International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 8.141
Peer Reviewed Bi-Monthly
Vol. 6 No. 22 March - April - 2025

Exploratory Data Analysis of Cardiovascular Diseases

Roshani Atar[1] & D. S. Jadhav[2]
[1]*Assistant Professor, Department of Computer Science, PVG's College of Science, Pune*
[2]*Assistant Professor, Yashavantrao Chavan Institute of Science, Satara*
Corresponding Author – D. S. Jadhav
DOI -10.5281/zenodo.15533330

Abstract:
Cardiovascular diseases (CVD's) are disorder of the heart and blood arteries that are becoming a major cause of death and a global issue. Several aspects of the CVD epidemic in India are particularly worrying, including its rapid increase, the early onset of disease in the population, and the high case fatality rate. Statistical analysis is essential for predicting and detecting the risk of cardiovascular diseases, helping to raise awareness among the public. In healthcare sector, there is wealth of available data, but is the need to extract correct information from it to enhance medication discovery, diagnosis, treatment and overall care. Exploratory Data Analysis (EDA) helps in understanding the insights of dataset. In this research paper, we have used publically available UCI (University of California, Irvine) heart disease dataset for study. These findings provide valuable insights into the epidemiology of CVD and can inform targeted interventions for its prevention and management.

Keywords: Cardiovascular Diseases, Statistical Analysis, Exploratory Data Analysis, Medical Dataset

Introduction:

In today's era millions of people around the world suffer from different types of cardiovascular diseases (CVD's). According to World Health Organization (WHO), CVD is the primary cause of death in both males and females. CVDs are estimated to account for approximately 31% of all global deaths. Ischemic heart disease is the top cause of CVD mortality worldwide[1]. Statistical analysis has been increasingly cost-effective and plays a crucial role in healthcare, informing new research discoveries, managing emergencies, and identifying disease outbreaks. The World Heart Federation (WHF) underlines the importance of global collaboration and action in addressing the rising burden of CVDs, particularly in low- and middle-income countries where their impact is increasing. Since 2005, World health statistics 2023 report is assembling the health and health-related indicator from the Sustainable Development Goals (SDGs) and Thirteenth general programme of Work (GPW13) which has been published by WHO[2] .

Cardiovascular disease symptoms vary majorly between individuals and even within the same patient. Many persons with different risk factors do not notice any symptoms and may be ignore the problem. Neck and Jaw pain, Shortness of breath, dizziness, tingling in one or both arm are some of the signs or symptoms of CVD's. Also, Vomiting, muscular tremors, anxiety, chest pain, nausea, disorientation, and fatigue are possible symptoms of more severe types. If these symptoms ignored and untreated, all these risk factors can result in heart attacks, heart diseases, irregular heartbeat, persistent chest

discomfort (angina) and heart failure. These conditions can all cause sudden death. The key factors that put you at risk for cardiovascular disease include gender, smoking, age, family history, poor diet, lipids, and lack of physical exercise, high blood pressure, weight gain, and high alcohol consumption etc[3]. All heart related diseases can be managed by reducing and managing the mental stress, regularly check-up and consulting with health professionals, treating risk factors and managing other medical conditions.

Literature Review:

The data from global surveillance was compiled mostly from WHO or United Nations partner databases, with additional analysis from peer-reviewed papers. The researches were useful to embed cohort studies within the framework of surveillance projects. Also, real-time medical datasets gathered from different countries can be used to model development.

Bataineh et al. [4] introduced an algorithm that outperforms other algorithms such as Gaussian NB, Logistic regression (LR) , Decision tree(DT), Random forest (RF), Gradient boosting classifiers, K-nearest neighbors (KNN), XGB classifiers, Extra trees classifiers, and Support vector machine(SVM) on Cleveland dataset. Authors provided improved accuracy and speed in the prediction of heart disease to the clinicians.

In this study, Juhola et al. [5] extracted the data from calcium transient signals. Authors utilized Machine Learning (ML) techniques to classify and detect the anomalies in healthy profiles and diseased cardiomyocyte profiles. By using the Matlab tool Zriqat et al.[6] compared five classification algorithms like Naïve Bayes(NB), DT, Discriminant Analysis, RF, and SVM on the large dataset. Authors analyzed the significant relation between risk factors and heart disease. In their study, DT classifier gave 0.99% accuracy which is the highest accuracy among all classifiers.

Usha Rani et al. [7] proposed an innovative framework for case study that combined class based clustering to impute missing values with imputation method for the reduction of dimensionality in medical records.

Alfadli et al.[8] investigated early prediction and prevention of most critical factors between normal and heart disease patients with minimum set of records from UCI dataset. Also, authors developed hybrid model which achieved 84.24% accuracy.

Chaurasia et al.[9] found the accuracy of detecting cardiac illness using Naive Bayes (Naive), J48, and bagging. The findings revealed that bagging provided an accuracy of 85.03%. Bagging exhibited higher predictive power than Naive Bayes.

Gadde [10] described the Naive Bayes, Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and Logistic Regression (LR) algorithms for heart disease prediction by using the WEKA tool on the UCI dataset.

Noorozi et al.[11] studied machine learning algorithms using UCI dataset and explained feature selection methods like filter and wrapper method. Authors found accuracy for LR is 91.65%, Bagging (85.03%), PSO-SSAE (96.1%), MLP-PSO (84.6%), Multilayer Perceptron (MLP)(87.28%) and RF is 85.01%.

Materials and Methodology Dataset Description:

In this research paper, we have used the data from the Cleveland Clinic Foundation. UCI dataset consists of 303 records, without any missing/unknown values [12]. This dataset contains 165

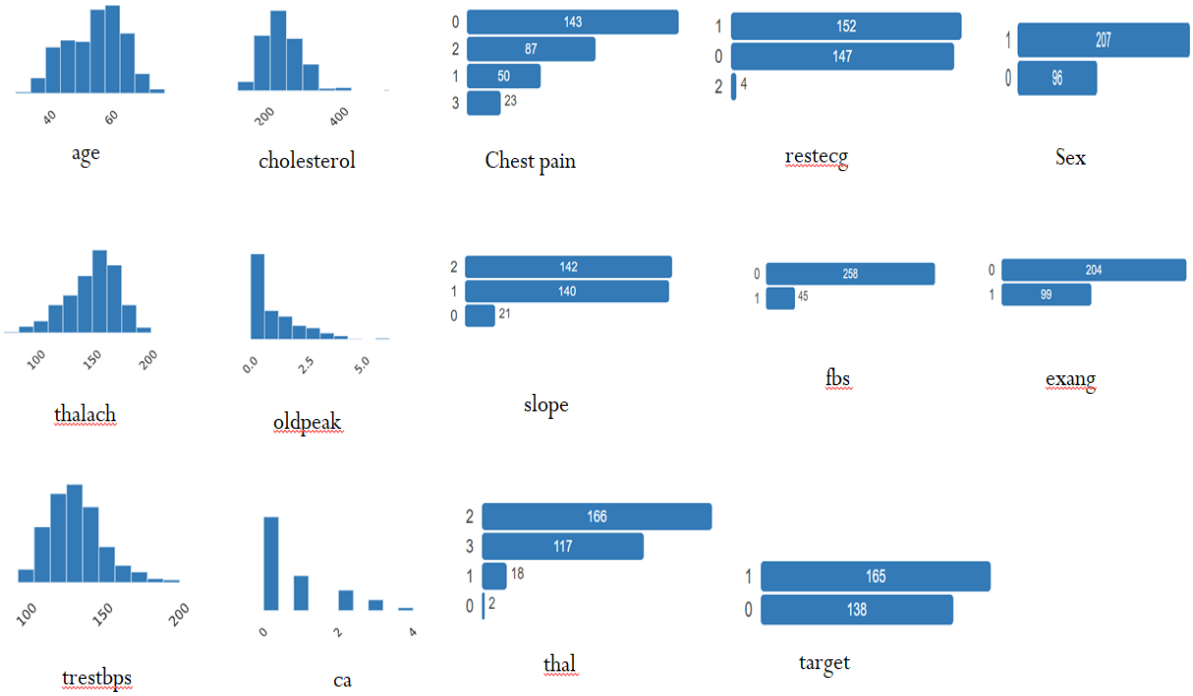
individuals with cardiovascular illness and 138 individuals with no cardiovascular history. This is a multivariate dataset, which involves a number of different mathematical and statistical variables, as well as multivariate numerical data analysis. The Cleveland UCI dataset includes a variety of studies on the prediction of heart disease. It is made up of 14 attributes: age, gender, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, old peak having ST depression caused by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels, and Thalassemia. Chest pain field indicates an integer that ranges from 0 to 4. The target field indicates whether the patient suffers from heart illness or not. Experiments with the Cleveland database have mostly attempted to distinguish between the presence and absence of disease.

Table: Heart disease data set with 14 attributes

Sr. No.	Name of Attribute	Datatype of Attribute	Description of Attribute
1	age	Quantitative	The patients age counted in years
2	sex	Qualitative	Counted in binary value(0=female,1=male)
3	cp	Qualitative	It shows Chest pain type having values ranges from 1 to 4. Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
4	trestbps	Quantitative	The patient's resting blood pressure (mm Hg at admission to the hospital).
5	chol	Quantitative	The person's cholesterol level in mg/dL
6	fbs	Qualitative	The fasting blood sugar level >120 mg/dl, 1 = true, 0 = false
7	restecg	Qualitative	Resting electrocardiogram findings. Value 0 indicates probable or definite left ventricular hypertrophy based on Estes' criteria. Value 1: normal Value 2: ST-T wave abnormality (T wave inversions, elevation, or depression of > 0.05 mV)
8	thalach	Quantitative	The individual's maximal heart rate
9	exang	Qualitative	Exercise-induced angina ,value can be 0 or 1 (1 = yes, 0 = no)
10	oldpeak	Quantitative	Exercise causes ST depression compared to rest ('ST' refers to places on the ECG plot).
11	slope	Qualitative	The slope of the peak exercise. ST segment: 0 (downsloping), ST segment: 1 (flat) ST segment: 2 (upsloping)
12	ca	Quantitative	Number of major vessels (0-3) colored by fluoroscopy. Attribute values can be 0 to 3

13	thal	Qualitative	A blood condition known as thalassemia. Value 0 indicates a previously dropped dataset, while Value 1 indicates a fixed deficiency (e.g., no blood flow in a specific area of the heart). Value 2 indicates regular blood flow. Value 3: reversible defect (blood flow is observed, although it is abnormal).
14	target	Quantitative	Patient's having heart disease or not 0 = no, 1 = yes

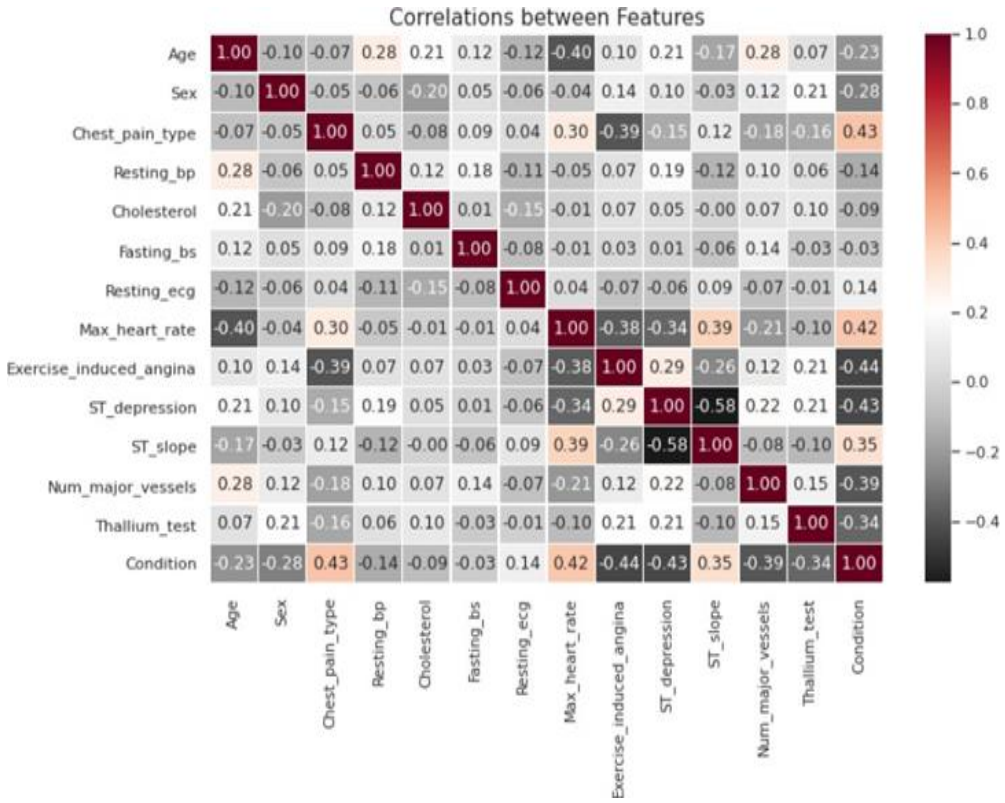
Exploratory Data Analysis of Dataset:



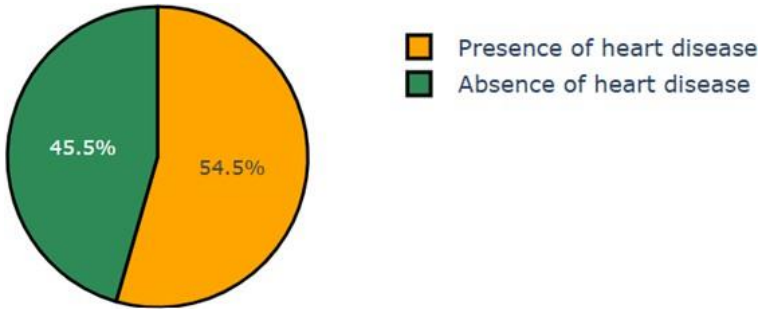
Due to the uniformity and global nature of the data set, only the missing value analysis are used as a pre-processing technique, and records with blank fields were eliminated from the data set. At this stage, the dataset has been filtered for missing data but there is no need to remove because dataset does not have the missing values.

Correlation Matrix:

Correlation indicates whether the attributes are related to one another or the goal variable. Correlation can be positive (an increase in one value improves the value of the objective variable) or negative (an increase in one value reduces the value of the target variable). Heatmap allows you to easily classify the features that are most important to the target variable.



We examined the quantity and proportion of each target value using a pie chart. It shows 138(45.5%) patients having absence of heart disease and 165(54.5%) patients having presence of heart disease.



The performance metrics used in this research work, namely, accuracy, mean absolute error (MAE), sensitivity (recall), precision, F-measure, and specificity are discussed. After applying ML algorithms we got the accuracy of Logistic Regression (LR) is 85.25%, Random Forest (RF) is 83.61%, and XGBoost gives 86.88%. In this research, we found Chi-square test help to assess the dependency between risk factors and heart disease presence. So, exercise induced angina and chest pain are the main causes of cardiovascular diseases.

Conclusion:

Exploratory Data Analysis (EDA) involves identifying errors and relevant data, validating assumptions, and assessing the relationships among explanatory variables. Also, investigate the contribution of risk factors for CVD. We analyzed the dataset and After applying the ML algorithms, XGBoost algorithm gives 86.88% accuracy and found this algorithm is better to predict disease compared with others. Future work can combine multivariate datasets to enhance the classification accuracy of other algorithms.

The health professional can help to assess any risks or associated conditions through statistical analysis in the future.

References:

1. C. H. Crisis, "20240502_World-Heart-Report_240628," 2024.
2. WHO, *World health statistics 2023: monitoring health for the sdgs, sustainable development goals*, vol. 27, no. 2. 2023. [Online]. Available: https://www.who.int/publications/book-orders.
3. A. Rezaianzadeh, L. Moftakhar, M. Seif, M. G. Johari, S. V. Hosseini, and S. S. Dehghani, "Incidence and risk factors of cardiovascular disease among population aged 40–70 years: a population-based cohort study in the South of Iran," *Trop. Med. Health*,
4. vol. 51, no. 1, 2023, doi: 10.1186/s41182-023-00527-7.
5. A. Al Bataineh and S. Manacek, "MLP-PSO Hybrid Algorithm for Heart Disease Prediction," *J. Pers. Med.*, vol. 12, no. 8, 2022, doi: 10.3390/jpm12081208.
6. M. Juhola, H. Joutsijoki, K. Penttinen, D. Shah, R. P. Pölönen, and K. Aalto-Setälä, "Data analytics for cardiac diseases," *Comput. Biol. Med.*, vol. 142, no. September 2021, 2022, doi: 10.1016/j.compbiomed.2022.105218.
7. I. A. Zriqat, A. M. Altamimi, and M. Azzeh, "A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods," vol. 14, no. 12, pp. 868–879, 2017, [Online]. Available: http://arxiv.org/abs/1704.02799
8. Y. Usharani and P. Sammulal, "A novel approach for imputation of missing values for mining medical datasets," *2015 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2015*, vol. 39, pp. 184–195, 2016, doi: 10.1109/ICCIC.2015.7435816.
9. K. M. Alfadli and A. O. Almagrabi, "Feature-Limited Prediction on the UCI Heart Disease Dataset," *Comput. Mater. Contin.*, vol. 74, no. 3, pp. 5871–5883, 2023, doi: 10.32604/cmc.2023.033603.
10. A. Vikas Chaurasia and I. Saurabh Pal, "Data Mining Approach to Detect Heart Dieses,"
11. *Int. J. Adv. Comput. Sci. Inf. Technol.*, vol. 2, no. 4, pp. 2296–1739, 2013, [Online].
12. Available: http://ssrn.com/abstract=2376653
13. H. Gadde, "Heart Disease Predictions Using Machine Learning Algorithms and Ensemble Learning," *Int. J. Eng. Trends Appl.*, vol. 7, no. 4, pp. 1–4, 2020.
14. Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," *Sci. Rep.*, vol. 13, no. 1, pp. 1–15, 2023, doi: 10.1038/s41598-023-49962-w.
15. I. Brandon Simmons, J. A. D Allagan, K. L. Jones, M. Talukder, and G. H. Del Villar, "Investigating Heart Disease Datasets and Building Predictive Models Committee Chair Committee Member," 2021.