Heart Disease Risk Prediction

Course Title**: PREDICTIVE ANALYTICS** | Course Code: **INT234**

Submitted by:

Sushant Kumar

Registration No: 12218023

Roll Number 53

Programme Name: B.Tech. CSE

Under the Guidance of

Sir Vikas Mangotra

School of Computer Science & Engineering

Lovely Professional University, Phagwara

DECLARATION

I, **Sushant**, declare that this program, titled **Skill Based Assignment**, is solely prepared by me as a requirement for **INT:234** at Lovely Professional University. The program is based on my original work and represents my own efforts, ideas, and analysis.

I further declare that the guidance provided by my instructor, **Vikas Mangotra** , was instrumental in shaping my understanding and analysis of the subject matter. However, the final program and its contents are my own and do not reflect the views of Lovely Professional University or any other organization.

I take full responsibility for the accuracy and authenticity of the information presented in this program. I have cited all sources and references used in this program in accordance with the guidelines provided by my instructor and the academic standards of Lovely Professional University.

I understand that any act of plagiarism or academic misconduct may result in severe consequences, including the revocation of my academic degree, and I affirm that I have not engaged in any such activities.

Date: 17 /11/2024

Name of Student: Sushant Kumar
Registration no: 12218023

Indexing

Introduction

Predictive modeling, indeed, has become a cornerstone in the evolving landscape of healthcare analytics, that is critically used in making preemptive medical diagnoses and treatment planning. The report hereby provides an all-inclusive analytical tool developed by use of an exhaustive dataset to predict the risk of heart disease in people. The ultimate aim of this tool is to arm healthcare professionals and researchers with action-driven insights that help improve decision-making processes, as well as patient outcomes.

Modern methodologies with the help of data have dramatically changed the mode in which the healthcare sector approaches predicting and hence managing disease. Seeing the critical need for promisingly accurate and accessible predictive analytics, this project comes bearing a dynamic dashboard-not only predicting risks for heart diseases but also dissecting all those elements of demographics and physiology that lead towards its development. This very powerful tool will pave the way for detail-heavy heart health data, revealing a deeper analysis of risk factors, patient profiles, and predictive outcomes.

The heart of the dashboard lies in its capability to give a granular view of individual risk elements with an array of interactivity visualizations, which include:

Demographic Analysis: Understand how age, gender, and ethnicity are risk influencers of heart disease risks.
Risk Factor Correlation: This speaks to association of lifestyle and genetic factors with the potential heart disease risk.
Accuracy of Performance Indicators: It can be measured using advanced metrics that represent predictive algorithms and model effectiveness.
By bringing these insights into a single, intuitive platform, it helps to enrich the understanding of heart disease dynamics and consequently plot a path forward for targeted interventions and more enhanced patient care strategies. Designed with modernity in mind and navigable with ease, the dashboard serves as a very critical tool for healthcare providers to better work towards reducing the risks associated with heart disease among their patients.

Scope of Analysis

The scope of the Heart Disease Risk Prediction dashboard's analysis is to systematically estimate how different clinical and lifestyle variables may predict heart disease risk in diverse populations. This project uses a rich dataset and then supplies detailed patient records and health metrics, which allows for multiple exploration of the dynamics of heart health.

This analysis combines Python-based data science tools and techniques to focus on the following crucial aspects.

Demographic and Physiological Insights: A descriptive analysis of how age, sex, cholesterol levels, blood pressure, and other factors lead to increasing heart conditions can be explored. Such an analysis could identify the profiles that are prone to heart disease and knowledge of how various demographics are affected by different heart health factors.

Lifestyle impact analysis: Analyze lifestyles, including smoking and exercise, determining direct and combined effects on heart health. The concept is to try to quantify the risk associated with different lifestyle factors and their interactions with physiological metrics.

Predictive Model Evaluation: It is a predictive model built for this dashboard using machine learning algorithms. Accuracy, precision, sensitivity, and specificity will be tested to ensure clinical reliability at any level.

Risk Stratification: Classify the patient in categories of risk statistics and machine learning techniques on developing heart disease. It will be represented with dynamic graphics and charts to make it easier for the healthcare provider to compare the patient risk level at hand.

Time-series Trend Analysis: Explain how the prevalence of heart disease risk factors has changed over time from the dataset. This can prove very useful in estimating the impact of interventions at the level of public health and in outlining changes over time in behavior and the practice of healthcare.

Then, these analyses will be incorporated into an interactive dashboard in dynamic form so that they can be used by healthcare providers as a tool that is informative yet actionable. The users of the dashboard will have the ability to create a variety of customized views of the data,

thereby rendering the information more useful in clinical and research settings. Ultimately, it would greatly help in data-driven decision-making of preventing and handling heart disease cases, thus leading to improved positive outcomes for patients along with optimizing health care.

Existing System

The current landscape of heart disease prediction primarily relies on traditional risk assessment models, such as the Framingham Risk Score and other clinical algorithms that utilize a limited set of patient data points. These models have been instrumental in clinical settings for many years, providing a baseline for assessing heart disease risk based on key variables like age, cholesterol levels, and blood pressure. However, these traditional models often do not incorporate a wider range of variables, including nuanced lifestyle and demographic factors that can significantly impact risk assessment.

Limitations of the Existing System:

- **Narrow Variable Scope**: Conventional models typically use a limited set of variables, which may not capture all the nuances of an individual's risk profile.
- **Static Analysis**: Many existing tools do not offer dynamic or interactive capabilities, limiting their usefulness in real-time decision-making.
- **Lack of Customization**: Traditional models do not allow for easy adjustments based on new research or regional health trends, which can hinder their accuracy over time.
- **One-size-fits-all Approach**: These models often apply the same risk factors across diverse populations without adjustments for demographic variations such as ethnicity, gender, or socioeconomic factors.

Improvements Offered by My Heart Disease Risk Prediction Dashboard:

- **Incorporation of a Broader Set of Variables**:
    - My ML models utilize a wider array of variables, including lifestyle factors such as smoking status and physical activity, which are often overlooked in traditional models. This comprehensive data input allows for a more holistic view of an individual's health.
- **Dynamic and Interactive Analysis**:
    - The dashboard I developed features interactive elements that allow users to manipulate variables and instantly see how these changes affect the risk predictions. This functionality is crucial for healthcare providers to simulate various scenarios and better understand patient risks.
- **Customizable and Adaptable Models**:

    - My machine learning models can be continuously updated and trained with new data as it becomes available. This adaptability ensures that the models remain accurate over time and reflect the latest research and health trends.
- **Demographic-Specific Risk Assessments**:
    - By including demographic factors in the risk assessment, my models can provide more personalized predictions. This is particularly important in tailoring health interventions to specific groups who may be at a higher risk due to genetic, lifestyle, or environmental factors.
- **Advanced Analytical Techniques**:
    - I employ sophisticated ML algorithms such as logistic regression, decision trees, or neural networks to uncover complex patterns and interactions among risk factors that traditional statistical methods might miss.

By addressing the shortcomings of existing systems, my Heart Disease Risk Prediction project significantly enhances the toolset available to healthcare providers, offering a more precise, personalized, and proactive approach to managing heart disease risk. This project not only improves individual patient care but also contributes to broader public health strategies by providing insights that can guide policy and education on heart health.

Source of Dataset

The dataset for the Heart Disease Risk Prediction was sourced from Kaggle, a platform renowned for its comprehensive repository of datasets used for data science competitions and research projects.

Dataset Overview:

1. **Number of Rows**: The dataset comprises approximately 303 entries.
2. **Number of Columns**: There are 14 columns in the dataset.
3. **Types of Data Contained**:
    - **Age**: The age of the individual.
    - **Sex**: The gender of the individual (1 = male; 0 = female).
    - **Chest Pain Type**: The type of chest pain experienced (values 1-4, representing different types of angina and non-anginal pain).
    - **Resting Blood Pressure**: The resting blood pressure of the individual in mm Hg (millimeters of mercury).
    - **Cholesterol**: The individual's serum cholesterol in mg/dl (milligrams per deciliter).
    - **Fasting Blood Sugar**: Whether the individual's fasting blood sugar is above 120 mg/dl (1 = true; 0 = false).
    - **Resting Electrocardiographic Results**: Resting electrocardiographic measurement (0 = normal; 1 = having ST-T wave abnormality; 2 = showing probable or definite left ventricular hypertrophy).
    - **Maximum Heart Rate Achieved**: The maximum heart rate achieved by the individual.
    - **Exercise Induced Angina**: Whether exercise induced angina (1 = yes; 0 = no).
    - **ST Depression**: ST depression induced by exercise relative to rest.
    - **Slope of the Peak Exercise ST Segment**: The slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping).
    - **Number of Major Vessels**: The number of major vessels (0-3) colored by fluoroscopy.
    - **Thalassemia**: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversible defect).
    - **Target**: Diagnosis of heart disease (angiographic disease status) (1 = disease; 0 = no disease).

Utility of Dataset:

This dataset serves as a pivotal tool for analyzing and predicting heart disease risks, furnishing healthcare professionals and researchers with vital data to assess the prevalence and predictors of heart disease across a diverse population. The variables included in the dataset allow for a detailed examination of the correlations between various risk factors and the presence of heart disease, supporting the development of targeted intervention strategies. Additionally, it is instrumental in training machine learning models to predict the likelihood of heart disease, ultimately aiding in proactive healthcare planning and management.

.

ELT

The ELT (Extract, Transform, and Load) process for my Heart Disease Risk Prediction dashboard began with me extracting a detailed dataset from Kaggle, renowned for its extensive collection of clinical and demographic data essential for heart disease research. After downloading the dataset, I moved on to the transformation phase to refine the data for comprehensive analysis:

Extract:

- Utilizing the R programming language, I started the process by loading the dataset from a CSV file into an R dataframe. This was efficiently accomplished using R's `read.csv` function, ensuring that the dataset was promptly ready for manipulation and analysis.

Transform:

- **Data Cleaning**: I applied various data cleaning techniques to prepare the dataset for analysis. This included addressing missing values and removing records that lacked essential information, which is crucial to maintain the robustness of the analyses performed.
- **Data Transformation and Feature Engineering**: Using R's `dplyr` package, I manipulated the dataset to format and create new variables that better represented the underlying patterns in the data. This included categorizing continuous variables into meaningful groups and encoding categorical variables for more effective analysis.
- **Visualization and Exploratory Data Analysis (EDA)**: I employed `ggplot2` from R's `tidyverse` to generate visualizations. These plots were instrumental in understanding the distributions and relationships within the data, which informed further data transformations and hypothesis generation.
- **Preparing Data for Modeling**: Utilizing the `caret` package, I prepared the dataset for modeling by setting up training and testing sets, ensuring that the data fed into the model training process was well-prepared and representative of the overall dataset.

Load:

- **Model Building and Evaluation**: With the data fully prepped, I used R's `caret`, `e1071`, and `klaR` packages to train and evaluate various machine learning models, including logistic regression, k-nearest neighbors (KNN), Support Vector Machines (SVM), and Naive Bayes. The models were rigorously evaluated using accuracy metrics and ROC curves, aiding in identifying the most effective predictors of heart disease.
- **Operationalization**: The final step involved loading the trained models into an R Shiny application, transforming them from static models into interactive, user-friendly tools that healthcare professionals can use to predict heart disease risk based on individual patient data.

This meticulous ELT process, supported by R's robust libraries and functionalities, ensured that my analysis was not only thorough but also adaptable and responsive to the needs of healthcare stakeholders aiming to mitigate heart disease risks

Analysis on Datasets

In my Heart Disease Risk Prediction project, a thorough analysis of the dataset was crucial for crafting robust predictive models. This detailed process encompassed several sophisticated techniques aimed at deciphering underlying patterns and interrelationships within the data, thus ensuring the reliability and accuracy of the predictive models.

Data Visualization and Exploration:

- **Utilizing R's `ggplot2` package**, I initiated data visualizations to explore the distribution of pivotal variables such as age and sex. Histograms were generated to examine the age distribution across the dataset, highlighting potential biases or notable skews in the data. This approach was instrumental in visually assessing the demographic makeup of the dataset.
- **For categorical variables like sex**, bar charts were crafted to provide clear visual insights into the composition of the dataset, which was vital for understanding the distribution and potential impact of these variables on heart disease.

Correlation Analysis:

- **Using the `corrplot` package**, I conducted correlation analyses to pinpoint which variables strongly correlate with the occurrence of heart disease. This step was key in guiding the feature selection for subsequent model building, as it identified significant predictors that could influence model accuracy.

Data Preprocessing:

- **Rigorous preprocessing steps were implemented**, including handling missing values and encoding categorical variables using R's factor levels. This preprocessing ensured the data was primed for modeling, enhancing the integrity and utility of the dataset.
- **The data was then split into training and testing sets**, a critical step for model validation. This split was facilitated by randomly sampling 70% of the data for training, ensuring a representative mix of instances for building and subsequently evaluating the models.

Model Building and Evaluation:

- Several machine learning techniques such as **k-nearest neighbors (KNN), Support Vector Machines (SVM), and Naive Bayes** were applied, coupled with use of R packages like **class, e1071 and caret**. Each model has been rigorously trained on the training dataset and its performance tested on the test set.
- **Model performance was assessed through confusion matrices and accuracy metrics**, which allowed me to compare the efficacy of each model and choose the best performer for deployment.

Operationalizing the Models:

- **The culmination of my analysis was the operationalization of the best-performing models** in an R Shiny application. This platform transforms static models into interactive, user-friendly tools that enable healthcare professionals to input patient data and receive real-time predictions on heart disease risk, significantly augmenting decision-making processes in clinical settings.

This in-depth analysis not only strengthened the prediction models but also brought to light a spectrum of variables relevant to heart disease, yet importantly emphasized the importance of data in health analytics. This project pushes the agenda for personalized patient care as well as more general healthcare approaches forward through actionable insights derived from the analysis.

Datasets Overview

The dataset employed for my Heart Disease Risk Prediction project encompasses a rich array of clinical and demographic variables critical for analyzing and predicting heart disease. Sourced from a reliable and extensive database on Kaggle, this dataset consists of structured data covering various aspects of patient health and characteristics.

Key Attributes and Statistical Overview:

Age: The patients' age ranges from 29 to 77 years, with a median age of 55, indicating a dataset primarily composed of middle-aged to elderly individuals, which is crucial for heart disease studies as age is a significant risk factor.

Sex: The dataset encodes sex as a binary attribute (0 for female and 1 for male). The majority of the dataset comprises male subjects (68.32%), reflecting common trends in heart disease prevalence among genders.

Chest Pain Type (cp): Encoded from 0 to 3, this variable represents the type of chest pain experienced, a vital sign in diagnosing cardiac conditions. The distribution across types suggests diverse clinical presentations among the subjects.

Resting Blood Pressure (trestbps): Blood pressure readings vary from 94 to 200 mm Hg, with a mean of approximately 131.6 mm Hg, highlighting a range from normal to significantly elevated levels.

Cholesterol (chol): Levels range from 126 to 564 mg/dL, with a median of 240 mg/dL, indicating variability in lipid profiles among the participants.

Fasting Blood Sugar (fbs): This binary variable indicates if fasting blood sugar is above 120 mg/dL (1 = true, 0 = false), with about 14.85% of subjects having elevated fasting sugar levels.

Resting Electrocardiographic Results (restecg): Ranging from 0 to 2, these results show normal to definitive signs of heart abnormalities, with the majority showing probable left ventricular hypertrophy.

Maximum Heart Rate Achieved (thalach): The values range from 71 to 202 beats per minute, with a median of 153, useful for assessing cardiac function during stress.

Exercise Induced Angina (exang): A key indicator of ischemia, 32.67% of the subjects experienced angina during exercise.

Oldpeak: ST depression induced by exercise relative to rest, ranging from 0 to 6.2, provides insights into myocardial oxygen consumption.

Slope: The slope of the peak exercise ST segment is an indicator of heart health during exercise, with a mean value of 1.399, suggesting varying levels of risk.

Number of Major Vessels Colored by Fluoroscopy (ca): Reflects the number of major vessels affected, ranging from 0 to 4.

Thalassemia (thal): A blood disorder included in the study, with values from 0 to 3, affecting the blood's ability to carry oxygen.

Target: The outcome variable (0 for no presence and 1 for presence of heart disease), with approximately 54.46% of the subjects diagnosed with heart disease.

Initial Data Exploration:

The **str(heart_data**) function provided insights into the structure of each variable, confirming their data types and completeness.
A summary **(summary(heart_data))** of the dataset offered a detailed statistical breakdown, facilitating an understanding of central tendencies and dispersion, critical for preliminary analyses.
Viewing the first few records **(head(heart_data, 5))** allowed for a quick verification of data integrity and preliminary insights into individual cases.

This comprehensive overview of the dataset forms the foundation of all subsequent analyses, enabling me to meticulously explore various dimensions such as demographic influences, physiological markers, and their collective impact on heart disease risk. Through this dataset, my project aims to deliver actionable insights that significantly enhance predictive accuracies and contribute to better preventive and diagnostic solutions in healthcare settings.

Snapshot

```
$ thalach : num [1:303] 150 187 172 178 163 148 153 173 162 174 ...
$ exang   : num [1:303] 0 0 0 0 1 0 0 0 0 0 ...
$ oldpeak : num [1:303] 2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
$ slope   : num [1:303] 0 0 2 2 2 1 1 2 2 2 ...
$ ca      : num [1:303] 0 0 0 0 0 0 0 0 0 0 ...
$ thal    : num [1:303] 1 2 2 2 2 1 2 3 3 2 ...
$ target  : num [1:303] 1 1 1 1 1 1 1 1 1 1 ...
- attr(*, "spec")=
 .. cols(
 ..   age = col_double(),
 ..   sex = col_double(),
 ..   cp = col_double(),
 ..   trestbps = col_double(),
 ..   chol = col_double(),
 ..   fbs = col_double(),
 ..   restecg = col_double(),
 ..   thalach = col_double(),
 ..   exang = col_double(),
 ..   oldpeak = col_double(),
 ..   slope = col_double(),
 ..   ca = col_double(),
 ..   thal = col_double(),
 ..   target = col_double()
 .. )
- attr(*, "problems")=<externalptr>
> summary(heart_data)
      age             sex               cp           trestbps         chol            fbs        
 Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0   Min.   :126.0   Min.   :0.0000  
 1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:0.0000  
 Median :55.00   Median :1.0000   Median :1.000   Median :130.0   Median :240.0   Median :0.0000  
 Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6   Mean   :246.3   Mean   :0.1485  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0   3rd Qu.:274.5   3rd Qu.:0.0000  
 Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0   Max.   :564.0   Max.   :1.0000  
    restecg          thalach          exang           oldpeak         slope             ca        
 Min.   :0.0000   Min.   : 71.0   Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:133.5   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
 Median :1.0000   Median :153.0   Median :0.0000   Median :0.80   Median :1.000   Median :0.0000  
 Mean   :0.5281   Mean   :149.6   Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294  
 3rd Qu.:1.0000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
 Max.   :2.0000   Max.   :202.0   Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000  
      thal           target      
 Min.   :0.000   Min.   :0.0000  
 1st Qu.:2.000   1st Qu.:0.0000  
 Median :2.000   Median :1.0000  
 Mean   :2.314   Mean   :0.5446  
 3rd Qu.:3.000   3rd Qu.:1.0000  
 Max.   :3.000   Max.   :1.0000  
>
```

Dashboard Overview

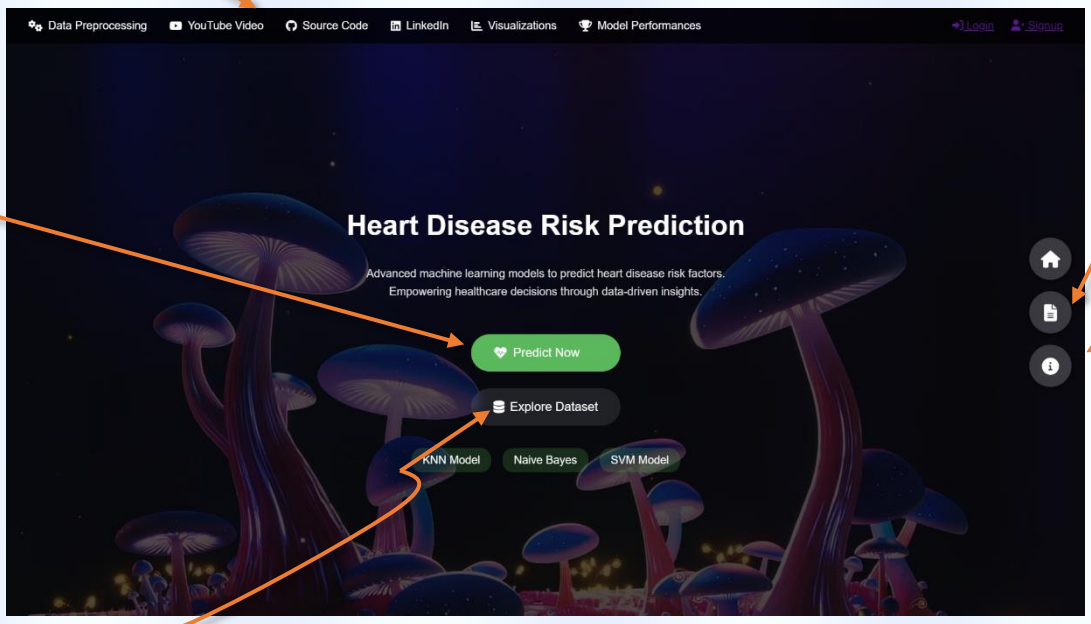
Main Dashboard

Watch Working Tutorial

Reports

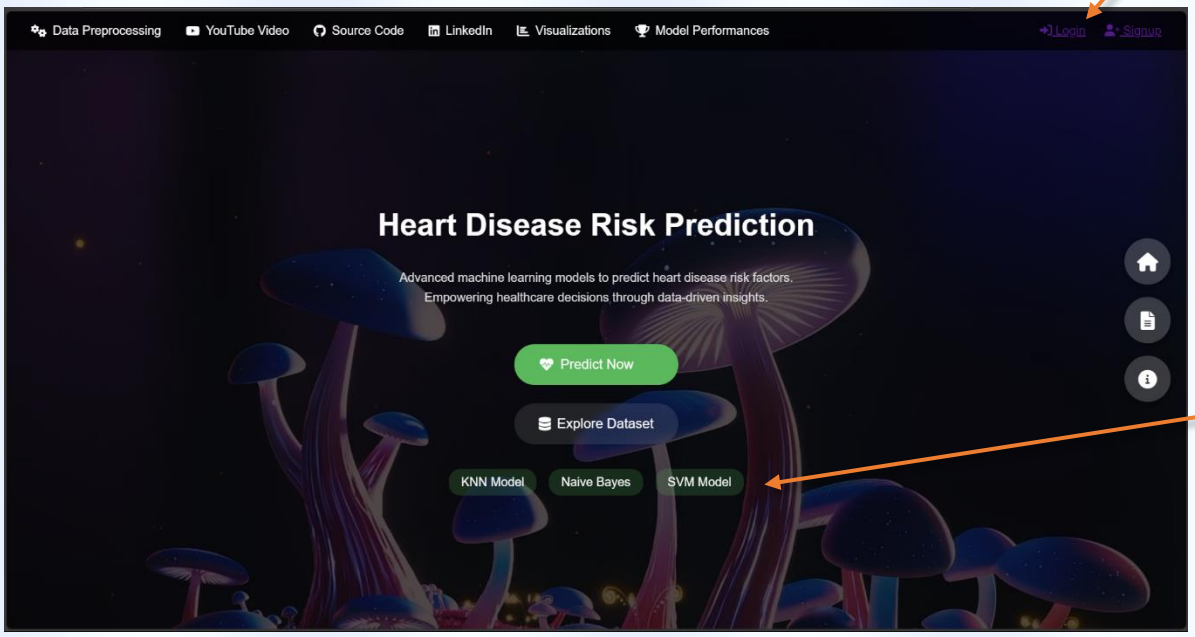
RealTime Prediction

About Section



Log in | Sign up

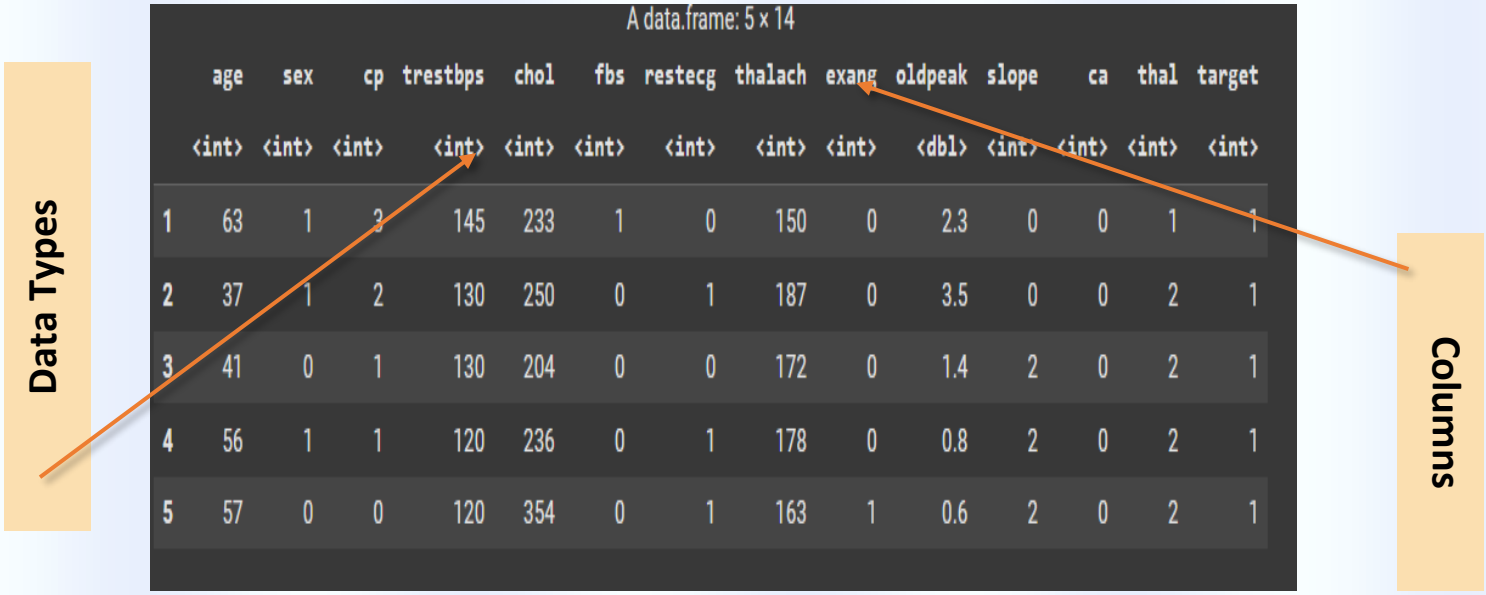
Explore The Dataset Live



Choose Your Model To Predict

DATA INSPECTION

Head



A data.frame: 5 x 14

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<dbl>	<int>	<int>	<int>	<int>
1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
5	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

```
str(heart_data)

summary(heart_data)

head(heart_data, 5)
```

Structure And Summary

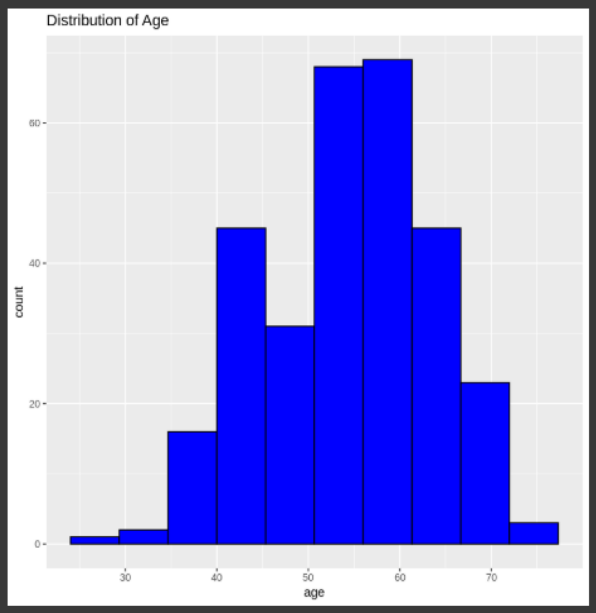
```
'data.frame':  303 obs. of  14 variables:
 $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
 $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
 $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
 $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
 $ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
 $ exang   : int  0 0 0 0 1 0 0 0 0 0 ...
 $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slope   : int  0 0 2 2 2 1 1 2 2 2 ...
 $ ca      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ thal    : int  1 2 2 2 2 1 2 3 3 2 ...
 $ target  : int  1 1 1 1 1 1 1 1 1 1 ...
      age             sex               cp           trestbps    
 Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0  
 1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0  
 Median :55.00   Median :1.0000   Median :1.000   Median :130.0  
 Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0  
 Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0  
      chol            fbs             restecg          thalach     
 Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0  
 1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5  
 Median :240.0   Median :0.0000   Median :1.0000   Median :153.0  
 Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6  
 3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0  
 Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0  
     exang           oldpeak          slope             ca        
 Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
 Median :0.0000   Median :0.80   Median :1.000   Median :0.0000  
 Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294  
 3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000  
      thal           target      
 Min.   :0.000   Min.   :0.0000  
 1st Qu.:2.000   1st Qu.:0.0000  
 Median :2.000   Median :1.0000  
 Mean   :2.314   Mean   :0.5446  
 3rd Qu.:3.000   3rd Qu.:1.0000  
 Max.   :3.000   Max.   :1.0000  
```

VISUALIZATION

Visualization Of Data Distribution

DATA VISUALIZATION

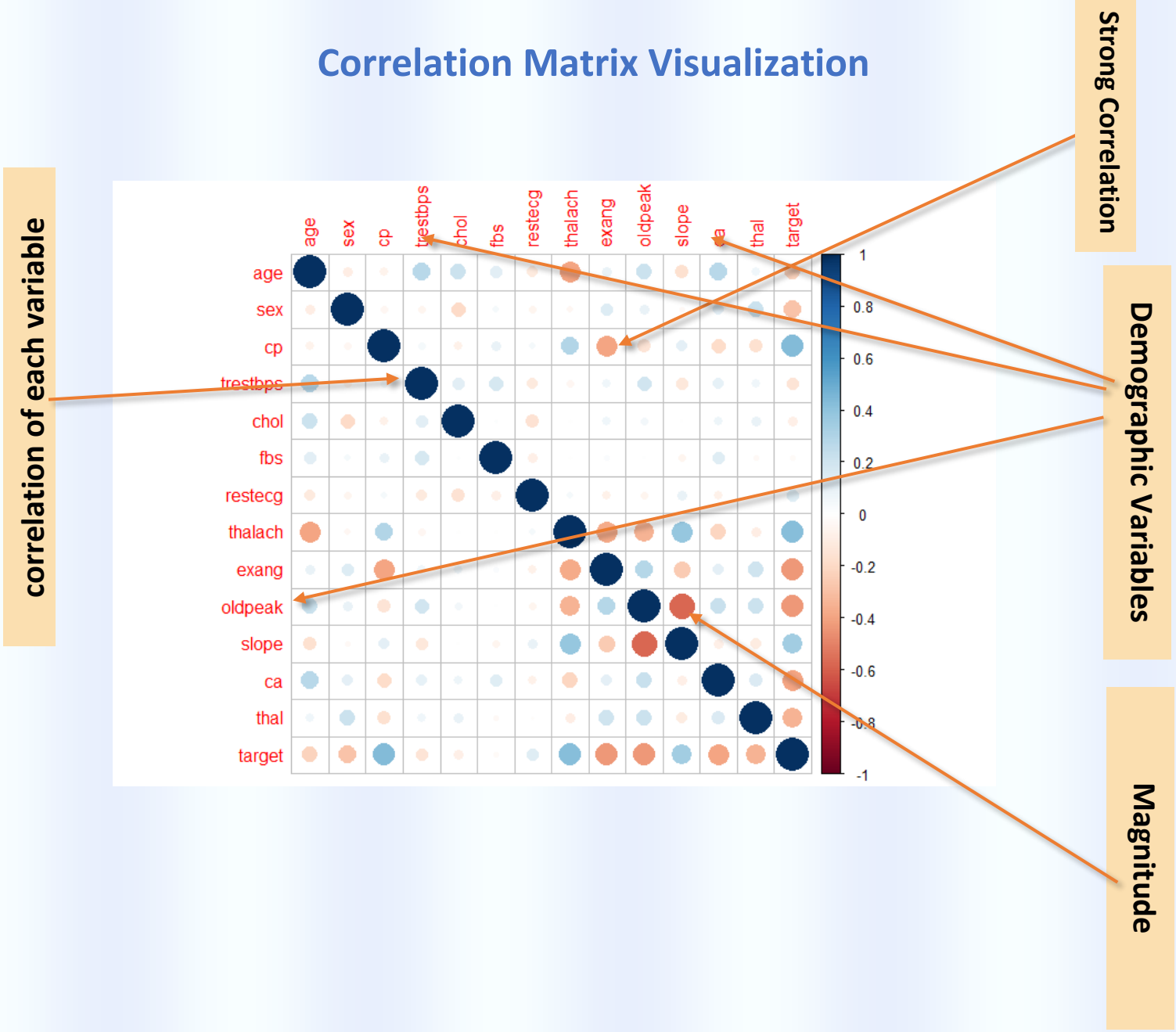
```
ggplot(heart_data, aes(x = age)) + geom_histogram(bins = 10, fill = "blue", color = "black") +
  ggtitle("Distribution of Age")
```




```
install.packages("ggplot2")

library (ggplot2)


ggplot (heart_data, aes(x = age)) + geom_histogram(bins = 10,  fill = "blue",

color = "black") + ggtitle("Distribution of Age")
```

VISUALIZATION

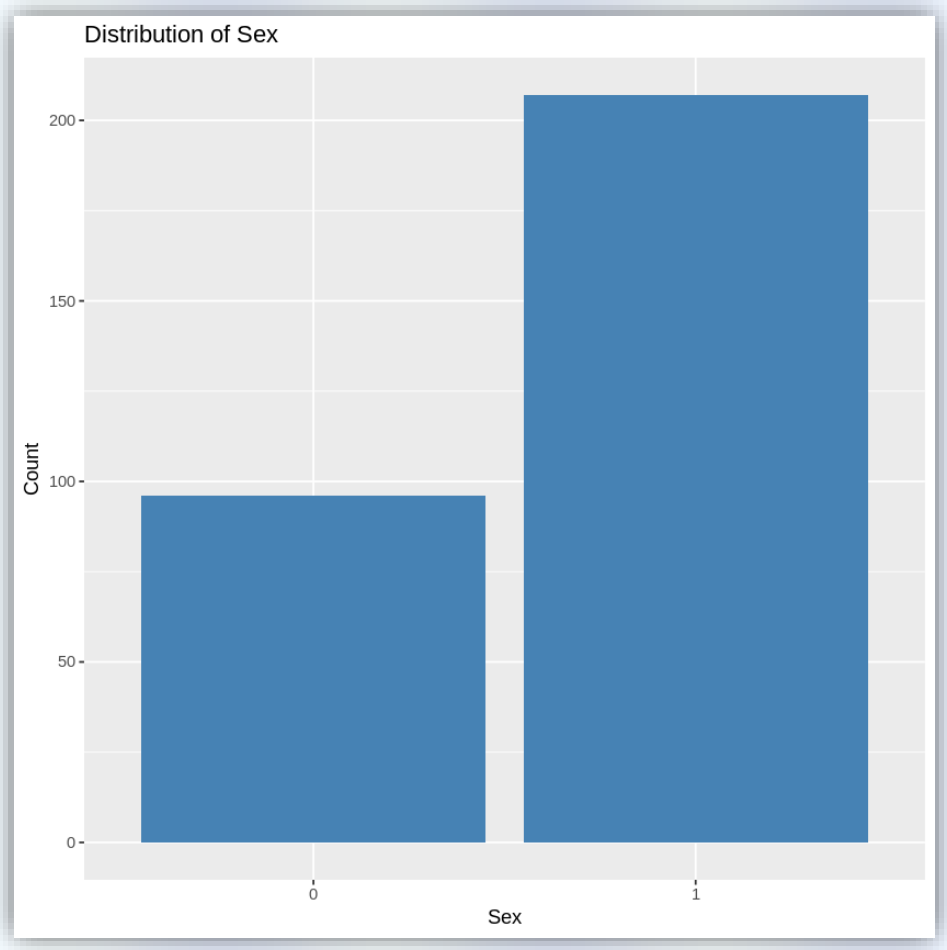
Correlation Matrix Visualization




```
install.packages("corrplot")
library(corrplot)
correlations <- cor(heart_data[, sapply(heart_data, is.numeric)])
corrplot(correlations, method = "circle")
```

VISUALIZATION

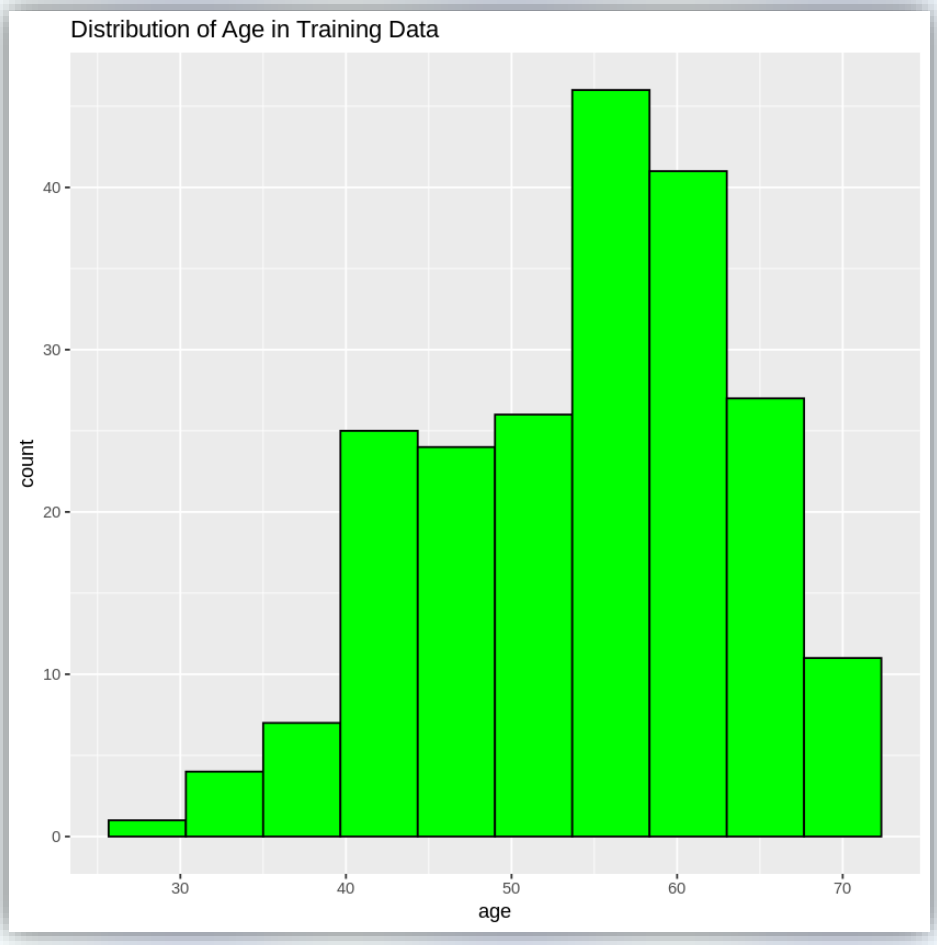
Categorical Variables




```
install.packages("ggplot2")
library(ggplot2)
ggplot(heart_data, aes(x = as.factor(sex))) + geom_bar(fill = "steelblue") +
labs(title = "Distribution of Sex", x = "Sex", y = "Count")
```

VISUALIZATION

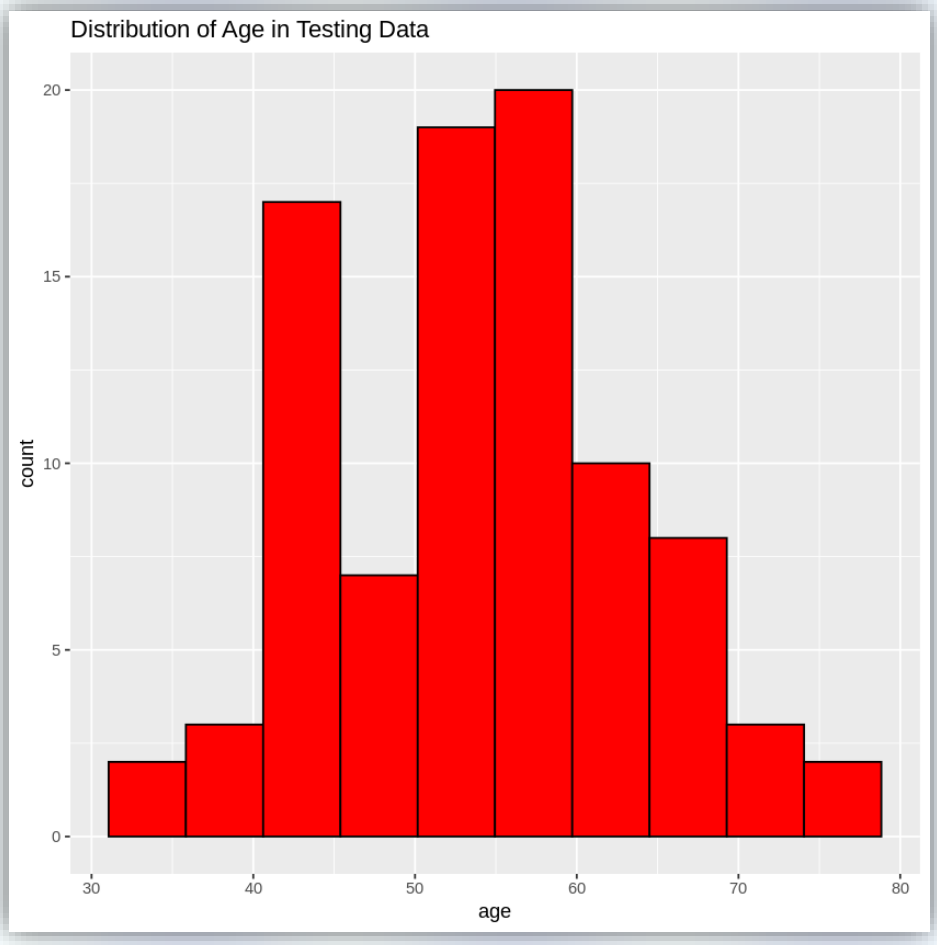
Histogram of Train Data




```
ggplot(train_data, aes(x = age)) +
geom_histogram(bins = 10, fill = "green", color = "black") +
ggtitle("Distribution of Age in Training Data")
```

VISUALIZATION

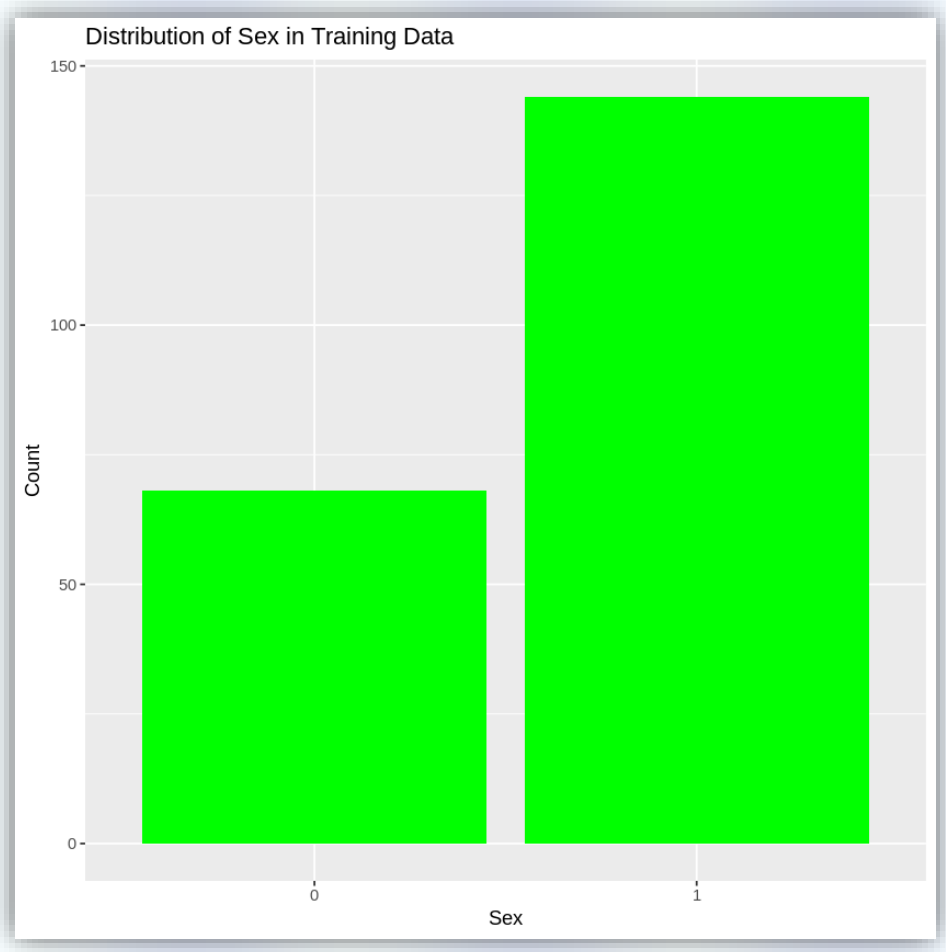
Histogram of Test Data




```
ggplot(test_data, aes(x = age)) +
geom_histogram(bins = 10, fill = "red", color = "black") +
ggtitle("Distribution of Age in Testing Data")
```

VISUALIZATION

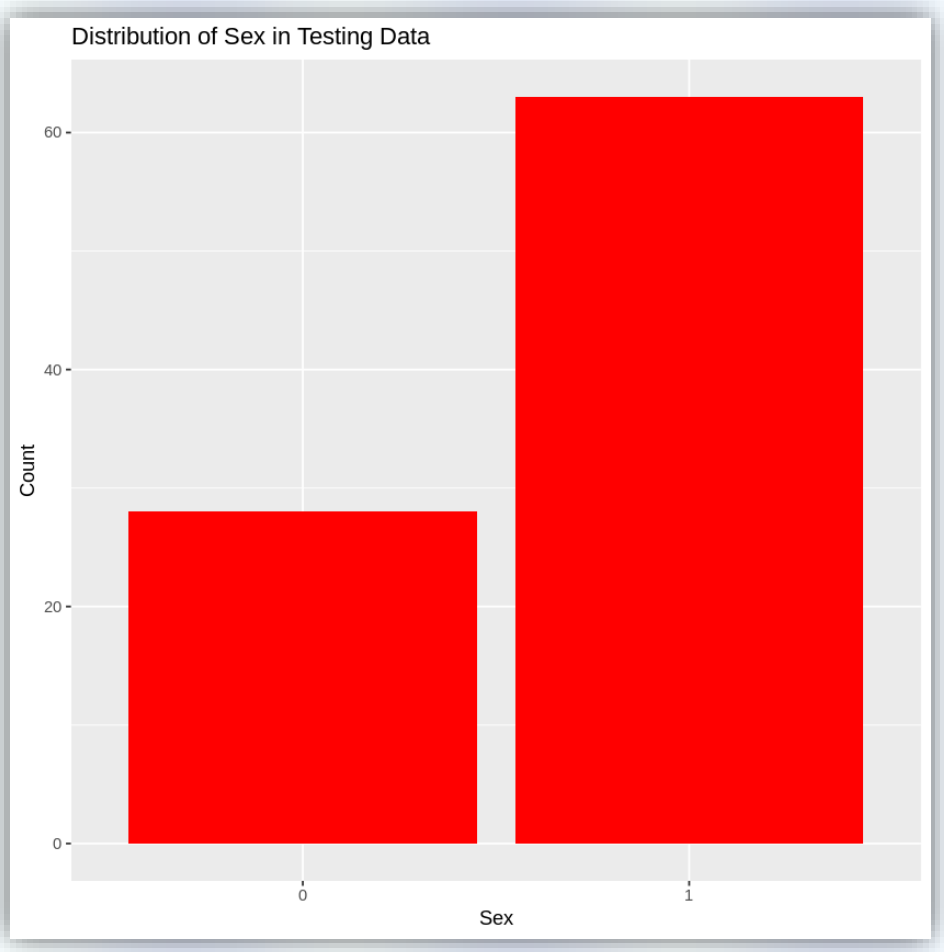
Bar Chart of Gendre In Train Data




```
ggplot(train_data, aes(x = as.factor(sex))) + geom_bar(fill = "green") +
labs(title = "Distribution of Sex in Training Data", x = "Sex", y = "Count")
```

VISUALIZATION

Bar Chart of Gendre In Test Data




```
ggplot(test_data, aes(x = as.factor(sex))) + geom_bar(fill = "red") +
labs(title = "Distribution of Sex in Testing Data", x = "Sex", y = "Count")
```

Algorithms

K-nearest neighbor algorithm(KNN)

The K-Nearest Neighbors (KNN) algorithm is a straightforward, yet powerful, supervised machine learning technique used primarily for classification tasks, but it can also be applied to regression. In the context of my Heart Disease Risk Prediction project, KNN is utilized to classify patients based on the likelihood of having heart disease, leveraging its capability to make predictions based on the proximity to known data points.

Implementation of KNN in R:

- **Library Usage**: For the KNN algorithm, I employed the class package in R, which is specifically designed to handle this type of algorithm. This package offers efficient tools for classifying datasets where the classification depends on the proximity of data points in a multidimensional feature space.
- **Data Preparation**: Before applying the KNN algorithm, the dataset was preprocessed to ensure optimal outcomes. This involved:
    - Encoding categorical variables into factors to ensure they are appropriately handled during distance calculation.
    - Splitting the data into training and testing sets to evaluate the model's performance accurately. Approximately 70% of the data was used for training, with the remaining 30% for testing, ensuring a robust sample for both model training and validation.
- **Model Training**: The KNN model was trained using the knn() function from the class package. This function requires the training dataset, testing dataset, and the classification labels of the training data. The number of neighbors (k) was set to 21, which was determined based on preliminary tests to optimize accuracy.
- **Model Evaluation**:
    - After training, the model was used to predict heart disease status in the testing set. The predictions were then compared against the actual data to evaluate the model's effectiveness.
    - A confusion matrix was generated to visualize the accuracy of predictions, helping to understand the model's performance in terms of false positives, false negatives, true positives, and true negatives.

Why KNN is Used:

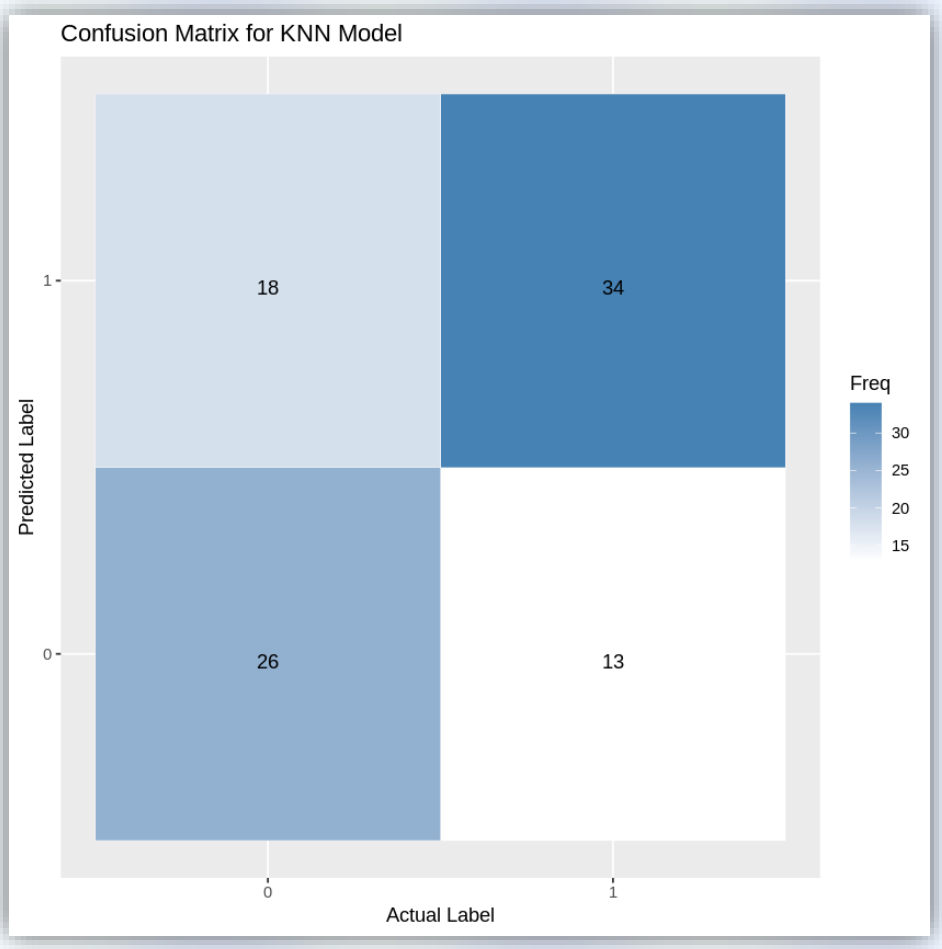
- **Simplicity and Efficacy**: KNN is known for its simplicity and effectiveness, especially in cases where the decision boundary is irregular. For heart disease prediction, where multiple factors interact in complex ways, KNN can capture these nuances without a priori assumptions about the data distribution.
- **Non-parametric Nature**: As a non-parametric method, KNN makes no assumptions about the underlying data distribution, which is advantageous in medical datasets where the variables may not follow a known or typical distribution.
- **Interpretability**: The results from KNN are relatively easy to interpret, allowing healthcare professionals to understand the factors influencing the model's predictions. This is particularly valuable in a clinical setting where the interpretability of a model can enhance trust and reliability in its use.

The utilization of KNN in this project was pivotal in achieving a nuanced understanding of heart disease risks, thereby aiding in the development of targeted interventions and improved patient outcomes. This approach underscores the importance of selecting appropriate machine learning techniques that align with the specific characteristics and requirements of the dataset and the research objectives.

```
library( class )

train_predictors <- train_data[ , - which( names( train_data ) == "target" ) ]
train_target <- train_data$target

test_predictors <- test_data[ , - which( names( test_data ) == "target" ) ]
test_target <- test_data$target

knn_model <- knn( train = train_predictors, test = test_predictors, cl = train_target, k = 21 )

knn_predictions <- knn( train = train_predictors, test = test_predictors, cl = train_target, k = 21 )
table( Predicted = knn_predictions, Actual = test_target )

library( ggplot2 )
conf_matrix <- table( Predicted = knn_predictions, Actual = test_target )
ggplot( as.data.frame( conf_matrix ), aes( x = Actual, y = Predicted ) ) +
  geom_tile( aes( fill = Freq ), colour = "white" ) +
  geom_text( aes( label = Freq ), vjust = 1 ) +
```

```
  scale_fill_gradient( low = "white", high = "steelblue" ) +
  labs( title = "Confusion Matrix for KNN Model", x = "Actual Label", y = "Predicted Label" )

saveRDS( knn_model, file = "/content/KNN_heart.rds" )
```

Actual Predicted 0 1 0 26 13 1 18 34



Algorithms

Naive Bayes

The Naive Bayes algorithm is a probabilistic classifier based on Bayes' Theorem with an assumption of independence among predictors. In the domain of medical diagnostics, where swift and reliable decision-making is crucial, Naive Bayes offers a straightforward and effective classification mechanism. For my Heart Disease Risk Prediction project, the Naive Bayes method was chosen due to its efficiency in handling large datasets with multiple attributes, making it ideal for determining the likelihood of heart disease in patients.

Implementation of Naive Bayes in R:

- **Library Usage**: The implementation of the Naive Bayes algorithm was facilitated by the e1071 package in R, which provides comprehensive functions for creating and evaluating Naive Bayes models. This package is well-suited for applications in medical statistics where model simplicity and interpretability are important.
- **Data Preparation**: Prior to applying the Naive Bayes classifier:
    - Categorical variables were encoded as factors to maintain the integrity of the dataset, ensuring that each categorical variable is appropriately considered in the probabilistic framework of Naive Bayes.
    - The dataset was divided into training and testing sets, with a 70/30 split, to provide a balanced approach to training and subsequently validating the model's predictive power.
- **Model Training**: Using the naiveBayes() function from the e1071 package, the model was trained on the processed training data. This function is adept at handling both continuous and categorical data, automatically adjusting calculations for the underlying distributions of features (Gaussian for continuous features).
- **Model Evaluation**:
    - The trained model was then used to predict heart disease occurrences in the testing set. Performance metrics were calculated to assess the accuracy and effectiveness of the model.
    - A confusion matrix was generated to visualize the classification results, providing insights into the true positives, true negatives, false positives, and false negatives.

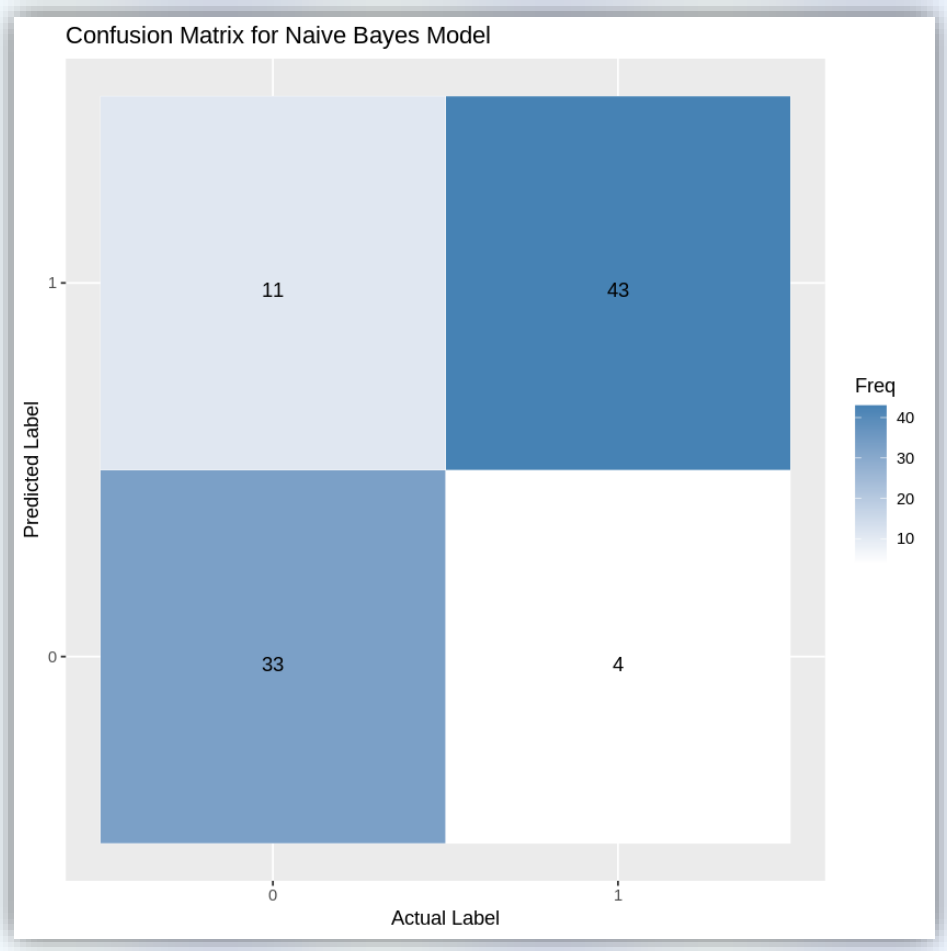
Why Naive Bayes is Used:

- **Efficiency**: Naive Bayes is known for its computational efficiency, especially in cases where the dimensionality of the data is high. Its simplicity allows for rapid model training and prediction, which is beneficial in clinical settings where timely diagnosis is critical.
- **Probabilistic Nature**: This algorithm provides not just classifications but also the probabilistic estimates of those classifications, offering valuable insights into the confidence of predictions. This aspect is particularly useful in medical diagnostics, where understanding the probability of disease presence can guide further testing and intervention strategies.
- **Strong Performance with Assumptions**: Despite its assumption of feature independence, Naive Bayes often performs well in practice even when the independence assumption does not hold, particularly in complex medical datasets where interdependencies are common.

The application of the Naive Bayes algorithm in this project was instrumental in efficiently identifying patterns and relationships in the data that are indicative of heart disease. This method's ability to quickly process and analyze extensive clinical data underscores its utility in predictive health analytics, contributing significantly to the project's goals of enhancing diagnostic accuracy and facilitating early intervention.

```
library( e1071 )

nb_model <- naiveBayes( train_predictors, as.factor( train_target ) )

nb_predictions <- predict( nb_model, test_predictors )
nb_conf_matrix <- table(
Predicted = nb_predictions, Actual = test_target )

ggplot( as.data.frame( nb_conf_matrix ), aes(
x = Actual, y = Predicted ) ) +
geom_tile( aes( fill = Freq ), colour = "white" ) +
geom_text( aes( label = Freq ), vjust = 1 ) +
scale_fill_gradient( low = "white", high = "steelblue" ) +
```

```
labs( title = "Confusion Matrix for Naive Bayes Model", x = "Actual Label", y = "Predicted Label" )

saveRDS( nb_model, file = "/content/NAIVE_BAYES_heart.rds" )
```



Support Vector Machine (SVM)

Support Vector Machine (SVM) is a robust supervised machine learning algorithm widely used for classification and regression tasks. In my Heart Disease Risk Prediction project, SVM was employed to classify patients based on the likelihood of having heart disease, utilizing its capability to find the optimal hyperplane that best separates the classes in a high-dimensional space.

Implementation of SVM in R:

- **Library Usage**: The SVM model was implemented using the e1071 package in R, which is based on the LibSVM library. This package is popular for its efficiency and effectiveness in handling various types of data, including complex medical datasets.
- **Data Preparation**: The data was carefully preprocessed before applying the SVM algorithm:
    - Continuous and categorical variables were appropriately formatted; categorical variables were encoded as factors to ensure they are correctly processed by the algorithm.
    - The data was divided into training and testing sets, maintaining a 70/30 split to ensure a robust evaluation of the model's predictive accuracy.
- **Model Training**: The svm() function from the e1071 package was used to train the model. This function allows for various kernel types to be specified; in this project, a linear kernel was chosen to maintain model simplicity and interpretability:
    - The linear kernel was selected due to its suitability for data with a linear boundary between classes, and because it helps in avoiding overfitting when the data dimensionality is high relative to the number of samples.
- **Model Evaluation**:
    - Post-training, the SVM model was used to make predictions on the testing dataset. The effectiveness of the model was evaluated by comparing these predictions against the actual outcomes.
    - A confusion matrix was constructed to provide a clear visual representation of the model's performance, highlighting the accuracy, and the numbers of false positives and false negatives.

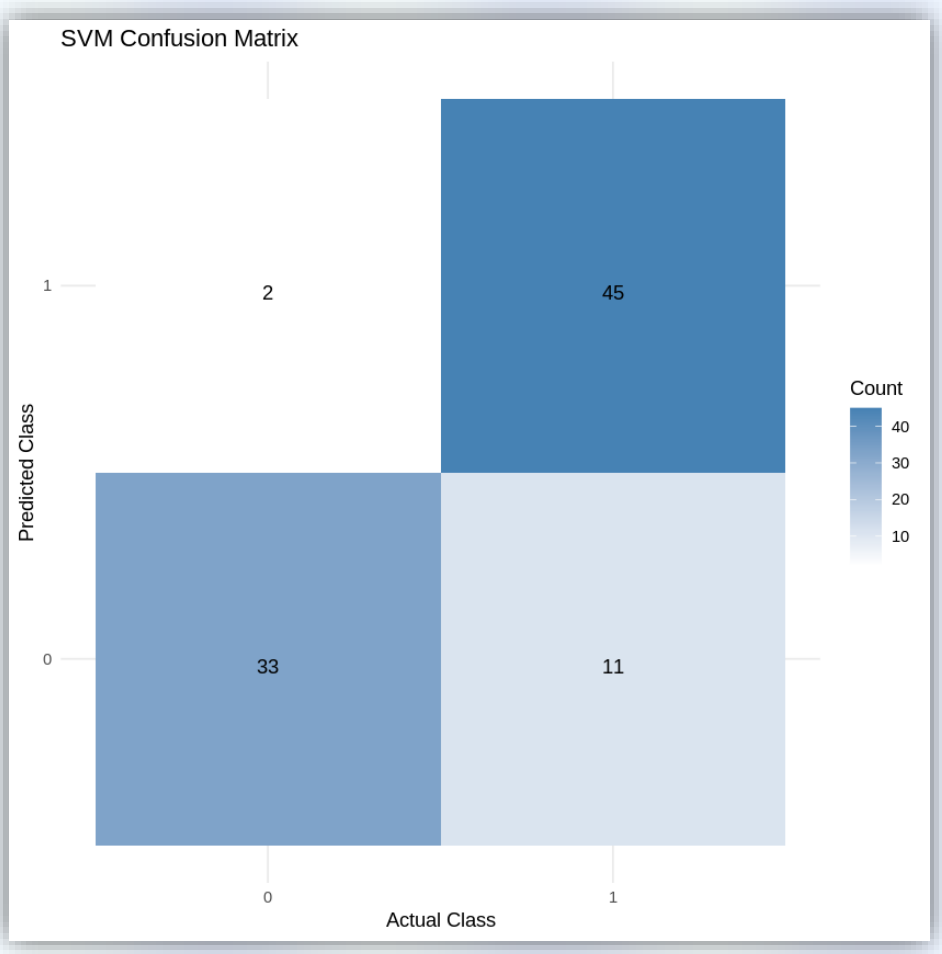
Why SVM is Used:

- **Maximal Margin Classifier**: SVM is known for its characteristic of maximizing the margin between the data points of the classes, which can be particularly useful in medical datasets where the distinction between classes (e.g., diseased vs. non-diseased) needs to be as clear as possible.
- **Handling High-Dimensional Data**: With heart disease data often encompassing a wide range of variables (from demographic data to detailed blood work results), SVM's ability to handle high-dimensional space effectively is highly beneficial.
- **Robustness to Overfitting**: Especially in scenarios where the number of dimensions exceeds the number of samples, SVM's regularization capabilities help prevent overfitting, making it a reliable choice for predictive modeling.

The application of the SVM algorithm in this project significantly enhanced the ability to discern subtle patterns in the data that indicate the presence of heart disease. This method's robustness and precision in classifying complex datasets ensured that the project achieved its objective of developing a predictive model with high accuracy and reliability, thereby aiding in the early diagnosis and better management of heart disease.

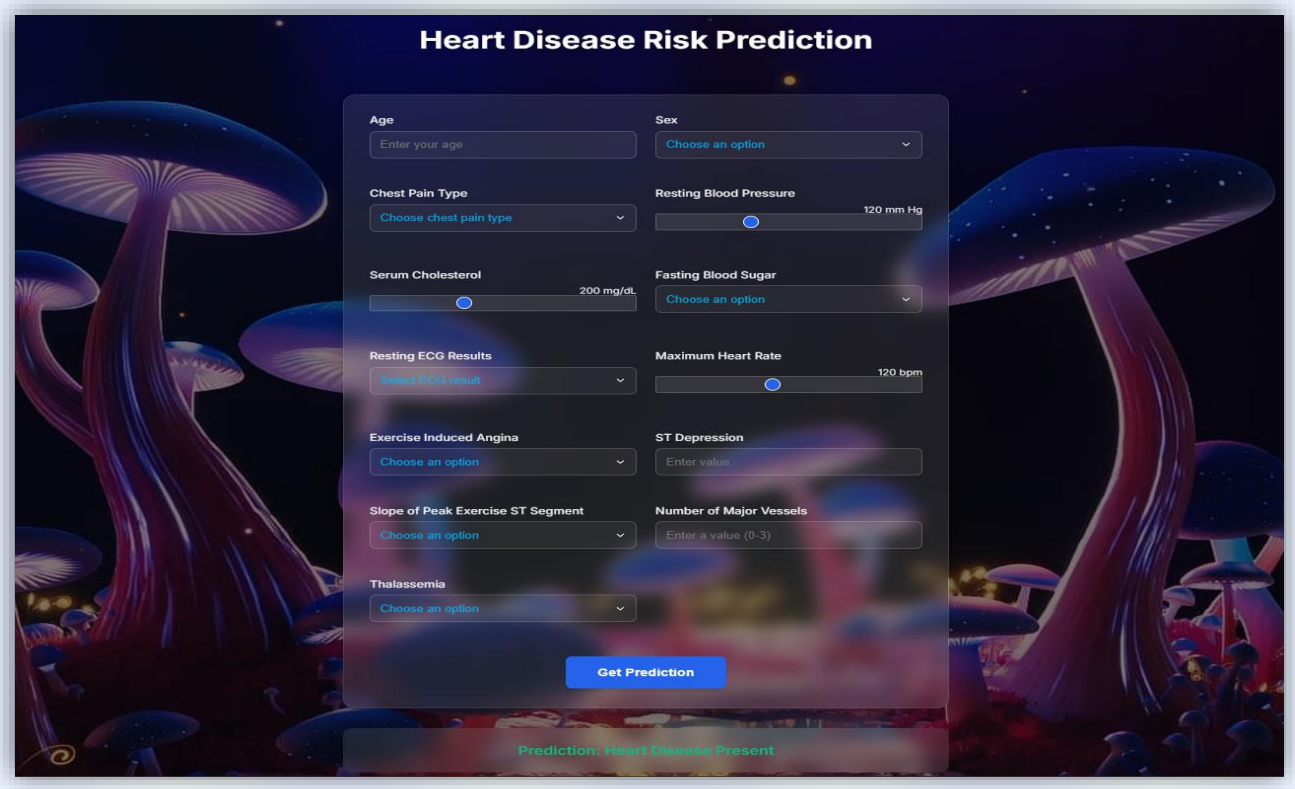
```
library( e1071 )

svm_model <- svm( train_target ~ ., data = data.frame(
train_predictors, train_target = train_target ), type = 'C-
classification', kernel = 'linear' )

svm_predictions <- predict( svm_model, newdata = test_predictors )

svm_cm <- table( Predicted = svm_predictions, Actual = test_target )

library( ggplot2 )
df_svm_cm <- as.data.frame( as.table( svm_cm ) )
colnames( df_svm_cm ) <- c( "Actual", "Predicted", "Count" )

ggplot( df_svm_cm, aes( x = Actual, y = Predicted, fill = Count )
) +
geom_tile() +
geom_text( aes( label = Count ), vjust = 1 ) +
```

```
scale_fill_gradient( low = "white", high = "steelblue" ) +
theme_minimal() +
labs( title = "SVM Confusion Matrix", x = "Actual
Class", y = "Predicted Class" )

saveRDS( svm_model, file = "/content/SVM_heart.rds" )
```



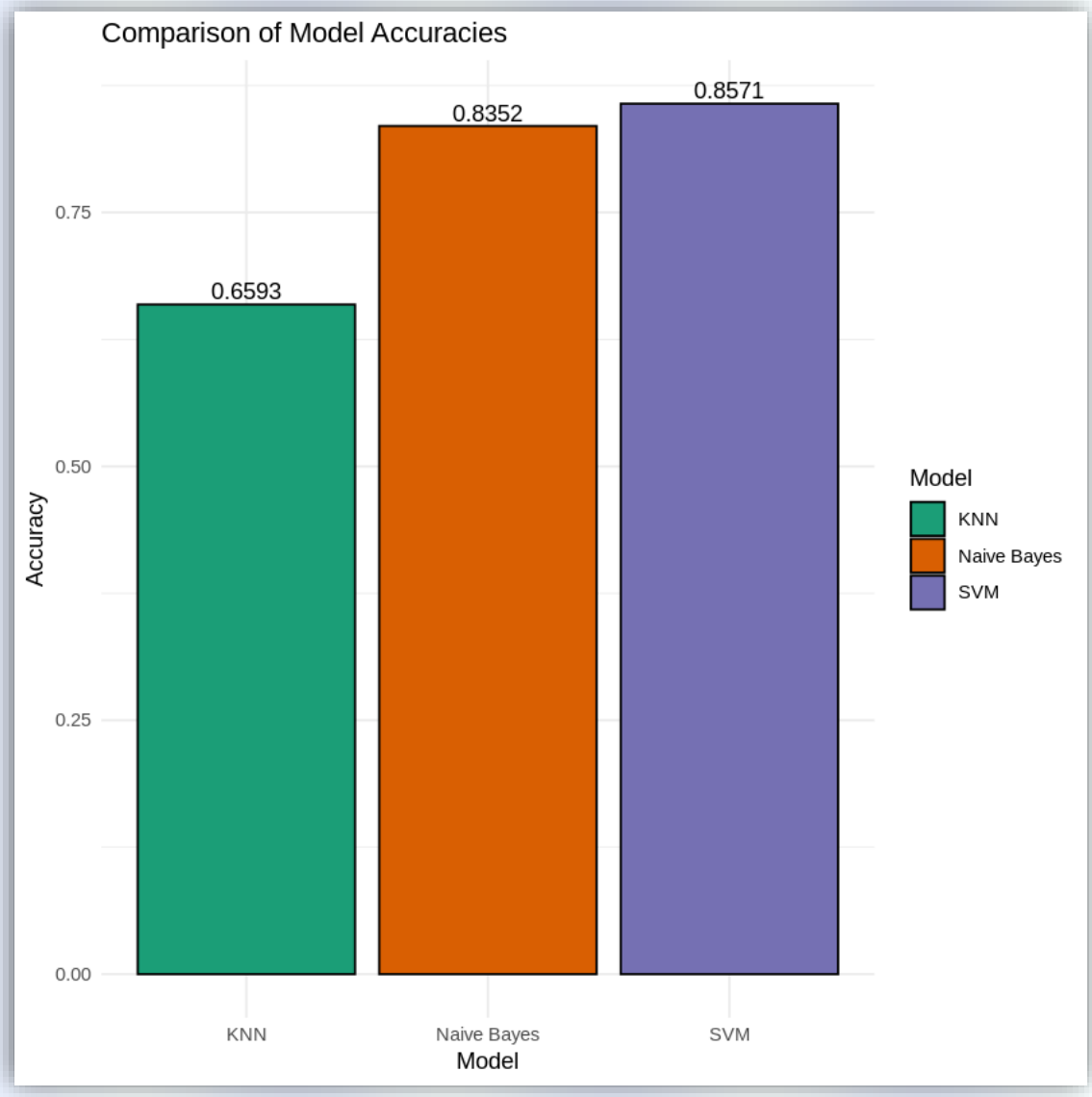
Realtime Prediction

Prediction on new data



MODEL PERFORMANCE COMPARISON

KNN VS NAÏVE BAYES VS SVM




```
knn_accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
nb_accuracy <- sum(diag(nb_conf_matrix)) / sum(nb_conf_matrix)
svm_accuracy <- sum(diag(svm_cm)) / sum(svm_cm)

model_accuracies <- data.frame(
  Model = c("KNN", "Naive Bayes", "SVM"),
  Accuracy = c(knn_accuracy, nb_accuracy, svm_accuracy)
```

```
)

 library(ggplot2)

ggplot(model_accuracies, aes(x = Model, y = Accuracy, fill = Model)) +
  geom_bar(stat = "identity", color = "black") +
  geom_text(aes(label = round(Accuracy, 4)), vjust = -0.3) +
  ggtitle("Comparison of Model Accuracies") +
  ylab("Accuracy") + xlab("Model") +
  theme_minimal() +
  scale_fill_brewer(palette = "Dark2")
```

Future Scope

The Heart Disease Risk Prediction project has laid a solid foundation for advancing the predictive analytics capabilities in healthcare, specifically targeting cardiovascular diseases. The successful implementation and evaluation of machine learning models like KNN, Naive Bayes, and SVM using the R programming environment have opened several pathways for future enhancements and expansions. Here's a detailed outlook on the future scope of this project:

1. **Model Enhancement and Optimization**:
    - **Algorithm Tuning**: Further tuning of the hyperparameters for each model could enhance their accuracy and overall performance. Techniques such as grid search or random search could be systematically applied to find the optimal settings.
    - **Advanced Machine Learning Models**: Integrating more complex models such as ensemble methods (e.g., Random Forests, Gradient Boosting Machines) or deep learning approaches could potentially provide improvements in predictive accuracy and model robustness.
2. **Feature Engineering**:
    - **Inclusion of More Variables**: Future iterations of the project could include additional variables, such as genetic markers, detailed lipid profiles, or patient lifestyle factors (diet, exercise, stress levels), to provide a more comprehensive analysis.
    - **Advanced Feature Selection Techniques**: Employing techniques like Principal Component Analysis (PCA) or Regularization methods could help in identifying the most impactful features and reducing dimensionality, thus improving model performance.
3. **Data Collection and Quality Improvement**:
    - **Larger Dataset**: Expanding the dataset to include a larger and more diverse population sample would enhance the generalizability of the models. Collaboration with healthcare providers to access more comprehensive datasets could facilitate this expansion.
    - **Data Quality Enhancements**: Efforts to improve the quality and granularity of data collected, such as more precise measurement techniques or standardized data entry processes, would enhance the reliability of the predictions.
4. **Real-Time Prediction Integration**:
    -

- **Integration with Healthcare Systems**: Developing an API that allows real-time data input and prediction output could be integrated into electronic health record (EHR) systems, enabling healthcare providers to assess heart disease risk as part of routine patient assessments.
- **Mobile App Development**: Creating a patient-friendly mobile application that allows users to input personal health data and receive instant risk assessments could empower individuals to manage their health proactively.

5. **Validation and Clinical Trials**:
   - **Prospective Validation**: Conducting prospective studies to validate the predictive models in real-world clinical settings would be crucial. This would involve using the model predictions to influence patient treatment plans and monitoring outcomes to assess impact.
   - **Ethical and Regulatory Considerations**: Ensuring that the use of AI in healthcare complies with ethical standards and regulatory requirements, including patient privacy concerns and model bias mitigation.
6. **Educational and Awareness Programs**:
   - **Patient Education**: Developing educational programs that leverage the insights gained from the data analysis to inform patients about heart disease risks and prevention strategies.
   - **Healthcare Provider Training**: Offering training for healthcare providers on using predictive analytics in their practice could improve the adoption and effectiveness of these technologies.

By pursuing these avenues, the Heart Disease Risk Prediction project can significantly evolve to not only enhance predictive accuracy but also to become a vital part of preventive healthcare strategies, ultimately leading to better patient outcomes and reduced healthcare costs.

References

1. **Kaggle**: The heart disease dataset utilized for this project was sourced from Kaggle, which is widely recognized for its rich repository of medical datasets. Kaggle's platform not only provided the data but also offered a community-driven resource for innovative data science approaches and peer insights that were instrumental in shaping the project's methodology.
2. **YouTube**: Educational videos on YouTube were invaluable in understanding complex statistical concepts and machine learning algorithms. Specific tutorials on implementing machine learning models in R, such as KNN, Naive Bayes, and SVM, helped clarify the practical aspects of the algorithms used in the project.
3. **Google Data Science Bootcamp**: Participating in Google's Data Science Bootcamp provided comprehensive training in data manipulation, visualization, and machine learning techniques. The bootcamp's sessions on R programming and predictive analytics directly influenced the data preprocessing and modeling phases of the project.
4. **R Documentation**: Official documentation for R packages such as ggplot2, e1071, class, and caret was crucial for understanding the functions and their applications. This documentation served as a guide for effectively employing various data analysis and machine learning techniques within the project.
5. **Data Science Blogs**: Numerous blogs and articles dedicated to data science, especially those focusing on healthcare analytics, offered advanced insights into feature selection, model evaluation, and the interpretation of machine learning results. These resources were pivotal in refining the project's analytical strategies and enhancing the overall model performance.
6. **Research Papers on Cardiovascular Disease Analytics**: Academic journals and research papers provided a deeper understanding of the trends and methodologies in cardiovascular disease analytics. These papers discussed the latest advancements in predictive modeling for heart disease and were instrumental in aligning the project with current scientific standards and practices.

Bibliography

1. **Kaggle:** "Heart Disease UCI Dataset." Kaggle, 2020. Available online: Kaggle Heart Disease Dataset. This dataset provides clinical and demographic data essential for predicting heart disease, sourced from a reliable community-driven platform for data science projects.

   YouTube Tutorials:
2. "R Programming Tutorial - Learning Data Analysis and Visualization." YouTube, presented by FreeCodeCamp, 2021. This tutorial series covers essential R programming skills necessary for data analysis and machine learning applications in healthcare.
3. "Implementing KNN with R." YouTube, presented by Data Science Dojo, 2020. A practical guide to implementing the KNN algorithm in R, providing insights into model setup, tuning, and evaluation.
4. Official R Documentation:
5. R Core Team. "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria, 2021. URL: R Project. The core documentation for R provides comprehensive guidelines on utilizing various packages and functions for statistical analysis.
6. Meyer, David, et al. "Package 'e1071'." CRAN, 2021. Documentation for the e1071 package which includes SVM implementation guidelines in R.
7. Data Science Bootcamps and Courses:
8. "Google Data Analytics Professional Certificate." Coursera, 2021. A detailed curriculum focusing on data analysis, including predictive analytics using machine learning, which directly influenced the project's approach
9. Web Resources:
10. "Data Visualization with ggplot2." RStudio, 2021. Web. Available online: RStudio ggplot2. Guides and tutorials on using ggplot2 for data visualization in R.

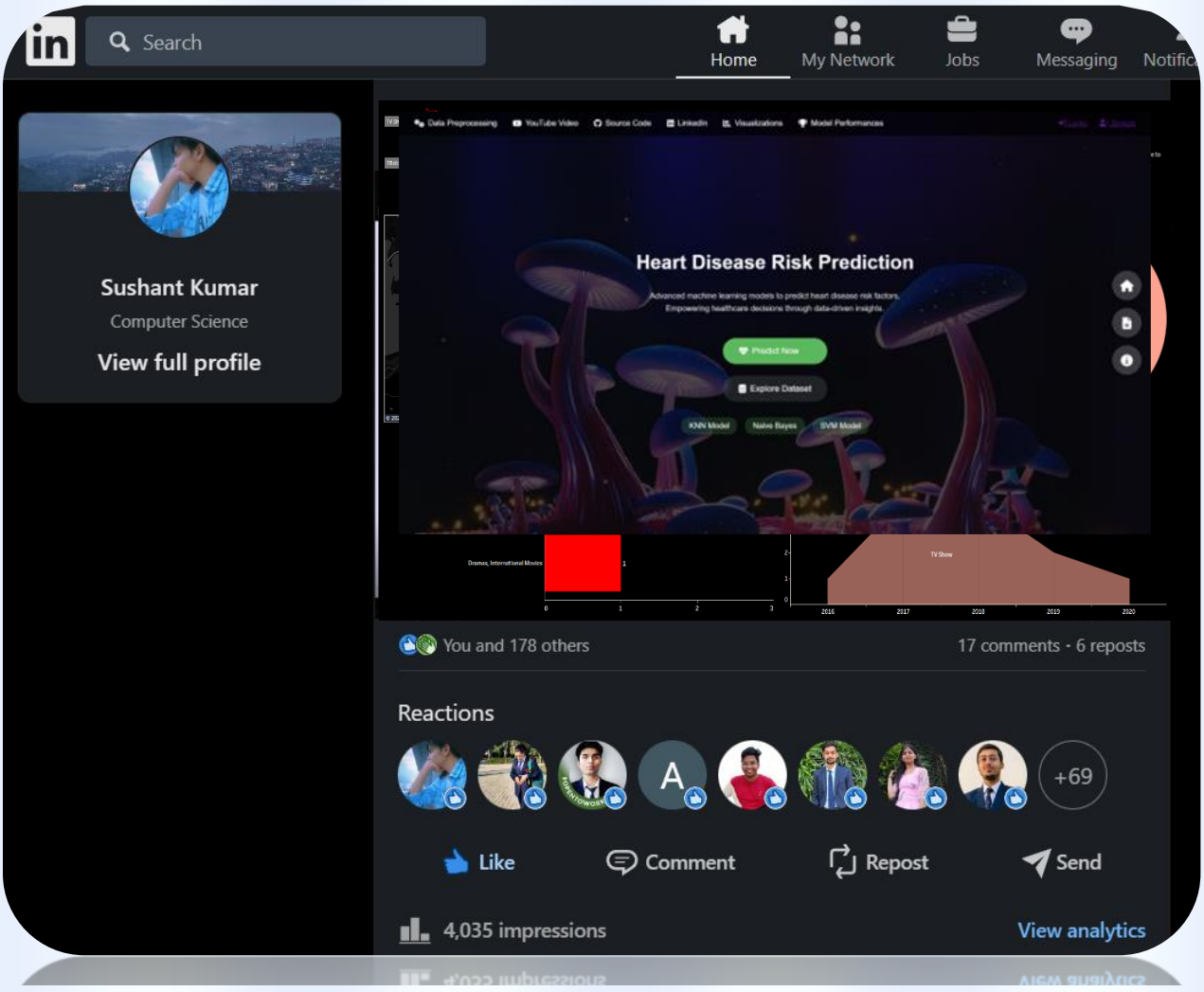
Social Media Insights

Student Profile

https://www.linkedin.com/in/07sushant/

Social Media Insights

Post Insights



https://www.linkedin.com/in/07sushant/

Thank You