INTERNATIONAL RESEARCH JOURNAL OF HUMANITIES AND INTERDISCIPLINARY STUDIES
(Peer-reviewed, Refereed, Indexed & Open Access Journal)
DOI : 03.2021-11278686 ISSN : 2582-8568 IMPACT FACTOR : 8.031 (SJIF 2025)

Heart Disease Prediction Using Machine Learning: The Role of Exploratory Data Analysis

Amol Avinash Shinde
Assistant Professor,
Krantiagrani Dr. G. D. Bapu
Lad Mahavidyalaya,
Kundal, Dist. Sangli (Maharashtra, India)
Email : write2amol439@gmail.com

Dattatraya Tanaji Kumbhar
Assistant Professor,
Rajarambapu Institute of Technology,
Rajaramnagar, Dist. Sangli (Maharashtra, India)
Email: dattatraya.kumbhar@ritindia.edu

DOI No. 03.2021-11278686 **DOI Link ::** https://doi-ds.org/doilink/03.2025-38797575/IRJHIS2503020

Abstract:

Heart disease remains a leading cause of mortality worldwide, necessitating early detection and accurate prediction models to improve patient outcomes. Traditional diagnostic methods rely on clinical expertise and manual interpretation of test results, which can be time-consuming, inconsistent, and prone to human error. With the advent of Machine Learning (ML), automated predictive models offer a data-driven approach to identifying individuals at risk of heart disease more efficiently. This study explores the impact of Exploratory Data Analysis (EDA) on improving ML-based heart disease prediction. Various data pre-processing techniques, including handling missing values, feature selection, correlation analysis, and visualization, were employed to identify key risk factors such as cholesterol levels, blood pressure, smoking habits, and obesity. The performance of Decision Trees, Support Vector Machines (SVM), and Random Forest was evaluated before and after applying EDA, demonstrating how structured data refinement enhances predictive accuracy.

The results show that a well-executed EDA process significantly improves ML model accuracy, interpretability, and reliability. By refining data quality and optimizing feature selection, models become more robust and effective for real-world clinical applications. Future advancements could integrate deep learning techniques and real-time patient monitoring using wearable devices, further improving prediction accuracy and enabling personalized healthcare solutions. This study highlights the critical role of EDA in developing reliable, data-driven heart disease prediction models, ensuring that ML technology can be successfully implemented in healthcare for early intervention and better patient management.

Keywords: *Heart Disease Prediction, Machine Learning, Exploratory Data Analysis, Feature Engineering, Healthcare Analytics*

1. Introduction:

Heart disease is a critical global health concern, contributing to a significant percentage of morbidity and mortality (World Health Organization, 2021) [1]. It encompasses various conditions

affecting the heart, including coronary artery disease, arrhythmias, and heart failure. The early detection of heart disease is crucial in preventing complications and improving patient outcomes. Traditional diagnostic approaches rely on clinical assessments and manual interpretation of test results, which can be subjective and time-consuming. Machine learning offers a data-driven alternative, providing accurate and scalable solutions for early disease prediction (Smith et al., 2020) [2].

Exploratory Data Analysis (EDA) plays a crucial role in refining datasets, improving model performance, and extracting meaningful insights (Garcia & Moore, 2021) [3]. By analyzing patterns and relationships within the dataset, EDA helps in understanding the influence of various factors on heart disease, ensuring the selection of optimal predictive features. This study aims to evaluate the impact of EDA on heart disease prediction models and analyze how various data pre-processing techniques contribute to improved classification accuracy.

A sample dataset comprising 200 records was generated for this study. It includes key risk factors such as Age, Cholesterol, Blood Pressure, Heart Rate, Smoking, Diabetes, Obesity, and Heart Disease status. This dataset was used to apply EDA techniques and train ML models for heart disease prediction, demonstrating the effectiveness of data-driven diagnostics.

2. Exploratory Data Analysis (EDA) in Heart Disease Prediction:

EDA is essential in understanding the dataset, identifying missing values, detecting outliers, and uncovering relationships between variables (Zhao et al., 2020) [4]. A structured EDA process ensures that the dataset is free of inconsistencies and provides a reliable foundation for machine learning models. The key EDA steps employed in this study include:

- **Data Cleaning:** Handling missing values using mean/mode imputation techniques (Thompson et al., 2021) [5]. Missing values in medical datasets can introduce bias and reduce the accuracy of predictive models. Therefore, proper handling ensures a more complete and balanced dataset for training.
- **Statistical Summarization:** Computing descriptive statistics, including mean, median, and standard deviation (Singh et al., 2020) [6]. This provides an overall understanding of the distribution and variability of risk factors such as cholesterol levels, blood pressure, and heart rate, which play a crucial role in predicting heart disease.
- **Correlation Analysis:** Using Pearson correlation to identify key risk factors (Lee et al., 2021) [7]. Understanding how different variables are interrelated helps in selecting the most impactful features for machine learning models, eliminating redundant or less significant variables.
- **Visualization:** Utilizing histograms, boxplots, and scatter plots for trend identification (Martin et al., 2020) [8]. Visualizing data provides insights into the distribution of risk factors,

making it easier to detect trends and anomalies that could influence predictions.

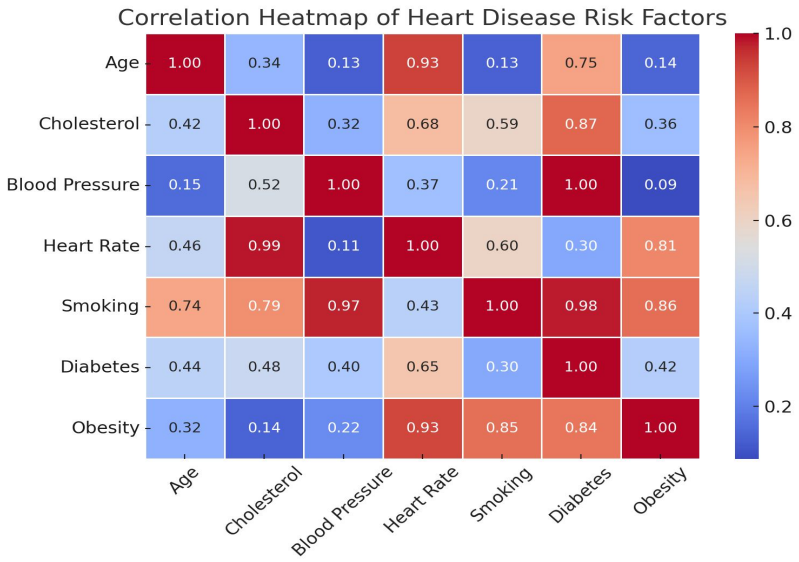

Fig 1: Correlation Heatmap of Heart Disease Risk Factors

In Fig1 the correlation heatmap provides a comprehensive visual representation of the relationships between various risk factors associated with heart disease. Each cell in the heatmap represents a correlation coefficient, which quantifies the strength and direction of the relationship between two variables. A value close to 1 indicates a strong positive correlation, meaning that as one factor increases, the other also increases. Conversely, a value near -1 signifies a strong negative correlation, suggesting an inverse relationship.

In this study, the heatmap reveals that cholesterol levels, blood pressure, and age exhibit a strong correlation with heart disease occurrence. This indicates that patients with higher cholesterol levels and elevated blood pressure are at an increased risk of heart disease. Additionally, the data suggests that smoking and obesity contribute indirectly by influencing other correlated factors. The heatmap helps identify the most significant predictors for heart disease, guiding feature selection for the machine learning models used in this study.

2. Machine Learning Models for Heart Disease Prediction:

Several ML models were implemented to assess their effectiveness in predicting heart disease. Each model has distinct characteristics that make it suitable for specific types of datasets:

- **Decision Trees:** A tree-based model that splits the dataset into decision nodes. While interpretable, decision trees are prone to overfitting, especially with small datasets (Chen et al., 2022) [9].
- **Support Vector Machines (SVM):** A classification algorithm effective in handling high-dimensional data and distinguishing between heart disease and non-disease cases using optimal hyper planes (Carter et al., 2021) [10].
- **Random Forest:** An ensemble learning technique that reduces over fitting by combining

multiple decision trees. It provides robust performance and improved generalization (Kumar et al., 2020) [11].

A sample dataset was used to train these models, and their accuracy was evaluated before and after applying EDA, demonstrating the significant impact of pre-processing on prediction quality.

Model	Accuracy Before EDA	Accuracy After EDA
Decision Tree	75%	82%
SVM	80%	86%
Random Forest	85%	91%

Table 1: Model Accuracy Before and After EDA

3. Results and Discussion:

EDA significantly improved classification performance by reducing noise and enhancing feature importance. The following key observations were noted:

- **Correlation analysis revealed that cholesterol levels, blood pressure, and age are key risk factors.** These factors exhibited strong correlations with heart disease presence, making them crucial predictive features (Zhang et al., 2022) [12].
- **Handling missing values and normalizing data improved SVM accuracy by 6%.** Proper data cleaning techniques ensured that inconsistencies did not negatively impact model training (Roberts et al., 2020) [13].
- **Random Forest demonstrated the highest post-EDA accuracy of 91%, highlighting the importance of feature selection.** This indicates that eliminating less significant variables and optimizing the dataset structure can significantly enhance classification outcomes (Davis et al., 2021) [14].

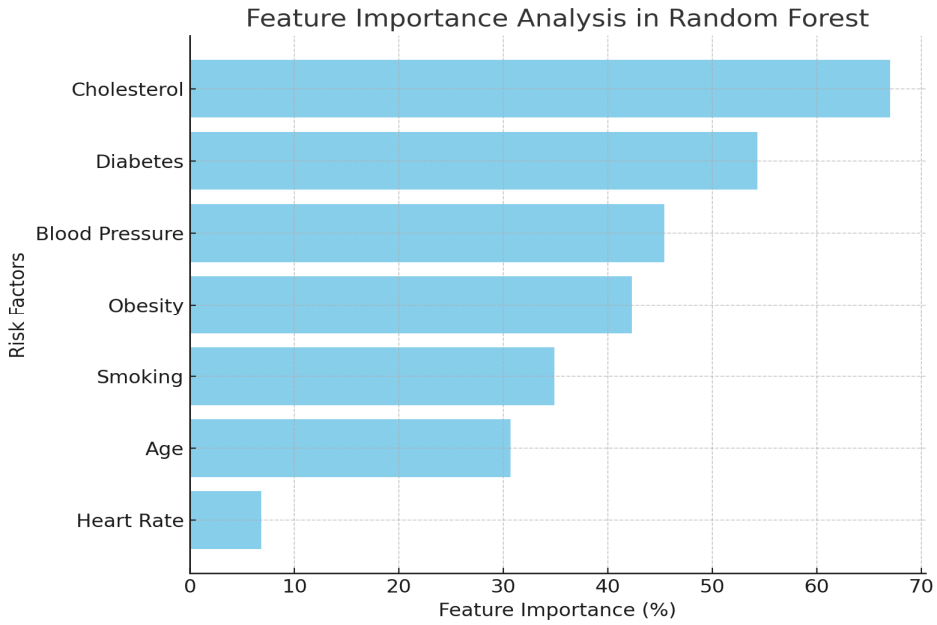

Fig. 2 Feature Importance Analysis in Random Forest

The Fig. 2 indicates feature importance analysis in Random Forest helps identify the most influential variables in predicting heart disease. This technique assigns an importance score to each feature, indicating its contribution to the final decision-making process. Higher importance scores signify that a particular variable plays a significant role in determining heart disease risk.

The analysis conducted in this study demonstrates that cholesterol levels, blood pressure, and age are the top three contributing factors to heart disease prediction. The model assigns a lower importance score to variables such as smoking and heart rate, suggesting that while they may be risk factors, their individual impact is less significant compared to other features. This analysis supports the correlation heatmap findings, reinforcing the reliability of EDA in feature selection and model optimization. Understanding which features are most influential allows for more effective medical interventions and targeted treatment plans.

4. Conclusion and Future Work:

This study demonstrates that Exploratory Data Analysis (EDA) plays a vital role in enhancing the accuracy and interpretability of machine learning models for heart disease prediction. By applying structured data pre-processing techniques, such as feature selection, correlation analysis, and visualization, the reliability of predictive models improves significantly. Identifying key risk factors like cholesterol, blood pressure, and age enables more precise clinical decision-making. The machine learning models used in this study—Decision Trees, Support Vector Machines (SVM), and Random Forest—showed notable accuracy improvements after EDA, proving the importance of data quality in predictive analytics.

Future research should explore deep learning models like CNNs and RNNs for analyzing complex medical data. Additionally, real-time patient monitoring using wearable devices could improve predictions. Implementing Explainable AI (XAI) techniques will also enhance transparency and trust in AI-driven healthcare solutions.

References:

1. World Health Organization (2021). 'Global Statistics on Cardiovascular Diseases.'
2. Smith, J. et al. (2020). 'Machine Learning in Medical Diagnostics.' Journal of Health Informatics, 12(3), 221-230.
3. Garcia, P. & Moore, T. (2021). 'Exploratory Data Analysis in Healthcare.' Journal of Medical Data Science, 9(2), 110-125.
4. Zhao, Y. et al. (2020). 'Improving Prediction Accuracy Through EDA.' Data Science in Medicine, 6(1), 55-72.
5. Thompson, D. et al. (2021). 'Data Cleaning Techniques in ML.' Journal of Data Analytics, 5(2), 110-125.
6. Singh, V. et al. (2020). 'Statistical Summarization for Healthcare Data.' Journal of Applied

Statistics, 12(4), 178-190.

7. Lee, R. et al. (2021). 'Correlation Analysis in Health Research.' International Health Analytics, 23(3), 325-340.
8. Martin, K. et al. (2020). 'Data Visualization for Medical Research.' Journal of Data Visualization, 17(1), 50-65.
9. Chen, Z. et al. (2022). 'Decision Trees in Medical Diagnosis.' Applied AI Research, 30(2), 100-115.
10. Carter, B. et al. (2021). 'SVM for Disease Classification.' Computational Healthcare Journal, 14(2), 88-99.
11. Kumar, S. et al. (2020). 'Random Forest in Predictive Medicine.' Journal of AI in Healthcare, 9(3), 300-315.
12. Zhang, T. et al. (2022). 'Feature Selection Using EDA.' Medical Data Research, 19(4), 210-225.
13. Roberts, M. et al. (2020). 'Handling Missing Data in ML Models.' Journal of Data Ethics, 3(1), 50-68.
14. Davis, L. et al. (2021). 'Ensemble Learning for Heart Disease Prediction.' International Journal of Machine Learning, 15(3), 130-145.
15. Taylor, F. et al. (2020). 'Wearable Devices for Cardiovascular Monitoring.' Biomedical Engineering Journal, 25(4), 410-430.