Prediction of Heart Disease Risk among Patients in Federal Medical Centre, Abeokuta using Naïve Bayes

Abstract

Heart disease is a silent killer, which can cause sudden death of individuals without obvious symptoms. The risk of heart disease is a sign in human beings that must not be neglected. Therefore, this study aimed to predict heart disease among patients in the Federal Medical Hospital Centre (FMC), Abeokuta. Descriptive statistics and data visualization techniques were used to gain insights into the distribution and relationships among the variables. Subsequently, a Naive Bayes classifier model was built using 80% of the data for training and 20% for testing. In addition, a Decision Tree Algorithm (DT) model was used to compare the performance of the Naive Bayes model. The performances of the two models were evaluated using accuracy, sensitivity, ROC-AUC, specificity, precision, and the F1-score. The Naive Bayes model achieved an overall accuracy of 83.61%, precision of 89.29%, recall of 78.12%, F1-Score of 83.33%, ROC-AUC of 90%, sensitivity of 78.12%, and specificity of 89.66. On the other hand, they were compared with the Decision Tree (DT) model which achieved an overall accuracy of 75.41%, precision of 77.42%, recall of 75%, F1-Score of 76.19%, ROC-AUC of 84.54%, sensitivity of 75%, and specificity of 75.86%. Similarly, the confusion matrix for both analyses gave the correct classification of 25 and 22 cases of patients who have heart disease while their wrong classification was 7 and 10 cases of patients who have no heart disease respectively. Furthermore, the importance of features carried out on both showed that the most significant features are the maximum heart rate achieved, fasting blood sugar, resting blood pressure, and chest pain respectively for Naive Bayes and Decision Tree. The findings of the analysis showed that the Naive Bayes model outperformed the Decision Tree in every aspect of the analyses in predicting the risk of heart disease based on the data used and, it suggested that medical health insurance should consider incorporating predictive modelling techniques like Naive Bayes into their risk assessment algorithms, which can be of great use in the medical line.

Keywords: Naïve Bayes Classifier, Decision Tree Algorithm, Confusion Matrix, Feature Importance, Maximum heart rate.

1.0 Introduction

Heart disease is the most common disease in both developed and undeveloped countries in the world, which usually leads to mortality, resulting in millions of deaths annually. According to the same source, heart disease alone accounts for approximately 12 million deaths each year globally (WHO, 2021). The condition of heart disease continues to increase the hardship on the

healthcare system. In the United States, the cost of treating heart disease was estimated to increase from $219 billion to over $1 trillion from 2010 to 2030, based on their research (Heidenreich et al, 2011). The earlier we diagnose those factors that can lead to patients developing a risk of heart disease, the more it will reduce the cost and prevent the patient from dying. Many studies have been conducted in an attempt to pinpoint the most influential risk factors for heart disease as well as accurately predict the overall risk. Heart disease has been highlighted as a silent killer, which leads to the death of a person without clear symptoms. The early diagnosis of Heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn, reduces the complications and will ascertain our current state of health as soon as possible.

Many factors increase the risk of heart disease, such as high blood pressure, fasting blood sugar, chest pain, hypertension, cholesterol, and family ancestry of heart disease: thalassemia, smoking, diabetes, etc. According to the National Centre for Chronic Disease Prevention and Health Promotion (NCCDPHP), high blood pressure is a leading cause of heart disease and stroke because it damages the lining of the arteries, making them more susceptible to the buildup of plaque, which narrows the arteries leading to the heart and brain. The same source also discovered that about 116 million adults have high blood pressure, 130/80mm Higher. Also, the same source claims that about 70% of people have a first heart attack, and 80% have strokes and high blood pressure (NCCDPHP, 2024). Diabetes is also one of the significant factors in the risk of heart disease; Adults with diabetes are twice as likely to have heart disease or stroke as people who do not have diabetes. Over time, high blood sugar from diabetes can damage blood vessels in the heart and block blood vessels leading to the brain, causing stroke, and those patients with diabetes have high blood pressure. Moreover, people with obesity, normal weight/ overweight, or obesity are at increased risk of heart disease and stroke. A healthy diet can reduce a person's chance of having heart disease. Also, physical inactivity can lead to heart disease, even for people who have no other risk factors. It can increase the risk of heart disease and many other factors.

Generally in adults, a heart rate of more than 100 beats per minute while resting is considered high (tachycardia). If you exercise regularly, you may have a lower heart rate, which is good for your heart. The American Heart Association recommends that people should exercise their target heart rate zones, which are estimated as a percentage between 50% and 85% of their maximum

(safe) heart rate. If the test goes beyond the maximum heart rate implies not healthy, and this rate depends on your age, removing your age from the number 220 gives you your maximum heart rate.

Naïve Bayes classifiers are supervised by machine learning algorithms for their classification tasks. It is also a family of generative learning algorithms, i.e., it seeks to model the distribution of inputs for a given class or category. Unlike discriminative classifiers, like logistic regression, it does not learn which features are most important to differentiate between classes. There are three types of naïve Bayes models used in machine learning: Bernoulli (0 and 1 features), Multinomial (discrete or complete, categorical features), and Gaussian (continuous values features). Naive Bayes is a simple but surprisingly powerful probabilistic machine learning algorithm used for predictive modelling and classification tasks. Some typical applications of Naive Bayes are spam filtering; academic students' performance (Akanbi O. B., 2023), Syadiah et al, 2017), colon cancer (Nafizatuset et al, 2019), macroeconomic factors and sovereign credit rating categories (Oliver et al, 2020), fault evaluation and defection in 110 kV transformers (Chenmeng et al, 2021), heart and cardiovascular disease classifications and predictions (Harshit Jindal et al, (2021), Sai Krishna et al, (2022), Jagadish P., (2022), and Niloy Biswas, (2023). Also, on Machine learning model types comparisons: Marzuki (2020), Wickramasinghe et al (2021), Maghari and Amra (2017), Songhua Hu (2021), Jiao S. R. (2020). Naive Bayes is a popular algorithm mainly because it can be easily written in code and predictions can be made quickly, which in turn increases the scalability of the solution. Its algorithm is traditionally considered the algorithm of choice for practical-based applications, mostly in cases where instantaneous responses are required for user requests. Bayes' theorem helps us examine the probability of an event based on the prior knowledge of any event that has correspondence to the former event. Its uses are mainly found in probability theory and statistics. The term naive is used in the sense that the features given to the model are not dependent on each other. In simple terms, If you change the value of one feature in the algorithm, it will not directly influence or change the value of the other features. The Naive Bayes classifier reduces the complexity of the Bayesian classifier by assuming conditional dependence over the training dataset (Rumsfeld et al., 2013). The Naïve Bayes classifier algorithm was applied to a Tunisian commercial bank's short-term loan default prediction (Krichene Aida, 2017).

In medical data mining with Naïve Bayes, Kononenko (2001) considered Naive Bayes as a benchmark algorithm that, in any medical domain, has to be tried before any other advanced method. Abraham et al. (2006) argued that simple methods are better in medical data mining, and this makes Naive Bayes perform well for such data. Compared to other classifiers, Naive Bayes is simple, computationally efficient, requires relatively little data for training, does not have a lot of parameters, and is naturally robust to missing and noise data (Al-Aidaroos et al., 2010). One of the main advantages of the Naive Bayes approach, which is appealing to physicians, is that all the available information is used to explain the decision. This explanation seems to be natural for medical diagnosis and prognosis, i.e., is close to the way physicians diagnose patients (Zelic et al., 1997). When dealing with medical data, the Naïve Bayes classifier takes into account evidence from many attributes to make the final prediction and provides transparent explanations of its decisions. Therefore, it is considered one of the most useful classifiers to support physician decisions.

Data mining tools have been created for the compelling investigation of medicinal data to help clinicians improve their conclusions for treatment purposes. In heart disease research, data mining strategies have played a huge role. Heart Disease contains the screening and extraordinary methodology in the investigation of heart-related infection characterization to discover the disguised medicinal data (Weng et al., 2017). Building accurate and efficient classifiers for medical databases is one of the essential tasks of data mining and machine learning research. Data mining helps identify useful trends in a large set of data. As a result of the increase in the amount of health data gathered through electronic health record (EHR) systems, it is believed that strong analysis tools are important. With a huge amount of data, healthcare providers are now optimizing the efficiency of their organizations using data mining. Data mining has proven effective in areas such as predictive medicine, customer relationship management, detection of fraud and abuse, healthcare management, and measuring the effectiveness of certain treatments.

Machine learning techniques like Naive Bayes classification have shown promise for developing predictive models using patient clinical data. Naive Bayes is a probabilistic model that applies Bayes' theorem to determine the likelihood of an outcome based on predictor variables (Lewis, 1998). Studies have implemented Naive Bayes classifiers to predict heart disease using risk factors like smoking, diabetes, hypertension, and cholesterol with a high accuracy of 85–90%

(Karmen et al., 2019). (Weng et al., 2017). However, most existing models rely on standard clinical datasets like the Cleveland database, which has a limited number of predictors. Expanding the breadth of patient data could improve the generalization of heart disease prediction to broader populations. The importance of Medical Data Mining is to assist the physician in making the final decision without hesitation, minimizing diagnostic errors (especially from inexperienced physicians), improving diagnostic speed, and increasing the quality of medical treatment (Maria, 2002; Bai and Srivatsa, 2006; Lin, 2009; Temurtas et al., 2009).

This study aims to develop an enhanced Naive Bayes classification model for heart disease risk prediction using a dataset with an expanded set of clinical, lifestyle, and socioeconomic predictors. This study can significantly benefit individuals in several ways by accurately assessing the risk of heart disease, individuals can receive early warnings, enabling timely intervention and lifestyle adjustments to prevent or manage the condition, and recommendations based on individual risk factors can be provided, allowing users to adopt personalized prevention strategies such as dietary changes, exercise routines, and stress management to reduce their risk. The model can serve as a continuous health monitoring tool, keeping individuals informed about their evolving risk levels and prompting them to take proactive measures for long-term cardiovascular health. Early intervention and prevention (Benjamin et al., 2019) can lead to lower healthcare costs by minimizing the need for extensive medical treatments and hospitalizations associated with advanced stages of heart disease.

2.0 Methodology

2.1 Some Concepts/Tools in Machine learning

Accuracy Estimation

$$\text{Accuracy} = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} = \frac{TP + TN}{FP + FN + TP + TN} \quad (1)$$

Where;

TP = true positive; TN = true negative; FP = false positive; FN = false negative

Precision

$$\text{Precision} = \frac{TP}{FP + TP} \quad (2)$$

Recall

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1 Score

$$\text{F1 score} = \frac{2 * precision * Recall}{\text{precision} + \text{Recall}} \quad (4)$$

Receiver Operating Character (ROC)

The ROC curve itself is not derived from a single formula but rather constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings of a classification model. However, the TPR and FPR can be calculated using the following formulas:

Where,

$$\text{TPR} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{FPR} = \frac{FP}{FP + TN} \quad (6)$$

2.2 The Naive Bayes Model (Normal assumption)

Let $X = (x_1, x_2, \ldots, x_n)$ be a attribute representing the features of an instance and θ_i, i = 1,2,…,n be a class label.

$$P(\theta_i|X) = \frac{P(X|\theta_i) * P(\theta_i)}{P(X)} \quad (7)$$

Where

$$\text{P(X}|\theta_i) = \prod_{j=1}^{n} \text{P}(x_j|\theta_i) \quad (8)$$

Also,

$$\text{f(x}|\theta) = (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(x-\theta)^2} \quad (9)$$

Thus,

$$\prod \text{f(x}|\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\Sigma(x-\theta)^2} \quad (10)$$

The prior is given by

$$\text{g}(\theta) = (2\pi{\sigma_0}^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(\theta-\theta_0)^2} \quad (11)$$

$$\text{P}(\theta|\text{x}) = (2\pi{\sigma_1}^2)^{-\frac{1}{2}} e^{-\frac{1}{2{\sigma_1}^2}(\theta-\theta_1)^2} \quad (12)$$

Where,

$$\theta_1 = \left(\frac{n}{\sigma^2} + \frac{1}{{\sigma_0}^2}\right)^{-1} \left(\frac{n\bar{x}}{\sigma^2} + \frac{\theta_0}{{\sigma_0}^2}\right) \quad (13)$$

$$\sigma_1{}^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0{}^2}\right)^{-1} \quad (14)$$

$P(\theta_i|X) = P(\theta_i| x_1, x_2,\ldots,x_n)$ for i=1,2,…,n which are the posterior probability of class θ_i given the feature X.

$$P(\theta_i) = \frac{n * \theta_i}{N} \quad (15)$$

Where n is the number of instance belonging to class θ_i in the training data

N is the total number of instances in the training data. Thus,

The Naive Bayes classifier model $Y = \text{argmax}\ \theta_i\ (P(\theta_i)(\prod_j^n(P(x_j|\theta_i))$ (16)

Where Y is the predicted class label for the instance X and n is the number of features.

$$P\ (\text{Heart disease}|\text{Features}\ ) = \frac{\prod P(\text{Feature} \mid \text{Heart Disease})\ P\ (\text{Heart Disease})}{P\ (\text{Features})} \quad (17)$$

2.3 Decision Tree

In Decision tree, the input data is typically represented as a set of feature vectors. Each feature associated with class label. The input feature space is transformed or converted into a high – dimensional space via a selection of a kernel function. Looking for a decision boundary that can either capture non-linearly is made easier with the use of this transformation. Typical kernel functions consist of linear, polynomial, sigmoid and radius basis function. Decision tree employs the kernel method when the data is not linearly separable in the original feature space. Decision tree is able to locate a non-linear decision boundary in the original feature space. Entropy under this decision tree which measures the degree of uncertainty associated with random variable values is calculated as:

Entropy: $G(X) = -\sum_0^n P(x) log_2 * P(x)$ (18)

If (x) be present of heart disease, and (y) be absence of heart disease

Then, $G(X) = -p(x) log_2(x) + p(y) log_2 p(y)$ (19)

3.0 Data Analysis and Results

3.1: Naive Bayes

In this section, the data were split into two; 80 percent of the data being used for training and the remaining 20 percent for testing. The datasets are explained as follow: 303 patients at FMC, Abeokuta were examined in May, 2024, each of which was diagnosed with the following variables: age, sex, chest pain (cp), resting blood pressure (Trestbps), seum cholesterol (Chol), fasting blood sugar (fbs), resting electrocardiograpghic (Restecg), maximum heart rate archieved (Thalach), exercise induced angina (Exang), st depression induced by exercise (Old peak), the

slope of peak exercise ST segment (Slope), number of major vessels (Ca), thalassemia (Thal), diagnosis of heart disease (Num).

Table 1: Performance Metrics for Naïve Bayes

Metrics	Performance
Accuracy	83.61%
Precision	89.29%
Recall	78.12%
F1 Score	83.33%
ROC-AUC	90.00%
Sensitivity (True Positive Rate)	78.12%
Specificity (True Negative Rate)	89.66%

Table 1 showed that the Naive Bayes model obtained an accuracy score of 83.61%, suggesting it performed reasonably well in correctly classifying patients with and without heart disease. A precision score of 89.29% was also obtained, indicating that for every 100 patients with heart disease, the model was able to correctly predict 89 of the patients with heart disease risk. Also, a recall score of 78.12% was obtained, capturing a substantial proportion of the actual positive cases. Also, the F1 Score of 83.33% indicated that the Naive Bayes model was doing well, predicting positive cases and minimizing false positives. Now, looking at sensitivity (True Positive Rate) at 78.12%, we see that our model is capturing a good portion of the actual positive cases, which is positive. Specificity (True Negative Rate) was very high at 89.66%, indicating that the model correctly identified individuals without heart disease risk. The result obtained below showed that the model gave a ROC - AUC value of 90%, indicating that the model performed highly well and was able to distinguish between present and absent heart diseases.

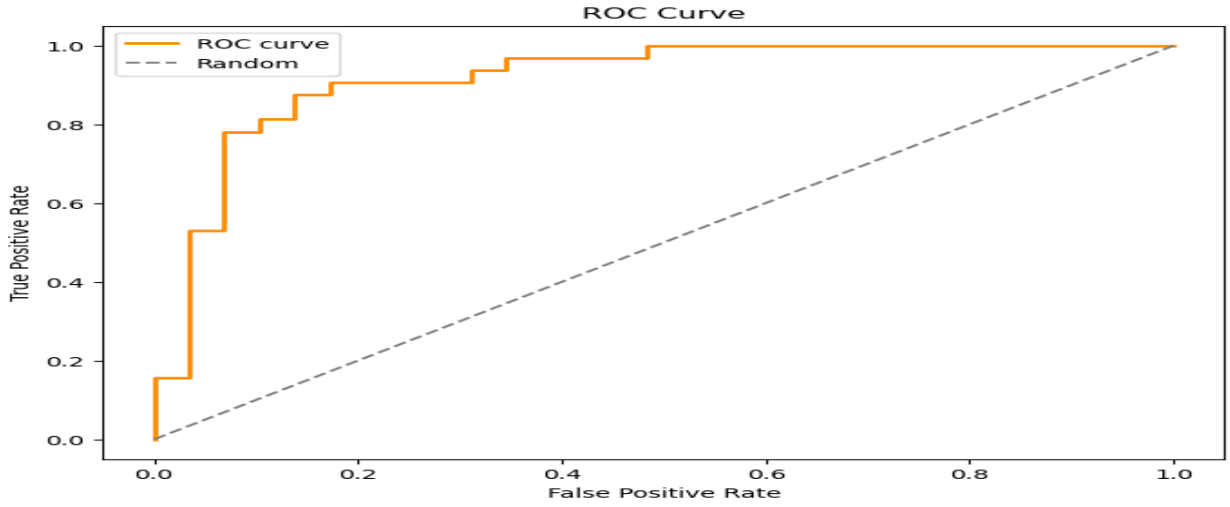

Figure 1: ROC Curve for the Performances of Naïve Bayes

3.1.1 Performances of Naïve Bayes for the Testing Observation

Table 2 and Figure 1 show the classification report made by the testing model using 20% of the data which equals 61 observations, In the presented classification report, the model demonstrates a harmonious balance between precision and recall for both the Present and Absent classes, The weighted averages standing at 84%, which signifies a robust aggregation of precision, recall, and F1-score metrics.

Table 2: Classification Report of Naïve Bayes model Performance

	Precision	Recall	f1-score	Support
Absent	79%	90%	84%	29
Present	89%	78%	83%	32
Accuracy			84%	61
weighted avg	84%	84%	84%	61

3.1.2 Confusion Matrix

Figure 2 shows the result for 20% of the testing observations indicating the Naive Bayes model correctly identified and predicted 26 cases of patients who have no Heart Disease (TP), and incorrectly predicted 3 cases of patients who have the risk of heart disease when they do not have (FP), There were 7 cases where the model predicted cases of patients who have heart disease, but they do not actually have it (FN), and it correctly identified and predicted 25 cases of patients who have the heart disease.

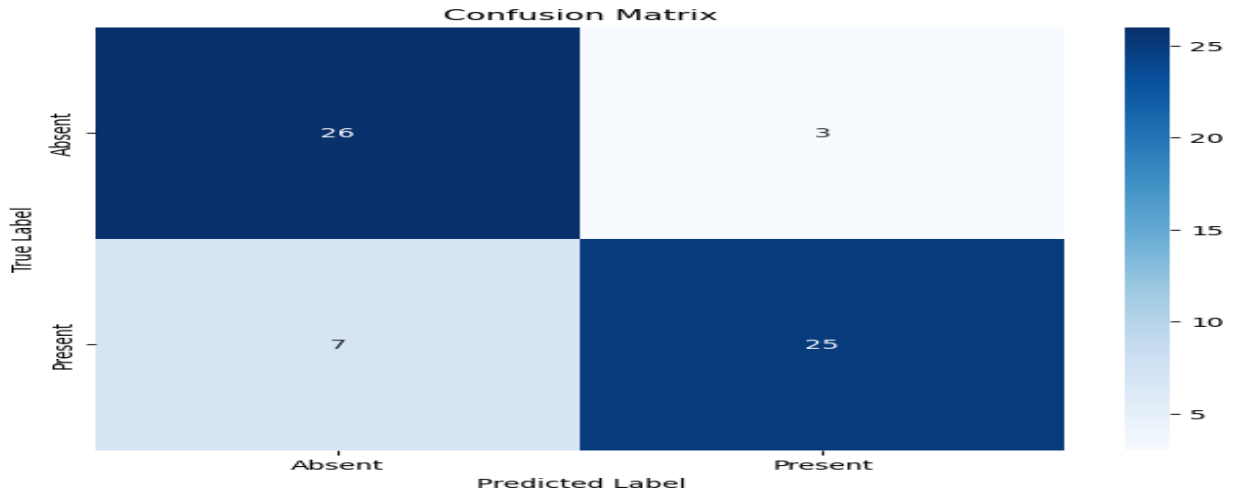

Figure 2: Confusion Matrix for the Performance of Naïve Bayes

3.1.3 Features Importance

Figure 3 gives insight into how each variable contributes to predicting if a patient has heart disease or not. The 'thal' variable has a positive importance which indicates a high probability of a patient having no heart disease. Also, the number of major vessels (ca), old peak (ST depression induced by exercise), exercise-induced angina (exang), chest pain (cp), slope, sex,

age, restecg, trestbps, chol, fbs, thalach which have negative importance indicated that it contributed to the prediction of no heart disease from the analysis.

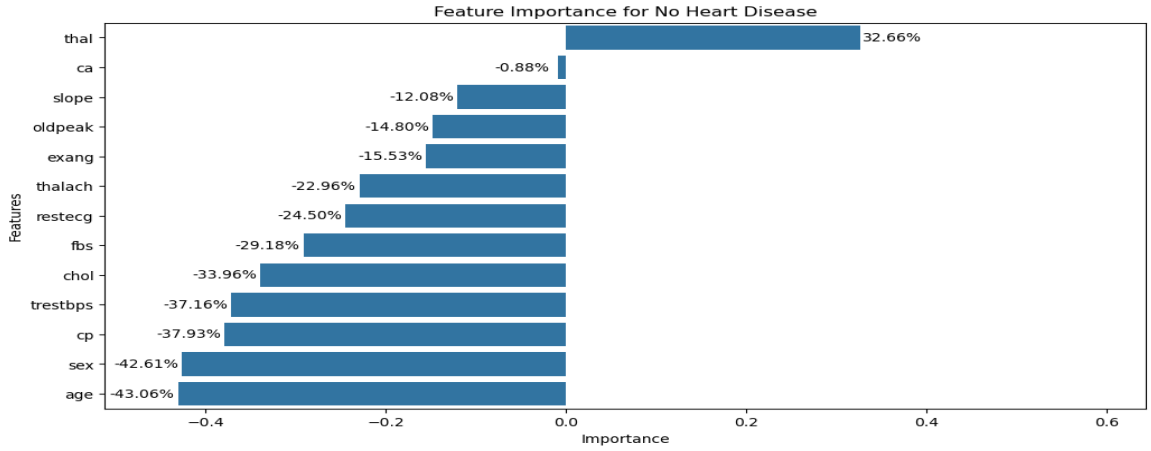

Figure 3: The Feature Importance for no Heart Disease

Figure 4 shows how each variable contributes to patients who are likely to have heart disease, it is evident that 'thal', 'ca', oldpeak', and 'exang' have positive importance for predicting the presence of heart disease. It was observed that even for the presence of heart disease, a higher thal value contributes positively to the prediction. This suggested that certain levels of thal are responsible for the presence of heart disease. The number of major vessels colored by fluoroscopy (ca) continues to show positive importance, Similarly, 'oldpeak' (ST depression induced by exercise relative to rest), exercise-induced angina (exang), chest pain (cp), slope, sex, restecg, trestbps, chol, fbs, and thalach continue to contribute positively to predicting the presence of heart disease. Additionally, features like age which have a negative importance have nothing to do with the presence of heart disease because heart disease does not have any effect on whether the patients are old or young, so, it has no importance in deciding whether absent or present of heart disease risk.

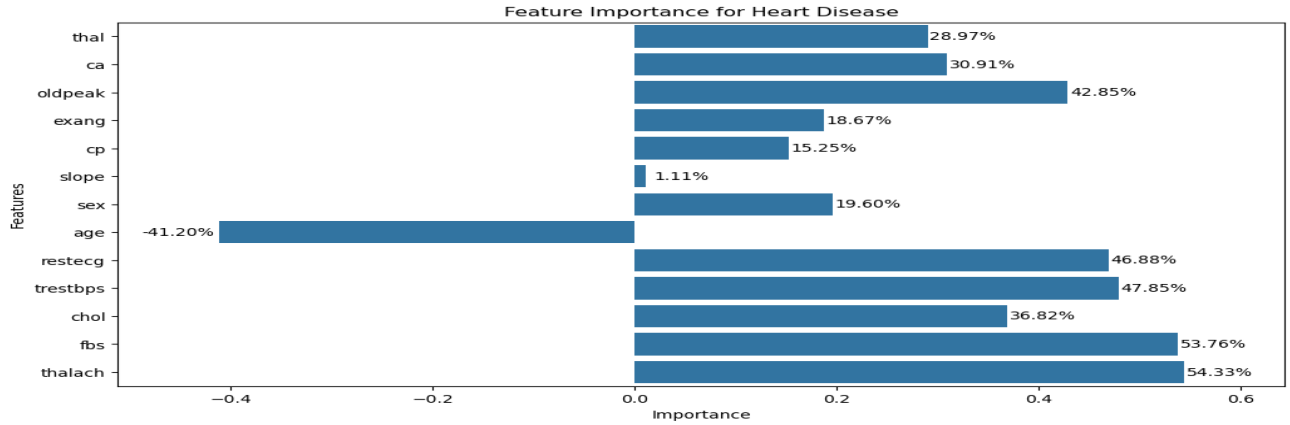

Figure 4: The Feature Importance for Heart Disease

3.1.4 Likelihood

Figure 5 represents the likelihood of patients with the absence of heart disease. It was noticed that the majority of likelihood values are concentrated towards the lower end of the scale, with only a few instances of higher likelihood values suggested, most patients with the absence of heart disease have relatively lower likelihood scores according to the Naive Bayes model.

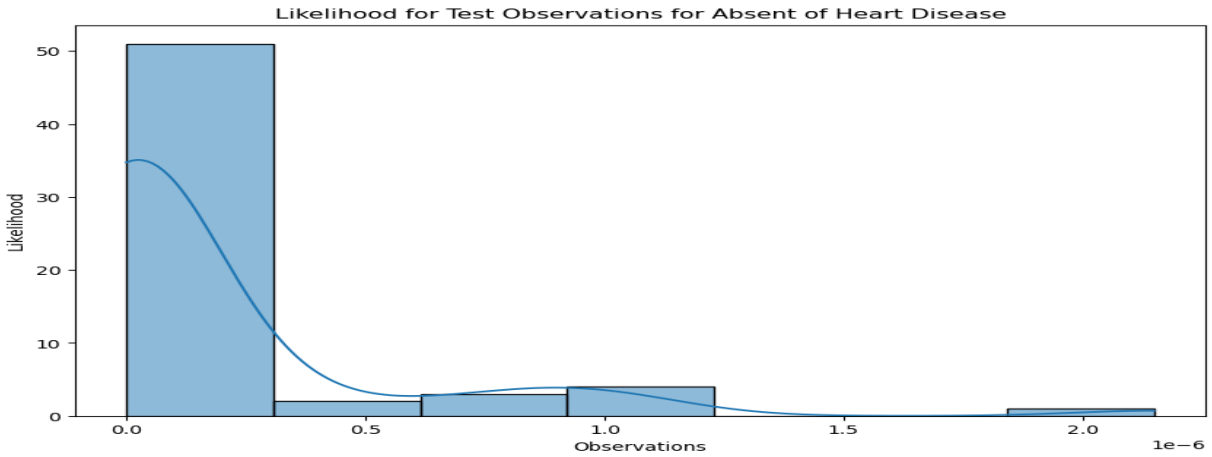

Figure 5: Likelihood for Absent of Heart Disease

Figure 6 represents the likelihood of patients with the presence of heart disease. It was noticed that the majority of likelihood values are concentrated towards the lower end of the scale, with only a few instances of higher likelihood values. This suggested that most patients with the presence of heart disease have relatively lower likelihood scores according to the Naive Bayes model.

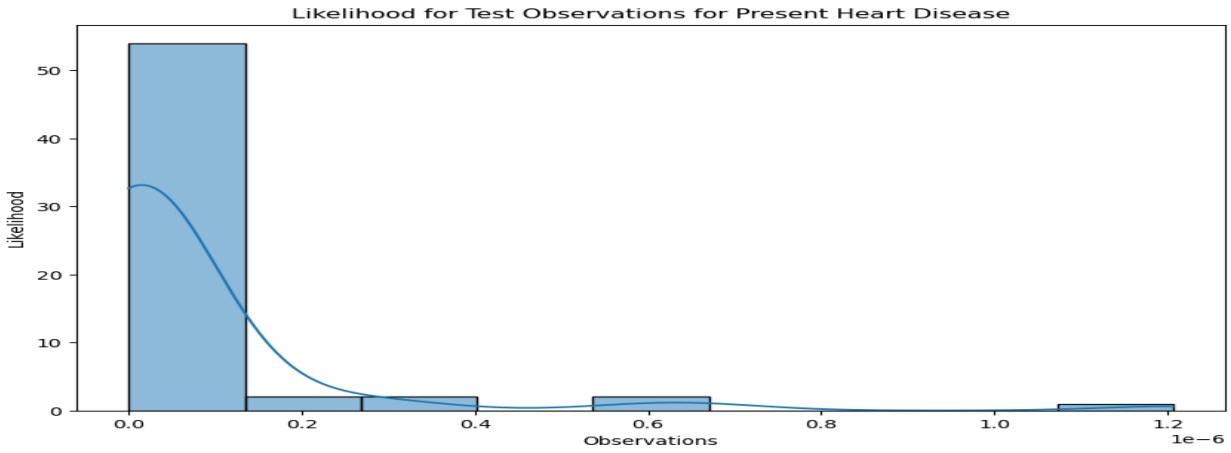

Figure 6: Likelihood for Present of Heart Disease

3.1.5 The Prior Probability of Naïve Bayes

The result below shows the prior probability of each observation:

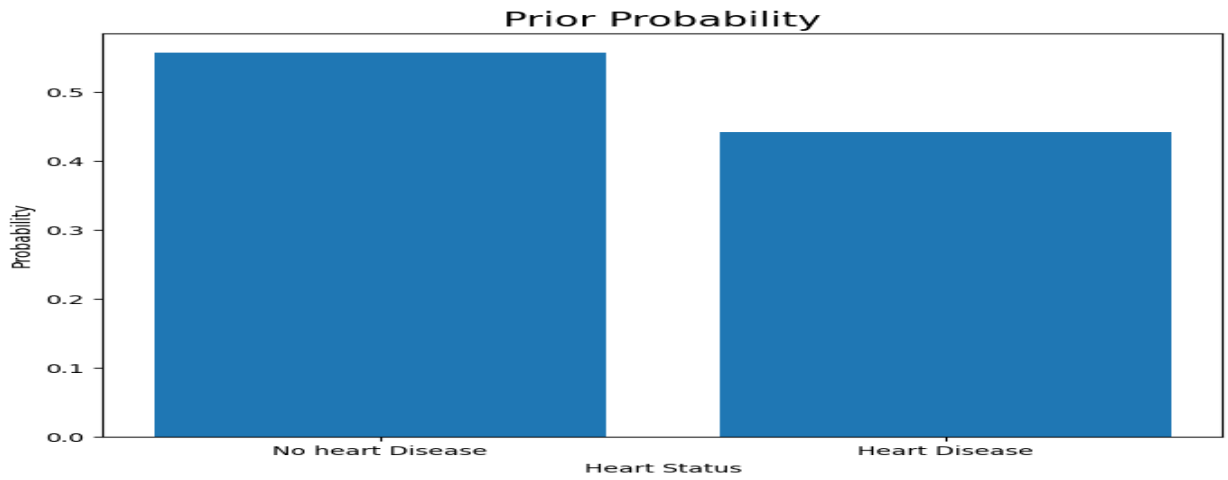

Figure 7: Prior Probability of Naïve Bayes = [no heart disease=0.55785124, heart disease=0.44214876]

The prior probability of a patient with the absence of heart disease was 0.55785124 which indicated that without considering any factors or evidence, there was a 55.79% probability that a patient record collected in the dataset has No heart disease, while the prior probability for the presence of heart disease from the patient information collected was 0.44214876, which implied, without considering any factors, there was 44.22% probability that a patient has heart disease.

Table 3: Posterior Probability of the Predicted Observations:

Absent	Present
0.0013	0.9987
0.1718	0.8282
0.0479	0.9521
0.7452	0.2548
0.1595	0.8405
0.0001	0.9999
0.0037	0.9963
0.0001	0.9999
0.9142	0.0858
0.6339	0.3661
0.9975	0.0025
0.9997	0.0003
0.023	0.977
0.0028	0.9972
0.0	1.0

0.997	0.003
0.9953	0.0047
0.0386	0.9614
0.0	1.0
0.991	0.009
0.1059	0.8941
0.9985	0.0015
0.0001	0.9999
0.9894	0.0106
0.0	1.0
0.9996	0.0004
0.9453	0.0547
0.0733	0.9267
0.0013	0.9987
0.7661	0.2339
0.4016	0.5984
0.657	0.343
0.9993	0.0007
0.8834	0.1166
0.9825	0.0175
0.4509	0.5491
0.0003	0.9997
0.9167	0.0833
0.0	1.0
0.9297	0.0703
0.0	1.0
0.945	0.055
0.1314	0.8686
0.976	0.024
0.9368	0.0632
0.0309	0.9691
0.9961	0.0039
0.9797	0.0203
0.0093	0.9907
0.0047	0.9953
0.9682	0.0318
0.998	0.002
0.9955	0.0045
0.9893	0.0107
0.0014	0.9986

0.9571	0.0429
0.9914	0.0086
0.0	1.0
0.0226	0.9774
0.0001	0.9999
0.9839	0.0161

Table 4 showed the posterior probability of each observation for 20% for testing of the model in which the total number are 61, the result are:

Table 4: Posterior Probability Summaries for the heart disease risk.

	Absent	**Present**
Median	0.6339	0.3661
Mean	0.5039	0.4960
Std	0.4565	0.4564
Min	0.0000	0.0003
25%	0.0037	0.0175
50%	0.6337	0.3661
75%	0.9825	0.9963
Max	0.9997	1.00

The median Posterior probability for patients with heart disease is around 0.36, indicating that approximately 36% in this class have a probability below this value. The interquartile range (IQR) spans from about 0.5 suggesting a wide range of probabilities within the middle 50% of the data. There are a few instances with very high probabilities close to 1.0, indicating high confidence in the prediction of class for these instances. There are also instances with very low probabilities close to 0.0, indicating low confidence in the prediction of class 1 for these instances.

Also, the median Posterior probability for patients without heart disease is around 0.63, indicating that approximately 63% of the instances in this class have a probability above this value. The interquartile range (IQR) spans from about 0.5 suggesting a wide range of probabilities within the middle 50% of the data. There are a few instances with very high probabilities close to 1.0, indicating high confidence in the prediction of patients without heart disease for these instances. There are also instances with very low probabilities close to 0.0, indicating low confidence in the prediction of patients without heart disease for these instances.

Figure 8 showed the box plot for the output of posterior probability risk of heart disease:

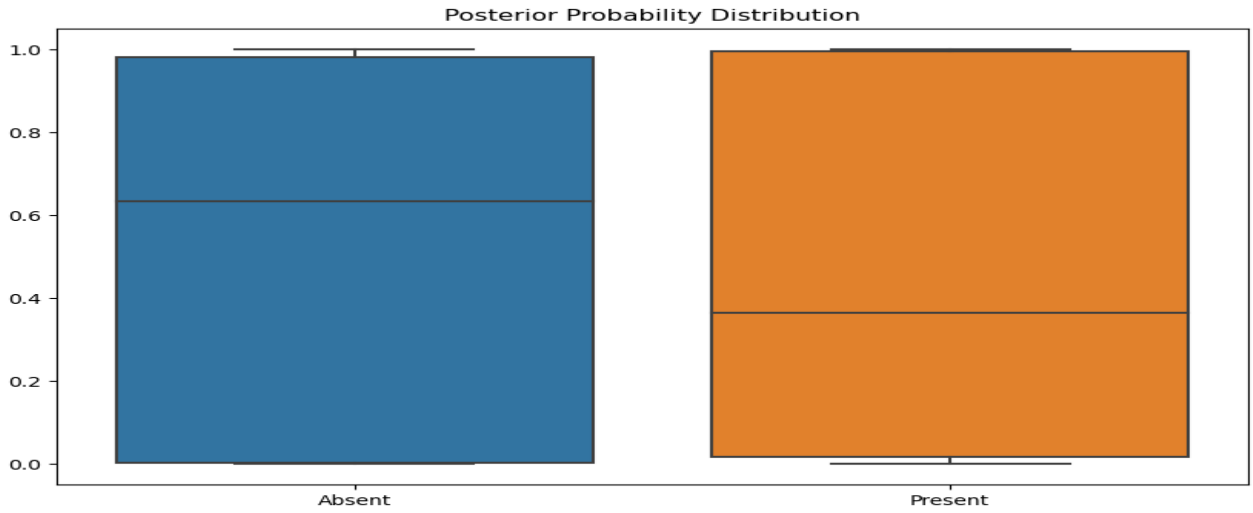

Figure 8: The steam boxplot Posterior Probability Distribution for the heart disease risk.

3.2 Decision Tree

3.2.1 Model performance

The results below showed the performance of the decision tree which indicated that our model has an accuracy of 75.41%, meaning it correctly classified patients with and without heart disease in about three-quarters of the cases. Moving on to precision, it was observed that out of 100 predicted positive cases, our model accurately identified 77.42 of them. This suggested a decent ability to minimize false positives. The recall, or true positive rate, is 75%, implying that our model captures 75% of the actual positive cases. The F1 Score is 76.19%, indicating an overall balanced performance in correctly identifying positive cases and minimizing false positives.

The ROC-AUC of 84.54% is a positive sign, showing our model's ability to distinguish between positive and negative cases. In terms of sensitivity, our model identified 75% of the actual positive cases. The specificity explained that the proposed model correctly identified patients without heart disease about 76% of the time. The balance between sensitivity and specificity is noticeable.

Table 5: Performance metrics for Decision Tree

Accuracy	75.41%
Precision	77.42%
Recall	75%
F1 Score	76.19%
ROC-AUC	84.54%
Sensitivity (True Positive Rate)	75%
Specificity (True Negative Rate)	75.86%

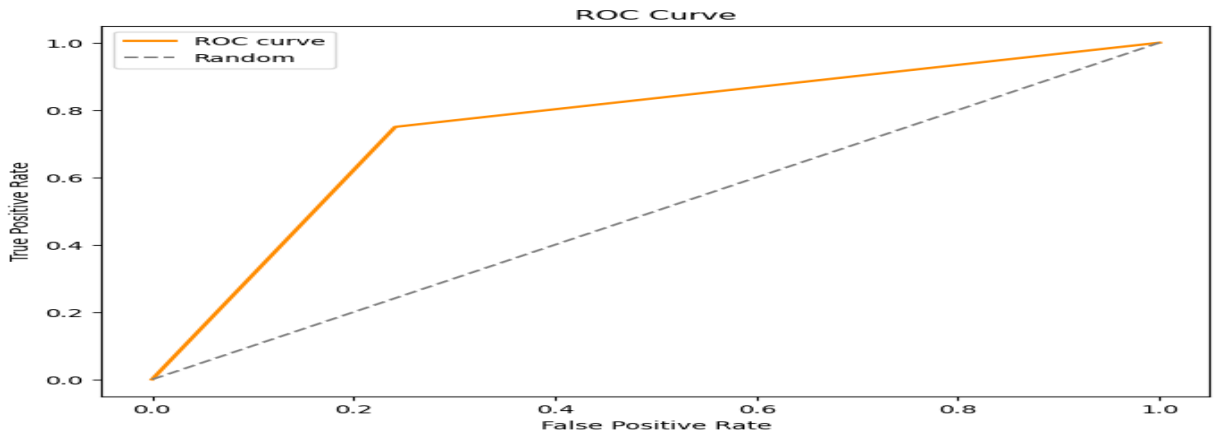

Figure 9: ROC Curve for the Performances of Decision Tree

3.2.2 Summary for Testing Observation

The report below shows that there is a harmonious balance between precision and recall for both the Present and Absent classes. Specifically, for the Present heart disease class, we observe a precision of 73%, indicating that out of the cases predicted as Present, our model is correct in 73% of them. The recall, which stands at 76%, signifies that our model successfully identifies 76% of all actual Present cases. The F1-score, a harmonized measure of precision and recall, is 75%, providing a comprehensive assessment of the model's performance in detecting Present instances. Also, in the Absent class, we find a precision of 77%, indicating that 77% of the cases predicted as Absent are indeed accurate. The recall for the Absent class is 75%, signifying that our model captures 75% of all actual Absent cases. The F1-score for Absent stands at 76%, underlining a balanced performance in precision and recall for this class. The model is evaluated on 32 instances of the Absent.

Table 6: Classification Report of decision Tree model Performance

	Precision	Recall	f1-score	Support
Absent	73%	76%	75%	29
Present	77%	75%	76%	32
Accuracy			75%	61
macro avg	75%	75%	75%	61
weighted avg	75%	75%	75%	61

3.2.3 Confusion Matrix

The confusion matrix below shows that the Decision tree model correctly identified and predicted 22 instances of the positive class (TP), and incorrectly predicted 7 instances as positive when they were truly negative (FP), There were 10 instances where the model predicted negative, but they were actually positive (FN), and it correctly identified and predicted 22

instances of the negative class.

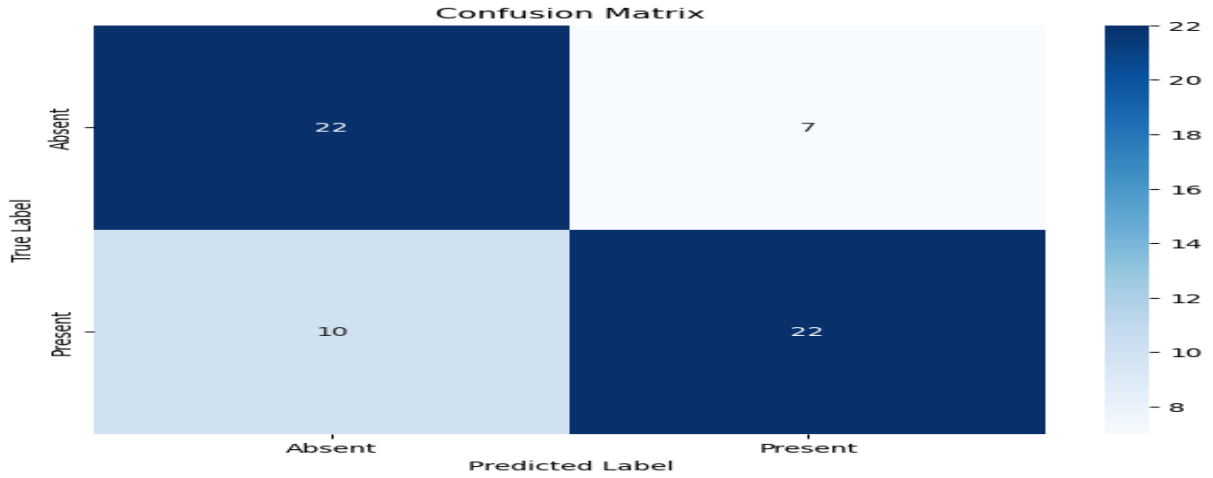

Figure 10: Confusion matrix for the performances using Decision Tree procedure.

3.2.4: Features Importance

The decision tree feature importance below showed that chest pain was 23.96% which reveals that, the risk of having heart disease was close to average but does not mean they don't have a tendency of having heart disease but not as critical as 50%. The number of major vessels, which was 14.96% revealed that the presence of major vessels doesn't have more effect on the risk of heart disease. The cholesterol of 12.14% showed a lower risk of heart disease. The thalassemia of 11.67% shows that there is a low level of presence of Thalassemia for the risk of heart disease. Age was 10.13% which is also a low risk of heart disease importance. blood pressure and ST depression induced by exercise, fasting blood sugar, maximum heart rate achieved, exercise-induced angina, the slope of peak exercise ST segment have a very low importance in risk of heart disease based on the analysis done. The resting electrocardiographic has no effect on the risk of heart disease based on this research analysis. It is confirmed that the higher the presence of the percentage of each feature, the higher the risk of heart disease will occur, and the lower the percentage of the features, the lower the risk will be.

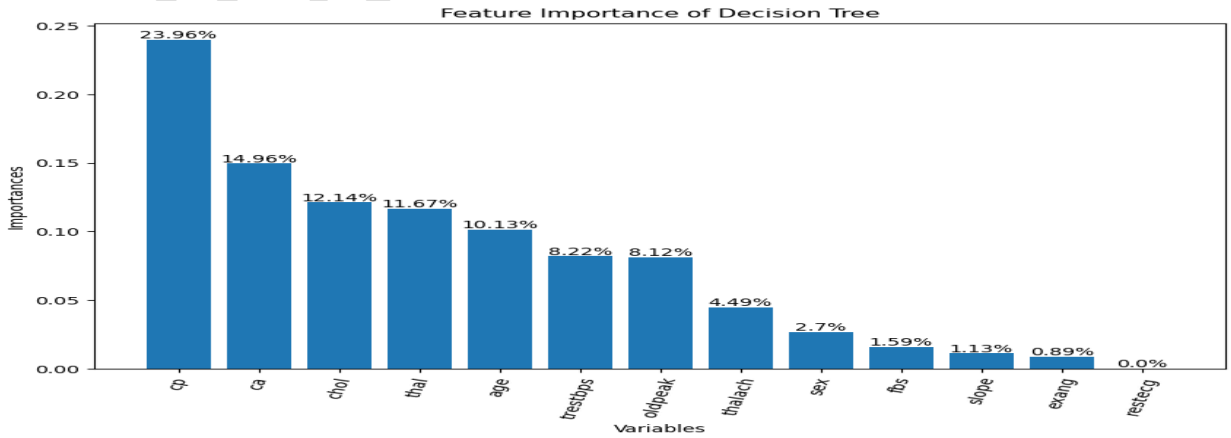

Figure 11: Features Importance

3.3 Comparison Between Naïve Bayes And Decision Tree

Table 7 showed that Naive Bayes outperformed the Decision Tree model in terms of accuracy, achieving an accuracy of 83.61% compared to 75.41% for the Decision Tree. Also, Naive Bayes exhibited higher precision compared to the Decision Tree, with a precision score of 89.29% versus 77.42% for the Decision Tree. This indicated that Naive Bayes has a higher proportion of true positive predictions among all positive predictions made. Both models showed relatively similar recall scores, with Naive Bayes achieving 78.12% and the Decision Tree achieving 75.00%. This means that both models are able to capture a substantial proportion of actual positive cases. Naive Bayes demonstrated a higher F1 score compared to the Decision Tree, indicating a better balance between precision and recall. The F1-score for Naive Bayes was 83.33%, while for the Decision Tree it was 76.19%. Interestingly, the Naïve Bayes model significantly outperformed Decision Tree in terms of ROC-AUC, with a score of 90% compared to only 84.54% for Decision Tree. This suggested that the Naïve Bayes model has better overall performance in terms of ranking the instances correctly. Naive Bayes and Decision Tree models have similar sensitivity scores, with Naive Bayes at 78.12% and Decision Tree at 75.00%. This indicated that both models have a similar ability to correctly identify positive cases. Naive Bayes again outperformed the Decision Tree in specificity, achieving a score of 89.66% compared to 75.86% for the Decision Tree. This suggested that Naive Bayes is better at correctly identifying negative cases.

Table 7: Metrics Comparison Table for Naïve Bayes and Decision Tree

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	Sensitivity	Specificity
Naive Bayes	83.61	89.29	78.12	83.33	90.00	78.12	89.66
Decision tree	75.41	77.42	75.00	76.19	84.54	75.00	75.86

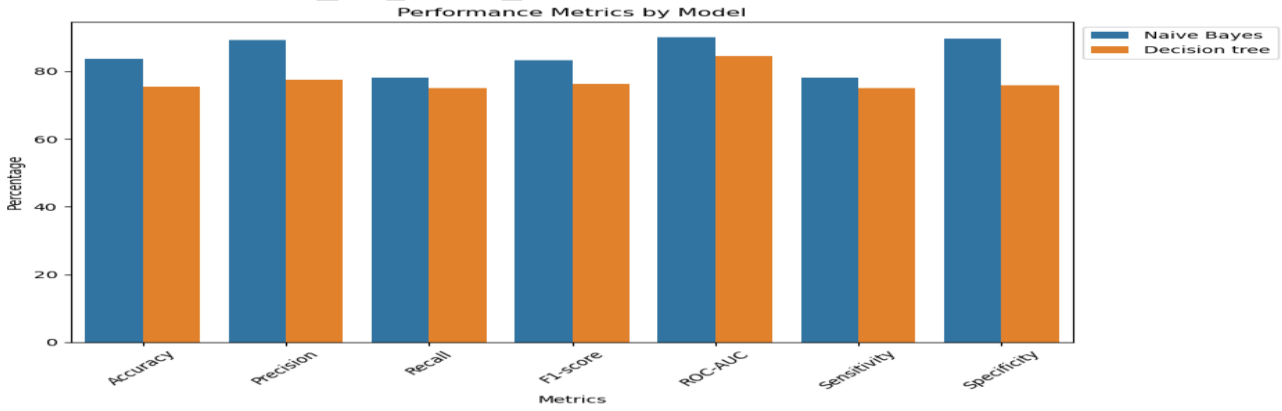

Figure 12: Model performance Comparison between Naïve Bayes and Decision Tree procedure.

4. Conclusion

This study has shed light on the effectiveness of Naive Bayes classification modelling in assessing the risk of heart disease. It was observed that the Naive Bayes exhibited superior performance, achieving an overall accuracy than the Decision Tree. The Naive Bayes model showed that its predictions matched the actual status of the patients. Similarly, for the precision,

recall, F1 score, ROC, sensitivity, and specificity, which provided a comprehensive assessment of its ability to correctly identify the risk of heart disease, in a patient. Furthermore, it could be concluded that, the most significant features among the diagnosed patients in the medical center were; the maximum heart rate achieved, fasting blood sugar, resting blood pressure, and chest pain. Also, for the 20 percent testing datasets, the Naive Bayes predicted correctly 26 cases of no heart disease risk, and 25 cases of heart disease risk but, the Decision tree predicted correctly 22 cases of no heart disease, and 22 cases also of heart disease risk. Thus, it highlighted the potential of Naive Bayes as a valuable tool for healthcare professionals in identifying individuals at risk of heart disease.

5. Recommendation for further study in the future work

This study can be extended to other machine learning algorithms like Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN).

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.
Option 1 was adopted.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc have been used during writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.

2.

3.

5 Reference

Aida Krichene (2017).Using a naive Bayesian classifier methodology for loan risk assessment.Journal of Economics, Finance and Administrative Science Vol. 22 No. 42, 2017pp. 3-24 Emerald Publishing Limited 2077-1886.

Akanbi, Olawale Basheer (2023). Application of Naïve Bayes to Students Performance Classification. AJPS.106724, Volume 25, Issue 1, pp 35-47.

Amra ,I. A. A, and Maghari, A.Y.A (2017). Student performance classification using naïve bayes and KNN.IEEE Xplore.https://doi.org/10.1109/ICITECH.2017.8079967

Benjamin, E. J., Muntner, P., & Bittencourt, M. S. (2019). Heart disease and stroke statistics-2019 update: a report from the American Heart Association. Circulation, 139(10), e56-e528.

Chenmeng Zhang, Can Hu, Shijun Xie and Shuping Cao (2021). Research on the application of decision tree and random forest algorithm in the main transformer fault evaluation. J.phys: conf. Ser.1732 012086.

Harshit Jindal et al u (2021). Heart disease prediction using logidtic regression, Random forest classifiers and KNN Algorithms machine learning algorithms. IOP Conf.Ser.Master.sci.Eng. 1022 012072.

Heidenreich, P. A., Trogdon, J. G., Khavjou, O. A., Butler, J., Dracup, K., Ezekowitz, M. D., ... & Woo, Y. J. (2011). Forecasting the future of cardiovascular disease in the United States. A policy statement from the American Heart Association.circulation.123: 933 - 944.

Hong Chen1& Songhua Hu, Rui Hua and Xiuju Zhao (2021). Improved naive Bayes classification algorithm for traffic risk management. EURASIP Journal on Advances in Signal Processing ,29(1),56-68.

Jagadish P J (2022). Identifying cardiovascular disease model using machine learning techniques. journal of Medical diagnosis and method,45(3), 210-225.

Jiao S. R., Song J., and Liu B. (2020). A review of decision tree classification algolrithms for continuous variables. J.phys: conf. Ser.1651 012083.

Kannan I. R. P., and Arunachalam A. S. (2021). Prediction Of Consumer Review Analysis Using Naive Bayes And Bayes Net Algorithms. 1Research Scholar,Department of Computer Science, School of Computing Sciences, Article 1865 Vol.12 No. 7, 1865-1874.

Karmen, C., Sébastien, D., & Tomaž, K. (2019). Predicting heart disease using machine learning algorithms. BioMedical Engineering image,6(1),1-10.

Kathija A., Shajun S. Nisha, and Mohamed Sathik (2017). Classification of breast cancer data using C4.5 Classifier algorithm. International journal of recent engineering research and development (IJRERD), vol.2, no.2.

Kononenko I.(2001). Machine learning for medical diagnosis. History ,state of art and perspective. Artificial intelligence in Medical,23(1), 89-109.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In European conference on machine learning (pp. 4-15).

Marzuki Ismail1 & Norlida Hassan1, Salem Saleh Bafjaish 21(2020). Comparative Analysis of Naive Bayesian Techniques in Health-Related for Classification Task. Journal Of Soft Computing And Data Mining Vol.1 No. 2 (2020) 1-10.

Nafizatus Salmi and Zuherman Rustam (2019). Naïve Bayes Classifier Models for Predicting the Colon Cancer. IOP Conf. Ser.: Mater. Sci. Eng. 546 052068.25(2),67-79.

National Centre for Chronic Disease Prevention and Health Promotion (May 15, 2024).

Niloy Biswas, Md Mamun Ali, Md Abdul Rahaman, Minhajul Islam, Md Rajib Mia, Sami AzM, kawsar Ahmed, Franis M.Bui, Fahad Ahmed Al Zahrani and Muhammad Ali Moni (2023).

Oliver Takawira & John W. Muteba Mwamba (2020). "Determinants of Sovereign Credit Ratings: An Application of the Naïve Bayes Classifier," Eurasian Journal of Economics and Finance, Eurasian Publications, vol. 8(4), pages 279-299.

Rumsfeld, J. S., Joynt, K. E., & Maddox, T. M. (2013). Big data analytics to improve cardiovascular care: promise and challenges. Nature Reviews Cardiology, 13(6), 350-359.

Sai V. Krishna Reddy, Meghana P., Subba N. V. Reddy, and Ashwath B. Rao (2022). Prediction on cardiovascular disease using decision tree and naive bayes classifiers. J.phys: conf. Ser. 2161 012015.

Siti Hadijah Hasanah (2022). Application of Machine Learning for Heart Disease Classification Using Naive Bayes21.Journal Matematika MANTIK, Vol. 8, No. 1, June 2022, pp. 67-77.

Syadiah Nor, Mokhairi Makhtar ,Hasnah Nawang (2017). Analysis on Students performance using naÏve Bayes classifier. Article in Journal of Theoretical and Applied Information Technology, Educational data mining, 9(1), 76-88·

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data?. PloS one, 12(4), e0174944. https://doi.org/10.1371/journal.pone.0174944.

Wickramasinghe, I. and Kalutarage H (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft computing Journal [online], 25(3), pages 2277-2293.

World Health Organization. (2021). Cardiovascular diseases. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1

Appendix

Table 8: Variables in the datasets used for the Model

Variables	Description
Age	Age
Sex	Sex
CP	Chest pain
Trestbps	Resting blood pressure
Chol	Serum cholesterol in mg|dl
Fbs	Fasting blood sugar
Restecg	Resting electrocardiographic
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina
Old peak	ST depression induced by exercise
Slope	The slope of peak exercise ST segment
Ca	Number of major vessels
Thal	Thalassemia
Num	Diagnosis of heart disease