

**TABLE OF CONTENTS**

[Background and Business Problem (Kunaal Gautam)](#_tbhpaprvjrrm)

[Exploratory Data Analysis (Zane Solender):](#_wm6u51xhqp0k)

[Variable Definition and Transformation Table:](#_5k0ncbtyuxjy)

[Variable Exploration:](#_eu7tf5tl3ceh)

[NAIVE BAYES Model (Tia Goyal)](#_ega38tiz4hnp)

[LOGISTIC REGRESSION (Sergio Avila)](#_9jw3x9o86xrm)

[CART Model (Akrur Khetan)](#_suf7rki9la70)

[COMPARATIVE DATA ANALYSIS (Kunaal Gautam):](#_3z5jnefrg6hl)

[CONCLUSION (Kunaal Gautam):](#_96j3n8ufkj9s)

# **Background and Business Problem (Kunaal Gautam)**

According to the Center for Disease Control and Prevention, heart disease refers to several types of heart conditions. Heart attacks - also known as myocardial infarctions - are primarily caused by heart disease; heart disease is the leading cause of death in the United States as almost 655,000 Americans die from it each year (accounting for 25% of deaths). In this detailed report, our group will analyze the Heart Disease UCI dataset found on Kaggle and build multiple models to predict the likelihood a patient will have heart disease.

First and foremost, it is imperative to know how the heart works to understand this dataset better. According to WebMD, the heart is a four-chambered, fist-sized organ responsible for nurturing every heartbeat that keeps us alive. It beats almost 100,000 times per day, pumping over 2,000 gallons a day. As the heart beats, it spreads blood through the circulatory system, a system of vessels that transport blood to every part of the body. Blood is essential for our body, as it carries fresh oxygen and removes waste products from our body tissue. This process of the heart is necessary to guarantee a healthy life. When the coronary arteries - the vessels that directly connect to the heart - get blocked due to buildup of fat, cholesterol or other substances, then the heart begins to receive less and less blood and therefore oxygen (this is known as Coronary Artery Disease - one of the major types of heart diseases that exist). When the blood constraints impede the blood flow into the heart, severe damage can occur to the heart muscle, which can cause a heart attack. Some symptoms of heart disease include heart attacks, arrhythmia, and heart failure. Heart disease is the leading cause of death in the United States and one of the most expensive ailments to treat (as it costs $219 billion each year, including the cost of healthcare services, medicines, and lost productivity due to death). Thus all people must familiarize themselves with the data to prevent the risk of having heart disease in the future. Because there are many factors that can cause heart conditions, the Heart Disease UCI dataset is an excellent tool for recognizing heart disease's leading factors. Therefore, our machine learning models can serve as tools to identify risk factors and assist doctors in diagnosing the presence of heart disease.

The Heart Disease UCI dataset provides records of 304 patients from Cleveland, Ohio. These patients were monitored and tested on various subjects relating to their health upon admission to the hospital, as seen in the dataset variables. There is no definitive reason why these patients underwent these tests; however, it is probable they were experiencing some symptoms that merited an analysis of the heart (ex. chest pain or shortness of breath). There are thirteen predictor variables and one response variable in the dataset. Due to a combination of numerical and categorical variables in our data, our group decided to create three models that could analyze our dataset efficiently: Naive Bayes, Logistic, and CART. To ensure our models were accurate, we eliminated any observations that contradicted the nature of our predictor variables (such as ca = 4 as that is an inconclusive value which will be explained further on). Furthermore, we converted nine out of our fourteen total variables using as.factor as these variables were categorical. Our overall goal is to know which of these predictor variables affects the response variable, which is the target variable; a value of 0 corresponds to not having heart disease, whereas a value of 1 corresponds to an individual having heart disease.

# 

# **Exploratory Data Analysis** (Zane Solender):

## Variable Definition and Transformation Table:

| **Variable** | **Description** | **Transformations** |
| --- | --- | --- |
| target | Heart Disease  0 = Heart Disease  1 = No Heart Disease  Diagnosis of heart disease (angiographic disease status) was defined and determined as follows based on UCI data. This refers to the diameter narrowing of arteries including the coronary artery.  -- Value 0: < 50% diameter narrowing  -- Value 1: > 50% diameter narrowing | as.factor |
| Age | Age of the patient in years (Numerical) |  |
| Sex | Sex of the patient (0 = Female; 1 = Male) | as.factor |
| Cp | Chest Pain Type (Angina)  *(0= Typical Angina (When heart is working harder than usual, usually due to constricted blood flow from heart disease, often occurs during exercise);*  *1= Atypical Angina (Much more dangerous as it may occur at rest and likely indicator of heart attack);*  *2= Non- Anginal Pain;*  *3= Asymptomatic)* | as.factor |
| tresbps | Resting SYSTOLIC blood pressure (in mm Hg on admission to hospital)   * *This measurement is taken before the stress test*   *For reference:*  *Normal: <120*  *Elevated: 120-129*  *Hypertension stage 1: 130-139*  *Hypertension Stage 2: ≥140*  *Hypertension Crisis: ≥180* |  |
| chol | Serum cholesterol level measured in mg/dl.   * *Measures total cholesterol including HDL (High Density) and LDL (Low Density) cholesterol.* * *For reference:*   + *Healthy = Less than 200*   + *Elevated = 200-239*   + *High = 240 or higher* |  |
| fbs | Fasting Blood Sugar ( when fbs> 120 mg/dl 1 = true; 0 = false)  *Fasting blood sugar is considered healthy when under 100mg/dl while 100-125mg/dl is considered pre-diabetic.*  *UCI dataset categorizes fasting blood sugar into greater than 120 (FBS=1) or less (FBS=0 )as a way to categorize risk from blood sugar.* | as.factor |
| restecg | Resting electrocardiogram results   * *What is the condition of the heart before a stress test?*   *0= probable left ventricular hypertrophy (Likely sign of heart disease as left ventricle grows to compensate for lack of O2 from artery narrowing)*  *1= normal*  *2= abnormalities in the T wave or ST segment (Heart abnormality not related to left ventricle hypertrophy)* | as.factor |
| thalach | Maximum heart rate achieved during stress test  *Maximum heart rate decreases with age and is usually defined as (220-Age). In the context of hospitalized patients other factors aside from age may impact this figure.* |  |
| exang | Exercise induced angina (1 = Yes; 0 = No)   * *Did the patient have chest pain during the stress test?* | as.factor |
| oldpeak | ST depression induced by exercise relative to rest   * *How much performance declined during the stress test?* |  |
| slope | The slope of the ST segment furong peak or most intense part of exercise. *Results from this test are nuances but slope is usually not as objective a predictor due to the nature of the stress test.*  0 = descending  1 = flat  2 = ascending | as.factor |
| ca | Number of major blood vessels colored by radioactive dye in heart fluoroscopy.  *The Value refers to the number of narrow blood vessels, which means the higher the number the more likely an individual is suffering from heart disease as defined by UCI.*  0-3 (colored by fluoroscopy)  4= inconclusive (Dropped from Datset) | as.factor  ca=4=NULL |
| thal | *Is thalassemia [blood disorder] found in observation?*  ( 0 = NULL  1 = fixed defect (no blood flow in some part of the heart)  2 = normal blood flow  3 = reversible defect (a blood flow is observed but it is not normal) | as.factor  thal=0 = NULL |

**Scrubbed Dataset Dimensions**

|  | ***Observations*** | ***Predictors*** |
| --- | --- | --- |
| ***mydata*** | **294** | **13 + Response** |
| ***Training Set*** | **235 (80%)** | **13 + Response** |
| ***Testing Set*** | **59 (20%)** | **13 + Response** |

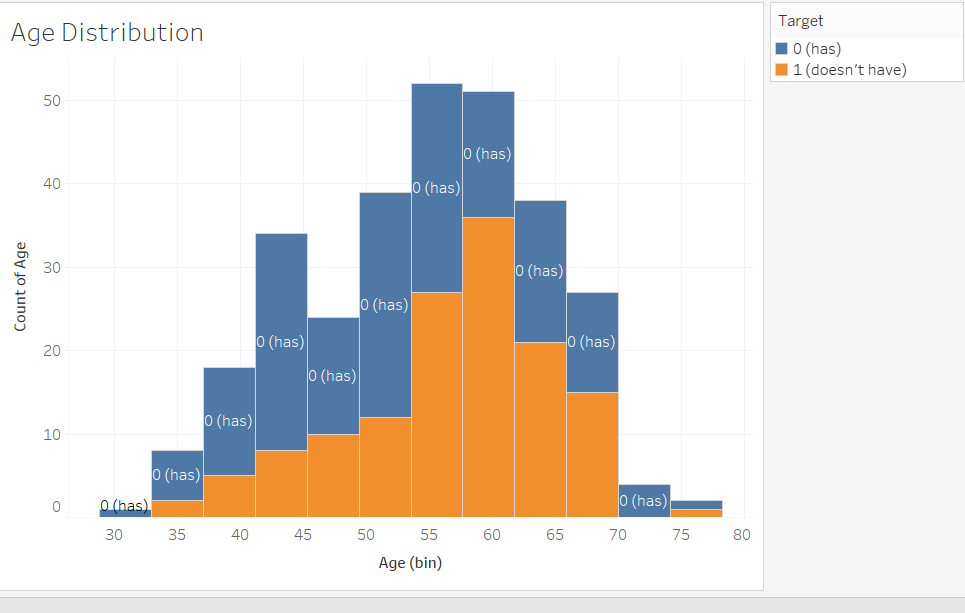
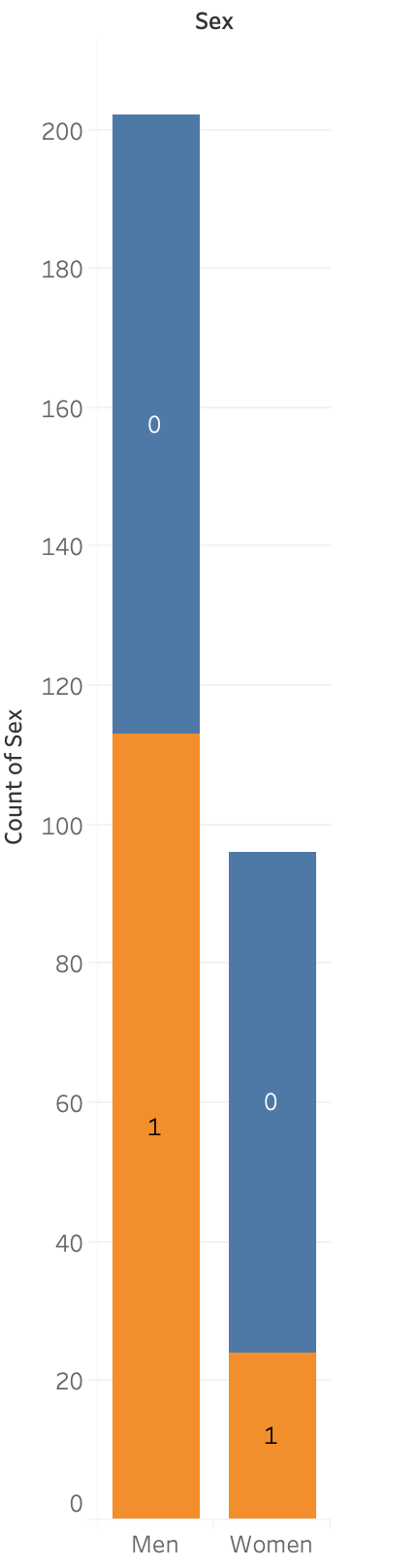
## 

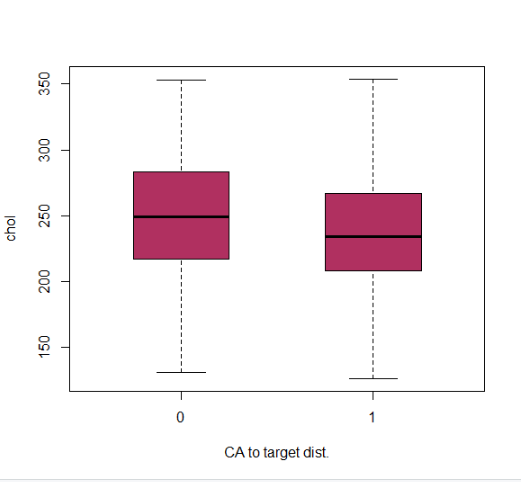
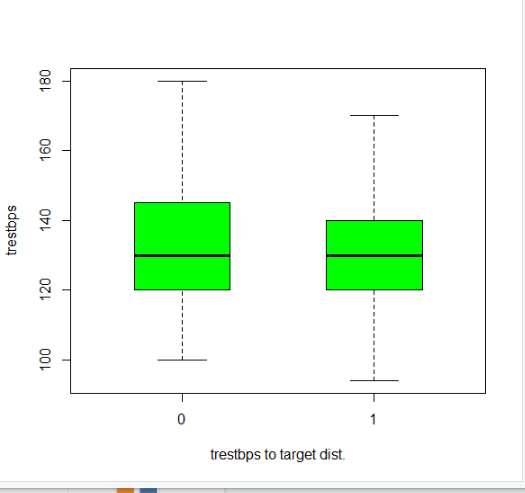
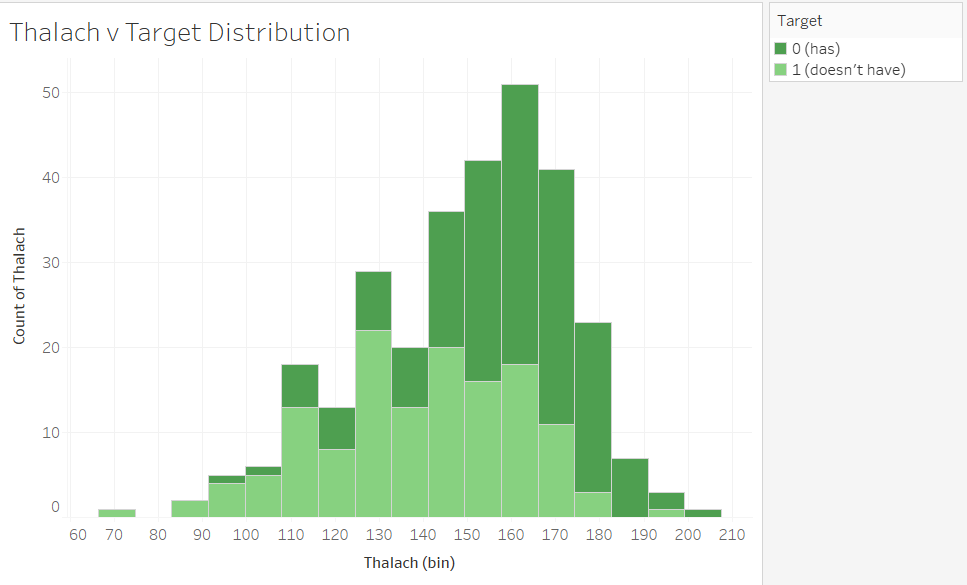
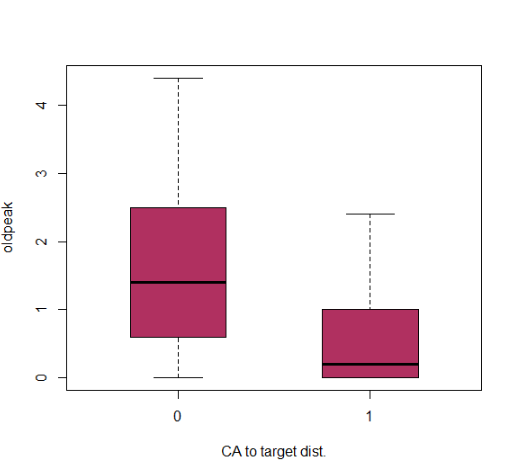
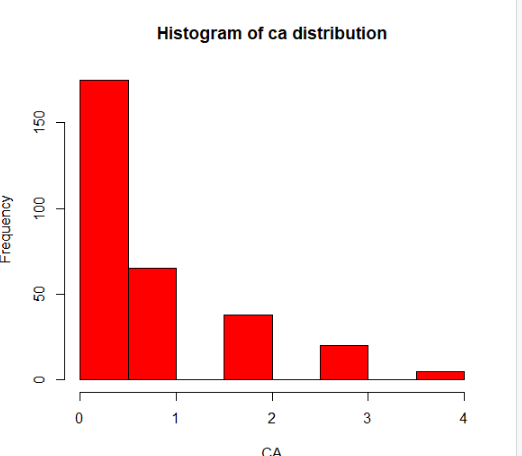
## Variable Exploration:

## 

***Not every variable will have a graph, instead we have chosen the most relevant and impactful variables to further explore graphically.***

1. **TARGET** : This is the main focus of our analysis as it is our response variable, labeling each patient as to whether or not they have a heart disease. The *UCI Heart Disease* dataset denotes two labels for this variable: 0 which indicates no presence of heart disease and 1 which indicates the presence of heart disease. The diagnosis of heart disease (angiographic disease status) was defined and determined as follows based on the UCI data. This refers to the diameter narrowing of arteries including the coronary artery where less than 50% diameter narrowing corresponds to a value of 0, and a greater than 50% diameter narrowing corresponds to a value of 1.
2. **AGE**: Statistically, heart disease is more prevalent in older populations and therefore impacts likelihood of heart disease.

**Graph 1** shows the age distribution while separating the color based on the target. Blue represents the portion of the age group with heart disease while organe represents the absence of heart disease. The mode is between 54-58 years showing the true age demographic of patients tested.

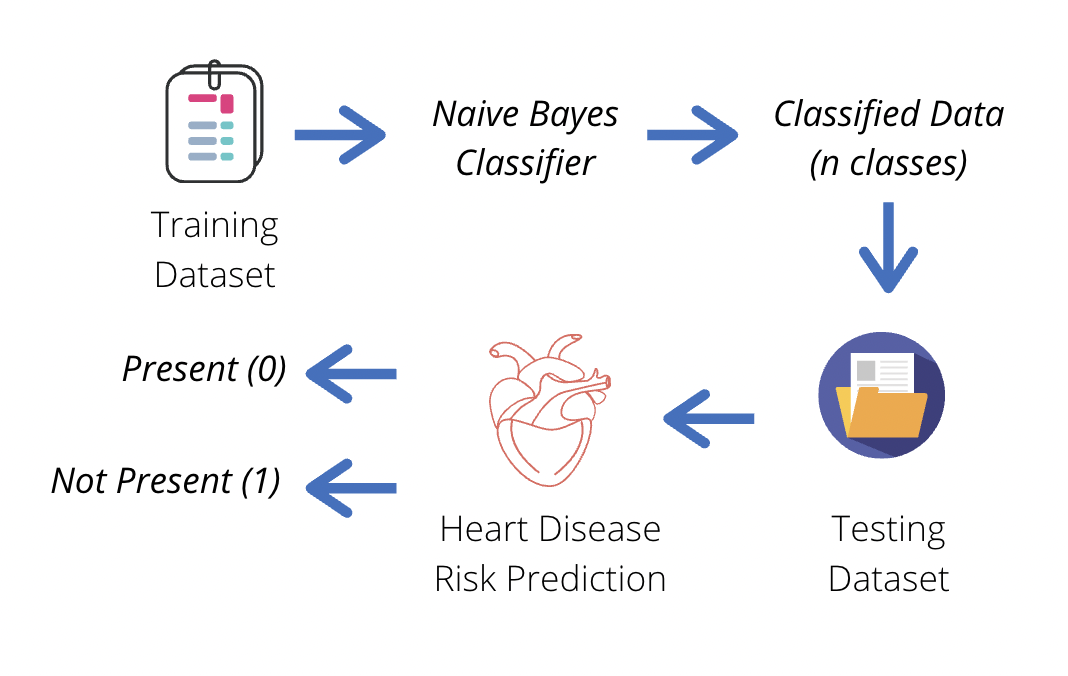
1. **SEX:** Variable was converted into a categorical variable with the as.factor function. First of all, there were more men recorded in the data showing they find themselves in these situations where they are being tested for heart disease. Meanwhile, proportionally women have a higher percentage of having heart disease than the men in the tested group. Meanwhile graph 2 above on the right shows the distribution of sex compared to the target.
2. **CHEST PAIN TYPE (Cp)** was converted into a categorical variable with the as.factor function. The dataset denotes 4 labels for this variable: 0 which represents typical angina, 1 which represents atypical angina, 2 which represents non-anginal pain, 3 which represents asymptomatic chest pain. Angina refers to a type of chest pain caused by lack of blood flow to the heart.
3. **TRES BPS** is the resting SYSTOLIC blood pressure. Graph 3 to the right shows the average TRESTBPS vs target distribution. The average is relatively the same between both patients labeled as a high risk of heart disease and low risk patients. Both means of high and low risk have a mean around 130, however low risk patients have a wider distribution of trestbps.
4. **CHOLESTEROL** is one of the few variables in which each patient has control over. Using Graph 4 to the right, we can see the boxplots distribution of people at high and low risk are relatively the same. This goes to show that cholesterol can be a factor, but it is not the best indicator of heart disease.
5. **FASTING BLOOD SUGAR** (fbs) Measures how fast the blood sugar can change as well as measures the current blood sugar levels. The baseline number used in the study was 120 mg/dl which is a little above the average human blood sugar level (100mg/bl). If the fbs was above 120mg/bl then it was labeled true and vice versa.
6. **RESTING ELECTROCARDIOGRAM RESULTS (restecg**) is the condition of the heart before a stress test. It reflects the results of the ECG at rest. Value = 0 represents a probable left ventricular hypertrophy, which is a symptom of heart disease as the left ventricle increases in size to compensate for lack of blood flow in arteries. Value =1 represents a normal baseline and finally Value =2 represents abnormalities in the T or ST wave, which are signs of issues in the heart.
7. **THALACH is** the maximum heart rate achieved during stress tests. This is a function of both heart condition and age, as the maximum bpm of the heart decreases with age. **Graph 5** shows the distribution of THALACH by whether or not they have a heart disease.
8. **EXANG *(exercised induced angina)*** identifies whether patients had angina during exercise. This variable was converted into a categorical variable with the as.factor function. Patients were classified under two labels: 0 being no and 1 being yes.
9. **OLDPEAK** measures the depression in the slope of the patient’s stress test. Specifically it measures the decrease of the “ST” segment during exercise in comparison to at rest. The “ST'' refers to part of the electrocardiogram and is usually level or at around a baseline in a healthy heart. A decrease can indicate the presence of heart disease. **Graph 6 (right)** shows the distribution of the oldpeak to and shows greater variations in ST segment in those with heart disease.
10. **SLOPE** is the peak exercise ST segment. When looking at a patient’s data and performance, you do not want a very strong descent of the slope. When there is a strong descent, it shows higher signs that the patient is at risk of a heart disease.
11. **CA** refers to the number of red blood vessels measured throughout the process.We had to change this variable into a categorical variable with as.factor. To the right is **Graph 7** which shows the distribution of CA. As you can see the majority of the patients reported 0 or 1 blood vessels.
12. **THAL**, or thalassemia, is an inherited blood disorder characterized by less oxygen-carrying protein (hemoglobin) and fewer red blood cells in the body than normal. We had to change this variable into a categorical variable with as.factor.

# 

# **NAIVE BAYES** Model(Tia Goyal)

Naive Bayes is an excellent method to use in order to classify the outcomes of heart disease given the predictors above. This method can be used because the heart disease data is using the modeling of a **categorical variable**. This method approximates the required probability based on simplifying assumptions made regarding the data. This type of model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful in the field of medical science for diagnosing heart patients.

In this model, all predictor variables except numerical variables: age, trestbps, chol, thalach, and oldpeak were utilized. The 8 predictor variables used were converted with the “as.factor” function in R accordingly, as Naive Bayes only functions with categorical variables. Additionally, the NULL values were removed from the set, and the data was adjusted as indicated in the above variable definition and transformation table in the exploratory data analysis section. As seen in Figure 1, the data was split using the **holdout method**: into a training and testing set at 80% and 20% respectively. The predictive model was built using Professor Davit Khachatryan of Babson College’s NAIVE.MACRO. This classified the data into two categories of either heart disease present (target = 0) or heart disease not present (target = 1).

**Figure 1: Naive Bayes Process**

In this classification algorithm, the model assumes the **conditional independence assumption** where every variable is independent of the other ones, so it is no longer looking at the larger picture, but at the individual predictors. In this case, this is the likelihood of heart disease. The model looks at variables like chest pain type and fasting blood sugar and considers how these features contribute independently to the probability that heart disease is present, regardless of any possible correlations between such predictor variables. This assumption, although strong, is very useful. It allows the predictive model to work well with different forms of data. Utilizing the 20% testing set, the predictive model tested its accuracy and its conclusion resulted in the below confusion matrix and accuracy calculations (seen in Figures 2 and 3).

**Figure 2: Naive Bayes Confusion Matrix Figure 3: Naive Bayes Accuracy Measures**

|  | ***Predicted Class*** | | |  | **Correct Classification Rate** | = (25+30)/59 | 93.22% |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***True Class*** | **0** | **1** | **Row Total** |  | **Sensitivity** | = (30/32) | 93.75% |
| **0** | 25 | 2 | 27 |  | **False Positive Rate** | = (2/27) | 7.41% |
| **1** | 2 | 30 | 32 |  | **Specificity** | = (25/27) | 92.59% |
| **Column Total** | 27 | 32 | 59 |  | **False Negative Rate** | = (2/32) | 6.25% |

The figures above depict the accuracy measures discussed below, using the confusion matrix to calculate these values. The **sensitivity** value (true positive rate) shows the ratio of the heart disease not present patients that were correctly labeled by the classifier heart disease not present. The predictive model is more accurate in identifying if an individual does not have heart disease, it is correct 93.75% of the time. The **specificity** value (true negative rate) shows the ratio of the heart disease present patients that were correctly labeled by the classifier heart disease present. The model correctly identifies if an individual has heart disease 92.59% of the time.

Similarly, the **false positive rate** indicates that among individuals with heart disease present, 7.41% of them were misclassified into categories of not having heart disease. On the other hand, the **false negative rate** indicates that 6.25% of individuals with heart disease not present were classified into categories of having heart disease. These are both relatively low percentages, indicating that Naive Bayes was helpful in classifying into these two classes.

This proposed algorithm worked effectively to classify the data set because Naive Bayes provides accurate results, and with these results heart disease among people is predicted successfully. The model provides a **93.22% accuracy** with minimum time.

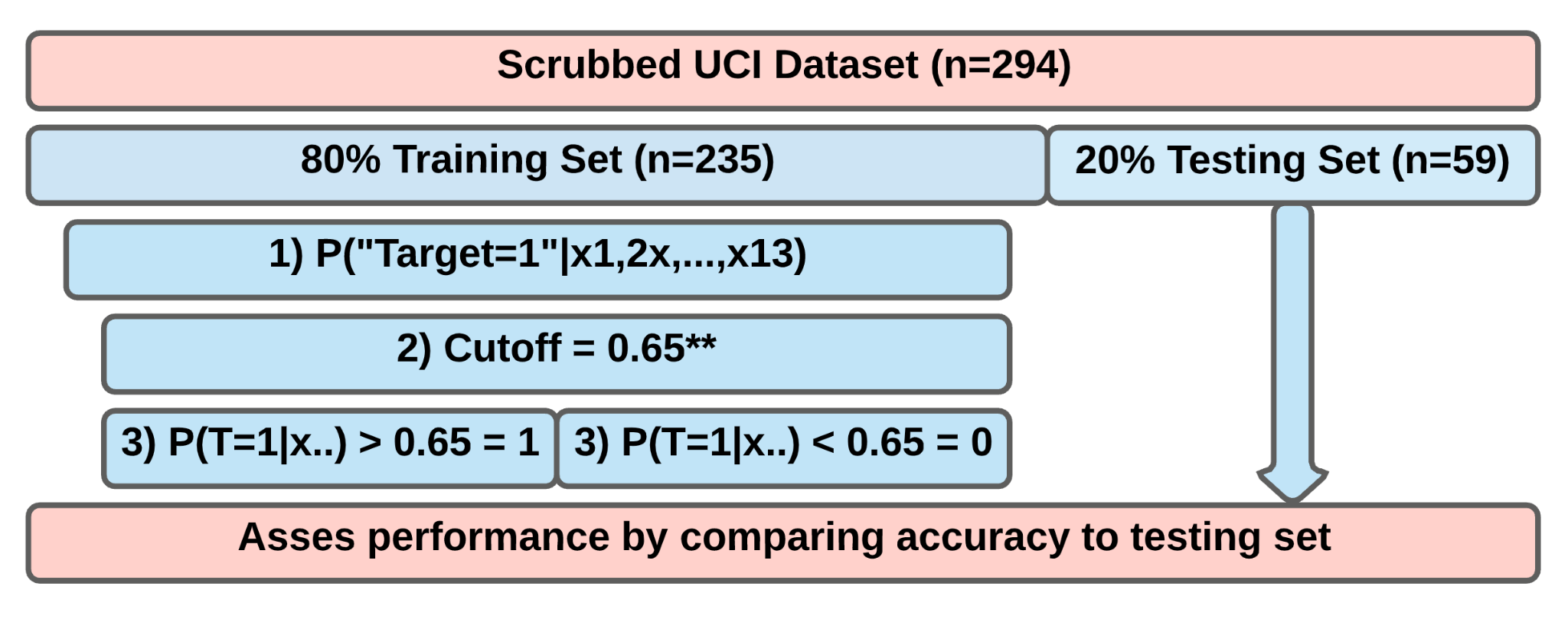
This method is very helpful since it is **computationally efficient** and tends to outperform other sophisticated algorithms like those that will be explained further in this report. However, the conditional independence assumption does serve as a disadvantage since it is not realistic to look at each predictor individually when heart disease encompasses a variety of conditions, and each predictor plays a role on eachother. Additionally, Naive Bayes assumes that each predictor carries the same value and weight over the final target outcome, which isn’t indicative of the actual results. In order to improve the predictive accuracy of this model, it may be necessary to collect more data, to ensure our classification is as accurate as possible.

In summary, the Naive Bayes is an excellent and simple model to use for the purpose of classifying patients into categories of heart disease present and not present despite the downfalls, it provides a relatively realistic conclusion that is helpful with medical diagnosis in heart patients. In other words, this system classifies the given data into different categories and predicts if heart disease is present. It can serve as a training tool for medical students and a helping hand for doctors.

# **LOGISTIC REGRESSION (Sergio Avila)**

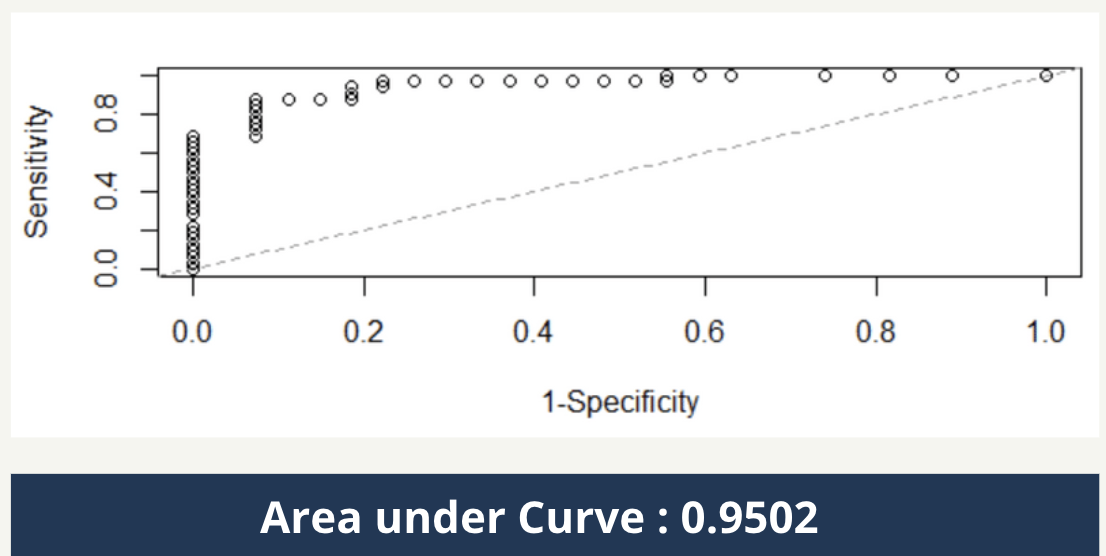
Another method in analyzing risk factors related to heart disease and potentially classifying the existence of heart disease was the Logistic regression method. In this method all 13 predictors were used in calculating the **probability** of a success event (success = 1 = Absence of Heart Disease).

Similar to our descriptive analysis and other models the data was adjusted and scrubbed. Categorical variables were converted with the “as.factor” function in R, while NULL values were removed from the data set. Then, as seen in Figure 4, the **holdout** method was performed by splitting data into an 80% training set and 20% testing. The predictive model, built from the LOGISTIC.MACRO provided by Professor Davit Khachatryan at Babson college, modeled the probability of **the absence of heart disease** and categorized its predictions into 0/1 based on a **0.65 Cutoff** that was set. This means that based on the 13 predictors and data from the training set, the model able to predict a probability of a success event as shown by “P(T=1|x…)” in the figure below. If the predicted probability is 65% or greater than the item is categorized as T=1 or as absence of heart disease. If the probability is less than 65% then the model categorizes the item as T=0, or as having heart disease.

**Figure 4: Logistic Regression Process**

Initially the cutoff probability was set at 0.5 (50%), meaning for the model to classify the testing set on a more “likely than not basis”. However, with the cutoff set at 0.5 the specificity was 81.48% and the sensitivity was the same, placing it at approximately (0.2, 0.9). This coordinate on Figure 5 below did not maximize the area under the curve as there was the possibility for an increase in specificity, which is related to correct diagnosis of heart disease. Therefore the cutoff was increased to 0.65 and no more as 0.70 decreased sensitivity. At this cutoff sensitivity remains at 87.50% while increasing the specificity to 92.59%. The implications of these figures will be explained in the following sections.

**Figure 5: Receiver Operating Characteristic (ROC) Curve**



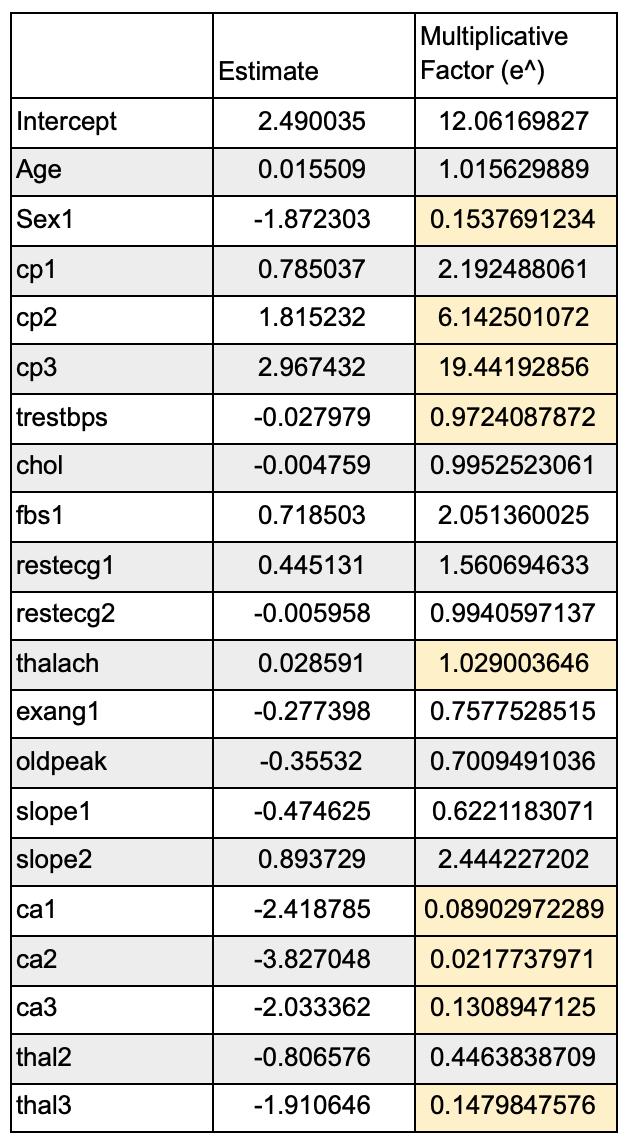
Lastly, the model’s accuracy was tested against the untouched 20% from the Testing Set. This resulted in the following **confusion matrix** and relevant accuracy measure presented in Figure 6.

**Figure 6: Logistic Model Confusion Matrix and Accuracy Measures**

|  | ***Predicted Class*** | | |  | **Correct Classification Rate** | = (25+28)/59 | 89.83% |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***True Class*** | **0** | **1** | **Row Total** |  | **Sensitivity** | = (28/32) | 87.50% |
| **0** | 25 | 2 | 27 |  | **False Positive Rate** | = (2/27) | 7.41% |
| **1** | 4 | 28 | 32 |  | **Specificity** | = (25/27) | 92.59% |
| **Column Total** | 26 | 33 | 59 |  | **False Negative Rate** | = (4/32) | 12.50% |

The figure above depicts the logistic regression model as having an overall correct classification rate of 89.83%. This means that if hospitals or doctors were to use this model they would on average be able to accurately predict the status of their patients heart conditions (diseased or not) with 89.83% accuracy. Specifically the Sensitivity (or true classification rate) was slightly higher than the specificity (true negative rate) at 87.50% and 92.59% respectively. This means the model is better at predicting the presence of heart disease than it is at predicting the absence of heart disease. In other words, out of 100 random patients, the model would correctly diagnose an average of 90 (89.83%) patients as either having or not having heart disease. Out of 100 patients with heart disease (Response = 0) the model would only correctly diagnose ~93 (92.59%) while out of 100 patients without heart disease (Response =1) it would correctly diagnose ~88 (87.5%) of them. While this may appear to be effective, given that human lives are on the line the model is simply not accurate enough to replace doctors for diagnosis. For instance, if the model were to be used instead of doctors, given the False positive rate of 7.41%, out of 100 patients with heart disease ~7 of them would incorrectly be diagnosed as not having heart disease. This would be negligent in terms of healthcare if the patients were sent home based on the model's prediction. On the other hand, out of 100 patients without heart disease the model would incorrectly classify 12.50% of them and ~13 patients would believe they had heart disease and incur subsequent treatment costs. Rather, the model is likely better at assisting in early diagnosis or in confirming a doctor's own diagnosis.

Nevertheless, while this model may not replace doctors, the logistic model has an advantage in that it can provide insight regarding the relationship among predictors and the outcome through odds ratio interpretations of the following parameters. The full macro printout can be found in the appendix while the following figure includes the “estimate” and the multiplicative factor (e^) of each variable. Additionally, the most significant variable interpretation, as measured by a P-Value of less than 0.05, were highlighted in yellow

**Figure 7: Logistic Regression Model Parameters** 

The general interpretation for the parameters can be broken into two categories, numerical and categorical. For instance, the interpretation of thalach, a significant variable showing the maximum heart rate achieved during a stress test, is as follows. For every one bpm increase the odds of not having heart disease increase by a multiplicative factor of 1.029, all else held equal. This same interpretation applied to other numeric variables like age. In the case of Thalach, it suggests that the higher the maximum heart rate the less likely the model is to predict heart disease.

In the case of categorical variables the model references the base value. For instance, Ca which referred to the number of narrowing blood vessels seen through special x-ray imaging when dye is injected into the blood (also known as heart fluoroscopy), can be interpreted as follows. Compared to patients with Ca=0, meaning there was no observed artery narrowing, the odds of a patient with Ca=3 (three observed instances of artery narrowings) is less by a multiplicative factor sof 0.13089 (e^-2.0233362), all else equal. This same referencing principle applies for all categorical variables. In the case of “Ca”, it reflects that generally the more artery narrowing is observed when compared to a healthy baseline the lower the odds of **not having heart disease**,and therefore the higher the odds of being diagnosed with heart disease are.

The nature of this model creates some advantages over the other predictive models we used. Most importantly, the logistic model can provide insight regarding the relationship among predictors and how these affect the outcome through the odds ratios. This was described above and is a feature unique to the logistic model. However, problems may arise when predictors are correlated with each other (multicollinearity), for instance the likely correlation between age and maximum heart rate. The current model assumes that there are no interactions between the predictors. A possible fix would be to create fixed variables and run the model through automatic model selection algorithms such as stepwise to exclude correlated predictors. However, for this analysis we acknowledge this possibility and compared the base logistic model against both the Naive Bayes and CART model to gauge raw predictive accuracy.

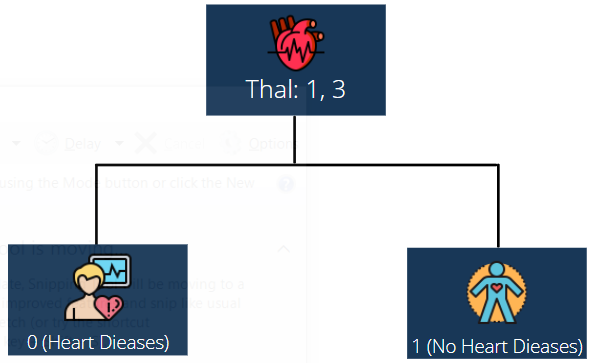
# **CART Model (Akrur Khetan)**

Another model to help predict the likelihood of an individual having or not having heart disease is the classification and regression tree model also known as CART. The CART model works in a way where it takes all the predictors, in this case there are 13, and then splitting the data in a fashion where the split results in the highest drop in **impurity** - in other words, the highest drop in the **deviance**. Each split becomes a node of the **decision tree** which is the main output of this model. R-Studio continues to run the model at a particular node until it is determined that the impurity cannot be significantly reduced further - that node becomes a leaf (**terminal node**).

The CART model was chosen as one of our models as it can handle a categorical response output which is the ***target*** variable; moreover, it has the ability to run a combination of numerical and categorical variables as seen in our predictor variables. The target and certain predictor variables that needed transformations were converted to a categorical variable using the “as.factor” function in R-Studio.

After transforming and scrubbing the dataset, the CART model was built using Professor Davit Khachatryan’s *CART.MACRO*. Before running the model, the minimum deviance was set to be 0, the minimum leaf size set as 30, and the tree was designated as a classification tree as our output was a **binary categorical variable.** The model used the **holdout method** where it ran the CART model on 80% of the data and then the model was tested on the remaining 20% of the data. The first output given by the model was a **pruned tree** which is shown in Figure 8.

**Figure 8: Pruned Tree Output**



|  |  |
| --- | --- |

The CART model’s pruned tree emphasizes which predictor variables have a significant influence on the response variable. The pruned tree considers the categorical variable *THAL*, which refers to thalassemia, which has three designated values as previously explained in our exploratory analysis section. The pruned tree displays the condition of ***THAL: 1,3****,* stipulating that if a person has a fixed defect (*THAL = 1*) or a reversible defect (*THAL = 3*), then the condition will be met; when the condition is met, you move to the left of the tree, resulting in a given target of 0, concluding that the individual likely has heart disease. However, if the condition is not met, which means if an individual has normal blood flow (*THAL = 2*), you will move to the right of the tree, giving you the target variable as 1, which concludes the individual will not likely have heart disease.As the model was tested on 20% of the data, it gave us a confusion matrix as seen in Figure 9 which can be used to assess the model’s performance relative to the other models.

**Figure 9: CART Model Confusion Matrix and Accuracy Measures**

|  | ***Predicted Class*** | | |  | **Correct Classification Rate** | = (20+24)/59 | 74.58% |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***True Class*** | **0** | **1** | **Row Total** |  | **Sensitivity** | = (24/32) | 75.00% |
| **0** | 20 | 7 | 27 |  | **False Positive Rate** | = (7/27) | 25.93% |
| **1** | 8 | 24 | 32 |  | **Specificity** | = (20/27) | 74.07% |
| **Column Total** | 28 | 31 | 59 |  | **False Negative Rate** | = (8/32) | 25.00% |

Figure 9 shows us a correct classification rate of 74.58% when the CART model was tested on the 20% data, 44 patients out of a total 59 patients were correctly diagnosed with the condition they had. This tells us that when this model is used for predictive purposes in real-life situations, there is a possibility that 25.48% of the time the model will incorrectly predict if a patient has heart disease or not. As we move on to the sensitivity and specificity, we can see that they were very low: 75.00% and 74.07% respectively. This means that the model has similar accuracy in predicting the likelihood of an individual having or not having heart disease; the difference between predicting an individual having heart disease is just 0.93% higher than the model predicting that they would not. As the sensitivity and specificity are low, the false positive rate and false negative rates are high, with values of 25.93% and 25.00% respectively.

As seen in the provided data, the CART model is not a desirable machine learning strategy to use independently as 26 out of 100 patients will be misdiagnosed as not having heart disease, when in reality, they do. Additionally, the other measures, including the correct classification rate, illustrate that this model is not a reliable tool that could be independently used to predict the likelihood of having or not having heart disease. To add on, the pruned tree has a size of only 2 as it only considers 1 variable; in reality, we know this not to be accurate as many factors affect an individual’s chances of having heart disease.

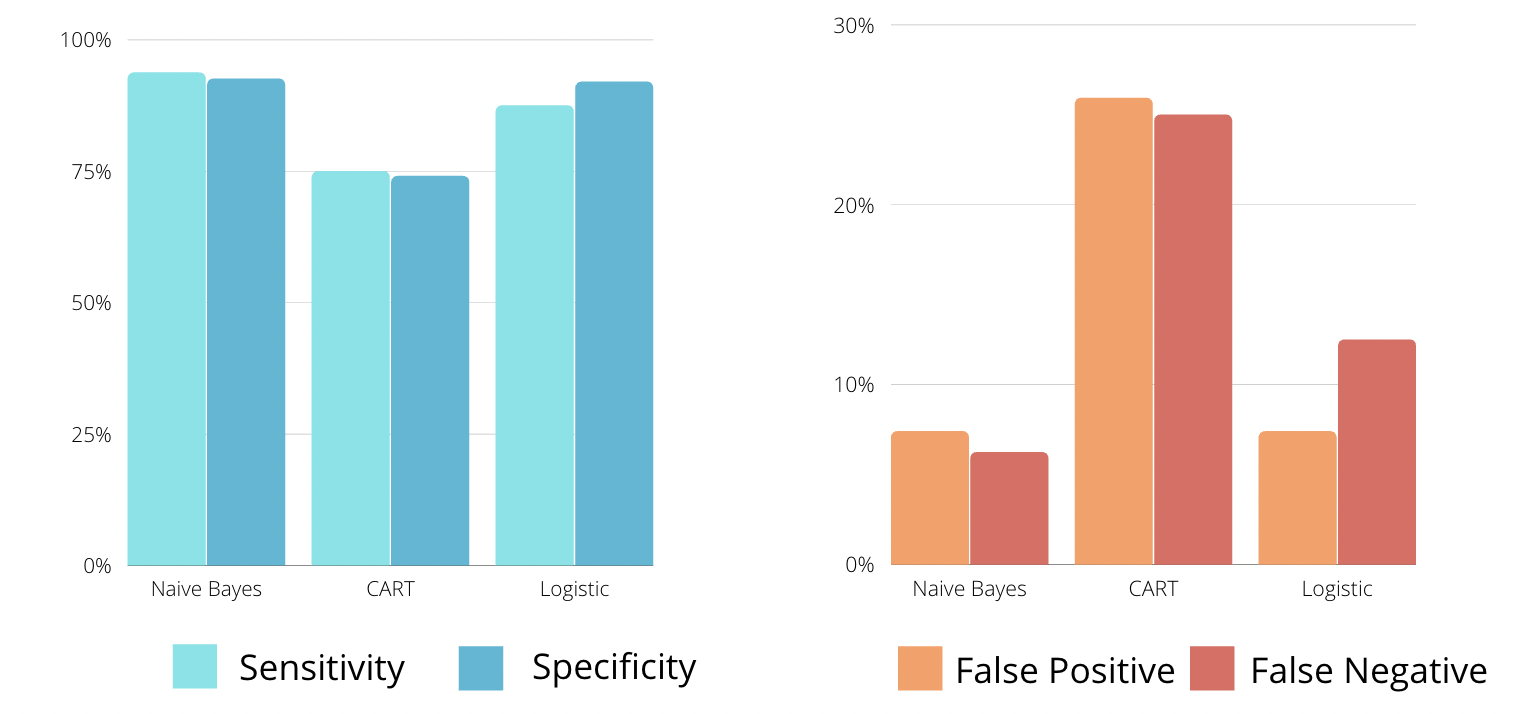
# **COMPARATIVE DATA ANALYSIS (Kunaal Gautam):**

# 

***Figure 10: Comparison of All Models***

|  | **Naive Bayes** | **CART** | **Logistic** |
| --- | --- | --- | --- |
| **Correct Classification Rate** | 93.22% | 74.58% | 89.83% |
| **Sensitivity** | 93.75% | 75.00% | 87.50% |
| **False Positive Rate** | 7.41% | 25.93% | 7.41% |
| **Specificity** | 92.59% | 74.07% | 92.59% |
| **False Negative Rate** | 6.25% | 25.00% | 12.50% |

After analyzing the three models, we can clearly make a few distinctions between the models’ results as elucidated by Figure 10. Foremost, the Naive Bayes model had the highest correct classification rate in which it correctly classified 93.22% of our observations found in the testing set. This was followed by the Logistic Regression and CART models, which had correct classification rates of 89.83% and 74.58%, respectively.

***Figure 11:***

A highly sensitive test signifies few false negative results, and thus fewer cases of diseases are missed; this contrasts with a highly specific test, meaning there are less false positive results, thus designating an individual who does not have a disease as negative. As seen in Figure 11, it is clear that the Naive Bayes model had the highest sensitivity of 93.57%, followed by the Logistic Regression model which had a sensitivity of 87.50%. Both the Naive Bayes model and Logistic Regression models had equal specificities of 92.59%. Lastly, the CART model had the lowest sensitivity and specificity of 75.00% and 74.07%, respectively. Based on the specificity and sensitivity of the models, if a doctor’s focus was to diagnose heart disease, then both Logistic Regression and Naive Bayes would be useful to use. Whereas, if they were trying to find the absence of heart disease, then they would be compelled to use only the Naive Bayes model.

Instantly, it is clear that the CART model had a significantly high false positive rate and false negative rate, both of which were around 25%. The Logistic Regression model had a slightly higher false negative rate and equal false positive rate when compared to the Naive Bayes model. It is crucial to avoid having high false positive rates and false negative rates as there are many consequences when making these diagnoses; infact, it is desirable to have these rates as low as 0%. A false positive rate implies that a patient is diagnosed with heart disease when they actually do not have it. This can cause severe emotional distress for the patients and their families and potential future medical costs for undergoing tests. Moreover, this can affect the hospitals providing care for these patients as seen in a study published by the Royal College of Physicians’ Clinical Medicine Department: patients that were identified through false positive rates “also negatively impacted [hospital] staffing levels to varying degrees, affected transfers in and out of the home, and caused a distraction from other elements of patient care [at hospitals]” (Healey). On the contrary, a false negative rate implies that a patient is not diagnosed with heart disease even though they have it. This is an execrable diagnosis to make as the consequences of it are quite burdensome on all people involved in the treatment and care for patients. According to a cumulative study - containing 420 previous studies that explained the implications of false negative rates - that was published by Health Technology Assessment,

“False-negatives are evident in all screening programmes, even when the quality of the service provided is high. They may have the potential to delay the detection of [various diseases], but there is little evidence to help assess their psychological consequences in these or other screening programmes. False-negatives are likely to lead to legal action being taken by those individuals affected, and potentially may reduce public confidence in screening.” (Petticrew).

# **CONCLUSION (Kunaal Gautam):**

# 

All points considered, our group came to a few conclusions about our models. To begin with, the Naive Bayes model may have shown significantly better results than its counterpart models due to its inherent nature. The Naive Bayes model converted every variable into categorical variables and individually tested each predictor variable on how it affected the response. Although it performed stronger relative to the other models, we think that the results would understate the true reality of how the predictor variables affect the likelihood of having heart disease.

The Logistic Regression model tells us a story as we could see whether each predictor variable affected each other. Additionally, each predictor variable told us how it would change the likelihood of having heart disease, assuming all other predictor variables were held constant. For instance, asymptomatic chest pain increased the likelihood of ***not*** having heart disease by a multiplicative factor of ~19 (Figure 7) when compared to typical angina, all else held equal. On the downside, using logistic regression can cause multicollinearity which can distort the perception of how a predictor variable may behave.

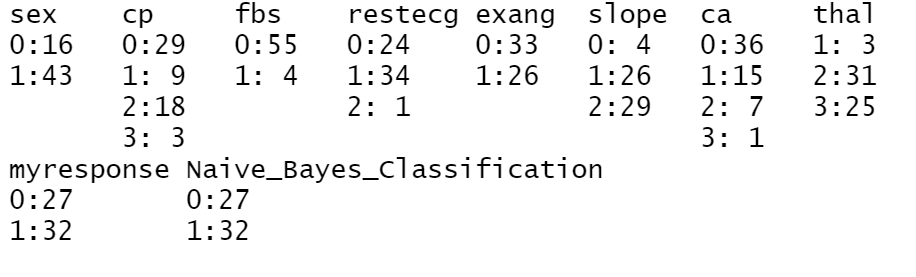
Inter alia, the CART model did not explain much compared to the other models as the pruned-tree contained only one leaf which focused on the thalassemia variable. According to the tree, if the patient had a reversible defect or a fixed defect, then they would have a higher likelihood of having heart disease. The model did not take any of the other predictor variables into account, underemphasizing the importance of the other variables. And as explained previously, CART performed the poorest in terms of correct classification rate, sensitivity, and specificity, and both false positive and negative rates when compared to the other models.

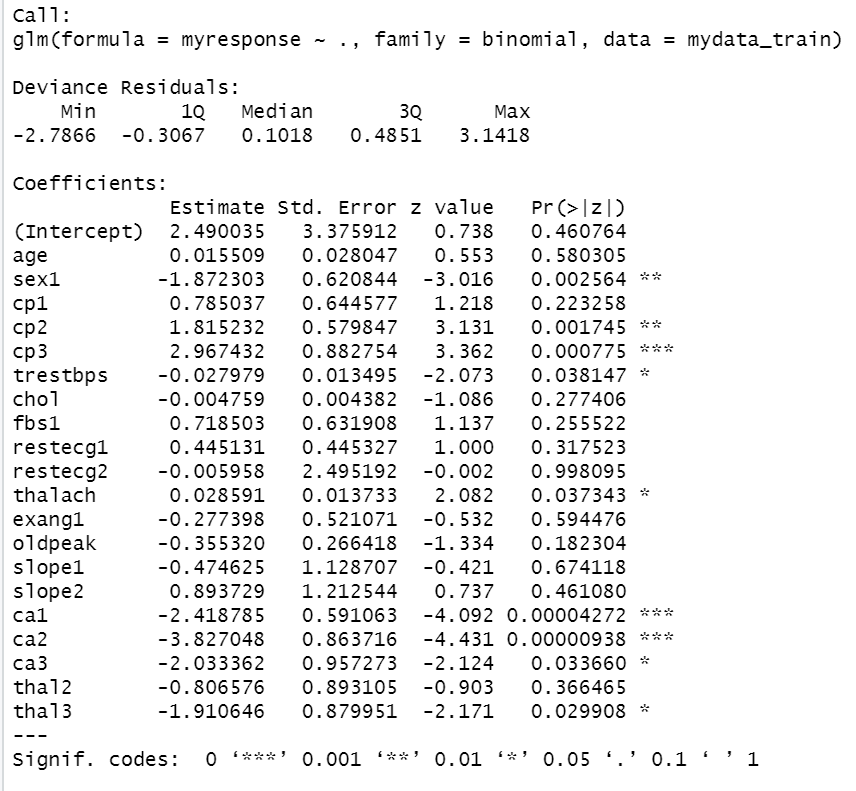
On a final note, our group decided that the Naive Bayes model was the best model to predict the likelihood of having heart disease due to its great performance in classification, sensitivity, and false negative rates. Although the specificity of both the Logistic Regression and Naive Bayes models were equal, it does not change the fact that the Naive Bayes model correctly classified more patients that did not have a disease as having no heart disease when compared to the other models. The Naive Bayes’ efficient and easy building strategy, as seen by its exception of parameter estimates, makes it a desirable model to use. Moreover, based on a cumulative study published by Acta Informatica Medica which included 23 studies with over 50,000 patients, the “[Naive Bayes model] is the simplest [machine learning strategy] and can be useful for predicting diseases and actually in some way can be better than other methods and this model can help health practitioners to make decisions more confidently” (Langarizadeh and Moghbeli).

It is important to state that these machine learning strategies are not designed to replace the need for doctors, but rather a tool that doctors can use to better their understanding of a patient's condition. As seen by the output of our models, every model could not correctly classify every patient with their corresponding condition, thus it would not make sense to use just the Naive Bayes model in the absence of a trained professional for predictive purposes.

Appendix

***I. Naive Bayes Summary Statistics (for\_export data)***



***II. Logistic Model Result Predictors Statistics (for\_export data)***

Works Cited

“Anatomy and Circulation of the Heart.” Edited by James Beckerman, WebMD,

WebMD, 4 Feb. 2021, www.webmd.com/heart-disease/high-cholesterol-healthy-heart.

Bradford, Alina. “What Is Blood Sugar?” LiveScience, Purch, 26 May 2018,

[www.livescience.com/62673-what-is-blood-sugar.html](http://www.livescience.com/62673-what-is-blood-sugar.html).

CDC. Heart Attack Symptoms, Risk, and Recovery. 11 Jan. 2021,

[www.cdc.gov/heartdisease/heart\_attack.html](http://www.cdc.gov/heartdisease/heart_attack.html).

Healy, Brendan et al. “The impact of false positive COVID-19

results in an area of low prevalence.” Clinical medicine (London, England) vol. 21,1 (2021): e54-e56. doi:10.7861/clinmed.2020-0839

Ishtake SH, Sanap SA. Intelligent heart disease prediction system using data mining techniques.

Int J Healthc Biomed Res. 2013;1(3):94–101. [Google Scholar]

Khachatryan, Davit. “Canvas Lecture 2 - Principles of Data Visualization.” Case Studies in

Business Analytics, Jan. 2021, Babson College, Wellesley Hills, MA. PowerPoint

Khachatryan, Davit. “Canvas Lecture 3 - Visual Analytics with Tableau.” Case Studies in

Business Analytics, Jan. 2021, Babson College, Wellesley Hills, MA. PowerPoint

Khachatryan, Davit. “Canvas Lecture 9 - Naive Bayes Method.” Case Studies in Business

Analytics, Jan. 2021, Babson College, Wellesley Hills, MA. PowerPoint

Khachatryan, Davit. “Canvas Lectures 10 and 11 - Logistic Regression Models.” Case Studies in

Business Analytics, Jan. 2021, Babson College, Wellesley Hills, MA. PowerPoint

Khachatryan, Davit. “Canvas Lectures 14 & 15 - CART.” Case Studies in Business Analytics,

Jan. 2021, Babson College, Wellesley Hills, MA. PowerPoint Presentation.

Khachatryan, Davit. “CART MACRO.R”, R Studio, Case Studies In

Business Analytics, 2016-2017

Khachatryan, Davit. “LOGISTIC.MACRO.R”, R Studio, Case Studies In

Business Analytics, 2016-2017

Khachatryan, Davit. “NAIVE MACRO FALL 2020”, R Studio, Case Studies In

Business Analytics, 2017-2018

Langarizadeh, Mostafa, and Fateme Moghbeli. “Applying Naive Bayesian Networks to Disease

Prediction: a Systematic Review.” Acta Informatica Medica : AIM : Journal of the

Society for Medical Informatics of Bosnia & Herzegovina : Casopis Drustva Za

Medicinsku Informatiku BiH, AVICENA, D.o.o., Sarajevo, Oct. 2016,

[www.ncbi.nlm.nih.gov/pmc/articles/PMC5203736/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5203736/).

Petticrew, M P et al. “False-negative results in screening programmes: systematic

review of impact and implications.” Health technology assessment (Winchester, England) vol. 4,5 (2000): 1-120.

Ronit. “Heart Disease UCI.” Kaggle, 25 June 2018, [www.kaggle.com/ronitf/heart-disease-uci](http://www.kaggle.com/ronitf/heart-disease-uci).

Shouman M, Turner T, Stocker R. Using data mining techniques in heart disease diagnosis and

treatment. Proceedings of Japan-Egypt Conference on Electronics; Communications and

Computers (JEC-ECC); 2012 Mar 6-9; Alexandria. Egypt: pp. 173–177. [Google

Scholar]

Software, R. (2017). Machine learning - classification and regression Trees

(CART).