M.Tech in Computer Science. Her Ph.D is in Incremental clustering. She is at present working as an Assistant Professor in College of Engineering (A), Andhra University. Her research interests include Data Analytics, Machine Learning and Data Mining.