

**Mini Project Report**

**Of**

**Data Warehouse and Data Mining (CSE 4060)**

**Heart Disease Detection Using Machine Learning Models**

SUBMITTED

BY

Arman Malik 210905200, SECTION-A

Vishal Puneyani 210905037, SECTION-A

Varun Anmol Sathaye 210905105, SECTION-A

**Department of Computer Science and Engineering Manipal Institute of Technology, Manipal.**

**TABLE OF CONTENTS**

**CHAPTER 1: INTRODUCTION**

**CHAPTER 2: PROBLEM STATEMENT & OBJECTIVES**

**CHAPTER 3: METHODOLOGY**

**CHAPTER 4: DATA SET DESCRIPTION**

**CHAPTER 5: RESULTS & DISCUSSION**

**CHATER 6: LIMITATIONS & FUTURE WORK**

**CHAPTER 7: CONCLUSION**

**CHAPTER 8: REFERENCES**

**INTRODUCTION**

The project focuses on developing predictive models to ascertain the likelihood of heart failure in patients based on a multitude of attributes. It is structured as a binary classification problem, incorporating both numerical and categorical features to predict the outcome. The dataset used comprises various parameters such as age, sex, type of chest pain, resting blood pressure, serum cholesterol levels, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise-induced angina, old peak, and the slope of the peak exercise ST segment, culminating in a determination of heart disease presence.

Through exploratory data analysis, the project identifies patterns, outliers, and key features within the data that might influence heart disease outcomes. It delves into the relationships between different variables and heart disease, utilizing various statistical and visual tools to uncover insights. For instance, it has been observed that patients with certain resting electrocardiogram results or those above a specific age bracket are more susceptible to heart diseases. Moreover, the analysis extends to examining the influence of cholesterol levels, fasting blood sugar, and maximum heart rate on heart disease prevalence.

The project employs a range of machine learning models, including Logistic Regression, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, and K Nearest Neighbors, to predict heart disease presence. Through a meticulous training and testing process, it was determined that the K Nearest Neighbors model achieved the highest accuracy, thereby underscoring its efficacy in heart disease prediction.

**PROBLEM STATEMENT**

With the proliferation of medical data and the advancement in Data Science technologies, there is a significant opportunity to tackle the challenge of developing predictive indicators for cardiovascular diseases (CVDs), which are among the leading causes of death globally. Heart failure, as a critical outcome of CVDs, necessitates early detection and effective management to mitigate its impact. This project seeks to leverage machine learning models to automate the detection and management of heart diseases, thereby contributing to the broader objective of addressing current health challenges and enhancing patient care.

**OBJECTIVES**

1. **Predictive Modeling**: To develop a predictive model that can accurately determine the likelihood of a patient being prone to heart failure based on a comprehensive set of attributes. This includes both numerical and categorical data, aiming to capture the multifaceted nature of heart disease risk factors.
2. **Binary Classification**: The project is framed as a binary classification problem, where the outcome variable indicates the presence or absence of heart disease. This approach simplifies the complex nature of heart disease diagnosis into a manageable computational task, facilitating the development of efficient and accurate predictive models.
3. **Data Utilization**: To effectively utilize the available medical data, encompassing a wide range of variables such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, and the slope of the peak exercise ST segment. The project aims to analyze these variables to identify patterns, correlations, and insights that contribute to the risk of heart disease.
4. **Model Evaluation and Selection**: To employ and assess various machine learning models to identify the most effective approach for predicting heart disease. The project will explore different algorithms, including Logistic Regression, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, and K Nearest Neighbors, to determine the model that offers the highest accuracy and reliability in prediction.
5. **Impactful Insights**: Beyond model development, the project aims to provide valuable insights into the risk factors and dynamics of heart disease, offering potential pathways for early intervention, preventive measures, and personalized patient care strategies.

**METHODOLOGY**

The methodology employed in the heart disease detection project using machine learning models encompasses several structured steps, from data preparation through model training and evaluation, to ensure the development of an accurate and reliable predictive system. Here is a detailed breakdown of the methodology used:

1. **Data Collection and Description**

* Data Source: The project utilizes a dataset that includes a range of attributes relevant to heart disease diagnosis, such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiogram results, maximum heart rate, exercise-induced angina, ST depression, and the slope of the peak exercise ST segment.
* Data Integrity Check: Initial steps involve verifying the dataset's integrity, including checks for missing values, outliers, and ensuring the data is correctly formatted for analysis.

**2. Data Preprocessing**

* **Cleaning:** Data cleaning steps include handling missing values, removing, or imputing outliers, and deleting any redundant or irrelevant features. For instance, rows with zero values in critical numerical attributes like resting blood pressure were removed to maintain the integrity of the analysis.
* **Feature Selection:** The project involved identifying the most significant features that influence heart disease prediction. This included analyzing the relationships between different variables and the outcome, using statistical methods to determine their relevance and impact.
* **Data Transformation**: Categorical variables were transformed into numerical formats using encoding techniques to facilitate their inclusion in machine learning models. Normalization and standardization techniques were also applied to numerical features to ensure consistent scale and distribution, enhancing model performance.

3**. Exploratory Data Analysis (EDA)**

* **Pattern Recognition**: EDA was conducted to uncover patterns, correlations, and insights within the data. This involved visual and statistical analysis to understand the distribution of various attributes and their relationship with heart disease presence.
* **Insight Generation**: The analysis helped identify critical factors affecting heart disease, such as the impact of cholesterol levels, fasting blood sugar, and maximum heart rate on the likelihood of disease presence.

4. **Model Development and Training**

* **Model Selection**: A range of machine learning models was explored, including Logistic Regression, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, and K Nearest Neighbors (KNN).
* **Training Process**: The dataset was split into training and testing sets, typically using an 80-20 ratio, to ensure both robust model training and the ability to evaluate model performance on unseen data.
* **Feature Importance**: During model training, feature importance was assessed to refine the models further and focus on the most impactful variables.

5. **Model Evaluation**

* **Performance Metrics**: The models were evaluated based on their accuracy, precision, recall, and F1 score, among other metrics, to determine their effectiveness in predicting heart disease.
* **Model Selection**: The model with the highest accuracy and best performance metrics was selected as the primary predictive tool. In this project, the K Nearest Neighbors model achieved the highest accuracy, indicating its superiority in this specific context.

**DATA SET DESCRIPTION**

The dataset utilized in the heart disease detection project is composed of various attributes that are crucial for analyzing and predicting the presence of heart disease in patients. Each attribute contributes to the comprehensive evaluation of potential risk factors associated with cardiovascular conditions. Below is the description of the dataset attributes:

**Dataset Attributes:**

1. **Age:** Represents the age of the patient in years, which is a critical factor in assessing the risk of heart disease.
2. **Sex:** Indicates the sex of the patient, categorized as Male (M) or Female (F). Sex is a significant factor in heart disease risk, with different prevalence rates and risk factors associated with each gender.
3. **Chest Pain Type (ChestPainType):** Describes the type of chest pain experienced by the patient, categorized into:

* TA: Typical Angina
* ATA: Atypical Angina
* NAP: Non-Anginal Pain
* ASY: Asymptomatic

1. Angina types offer insights into the nature of chest pain and its potential correlation with heart disease.
2. **Resting Blood Pressure (RestingBP)**: Measures the resting blood pressure in millimeters of mercury (mm Hg). High resting blood pressure is a known risk factor for heart disease.
3. **Serum Cholesterol (Cholesterol):** The serum cholesterol level in milligrams per deciliter (mg/dl). Elevated cholesterol levels can indicate a higher risk of heart disease.
4. **Fasting Blood Sugar (FastingBS):** Indicates whether fasting blood sugar is greater than 120 mg/dl (1) or not (0). High fasting blood sugar is a risk factor for cardiovascular diseases.
5. **Resting Electrocardiogram Results (RestingECG)**: Shows the results of the resting electrocardiogram, categorized into:

* Normal: Normal results
* ST: Having ST-T wave abnormality
* LVH: Showing probable or definite left ventricular hypertrophy by Estes' criteria.

1. **Maximum Heart Rate Achieved (MaxHR):** The highest heart rate achieved by the patient, measured in beats per minute. The maximum heart rate can reflect the heart's health and exercise capacity.
2. **Exercise-Induced Angina (ExerciseAngina):** Indicates whether exercise induces angina (Y: Yes, N: No). Exercise-induced angina is a symptom of potential underlying heart disease.
3. **Oldpeak**: The ST depression induced by exercise relative to rest, measured in millimeters. ST depression can indicate myocardial ischemia.
4. **ST Slope (ST\_Slope):** The slope of the peak exercise ST segment, categorized into:

* Up: Upsloping
* Flat: Flat
* Down: Downsloping
* The ST segment slope can provide insights into the heart's response to exercise and potential ischemic conditions.

1. **Heart Disease (HeartDisease):** The outcome variable indicating the presence (1) or absence (0) of heart disease. This binary variable is the primary target for predictive modeling efforts.

The dataset provides a multidimensional view of factors that are potentially related to heart disease, enabling the development of predictive models to assess the risk of heart disease in individuals based on their medical profiles and test results.

**RESULTS**

After processing the given dataset and employing various machine learning models to predict heart disease, the project yielded several notable results. Here is a summary of the outcomes obtained from the analysis and model evaluation:

**Model Performance**:

* A variety of machine learning models were explored to assess their effectiveness in predicting heart disease. These included Logistic Regression, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, and K Nearest Neighbors (KNN).
* K Nearest Neighbors (KNN) emerged as the model with the highest accuracy, achieving an accuracy of 88.59%. This indicates that the KNN model was most effective in correctly classifying individuals as having heart disease or not based on the given attributes.

**Insights from Exploratory Data Analysis (EDA):**

* The EDA provided several insights into the relationships between different attributes and the presence of heart disease. For instance, it was found that:
* Patients with certain types of chest pain (specifically, asymptomatic chest pain) showed a higher prevalence of heart disease.
* Higher cholesterol levels and specific resting electrocardiogram (RestingECG) results (Normal, ST, and LVH) were associated with an increased risk of heart disease.
* Age was a significant factor, with individuals above the age of 50 being more prone to heart disease irrespective of their RestingECG values.
* The analysis also revealed that resting blood pressure and maximum heart rate values have a consistent relationship with heart disease presence across different levels.

**Additional Findings:**

The project identified specific categorical and numerical features that had more pronounced effects on heart disease outcomes. For example:

Among categorical features, sex (male > female), chest pain type, fasting blood sugar levels, exercise-induced angina, and the slope of the peak exercise ST segment (ST\_Slope) were significant predictors.

In terms of numerical features, age, resting blood pressure, cholesterol, maximum heart rate, and old peak values provided valuable insights into heart disease risk.

**Conclusion:**

The project's methodology, encompassing data preprocessing, exploratory data analysis, and the application of machine learning models, effectively identified significant predictors of heart disease.

The superior performance of the K Nearest Neighbors model suggests its potential utility in clinical settings for early heart disease detection. However, the results also highlight the importance of considering a broad range of factors in heart disease prediction, underlining the complexity of cardiovascular health.

Future work could explore the integration of more sophisticated models, additional datasets, and newer attributes to further refine the predictive accuracy and understand the underlying mechanisms of heart disease.

**LIMITATIONS**

Some identified limitations:

1**. Dataset Scope and Diversity**

* **Limited Demographic Representation**: If the dataset primarily represents a specific population segment, the model's predictions might not generalize well to other demographics, including different ages, ethnicities, and geographic locations.
* **Data Size and Imbalance**: A relatively small or imbalanced dataset could limit the model's ability to learn complex patterns, particularly for underrepresented classes or outcomes.

2. **Feature Selection and Engineering**

* **Omission of Relevant Features**: The exclusion of potentially relevant features that could influence heart disease risk, such as lifestyle factors, genetic predispositions, and detailed medical histories, might reduce the model's predictive accuracy.
* **Feature Engineering Depth**: Limited feature engineering might not fully capture the complexities and interactions between different variables, possibly oversimplifying the relationships that contribute to heart disease.

**3. Model Selection and Evaluation**

* **Model Complexity**: The project utilized a range of models, but more complex models or ensemble methods might offer improved accuracy or insights at the cost of interpretability and computational efficiency.
* **Evaluation Metrics**: Relying primarily on accuracy as the performance metric might not adequately reflect the model's performance in scenarios where precision, recall, or the area under the receiver operating characteristic (ROC) curve are more relevant.

4**. Real-World Application**

* **Clinical Validation**: The models were evaluated in a controlled setting using historical data. Real-world applicability requires rigorous clinical validation to ensure models can effectively predict heart disease in live healthcare settings.
* **Integration with Healthcare Systems**: Practical deployment challenges, including integration with existing healthcare IT systems and workflows, user training, and patient privacy concerns, were not addressed.

**5. Ethical and Privacy Considerations**

* **Data Privacy**: Handling sensitive health data raises significant privacy concerns, requiring strict adherence to data protection regulations and ethical guidelines, which might limit data accessibility for model training.
* **Bias and Fairness**: Machine learning models can inadvertently perpetuate or amplify biases present in the training data, potentially leading to unfair predictions across different patient groups.

**FUTURE WORK**

The successful outcomes of this project pave the way for several avenues of future work aimed at further improving heart disease prediction and expanding the scope of machine learning applications in healthcare:

* Integration of Additional Data Sources: Incorporating more diverse datasets, including longitudinal patient data and more granular health indicators, could enhance the models' accuracy and generalizability.
* Advanced Model Exploration: Exploring more sophisticated machine learning algorithms and deep learning models could offer improvements in predictive accuracy and insights into complex patterns within the data.
* Feature Engineering and Selection: Further analysis and engineering of features could reveal more nuanced relationships between variables and heart disease, potentially identifying new predictive markers.
* Real-World Application and Testing: Deploying the model in clinical settings for real-world testing could provide valuable feedback on its practical utility and areas for improvement, facilitating iterative enhancements to the model's performance.
* Personalized Medicine: Leveraging machine learning models for personalized risk assessments could enable tailored prevention and treatment strategies, improving patient outcomes and healthcare efficiency.
* Interdisciplinary Collaboration: Collaborating with healthcare professionals, data scientists, and policymakers could lead to innovative solutions that address broader healthcare challenges, beyond heart disease prediction.

**CONCLUSION**

The project on heart disease detection using machine learning models successfully demonstrated the capability of data science and machine learning technologies to contribute significantly to the field of medical diagnostics. By employing a dataset encompassing a wide range of attributes associated with heart disease, the project offered a comprehensive analysis and predictive modeling approach to identify individuals at risk of heart disease.

The exploratory data analysis (EDA) phase provided deep insights into the dataset, revealing crucial relationships between various factors and the occurrence of heart disease. This foundational understanding guided the subsequent modeling phase, where several machine learning models were evaluated for their predictive performance.

Among the models tested, the K Nearest Neighbors (KNN) model emerged as the most accurate, achieving an 88.59% accuracy rate in predicting heart disease presence. This result underscores the potential of KNN and similar machine learning models in enhancing diagnostic processes and patient care.