ABSTRACT

Heart disease is a major cause of death globally and machine learning techniques have the potential to improve the diagnosis, treatment, and prevention of this condition. In this context, the project work aimed to investigate several heart disease datasets that are commonly available on popular data sites such as Kaggle.

During this project work, the team encountered several issues related to human errors and negligence while attempting to authenticate the medical datasets. These issues included encoding errors and duplicates, which could undermine the reliability of inferences or predictive models built on some of the published datasets. Therefore, the project team conducted feature analysis and used Logistics Regression, Bagging, Boosting, Decision Tree, Naïve Bayes models to identify the best dataset for machine learning and statistical analysis.

The features identified as statistically significant in explaining or classifying patients as having heart disease included maximum heart rate (thalach), chest pain (cp), and oldpeak.

The research work sheds light on the limits of current heart disease datasets, emphasizing the need of data authentication and feature analysis in improving the reliability and accuracy of machine learning and statistical analysis. The accuracy of prediction models in identifying persons at risk of heart disease can be enhanced by choosing the best dataset and statistically important characteristics.

Furthermore, the technique taken by the project team to addressing data concerns through feature analysis and machine learning models may be applied to other medical datasets to improve their reliability and accuracy.

Overall, this project work contributes to current efforts to enhance heart disease diagnosis, treatment, and prevention using data-driven approaches. The project's findings may have significant ramifications for the development of machine learning-based decision support systems for heart disease diagnosis, treatment, and prevention.

❖ Problem Description

Everything from organic to inorganic has been designed with a heart structure to aid in fulfilling daily functions. Humanity has utilized this same format with advancing technology today. In fact, the Center for Disease Control and Prevention (CDC) in 2018 classifies heart disease as the leading cause of mortality in the United States and remains the leading cause of mortality to date. Previous research invested in developing information surrounding the heart has sought to improve the methods of managing our heart's condition. With all the data being produced, machine learning aids with patient-level observations, where algorithms sift through vast numbers of variables, looking for combinations that reliably predict outcomes. Since the conclusion of a 2018 research study, the American Heart Association website posted that there has been a 15.1 % decrease across the United States. Further, according to CDC, in the United States, someone has a heart attack every 40 seconds and every year, about 805,000 Americans have a heart attack. Further, about half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease namely high blood pressure, high cholesterol, and diabetes.

Due to the complexity and the variations of the increasing number of risk factors, modern researchers are relying on Data Mining and Machine Learning techniques. Because of privacy concerns and other issues related to accessing public data, there are very few heart disease data sets available for the public to analyze. One of these sites is Kaggle. Kaggle is a very popular site which allows users to find and publish data sets, explore, and build models in a web-based data-science environment, and collaborate with other data scientists and machine learning engineers, and even enter competitions to solve data science challenges. In our quest for health-related data, we found on the site that heart.csv data has been one of the most analyzed by avid data analysts/scientists around the world.

We explored some of the reported analysis results on the data. We found that most of the analysis was not done on well-processed data. For instance, the variables 'sex' and 'cp' (Chest Pain) were kept as numeric instead of converted to categorical. Further, these analyses often involve machine learning codes which return a classifier output without enough information for the reader to accurately gauge the features' importance or the classification's criteria.

❖ Data Descriptions

The initial dataset came from Kaggle

(Dataset Link : https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset), a popular repository for data, which referenced a data file on Heart disease. This process is the beginning stages of data cleaning.

❖ Proposed Work

Through this project we aim to build an ML model that will predict heart disease. We have built five models using Logistics Regression, Bagging, Boosting, Decision Tree, Naïve Bayes algorithms and check which of these five gives us higher accuracy.

This project is being built on WORKBENCH using the R programming language. We are making use of functions stored in the following packages:

- ✓ library(ggplot2)
- ✓ library(dplyr)
- ✓ library(ggcorrplot)
- ✓ library(RColorBrewer
- ✓ library(randomForest)
- ✓ library(caret)
- ✓ library(factoextra)
- ✓ library(rpart)
- ✓ library(adabag)
- ✓ library(e1071)

The following order is being followed to achieve are result:

1. Loading Data

2. Cleaning Data

3. Data Visualization

4. Factoring to differentiate categorical data.

5. Feature Selection using "Correlation."

6. Splitting data in "Train" & "Test"

7. Model Fitting (Logistics Regression, Bagging, Boosting, Decision Tree, Naïve Bayes)

8. Testing & Evaluating Accuracy

❖ Attribute Information

AGE: age in years

SEX: (1 = male; 0 = female)

CP (Chest Pain Type):

--Value 0: typical angina (most serious)

--Value 1: atypical angina

--Value 2: non-anginal pain

--Value 3: asymptomatic (least serious)

TRESTBPS: resting blood pressure (in mm Hg on admission to the hospital)

CHOL: serum cholesterol in mg/dl

FBS: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

A fasting blood sugar level of less than 100 mg/dL is normal. From 100 to 120 mg is considered prediabetes. If it is 125 mg/dL or higher on two separate tests, you have diabetes.

RESTECG (Resting Electrocardiographic Results):

--Value 0: normal

--Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

--Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

THALACH: maximum heart rate achieved.

EXANG: exercise induced angina (1 = yes; 0 = no)

OLDPEAK: ST depression induced by exercise relative to rest.

SLOPE (the slope of the peak exercise ST segment):

--Value 0: upsloping

--Value 1: flat

--Value 2: downsloping

CA: number of major vessels (0-3) colored by fluoroscopy

THAL:

--Value 0- No Thalassemia

--Value 1- Normal Thalassemia

--Value 2- Fixed Defect Thalassemia

--Value 3- Reversible Defect Thalassemia

TARGET: diagnosis of heart disease (angiographic disease status)

-- Value 0 - Absence of heart disease

-- Value 1 - Presence of heart disease

❖ Attributes Descriptions

- **Age**:
  Because heart health risk increases with age, the screening age range for participants is greater than or equal to 35. Although heart risk is not typical for younger age groups, Heart Foundation (HF) doesn't detour younger participants especially family's that have a history with heart health risks.

- **Sex**:
  The standing statistic for heart disease it's the primary cause of death for both men and women in the U.S. The nominal values 0 and 1 are assigned to female and male, respectively, to facilitate concise binary evaluation. Assigning these values is necessary for certain types of machine learning which will be used in the evaluation later after more critical attributes have been considered.

- **Chest Pain**:
  The most important indicator of heart disease is chest discomfort or chest angina. Because discomfort provides the first indication of beginning symptoms for many diseases, chest angina establishes there is a problem present. The data tables below angina levels may be case by case due to pain tolerance. Although this is one of the most dangerous symptoms of many diseases, this attribute can act as a control for initial diagnoses verse actual diagnoses.

- **Blood Pressure**:
  Blood pressure contributes to the systematic structure for indicating the heart's proper functioning. As with other fields of study the pressure measures the stress of the blood by the contraction of the surrounding dimensions. As blood flows through blood vessels, the Blood pressure rises and falls naturally throughout the day. By analyzing this innate movement, research data collected may provide an more insight to medical professionals

about the patient’s heart condition. When the pressure remains too high for an extensive time period, the repercussions may result in high risk for CAD, heart attack stroke, and a series of other indicators for heart failure. The CDC has recorded that High Blood Pressure, also known as Hypertension, affects about 1 in 4 adults (24%) with hypertension have their condition under control in the US. Which infers American adults may not even be aware of they have it because the symptoms aren’t prevalent as diseases. Resting blood pressure in millimeters of mercury (mm Hg) when the patient was admitted to the hospital.

- **Cholesterol**:
  Cholesterol is a fat-like substance called a lipid that’s found naturally in the blood. Lipids are vital for the normal functioning of the body. The human body manufactures all the cholesterol it needs from diet. Cholesterol can be measured with a simple blood test.

- **Fasting Blood Sugar**:
  Blood Glucose or commonly recognized as Blood Sugar Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Milligrams per deciliter, is a measurement that indicates the amount of glucose in a specific amount of blood. According to the CDC, about 1 in 4 people with diabetes don’t know they have the disease. For the data sets, blood sugar is distinguished whether the patient’s blood sugar is higher than 120 mg/dl or not.

- **Electrocardiogram**:
  The electrocardiogram results are accepted as the current standard for evaluation of patients. Results cater to the patient’s body during the exercise. A patient with stable angina occurring through exercise which happens even on rest the disease got worse. This must be why there are so few patients that show an abnormality in the heart rate on rest, and it is also why seeing this abnormality is very indicative of a presence of a heart disease. On the other hand, the value 0, probable presence of a hypertrophy, doesn’t seem to be very indicative of the presence of a heart disease by itself.

  "Thallium Stress Test" also known as "Nuclear Stress Test" or "Cardiac Test" benefits heart disease research by analyzing the condition of blood flow. The gamma camera through nuclear imaging tracks the participant’s blood flow which carries a sample amount of Thallium, radioactive isotope.

- **Heart Rate**:
  The maximum heart rate was recorded during Thallium stress test. The data set showcases the optimal maximum healthy heart rate depends on the age (220 - age). Thus, higher rates tend to be from younger patients.

- **Exercise Pain**:
  This attribute is the patient’s level of angina or induced pain during exercise which is a necessary input for the presence of heart disease.

- **Old Peak**:
  The resting Stress Test segment depression is the marker for adverse cardiac events. The old peak monitors a certain level in a normal heartbeat which indicates a displacement for the presence of a heart disease.

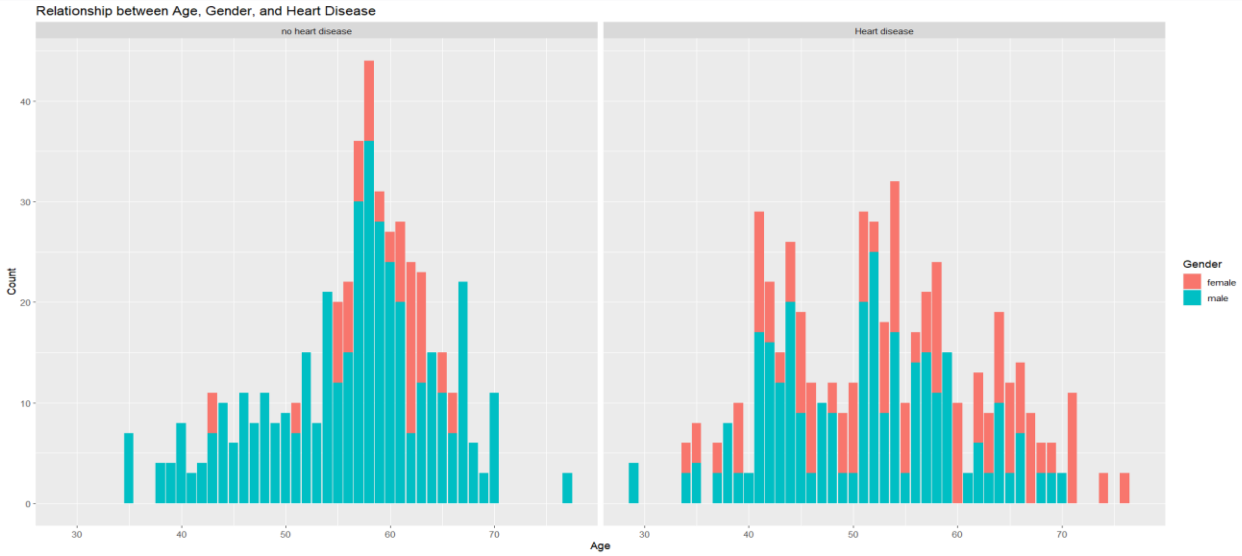
- **Slope**:
  The part of the Stress Test for indicators of exercise. The slope by itself can help determine whether there is heart disease or not if it is flat or ascending. Adding a third variable where we can see if the slope is descending, the depression of the ST segment can help to determine if the patient has a heart disease.

- **Target**:
  The target is designated as the condition of the patient for heart disease after conducting stress testing indicators. The results of the testing had an extensive range that determined the disease presence. For the sake of simplicity and design of approach, the patients which had any indication of disease present then in this analysis it was considered diagnosed as heart disease.

❖ Statistics and Inferences:

Using few Statistics and visualization tools we can be able to answer the below questions,

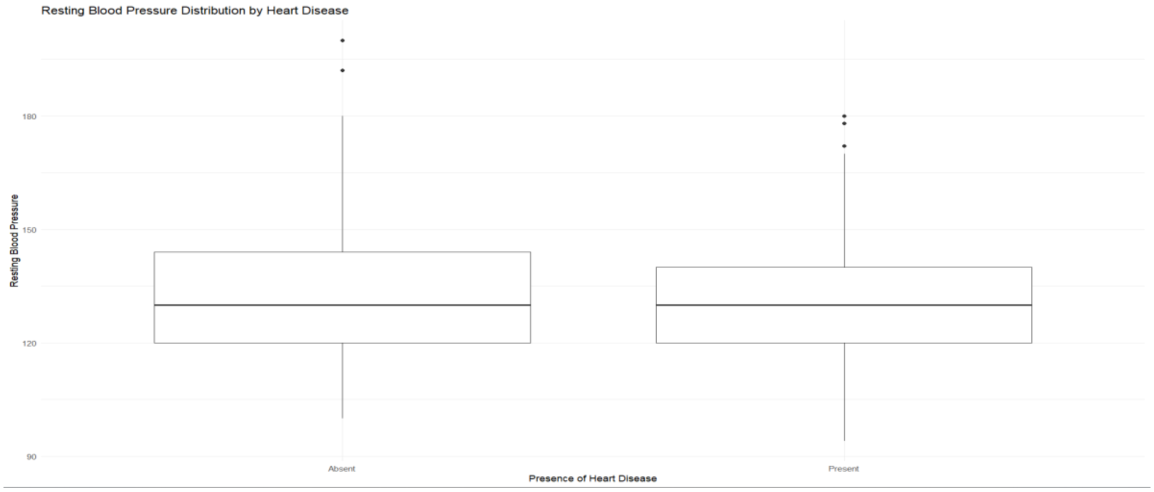
➢ How does age and gender affect the likelihood of having heart disease?



The inferences about age and gender of having a heart disease are:

1. People from the age group 40 to 65 years are more prone to have heart disease. The age group at 54 years has the highest count of people to have heart disease.
2. By looking at the above graph, males are more prone to the heart diseases in early stages of life (i.e., 30-60 age group). Further, females are more prone to the heart diseases in the later stages of life (i.e., 40-75 age group).
3. Finally looking at the no heart disease graph. We can infer that the males are less likely to be prone to heart disease than that of females. And the age group at 58 years has the highest count of people with no heart diseases.

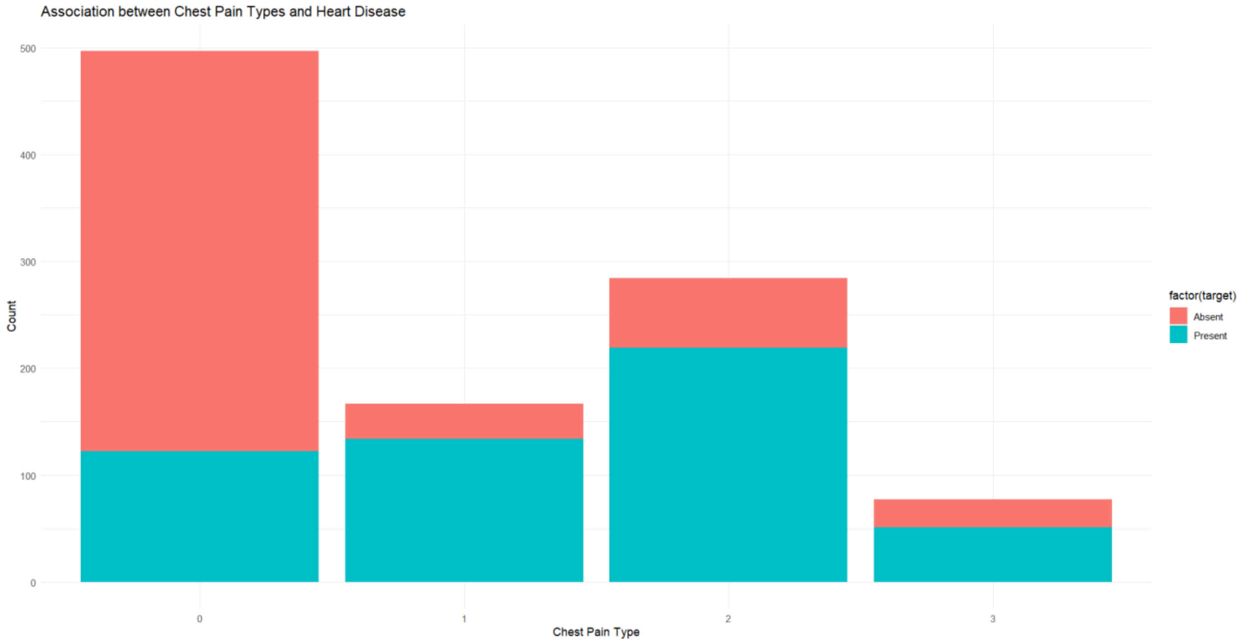
➢ Is there a correlation between resting blood pressure and the presence of heart disease?



By looking at the box plot we can conclude that:

1. Although people with heart disease seem to have a median resting blood slightly higher than that of people without heart disease, the difference is negligible.
2. Hence, we can conclude that there is no correlation between resting blood pressure and the presence of heart disease.

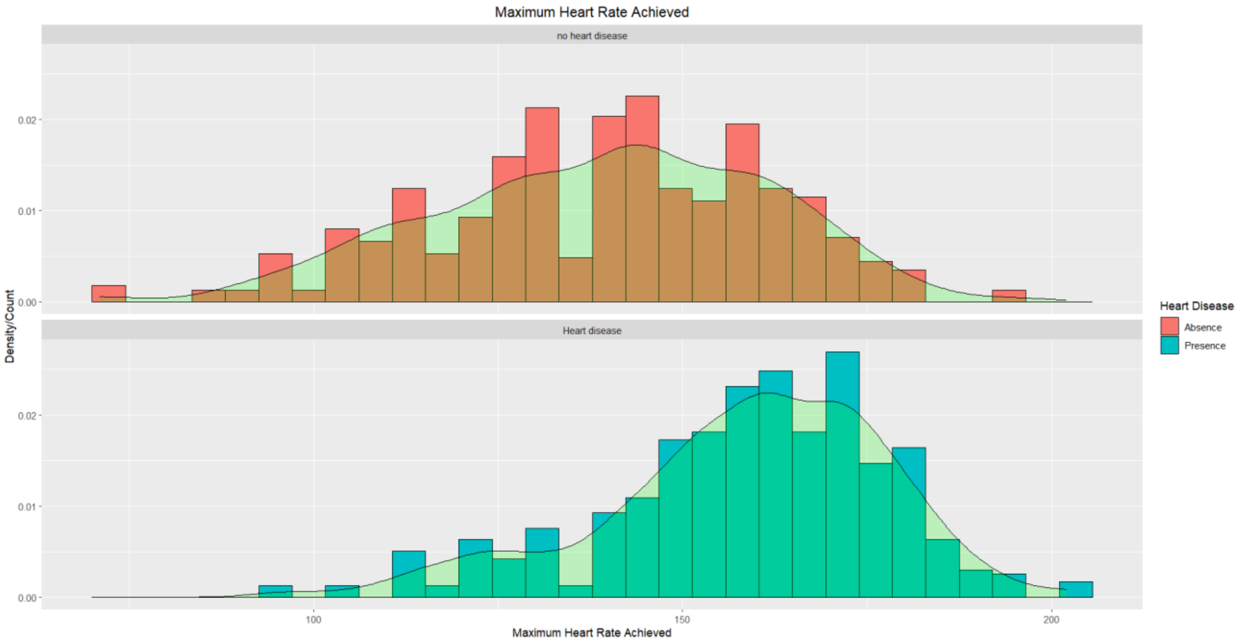
- Are certain types of chest pain (cp) more strongly associated with heart disease than others?



Yes, certain types of chest pain (cp) are more strongly associated with heart disease than the others. They are:

1. Non-Anginal pain (cp type 2) is more strongly connected with heart disease than typical anginal, atypical angina and asymptomatic pains (i.e., cp type – 0,1 and 3).
2. Typical anginal pain (cp type 0) is more strongly connected with no heart disease than atypical angina, asymptomatic and non-anginal pains (i.e., cp type – 1,2 and 3).
3. Atypical angina pain (cp type 1) is more strongly connected with heart disease than asymptomatic pain. (i.e., type 3)

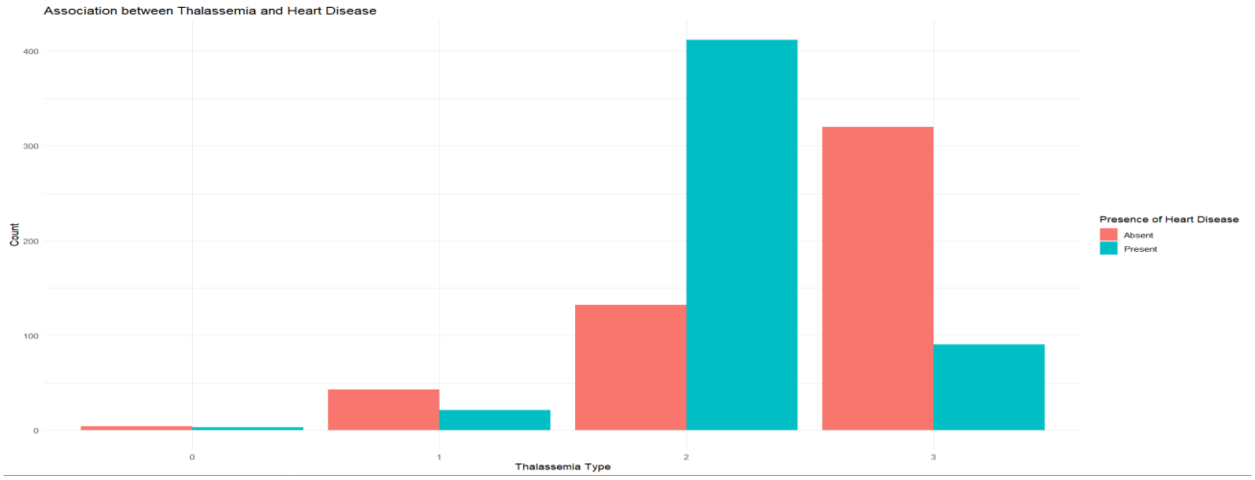
- ➢ How does the presence of heart disease relate to the maximum heart rate achieved during exercise (thalach)?



Yes, the presence of heart disease is strongly related to the maximum heart rate achieved during exercise (thalach).

1. Individuals who suffer from heart disease seem to attain a heart rate of 150-200 (thalach).
2. Whereas, Individuals who do not suffer from a heart disease have a maximum heart rate of around 150 (thalach).
3. Thus, it can be said that an elevated pulse rate may be indicative of the presence of heart disease.

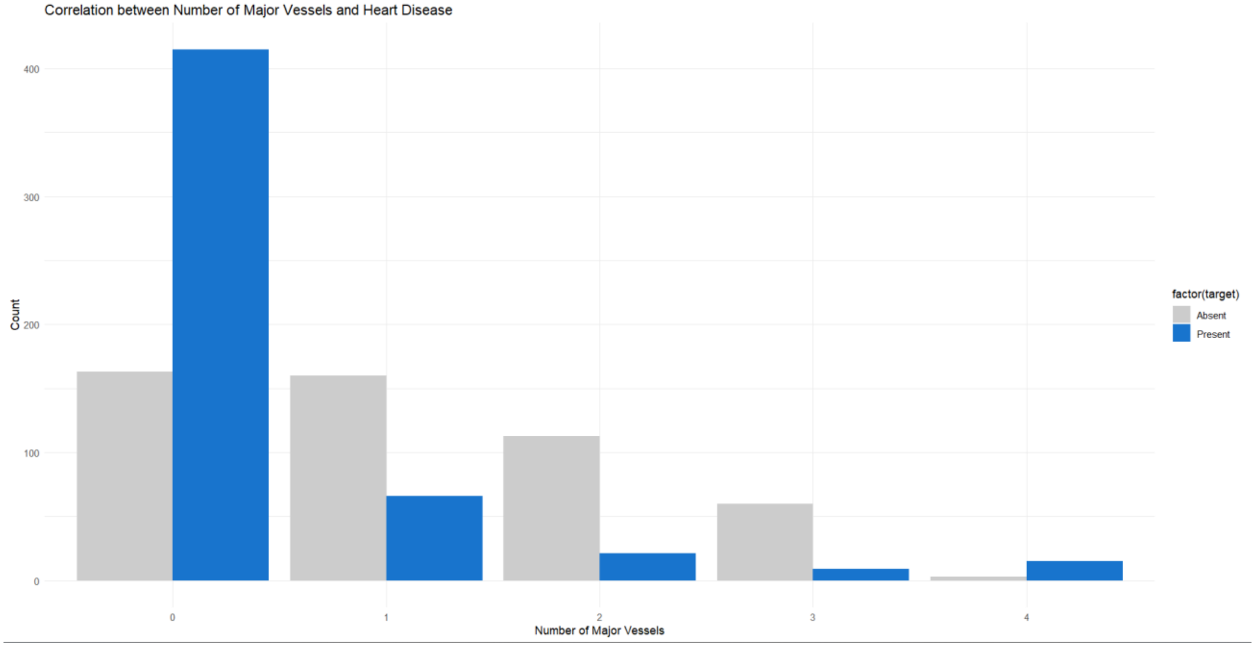
- ➢ Are there any values of the thalassemia blood disorder (thal) that are more strongly associated with heart disease?



Yes, some types of thalassemia blood disorder (thal) are more strongly associated with heart disease.

1. Fixed Defect Thalassemia type (i.e., thal type 2) is the most strongly connected with the presence of heart disease.
2. Reversible Defect Thalassemia (i.e., thal type 3) is the most strongly connected with the absence of heart disease.
3. Finally, by looking at the graph No Thalassemia & Normal Thalassemia (i.e., thal type 0 & 1) has no effect on the presence of heart disease.

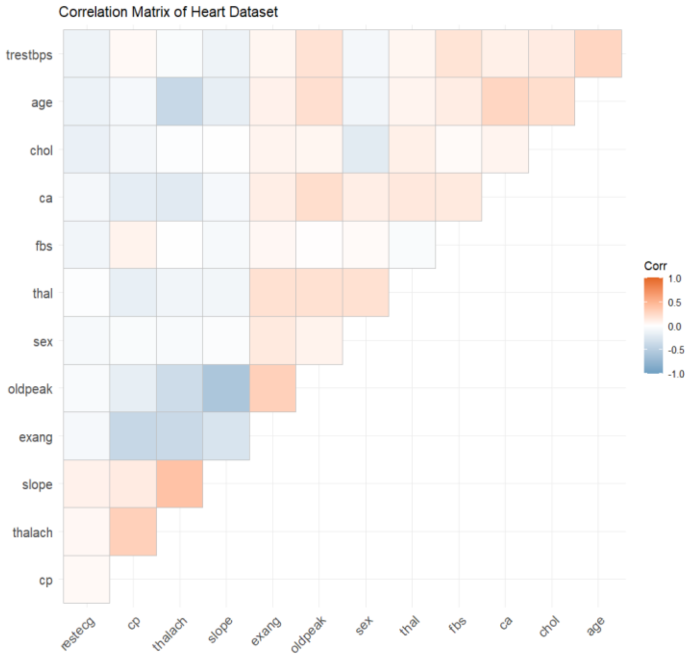
- Are there any correlations between the number of major vessels (ca) colored by fluoroscopy and the presence of heart disease?



Yes, there are correlations between the number of major vessels (ca) colored by fluoroscopy and the presence of heart disease. They are:

1. If there are no major vessels colored by fluoroscopy, then there is a strongest possibility of heart disease.
2. The possibility of having heart disease decreases as the number of major vessels colored by fluoroscopy increases (i.e., ca 0-4).
3. The possibility of heart disease increases slightly at 4 major vessels colored by fluoroscopy (i.e., ca 4) than that of 3 major vessels (i.e., ca 3).

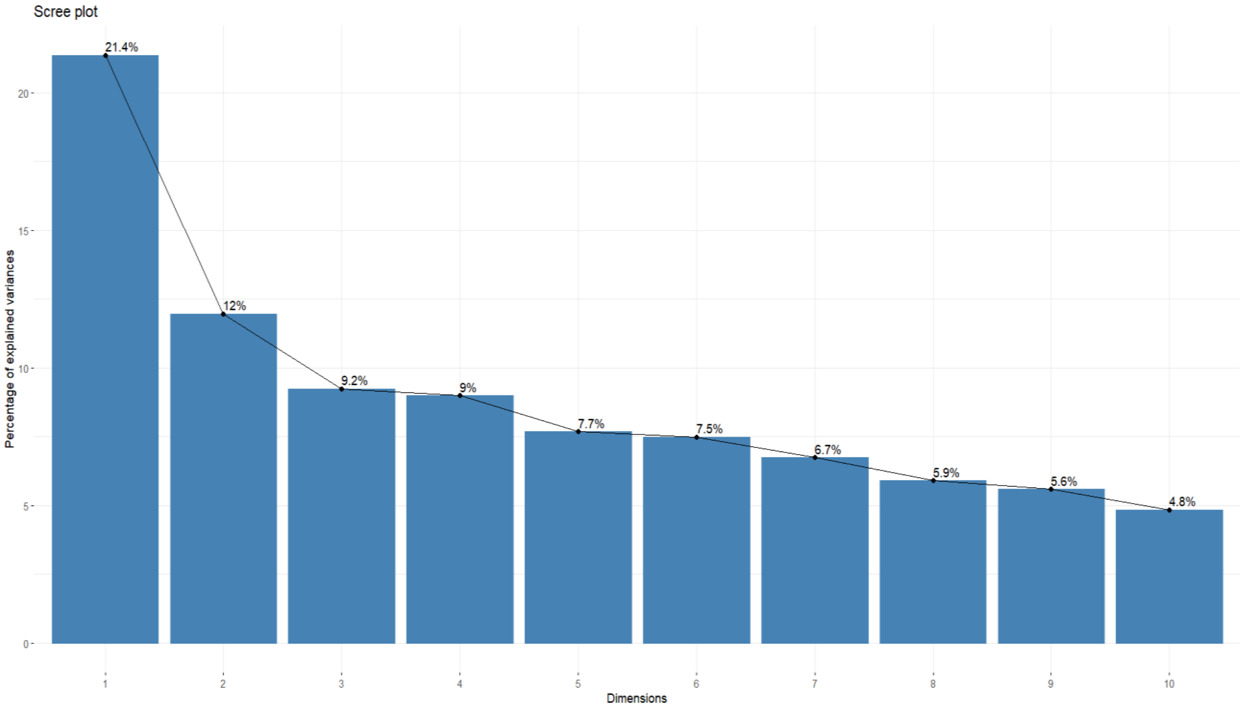
- How do different features of the dataset relate to one another?



A correlation matrix and heat map to identify the correlations between different variables in the dataset.

1. Slope and thalach have high positive correlation.
2. Slope and old peak have high negative correlation.

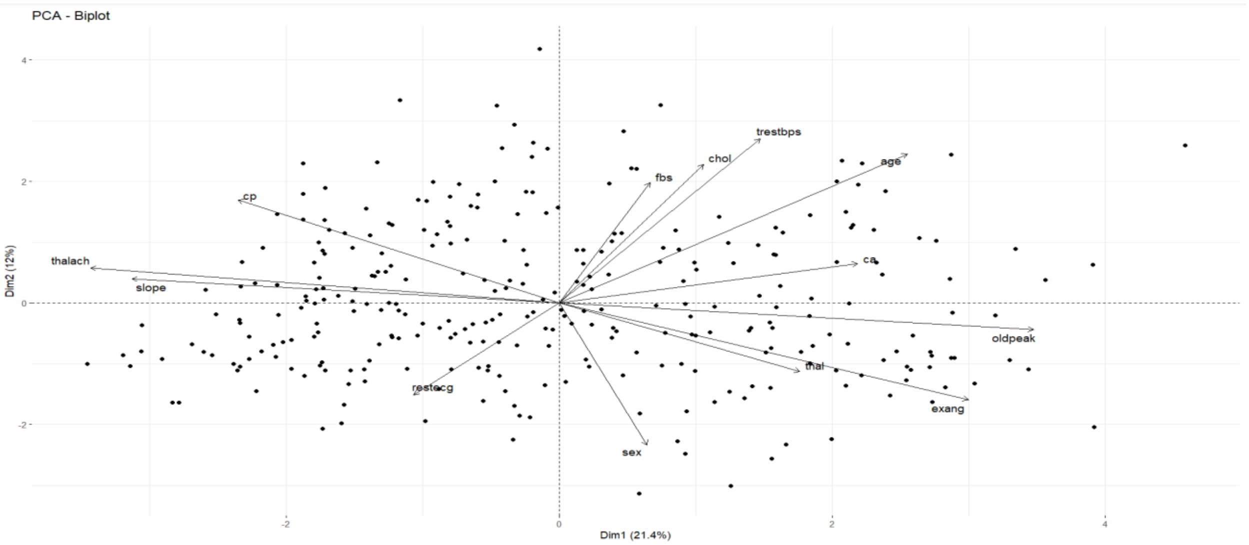
- What are the strongest predictors of heart disease?



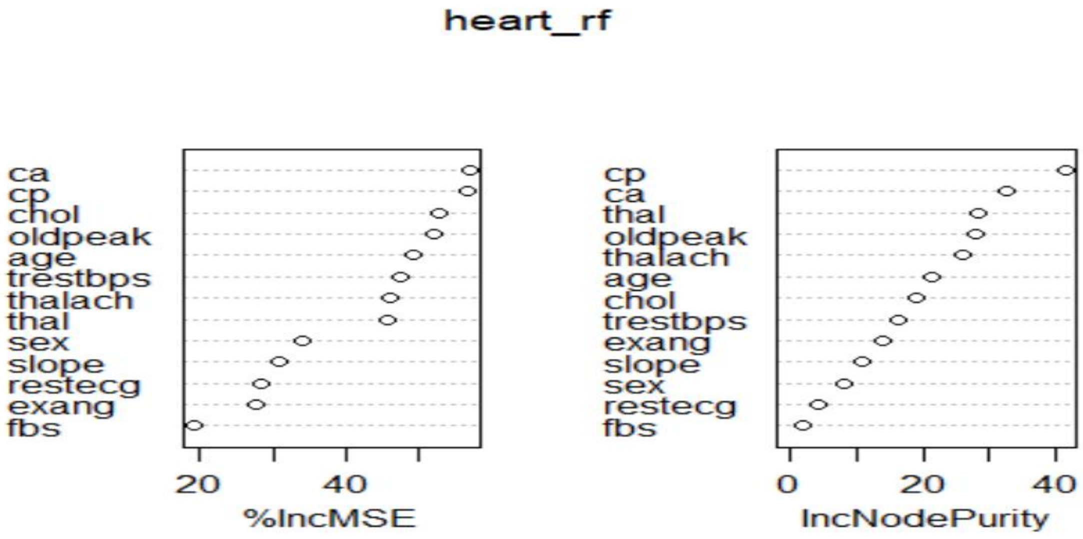
Principal Component Axis (PCA):

1. Since there is a clear elbow in the scree plot, it may be easy to decide on the number of principal components to keep.

.



1. The biplot displays the relationships between the variables and the principal components in a two-dimensional space.
2. It shows the relationships between the variables and the principal components. Variables that are located close to each other on the biplot are positively correlated, while variables that are located far apart are negatively correlated. The direction and length of the arrows indicate the contribution of each variable to the principal components.
3. Overall, the biplot can provide insights into the underlying patterns and relationships in the data and can be used to identify important variables and potential outliers.



1. The resulting graph shows a horizontal bar plot with bars representing the importance of each predictor variable, sorted in decreasing order. The y-axis shows the names of the predictor variables, and the x-axis shows the percentage increase in the mean squared error (MSE) of the model when that variable is removed.

2. Variables with higher importance are those with longer bars on the right-hand side of the plot. In this case, you can see which variables have the most significant impact on the model's accuracy. These variables are usually the most important for making predictions in the dataset.

Machine Learning Models

❖ Ensemble Methods:

1. Boosting:

Boosting is a machine learning ensemble technique used for improving the accuracy of a model by combining several weak models into a strong one. The idea behind boosting is to train a series of models sequentially, where each subsequent model learns to correct the mistakes of the previous model.

Boosting can be used with various machine learning algorithms, such as decision trees, neural networks, and support vector machines. The most popular boosting algorithm is AdaBoost, which stands for Adaptive Boosting. AdaBoost assigns higher weights to misclassified samples in each iteration, so that subsequent models are trained to focus more on those samples.

The general process of training a boosting model involves the following steps:

1. Train a weak model on the training data

2. Identify misclassified samples from the training set.

3. Increase the weights of the misclassified samples.

4. Train another weak model on the updated training data

5. Repeat steps 2-4 until a pre-determined number of weak models are trained or until the desired accuracy is achieved.

6. Combine the weak models into a strong one by assigning weights to each model based on their accuracy.

Boosting models tend to perform better than single models, as they can capture complex patterns and make better predictions. However, boosting models are more computationally expensive and can be prone to overfitting, especially if the number of weak models is too large.

2. Bagging:

Bagging, short for Bootstrap Aggregation, is a machine learning ensemble technique used for improving the accuracy and stability of a model by combining multiple independent models trained on different subsets of the training data. Bagging can be used with various machine learning algorithms, such as decision trees, random forests, and support vector machines.

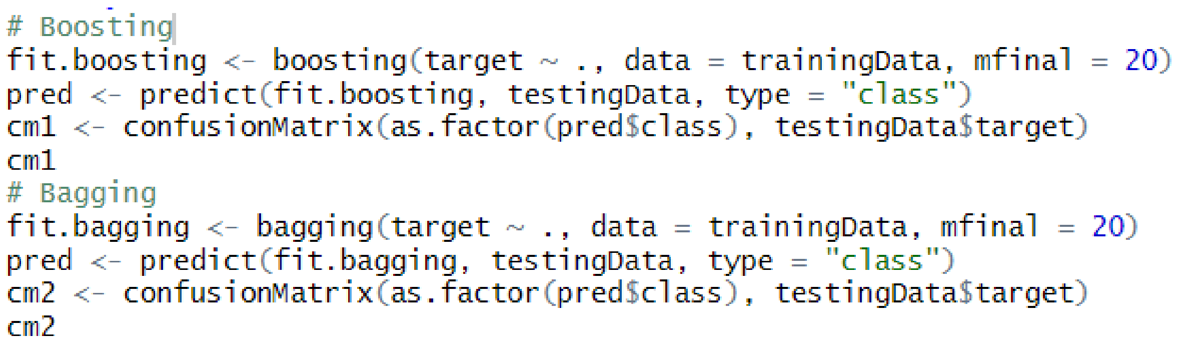
The general process of training a bagging model involves the following steps:

1. Randomly sample the training data with replacement to create multiple subsets (called bags) of equal size.
2. Train a weak model on each bag independently.
3. Combine the predictions of all weak models through averaging or voting.

Bagging models tend to perform better than single models, as they can reduce variance and improve stability. This is because each weak model is trained on a different subset of the training data and can capture different patterns and relationships. In addition, bagging models can help reduce overfitting, especially if the weak models are prone to overfitting.

One popular bagging algorithm is the Random Forest, which is a collection of decision trees trained on random subsets of the training data and features. The Random Forest algorithm improves upon bagging by introducing randomness in the feature selection and node splitting process, which can reduce correlation among the trees and improve the accuracy and robustness of the model.

Here we are using boosting and bagging techniques which improve the accuracy of models by combing multiple weaker models into a single strong model.



```
# Boosting
fit.boosting <- boosting(target ~ ., data = trainingData, mfinal = 20)
pred <- predict(fit.boosting, testingData, type = "class")
cm1 <- confusionMatrix(as.factor(pred$class), testingData$target)
cm1
# Bagging
fit.bagging <- bagging(target ~ ., data = trainingData, mfinal = 20)
pred <- predict(fit.bagging, testingData, type = "class")
cm2 <- confusionMatrix(as.factor(pred$class), testingData$target)
cm2
```

❖ Logistic Regression:

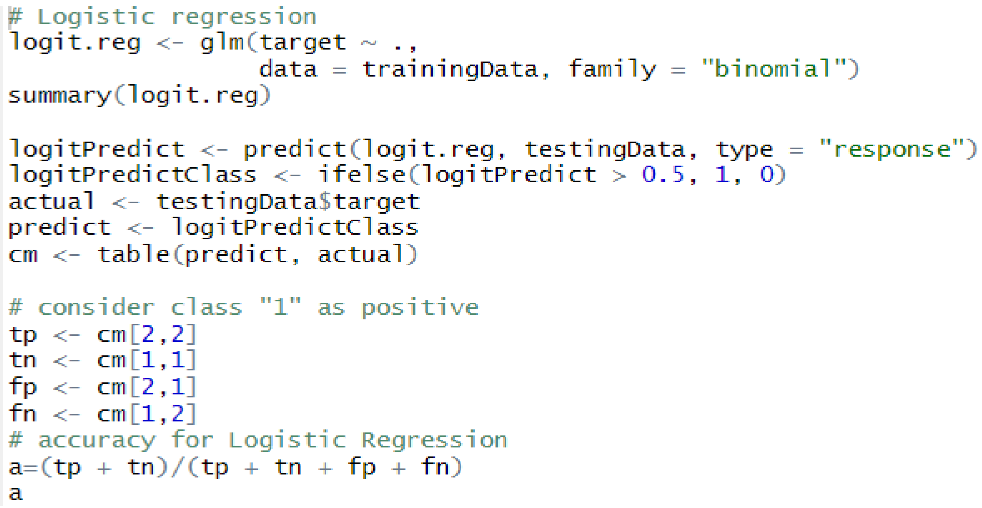
This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Logit(pi) = 1/(1+ exp(-pi))

ln(pi/(1-pi)) = Beta_0 + Beta_1*X_1 + … + B_k*K_k

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than .5 will predict 0 while a probability greater than 0 will predict 1. After the model has been computed, it's best practice to evaluate how well the model predicts the dependent variable, which is called goodness of fit. The Hosmer–Lemeshow test is a popular method to assess model fit.

We are using logistic regression to model the relationship between the dependent variable and the predictor variables. glm() can be used to fit logistic regression models, where the response variable is binary and follows a binomial distribution.



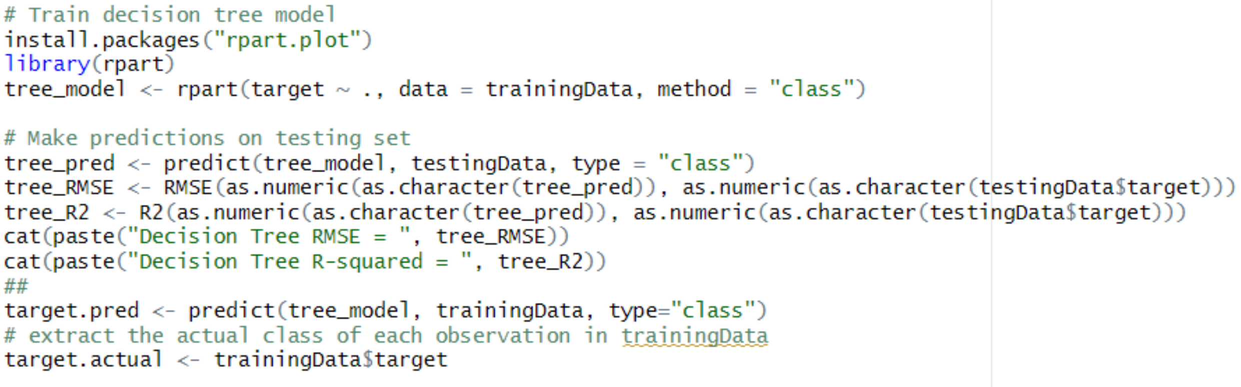
```r
# Logistic regression
logit.reg <- glm(target ~ .,
                 data = trainingData, family = "binomial")
summary(logit.reg)

logitPredict <- predict(logit.reg, testingData, type = "response")
logitPredictClass <- ifelse(logitPredict > 0.5, 1, 0)
actual <- testingData$target
predict <- logitPredictClass
cm <- table(predict, actual)

# consider class "1" as positive
tp <- cm[2,2]
tn <- cm[1,1]
fp <- cm[2,1]
fn <- cm[1,2]
# accuracy for Logistic Regression
a=(tp + tn)/(tp + tn + fp + fn)
a
```

❖ Decision tree:

A decision tree is a popular machine learning algorithm used for both classification and regression tasks. It is a tree-structured model that recursively partitions the data into subsets based on the values of input features, with the goal of predicting the value of a target variable.

At the top of the decision tree, there is a root node that represents the entire dataset. The tree is then split into branches based on the values of a particular feature that best separates the data into two or more subsets. Each branch then leads to a new node, which represents a subset of the data. This process of splitting continues until the subsets are pure (i.e., all instances in a subset belong to the same class in classification or have similar values in regression), or a stopping criterion is met.

Decision trees can be created using the rpart package. The package provides functions for building decision trees, pruning trees to prevent overfitting, and predicting outcomes for new data.



```r
# Train decision tree model
install.packages("rpart.plot")
library(rpart)
tree_model <- rpart(target ~ ., data = trainingData, method = "class")

# Make predictions on testing set
tree_pred <- predict(tree_model, testingData, type = "class")
tree_RMSE <- RMSE(as.numeric(as.character(tree_pred)), as.numeric(as.character(testingData$target)))
tree_R2 <- R2(as.numeric(as.character(tree_pred)), as.numeric(as.character(testingData$target)))
cat(paste("Decision Tree RMSE = ", tree_RMSE))
cat(paste("Decision Tree R-squared = ", tree_R2))
##
target.pred <- predict(tree_model, trainingData, type="class")
# extract the actual class of each observation in trainingData
target.actual <- trainingData$target
```

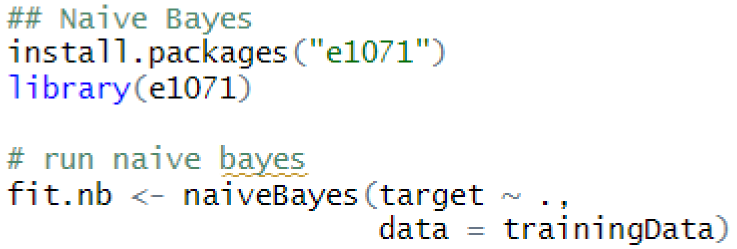
❖ Naive Bayes:

The Naive Bayes algorithm is a probabilistic machine learning algorithm used for classification tasks. It is based on the Bayes' theorem, which describes the probability of an event based on prior knowledge of conditions that might be related to the event.

In the context of Naive Bayes, we use Bayes' theorem to calculate the probability of a certain class label given a set of input features. The "naive" part of the algorithm comes from the assumption that the input features are independent of each other, which simplifies the calculations and makes the algorithm computationally efficient.

To classify a new instance using Naive Bayes, the algorithm first calculates the prior probability of each class based on the frequency of instances in the training data. It then calculates the likelihood of each input feature given each class, which is the conditional probability of that feature occurring in instances of that class. Finally, the algorithm applies Bayes' theorem to combine the prior probabilities and likelihoods to calculate the posterior probability of each class for the new instance. The class with the highest posterior probability is then chosen as the predicted class label.

Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. We use the naiveBayes() function to fit a Naive Bayes classifier to the training data.



```
## Naive Bayes
install.packages("e1071")
library(e1071)

# run naive bayes
fit.nb <- naiveBayes(target ~ .,
                     data = trainingData)
```

❖ Challenges faced:

Some of the common challenges faced while working on this heart disease dataset are:

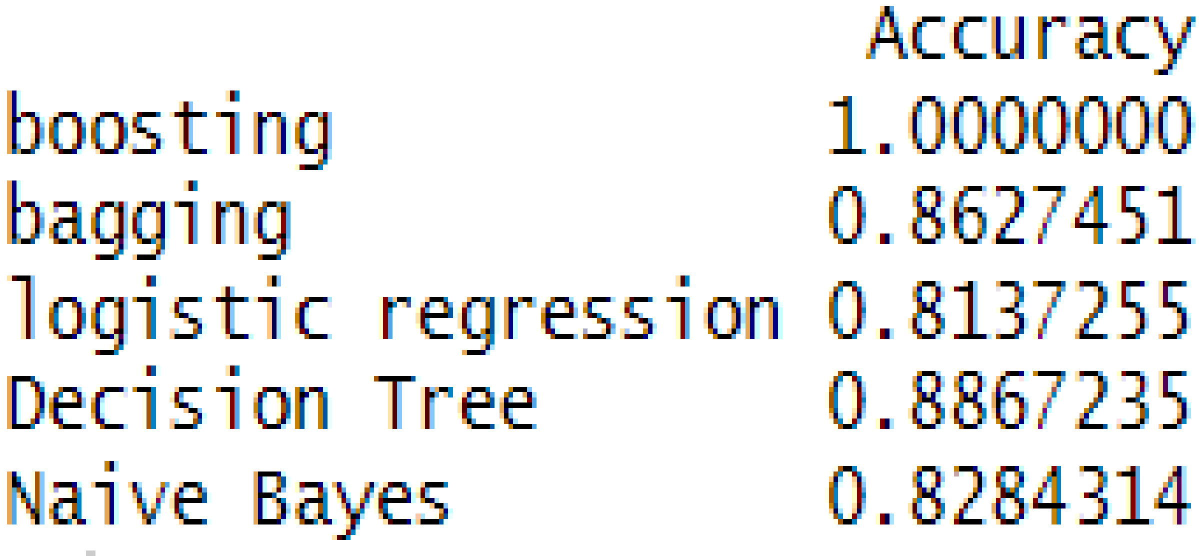
Imbalanced data: The dataset has a higher number of instances with no heart disease (class 0) compared to instances with heart disease (class 1). This can cause a bias towards the majority class during model training and affect the accuracy of the model in predicting the minority class.

Missing data: The dataset contains missing values, which can impact the accuracy of the model if not handled properly.

Algorithm selection: There are various classification algorithms available for building machine learning models. Selecting the right algorithm that best suits the dataset can be challenging and requires experimentation with different algorithms to find the best performing one.

Conclusion

Comparing Accuracies:



```
                     Accuracy
boosting            1.0000000
bagging             0.8627451
logistic regression 0.8137255
Decision Tree       0.8867235
Naive Bayes         0.8284314
```

1. The boosting model achieved the highest accuracy of 100%, which suggests that it is a strong predictor and has learned the underlying patterns and relationships in the data very well. However, it is important to note that achieving 100% accuracy on a test dataset does not necessarily mean that the model is perfect or that it will perform equally well on new, unseen data.

2. The bagging model achieved an accuracy of 86%, which is lower than the boosting model but still quite good. Bagging is a type of ensemble learning technique that combines the predictions of multiple models to make the final prediction. This helps to reduce overfitting and improve the stability of the model.

3. The logistic regression model achieved an accuracy of 81%, which suggests that it is a simple yet powerful model for classification tasks. However, it may not perform well when there are complex nonlinear relationships between the features and the target variable.

4. The decision tree model achieved an accuracy of 88%, indicating that it is a good model for handling both categorical and numerical features. However, decision trees are prone to overfitting and may not perform well when there are many features or when the relationships between the features and the target variable are complex.

5. The Naive Bayes model achieved an accuracy of 82%, which suggests that it is a simple yet effective probabilistic model that works well for text classification and other tasks where the features are independent. However, it may not perform well when there are strong correlations between the features.

6. After evaluating all the algorithms, we have decided to check their FIT for our model using the “ACCURACY” test. As we can see clearly, the accuracy provided by boosting is 100, which is higher than the others. Therefore, this demonstrates that boosting algorithm is more efficient when compared to remaining algorithms and hence we decided to go with boosting for our data set.

In conclusion, the boosting model is the strongest predictor among the 5 models, but the other models have also performed well and may be more appropriate for certain types of data or applications.