*International Journal of Engineering Applied Sciences and Technology, 2022*
**Vol. 7, Issue 4, ISSN No. 2455-2143, Pages 121-126**
Published Online August 2022 in IJEAST (http://www.ijeast.com)

# HEART DISEASE DATA ANALYSIS USING EXPLORATORY DATA ANALYSIS METHOD

Uma K
Research Scholar
Department of Computer Science and Applications
Bangalore University, Bangalore, Karnataka, India.

M Hanumanthappa
Professor
Department of Computer Science and Applications
Bangalore University, Bangalore, Karnataka, India.

***Abstract*— Health care industry companies generate a large volume of raw data, often called big data, which reveals hidden patterns and knowledge for decision making. Data-driven decisions are more accurate than intuition because they use massive data. In exploratory data analysis, mistakes are detected, data is identified, assumptions are checked, and the correlation between the variables is determined. Analyzing data without inferences or statistical modeling is considered Exploratory Data Analysis in the context. Analysts forecast the future and reveal hidden patterns in any profession. In the recent past, data analytics can be seen as a technology that is accessible, and it is essential to healthcare, especially current findings in outbreaks and emergency circumstances. Exploratory data analysis is a crucial step when analyzing data, and the application of analytics in healthcare improves treatment by enabling preventive care. The elements that lead to heart disease are predicted in this paper. The research uses two sets of publicly accessible heart disease data. Two datasets contain 303 records with 13 attributes and 4238 with 16 attributes.**

***Keywords*— Data Analysis, Exploratory Data Analysis, Data Mining, Heart Disease.**

## I. INTRODUCTION

The Healthcare industry generates an abundance of data through daily activities. Though data is voluminous, required information for the decision-making system is needed to discover from voluminous data. Data mining is one technique to explore the information required from big data. In data mining, data analysis is one of the main phases of the process. Analyzing the data precisely helps with data mining tasks and improves the model accuracy. One of the most dangerous and silent killer diseases is heart disease all over the globe. Every hospital and clinic collects patient data arbitrarily based on their visits. The data consists of patient demographical data, lab reports, medicine details, disease, treatments, and more.

### *A.* Heart Disease

According to the Health Organization's factsheet, heart disease has been among the leading causes of death globally since 1999. There is no end to the spread of this deadly disease. Heart disease has caused more deaths than any other disease [1]. In addition to the nature of significant risk factors such as blood pressure, cholesterol, smoking, physical activity, and diet, different countries also differ in the death rate from heart disease. 80% to 90% of the people who die of heart disease have at least one risk factor that results from lifestyle choices, regardless of genetics. Early detection and treatment can result in significant life savings when done correctly. People in older age groups are at greater risk for heart disease. A recent study shows that heart disease can be effectively controlled if detected early. Due to many complicated factors associated with heart diseases, it is hard to make a perfect analysis. In addition to heart diseases, other syndromes may result from heart disease. This condition complicates the diagnosis process, thus requiring an automated solution to aid in the diagnosis process.

### *B.* Exploratory Data Analysis (EDA)

EDA is the procedure for reviewing a dataset's fundamental properties using visual techniques. EDA is used to see what the data might tell us before we start the modeling effort. EDA can be categorized in two different ways. The first difference is whether each method is graphical or non-graphical. Second, every tactic is either univariate or multivariate. The exploratory data aims to find patterns, outliers, and abnormalities in the data. It also provides tools for creating hypotheses by understanding data through graphical representation and graphically visualizing it [2] [3].

After data collection and preprocessing, an essential phase called EDA occurs where data is displayed, plotted, and updated without any restrictions to aid in assessing the data quality and building models. Non-graphical strategies typically require the computation of the quantitative variables; however, graphical methods convey the facts graphically or visually. While multivariate techniques look at relationships between two or more variables simultaneously, univariate approaches focus on relationships between one variable at a time. Bivariate is the norm for multivariate EDA, but three or more variables may be included in rare times. Run a univariate EDA on each of the multivariate EDA's constituent parts before the multivariate EDA is complete.

The main motives for applying EDA are as follows:
Spotting errors; testing assumptions; preliminary model selection; determining relationships between explanatory factors, and analyzing the direction and estimating the size of links between explanatory and outcome variables [4].

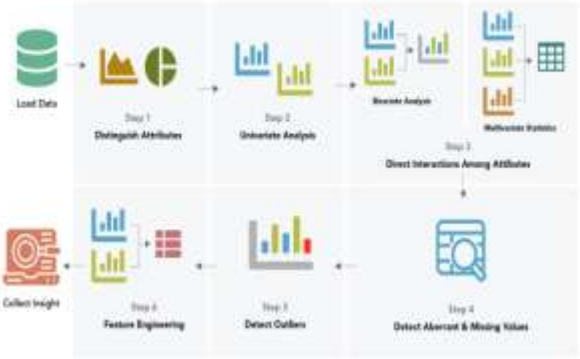


Figure.1. Exploratory Data Analysis process.

## II. RELATED WORK

Chih-Wei Huang et al. [2015] has proposed data analysis work on Chronic Kidney Disease using EDA and visualization technique [4]. The authors used a divide and conquer approach to group patients into homogenous subsets. The proposed model shows the automated correlation and visual analysis for kidney disease. And it also explores the patients suffering from other conditions over time. For visualization Sankey diagram is used to evaluate the knowledge. The researchers defined the factors for patient trajectories and preprocessed and partitioned the patient group. Finally, the cohort trajectory network filters the edges and visualization of knowledge.

R Indrakumari et al. [2020] have done exploratory data analysis utilizing K-means clustering to study heart disease with the help of the Tableau tool [5]. The authors chose the publicly accessible heart disease data used in the analysis. The data consists of 209 records with eight characteristics, including age, blood pressure, blood glucose level, resting ECG, heart rate, and four different types of chest pain. A K-means cluster analysis is performed on the dataset using the visualization tool tableau.

Amarpreet Singh Arora et al., [2021] has worked on the analysis of coronavirus spread in South Korea using COVID-19 data [6]. The dataset for analysis was collected from different sources. The authors study the data on the pattern of South Korea COVID cases spread and other countries. The research is visualized using OriginPro 2016 software tool to understand better the result, including overall confirmed, active cases, death cases, and recovered patient numbers. The distribution of patients gender-wise and age-wise was also plotted to get knowledge about how gender and age could be affected by COVID. Lastly, they took over the cases based on the traveler's history and shoed at the graph.

## III. METHODS

### *C.* Data Collection:

- The two heart disease datasets were collected online from UCI (University of California, Irvin).
- Machine Learning Repository called Cleveland heart disease dataset.
- Kaggle repository data called Framingham heart disease dataset.

The dataset contains diverse instance formats and attributes. Cleveland dataset consists of 303 records with 76 raw attributes, including one predicted attribute. Out of 76 attributes, only 13 attributes are important to predict the disease. Framingham dataset comprises 4238 records with 14 attributes. After data collection, the data should be put into the data warehouse as target data that are domain specific.

### *D.* Data Preprocessing:

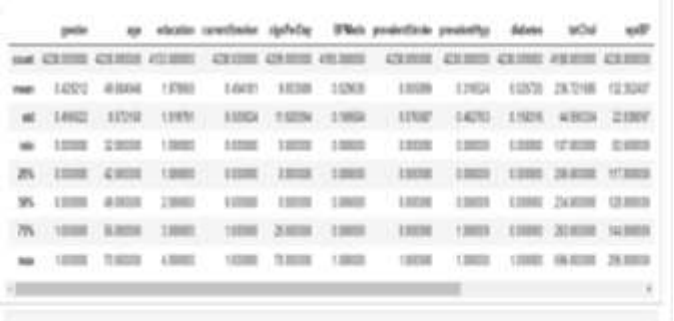
Data preprocessing is a data mining step that covers data reduction, integration, transformation, and cleansing. The collected heart disease dataset may contain missing, noisy, inconsistent values. Some entries in the dataset used for analysis were missing; these missing records were found and replaced with fair values using the proper techniques. Cleveland data does not contain null values but includes some noisy data. Framingham data has 582 null values. Since it is only 14 % of the data, we deleted it.

### *E.* Exploratory Data Analysis for Heart Disease data

Heart disease has been the top cause of death globally over the past 20 years. But right now, it's killing more people than in the past. Nearly 2 million more people have died from heart disease since 2000, bringing the total to almost 9 million in 2019 [who]. Heart disease prediction using data mining techniques produces the desired output for the decision support system. Before applying the data mining task to any dataset, one must do the analysis [7] [8]. Once the data is understood and analyzed precisely, the next phase is easy. Evolutionary data analysis was performed on heart disease data in this research work. This research explores the complete heart data analysis using evolutionary data analysis. The study has been done for each attribute of the dataset and how they relate to heart disease.

Dataset 1: Framingham Dataset

The Framingham dataset comprises 4238 records and 16 columns. The features include gender, age, education, current smoker, cigsperday, BP medication, Stroke, hypertension, diabetes, total cholesterol, systolic and diastolic BP, body mass index, heart rate, and blood sugar levels are all familiar. Overall, 14% null values are found in the dataset, i.e., 582 records. Since this study mainly focuses on the analysis phase, null or missing values are not treated, just deleted. The primary data analysis is described to identify the statistical analysis by finding the total number of records presenting min, max, mean, and standard deviation values for all attributes.



| | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |

Fig.2. Statistical summary of Framingham dataset.

The basic study of the dataset shows the mean value for age is 49 and the age range between 32 and 70. The target variable is showing 85-15 variations in each class.

The risk of cardiovascular disease among adults and older, which affects the heart, blood vessels, or both, is higher than it is among younger people. Developing cardiovascular disease may increase a person's risk as their heart and blood vessels change with age. The following bar chart indicates the age in data has more people in the age of 40, least in the age of 52 and 43, 44, 45 are the moderately distributed. Grouping people based on age for more straightforward analysis. Heart disease affects young individuals the least and middle-aged persons the most [9].

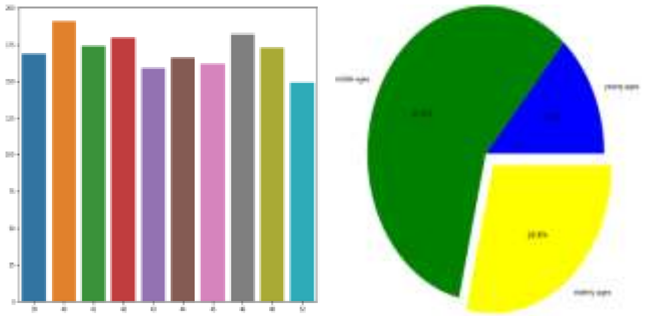


Fig.3.Age distribution.

Gender distribution in the dataset shows that men are affected more than women. In the following bar chart, gender '0' indicates female, and '1' means male.

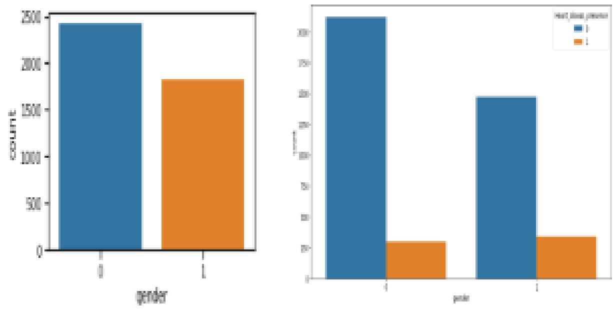


Fig.4.Gender Distribution

Additionally, diabetic patients are more likely to have additional heart disease risk factors: Blood flows through arteries more forcefully with high blood pressure, which can harm arterial walls.

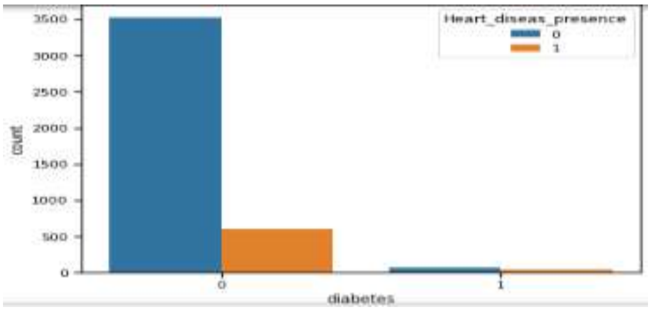


Fig.5. Diabetes v/s target variable

The prevalence of obesity plays a significant role in atherosclerosis and coronary artery disease. Heart failure is caused by structural and functional changes in the heart caused by obesity. Body Mass Index also affects heart disease badly. In the given data, the BMI falls 20-30 involved more.

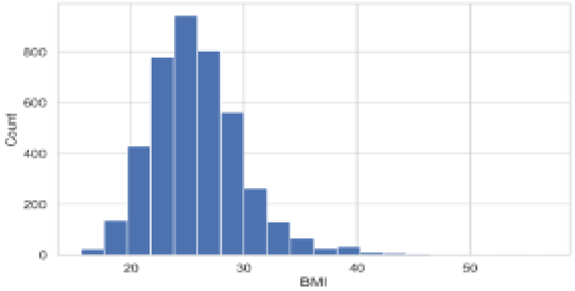


Fig.6. BMI analysis

When arteries become less elastic due to high blood pressure, they become less able to deliver blood and oxygen to the heart, which can lead to heart disease. From the scatter plot, we can see a positive relationship between the Systolic BP and the diastolic BP of the patient. In other words, the larger the sysBP diaBP also increases.

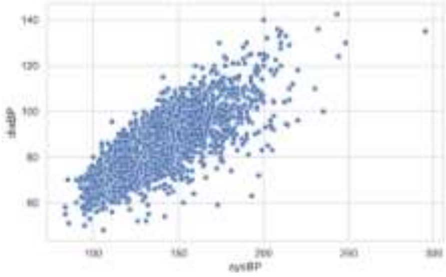


Fig. 7. Relation b/w sysBP and diaBP.

By looking at the scatter plot, we can see that heart rate is not much affected.

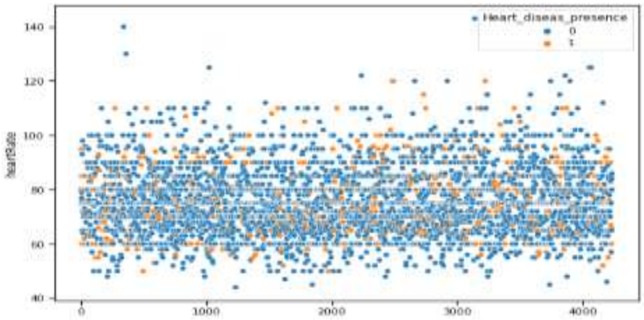


Fig.8. Heart rate v/s heart disease

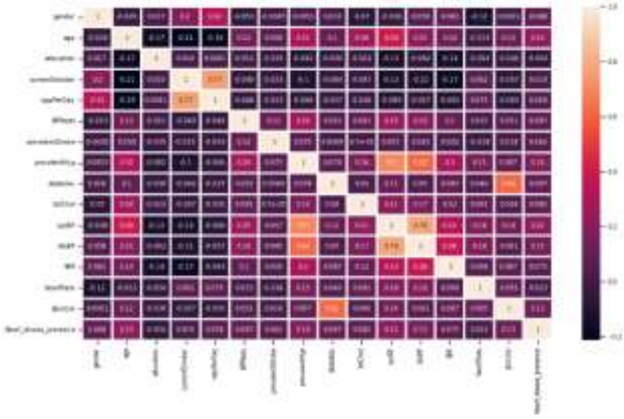


Fig.9. Heatmap.

The heatmap clearly shows that the data is not showing multicollinearity. Also, there is both positive and negative correlation with the data. BP and current smokers show a negative correlation where one increases and the other decreases. sysBP and diaBP are showing a positive relationship which means the more the sysBP more the diaBP of the patient.

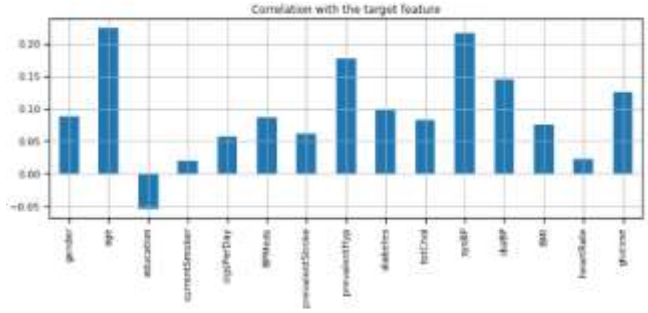
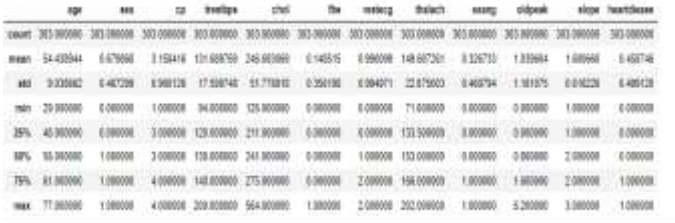


Fig.10. Correlation table

By the correlation with the target variable, it is seen that education has the highest negative correlation, i.e., education is not any chance to affect the heart disease. Age and SysBP has the highest positive correlation with the data. i.e., more the higher the sysBP more the chances of heart disease.

Dataset 2: Cleveland Data

Which is comprises 303 records and 13 columns. The attributes of heart the dataset includes gender, age, type of chest pain, resting BP, fasting BP, resting ECG result, cholesterol level, slope, thalach, ca, old peak, thal, and exang. The primary data analysis is described to identify the statistical analysis by finding the total number of records presenting Minimum, Maximum values with, Average, and standard deviation for all attributes. The mean value age attribute is 54, and the data range between 29 and 77. 50% of the data is facing non-anginal pain, and another 50% are asymptomatic, which implies they are not facing any chest pain. The target variable is showing 50- 50 variations in each class.



| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | heartdisease |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 |
| mean | 54.438944 | 0.679868 | 3.158416 | 131.689769 | 246.693069 | 0.148515 | 0.990099 | 149.607261 | 0.326733 | 1.039604 | 1.600660 | 0.458746 |
| std | 9.038662 | 0.467299 | 0.960126 | 17.599748 | 51.776918 | 0.356198 | 0.994971 | 22.875003 | 0.469794 | 1.161075 | 0.616226 | 0.499120 |
| min | 29.000000 | 0.000000 | 1.000000 | 94.000000 | 126.000000 | 0.000000 | 0.000000 | 71.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 |
| 25% | 48.000000 | 0.000000 | 3.000000 | 120.000000 | 211.000000 | 0.000000 | 0.000000 | 133.500000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 |
| 50% | 56.000000 | 1.000000 | 3.000000 | 130.000000 | 241.000000 | 0.000000 | 1.000000 | 153.000000 | 0.000000 | 0.800000 | 2.000000 | 0.000000 |
| 75% | 61.000000 | 1.000000 | 4.000000 | 140.000000 | 275.000000 | 0.000000 | 2.000000 | 166.000000 | 1.000000 | 1.600000 | 2.000000 | 1.000000 |
| max | 77.000000 | 1.000000 | 4.000000 | 200.000000 | 564.000000 | 1.000000 | 2.000000 | 202.000000 | 1.000000 | 6.200000 | 3.000000 | 1.000000 |

Fig.11. Statistical summary of Cleveland dataset.

Following figure.12. shows age distribution in data has more people in the age of 58, most minor in the age of 56 and 54,59,60 are the moderately distributed age.

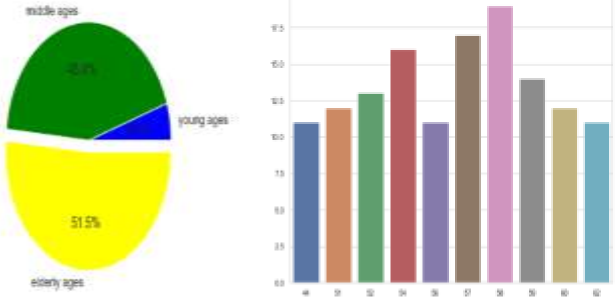


Fig.12.Age distribution

The Elderly are more prone to heart disease than the middle age and the young, whereas the young display a minute percentage of heart disease in the data. By this, we can conclude that age is one of the critical factors for heart disease.

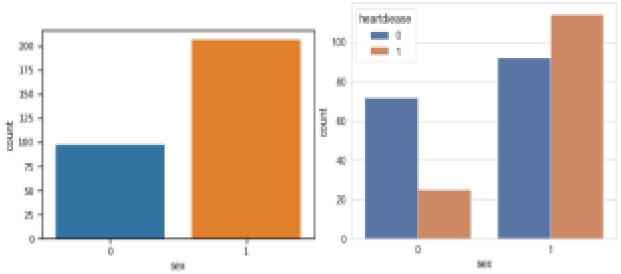


Fig.13. Gender distribution (0-female,1-male)

Sex distribution in the data is more of 70 - 30 split where the female percentage is less than the male portion. Heart disease analysis with sex implies that heart disease rates are more common in men than women.

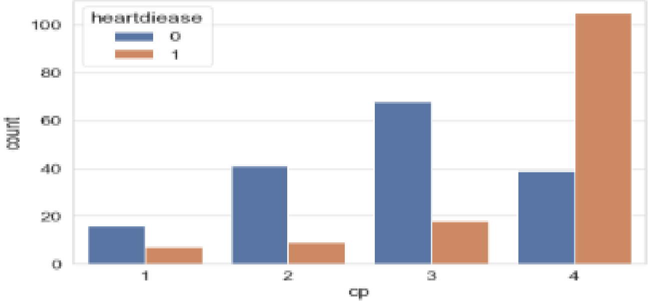


Fig.14. Chest pain type v/s target variable

There are four categories of chest discomfort.
1: Typical angina
2: Atypical angina
3: Non-angina pain
4: Asymptomatic

The above graph indicates clearly that the asymptomatic are showing the presence of disease more than once with the chest pain.

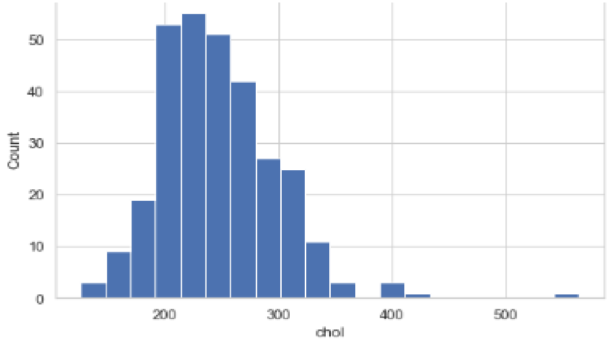


Fig.15. Cholesterol level distribution

The distribution for a continuous variable is seen above with variations from 0 to 600, and more people show cholesterol levels of 230 and 250.

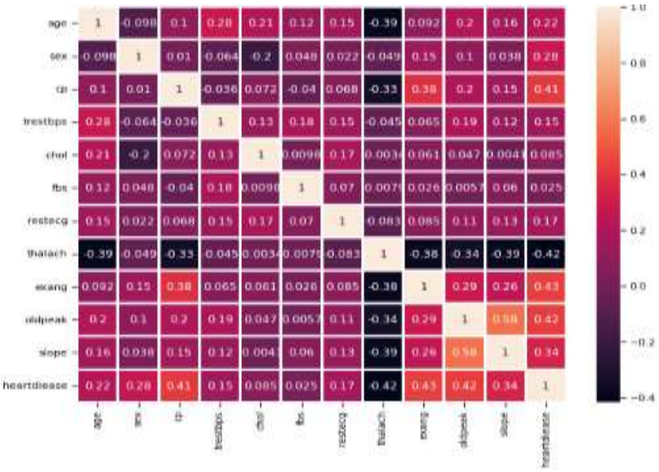


Fig.16. Heatmap.

The heatmap clearly shows that the data is not showing multicollinearity. Also, there is both positive and negative correlation with the data. Slope and thal show the negative correlation where one increases the other decreases, cp and exang show the positive relationship that means the more the exang more the chest pain.

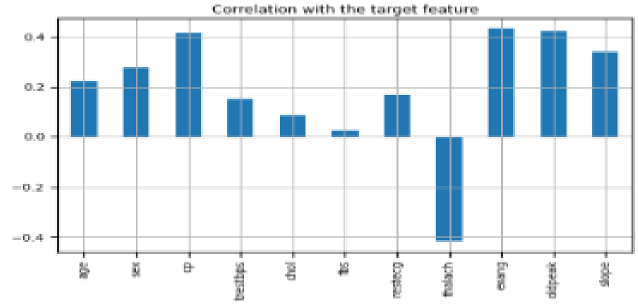


Fig.17. Correlation with the target variable.

The correlation with the target variable shows that the thalach has the highest negative correlation, i.e., if the thalach increases, the chance of heart disease is more diminutive. Exang has the highest positive correlation with the data, i.e., the higher the exang, the more cases of heart disease.

## IV. DISCUSSION

Heart disease is a silent killer disease that attacks with and without prior symptoms. It varies differently among different people. Heart disease risk is introduced with the following health conditions; High blood pressure, age, family history, smoking, poor diet, cholesterol level, diabetes, and obesity. The dataset chosen for this study is Cleveland data and Framingham heart disease data. Each dataset contains different attributes except age and gender. Both datasets are different from each other to indicate heart disease. The age distribution of Framingham data shows middle age people are affected more than young and older people. This condition suggests data in some way is unbalanced. Cleveland data shows older people are affected more than young middle age people. This dataset data is somehow balanced well. The gender distribution of Framingham data illustrates that female patients are more in number, and both men and women have diseases almost equal in ratio. In Cleveland data, male patients are in more number, and men are more prone to heart disease than women. Compared to the remaining features affecting heart disease in Framingham data are blood pressure, glucose, and diabetes. Similarly, in Cleveland, heart disease data features are chest pain, angina induced by exercise, old peak (ST depression), slope (ST segment), and restecg.

## V. CONCLUSION

This study mainly concentrates on the preprocessing and evolutionary data analysis process. Analyzed data will give a better result by improving the accuracy and performance of the model. Evolutionary data analysis will make processing data mining tasks easy and convenient.

## VI. REFERENCES

[1]. World Health Organization (2021). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2]. Komorowski M, Marshall D. C , J, Salciccioli J D and Crutain Y, (2016). Chapter 15- Exploratory Data Analysis - Secondary Analysis of Electronic Health Records. DOI: 10.1007/978-3-319-43742-2_15.

[3]. Valdiviezo-Diaz, P., Reátegui, R., Barba-Guaman, L., Ortega, M. (2022). Exploratory Data Analysis on Cervical Cancer Diseases. In: Botto-Tobar, M., Montes León, S., Torres-Carrión, P., Zambrano Vizuete, M., Durakovic, B. (eds) Applied Technologies. ICAT 2021. Communications in Computer and Information Science, vol 1535. Springer, Cham. https://doi.org/10.1007/978-3-031-03884-6_32.

[4]. Huang, CW., Lu, R., Iqbal, U. et al. (2015). A richly interactive exploratory data analysis and visualization tool using electronic medical records. BMC Med Inform Decis Mak 15, 92. https://doi.org/10.1186/s12911-015-0218-7.

[5]. R. Indrakumaria, T Poongodi and Sowmya Rajnan Jena, (2020). Heart Disease Prediction using Exploratory Data Analysis, International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020, Procedia Computer Science 173 (2020) 130–139.

[6]. Arora, A.S., Rajput, H. & Changotra, R. (2021). Current perspective of COVID-19 spread across South Korea: exploratory data analysis and containment of the pandemic. Environ Dev Sustain 23, 6553–6563. https://doi.org/10.1007/s10668-020-00883-y.

[7]. Katarya, R., Meena, S.K. (2021). Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. Health Technol. 11, 87–97. https://doi.org/10.1007/s12553-020-00505-7.

[8]. Valdiviezo-Diaz, P., Reátegui, R., Barba-Guaman, L., Ortega, M. (2022). Exploratory Data Analysis on Cervical Cancer Diseases. In: Botto-Tobar, M., Montes León, S., Torres-Carrión, P., Zambrano Vizuete, M., Durakovic, B. (eds) Applied Technologies. ICAT 2021. Communications in Computer and Information Science, vol 1535. Springer, Cham. https://doi.org/10.1007/978-3-031-03884-6_32.

[9]. Malik A M, Sagar A. K, and Sahana S, (2021). Prediction of Cardiopathy Using Exploratory Data Analysis, IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), pp. 117-122, doi: 10.1109/ICCCA52192.2021.9666241.