Comparative Analysis of Supervised Machine Learning Algorithms for Predicting Cardiovascular Disease

Shreyans Jain[1], Mr. Lakhan Bhaskar Kadel[2]

[1]Class 10 Student, Crystal Springs Uplands Upper School; Hillsborough, California 94010
[2]Research Scholar, Dept. of Computer Science; ICFAI University Jaipur, Rajasthan 302031

ABSTRACT

Heart disease is one of the most dangerous and largest killers in the world. Identifying heart disease early has a vast positive effect on patient outcomes and their quality of life. In this research, we try to identify heart disease using machine learning (ML) algorithms. ML algorithms have the highest probability of success if they work on a data set with extensive and diverse information about the given problem - in this case heart disease. There are multiple types of ML algorithms to test, so we can try many different ones on the data, making the results more precise. Even if there are many different algorithms to test, each machine-learning solution can yield different results depending on the dataset used and our target goals. The main goal of this study is to find the differences between individual ML algorithms being used in our specific case: which ML algorithm, or combination of algorithms, is appropriate to detect heart disease with high accuracy? The ML algorithms used in this research are the Naive Bayes Classifier, the Random Forest classifier, and the Support Vector Machine (SVM) algorithm. Furthermore, this study generates insights into these ML algorithms – if a particular algorithm's model performs better than another on the dataset, analyzing this difference can help us understand what makes the model more suitable for diagnostic screening. Changing models' hyperparameters or their pre-processing techniques allows for a more robust and reliable model that can be readily incorporated into a healthcare environment.

Keywords: ML,SVM,RF,NB

INTRODUCTION

The heart is one of the main organs in the human body, responsible for keeping the body alive and functioning. Keeping a person's heart healthy is crucial. Heart disease, according to the World Heart Federation (WHF) **[10]**, more than 20.5 million people died from cardiovascular (heart) disease in 2021, more than 19.8 million people in 2022 according to the American College of Cardiology **[5]**, and more than 20 million people in 2023 according to the World Health Organization (WHO) **[9]**. In the last 3 years, heart disease has accounted for more than 60 million deaths around the world.

Identifying heart disease accurately, and efficiently can save millions of lives moving forward. Cardiovascular disease can easily be prevented with proper medical care and support if it is identified early. This research aims to create a machine-learning model that accurately identifies heart disease patients. Using multiple machine learning algorithms, like binary trees/Random Forest classifiers **[3]**, SVMs (Support Vector Machines) **[1]**, and types of Naive Bayes classifiers **[2]**, a highly efficient model can be found by ruling out ML algorithms that are less accurate than others.

Large datasets, like the one used in this research **[11]**, can be hard to understand for humans and do not serve much purpose for manual diagnosis. However, a machine learning algorithm is able to take advantage of this large amount of data to achieve a high accuracy in identifying heart disease. After using 83% of the dataset to train the model, 17% can be used to test the model effectively. After testing, algorithms' results can be compared to eventually find the most accurate algorithm.

LITERATURE REVIEW

This research was inspired by similar papers focused on heart-disease-detecting machine learning. Papers reviewed researched algorithms like SVM and Random Forest in a similar fashion to how we did. However, the results obtained by other researchers were somewhat different compared to our research and others' research as well. Harshit Jindal et al. **[4]** created a model using supervised algorithms like regression, K-nearest neighbors (KNN), and random forest classification. This model was a hybrid of the three, with an overall accuracy rating of 87.5% on the same dataset that was used in this research. Similarly to us, Harshit Jindal et al. **[4]** argued that a hybrid model is best for predicting heart disease: in their model, they used algorithms that work similarly to the ones we used. KNN is similar (in some ways) to Naïve Bayes because they are simpler models and both make decisions based on the locality of a point. The SVM and logistic regression are also similar because they both are linear and try to find hyperplanes to separate data.

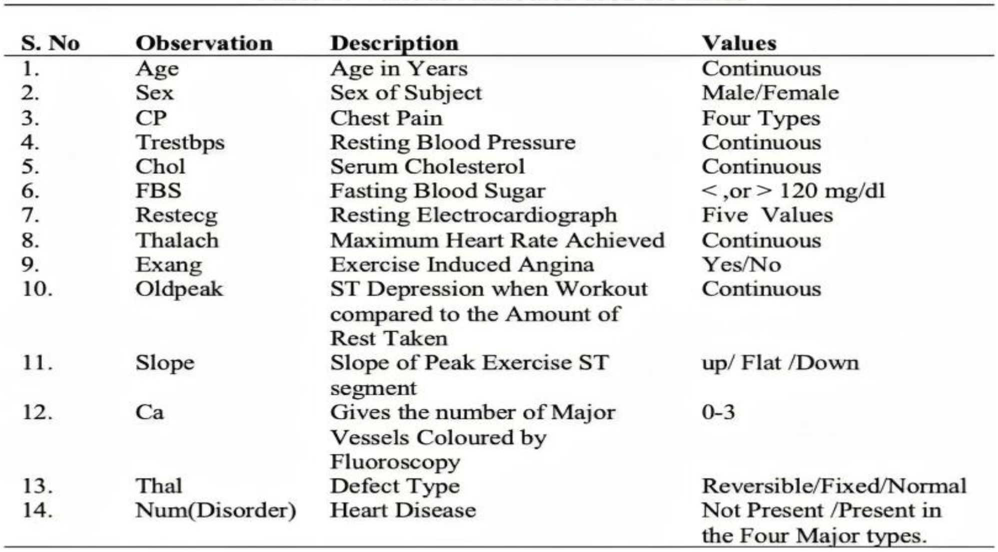
Another paper that supported our research was by Nikita Ahire et al. **[6]**. The algorithms used in their research were the SVM (achieving an accuracy rating of 84%), an artificial neural network (ANN - achieving an accuracy rating of 83.5%), and a random forest classifier, surprisingly only achieving an accuracy rating of 80%. Even though the achieved data accuracy ratings are different from ours, the pattern of algorithms' accuracy is similar. The random forest classifier achieved the lowest accuracy rating in both papers. The SVM and neural networks achieved very similar accuracies because they work somewhat similarly: ANNs use optimization techniques similar to SVMs' hyperplane-finding calculations.

Finally, the last paper we reviewed, by Reldean Williams et al. **[8]**, used the same UCI data library that we used. They researched several algorithms, along with SVM, Naïve Bayes, and Decision trees (random forest) like us, achieving similar accuracy results and ratings compared to us. However, the researchers had a higher accuracy rating for their random forest algorithm compared to our highest rating of Naïve Bayes. This could be because of their use of more pre-processing and specific hyperparameters that allowed the random forest classifier to more accurately predict heart disease.

USED DATASET

The dataset used in this research, as shown in 'Table 1' is from the UCI library **[11]** and is quite large and contains many different variables to allow for accurate identification of heart disease. The dataset used was beneficial to our specific ML models due to its data being easily sortable and recognizable. The dataset contains important attributes of patients like age, blood pressure, sugar level, peak blood pressure, and more. ML algorithms are able to identify specific traits with higher chances of having heart disease easily due to the dataset containing values from 303 people with 14 columns, each representing a different attribute. As stated before, the data is split for training and testing with training getting 83% of the data while 17% of the data is used for testing. These specific values were found after tweaking them to find which specific values would result in the overall highest accuracy among the three tested algorithms (SVM, Random Forest, and Naive Bayes).

Table 1. Various Attributes used are listed



S. No	Observation	Description	Values
1.	Age	Age in Years	Continuous
2.	Sex	Sex of Subject	Male/Female
3.	CP	Chest Pain	Four Types
4.	Trestbps	Resting Blood Pressure	Continuous
5.	Chol	Serum Cholesterol	Continuous
6.	FBS	Fasting Blood Sugar	< ,or > 120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five Values
8.	Thalach	Maximum Heart Rate Achieved	Continuous
9.	Exang	Exercise Induced Angina	Yes/No
10.	Oldpeak	ST Depression when Workout compared to the Amount of Rest Taken	Continuous
11.	Slope	Slope of Peak Exercise ST segment	up/ Flat /Down
12.	Ca	Gives the number of Major Vessels Coloured by Fluoroscopy	0-3
13.	Thal	Defect Type	Reversible/Fixed/Normal
14.	Num(Disorder)	Heart Disease	Not Present /Present in the Four Major types.

IV-I. Experiment & Methodology.

In all of the researched ML algorithms, the main steps remain the same to have accurate results that can be compared. ‘Figure 1’ below represents the cycle that an algorithm takes before reaching a conclusion if a patient has heart disease (1) or doesn’t have heart disease (0). All algorithms operate differently. All tests were run on a Macbook Pro with an M2 Pro processor with 16 GBs of RAM. The Python interpreter used was Python 3.12.2 (64 bits).

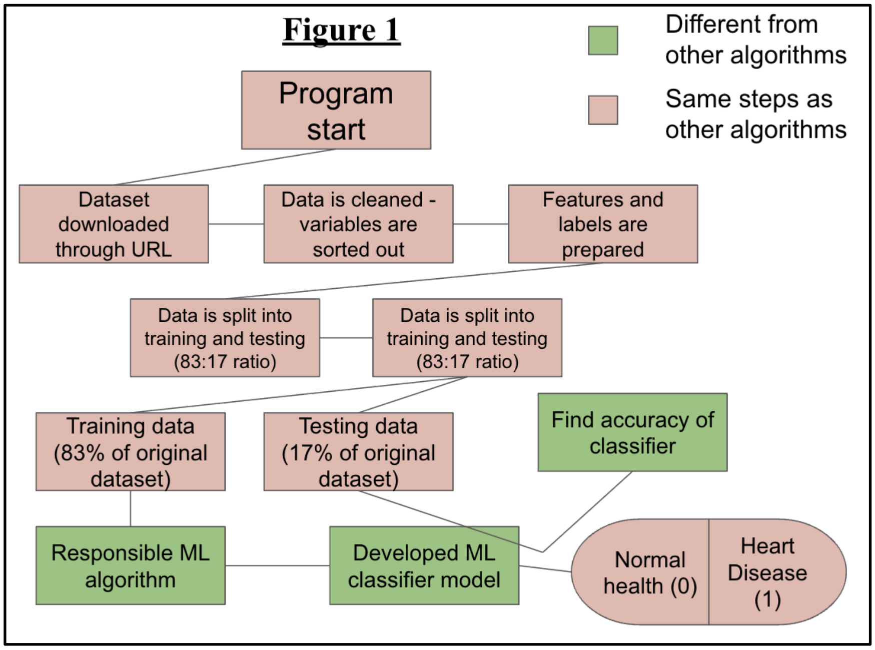


Figure 1

🟩 Different from other algorithms
🟥 Same steps as other algorithms
Program start
Dataset downloaded through URL
Data is cleaned - variables are sorted out
Features and labels are prepared
Data is split into training and testing (83:17 ratio)
Data is split into training and testing (83:17 ratio)
Training data (83% of original dataset)
Testing data (17% of original dataset)
Find accuracy of classifier
Responsible ML algorithm
Developed ML classifier model
Normal health (0)
Heart Disease (1)

IV-II. Support Vector Machine.

A support vector machine (an SVM) **[1]** works by plotting points on a graph. Before creating this graph, the data is organized and preprocessed, as shown in ‘Figure 1’. Each point represents a patient with different characteristics in the dataset (i.e. the variables associated with them are different from others). When all the points are graphed, the SVM algorithm separates the points that are associated with having heart disease from the ones that don’t. Finally, it draws a “line” that represents an average of all of the points, keeping both groups as far apart as possible. These groups are kept as far as possible because the algorithm maximizes the gap between the two closest points, or support vectors, from each group. This gap is called a “hyperplane” or the distance between those two points, as seen in ‘Figure 2’.

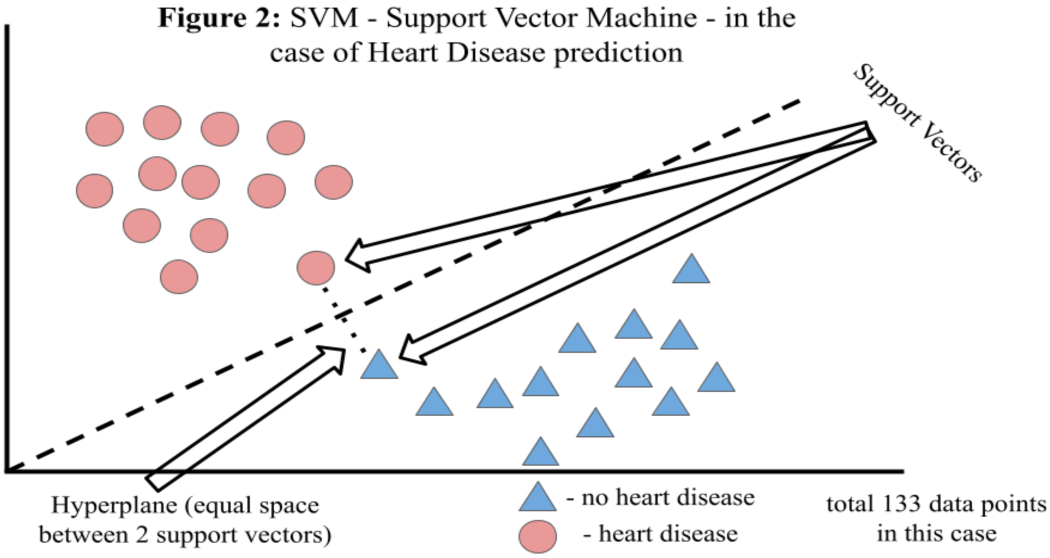


Figure 2: SVM - Support Vector Machine - in the case of Heart Disease prediction

Support Vectors
Hyperplane (equal space between 2 support vectors)
▲ - no heart disease
● - heart disease
total 133 data points in this case

IV-III. Random Forest Classifier.

The random forest classifier **[3]** consists of many binary trees (hence its name - forest). Each binary tree is created in the training process of the algorithm, where data is viewed after preprocessing and organization (as seen in 'Figure 1'). Each unique tree represents one instance, in this case, patients, in the dataset. Each tree consists of branches, as seen in 'Figure 3', that extend out from an original branch. Each branch consists of a binary-like-question: if the question is answered 0 (no) by the algorithm, it will move to a different branch than if the question was answered 1 (yes). Finally, when the final branch is reached, a conclusion is deduced. This "forest" of binary trees works together: first, each tree comes to its own conclusion. Then, majority voting takes place, resulting in a final decision of whether or not the patient with inputted characteristics has heart disease or not.

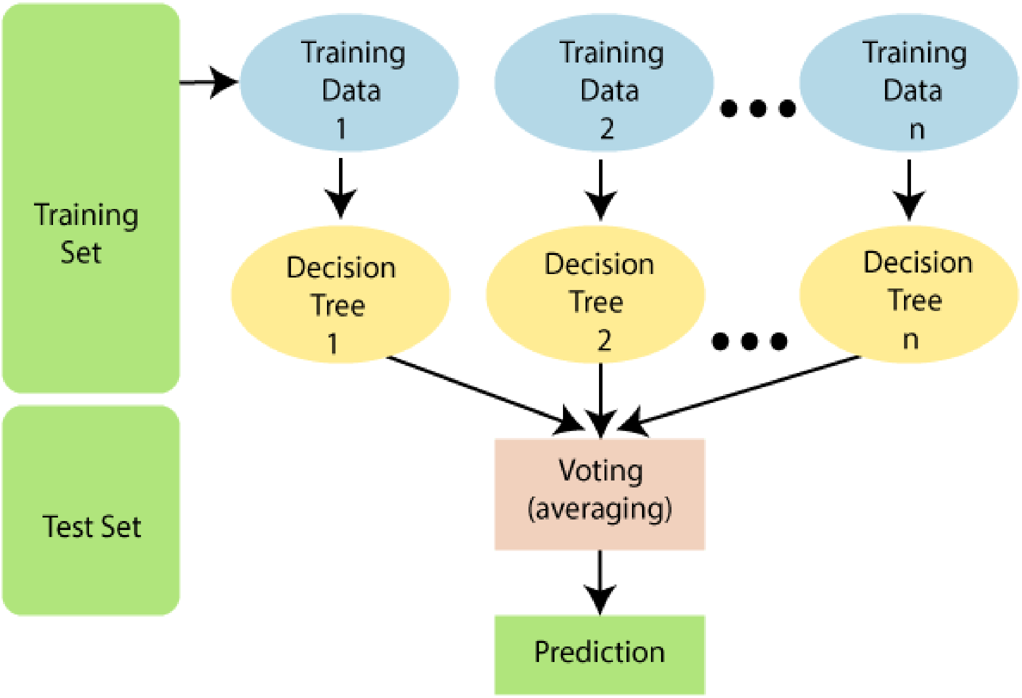
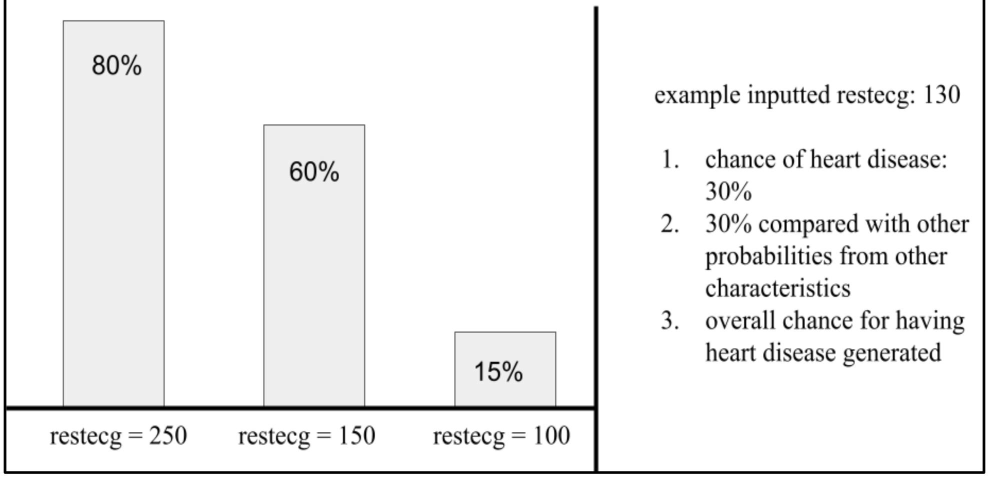


Figure 3

Figure 4: Naïve Bayes Classifier - predicting heart disease - using the "restecg" value from the UCI dataset
*values on the graph are just for explanation: they may be inaccurate



example inputted restecg: 130

1. chance of heart disease: 30%
2. 30% compared with other probabilities from other characteristics
3. overall chance for having heart disease generated

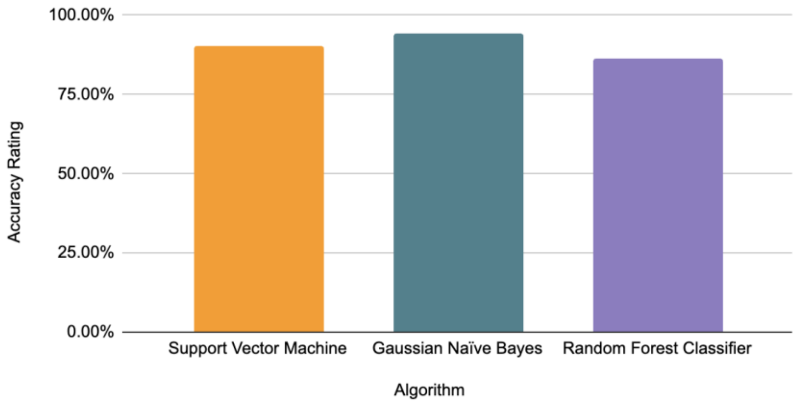
IV-IV. Naïve Bayes Classifier.

The Naïve Bayes Classifier **[2]** is quite similar to the SVM **[1]**. The Naïve Bayes classifier looks at all of the features of the dataset and calculates probabilities of certain traits being linked to certain outcomes. One way to think about it is using a graph. Instead of making combinations like the SVM, the Naïve Bayes algorithm treats all features independently of each other, as seen in 'Figure 4'. This method makes the algorithm run efficiently and simply. Once data is plugged into the algorithm from a testing dataset, it calculates a probability average based on all of the individual probabilities it has calculated for each trait from the training data of a patient having heart disease.

RESULTS & CONCLUSION

The heart disease prediction model created after testing each individual algorithm consisted of a Gaussian Naïve Bayes algorithm **[7]** as it performed the highest compared to any other algorithm, even after trying the three in combinations. Gaussian Naïve Bayes is a type of Naïve Bayes algorithm that focuses on continuous data instead of separate and independent data points. After running the SVM **[1]**, Random Forest **[3]**, and Gaussian Naïve Bayes algorithms **[7]**, each algorithm's individual accuracy rating was obtained. They were all tested under the same environment, with a training and testing data ratio of 83:17 which was used because it resulted in the overall highest accuracy for each of the three algorithms. The algorithms' overall accuracies were tested after measuring their precision, recall, and f1-score over their support. Then, their weighted averages were calculated which were then averaged (in regards to their corresponding support ratings) to create the final overall accuracy rating. The SVM achieved an accuracy rating of 90.2% with weighted averages of 90.16%, 90.2%, and 90.14%. The Random Forest Classifier received an overall accuracy rating of 86.27% with weighted averages of 86.48%, 86.27%, and 86.34%. Finally, the Gaussian Naïve Bayes algorithm obtained an overall accuracy rating of 94.12% on the data with weighted averages of 94.62%, 94.12%, and 94%. The Gaussian Naïve Bayes had a higher accuracy rating than the SVM and Random Forest classifier because it calculates features having normal (around equal) distributions: the used dataset has an equal distribution of all kinds of heart disease/healthy patients with similar traits which is why the Gaussian Naïve Bayes does so well. Both Naïve Bayes and the support vector machine are simpler and faster algorithms that work with linear relationships; both can be explained with a graph. However, the random forest algorithm uses a more complex approach with large amounts of binary trees, which in this case result in being inaccurate due to the limited data. Thus, with this specific dataset, an algorithm like Naïve Bayes or SVM is more likely to produce accurate data, but with a large medical dataset, a Random Forest Classifier could potentially produce accurate results in a hybrid with other algorithms. Each algorithm's accuracy rating is shown in 'Figure 5'.

Figure 5: Accuracy of heart disease predicting Machine Learning models using Python and a medical dataset



REFERENCES

[1]. IBM, editor. "How SVM Works." *IBM*, International Business Machines, 20 Dec. 2022, www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works. Accessed 12 July 2024.

[2]. ---, editor. "What are Naïve Bayes Classifiers?" *IBM*, International Business Machines, 18 Mar. 2023, www.ibm.com/topics/naive-bayes. Accessed 12 July 2024.

[3]. ---, editor. "What is random forest?" *IBM*, International Business Machines, 7 Mar. 2023, www.ibm.com/topics/random-forest. Accessed 12 July 2024.

[4]. Jindal, Harshit, et al. "Heart Disease Prediction Using Machine Learning Algorithms." *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, 1 Jan. 2021, p. 012072. *ResearchGate*, https://doi.org/10.1088/1757-899x/1022/1/012072. Accessed 12 July 2024.

[5]. Magazine, Cardiology. "New Study Reveals Latest Data on Global Burden of Cardiovascular Disease." *American College of Cardiology*, 2024 American College of Cardiology Foundation, 11 Dec. 2023, www.acc.org/latest-in-cardiology/articles/2024/01/01/01/42/feature-new-study-reveals-latest-data-on-global-burden-of-cardiovascular-disease. Accessed 10 July 2024.

[6]. Rindhe, Baban U., et al. "Heart Disease Prediction Using Machine Learning." *International Journal of Advanced Research in Science, Communication and Technology*, 12 May 2021, pp. 267-76, https://doi.org/10.48175/ijarsct-1131.

[7]. Sci-kit team. "1.9 - Naive Bayes: 1.9.1 - Gaussian Naive Bayes." *Scikit Learn*, Scikit-learn Developers, 2007, scikit-learn.org/stable/modules/naive_bayes.html. Accessed 12 July 2024.

[8]. Williams, Reldean, et al. "Heart Disease Prediction using Machine Learning Techniques." International Conference on Data Analytics for Business and Industry, 2021, Oct. 2021, ieeexplore.ieee.org/document/9655783/. Accessed 13 July 2024.

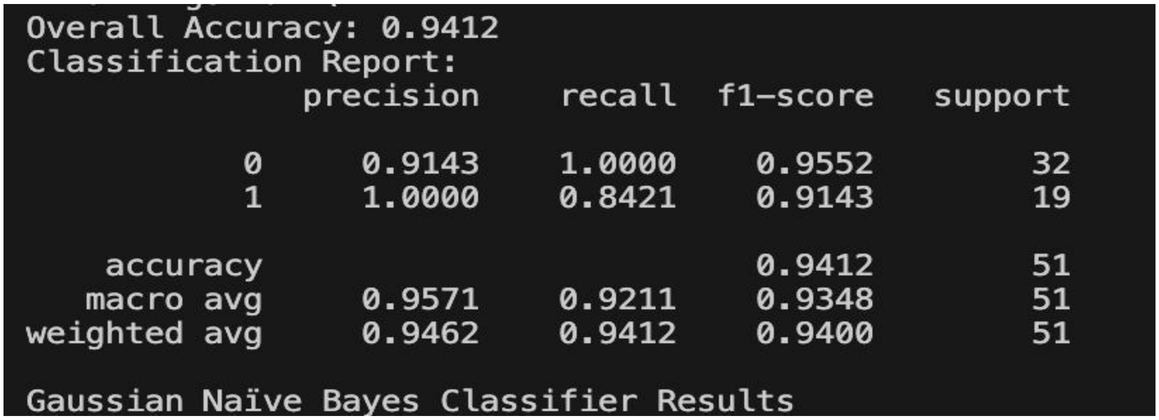
[9]. WHO, editor. "Cardiovascular diseases (CVDs)." *WHO*, World Health Organization, 11 June 2021, www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed 13 July 2024.

[10]. Aboyans, Victor, et al. "WORLD HEART REPORT 2023: CONFRONTING THE WORLD'S NUMBER ONE KILLER." Edited by Edward Fox. *World Heart Federation*, 2023, world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf.Accessed 13 July 2024.

[11]. "Heart Disease." *UC Irvine: Machine Learning Repository*, UCI, 30 June 1988,archive.ics.uci.edu/dataset/45/heart+disease. Accessed 13 July 2024. This is the UCI heart disease dataset used in our research.

VII. Appendix:

OUTPUT SCREENSHOTS:

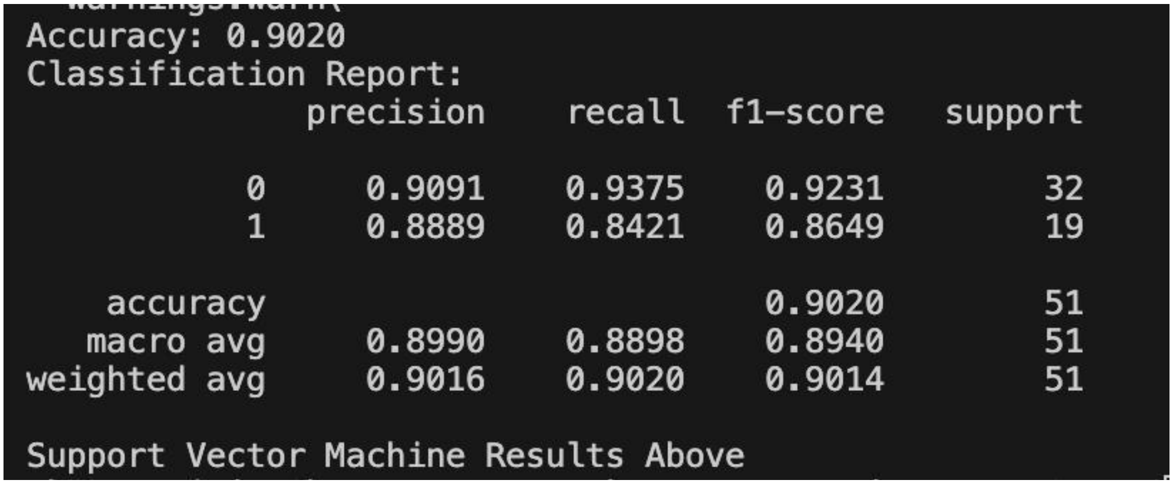


```
Overall Accuracy: 0.9412
Classification Report:
              precision    recall  f1-score   support

           0     0.9143    1.0000    0.9552        32
           1     1.0000    0.8421    0.9143        19

    accuracy                         0.9412        51
   macro avg     0.9571    0.9211    0.9348        51
weighted avg     0.9462    0.9412    0.9400        51

Gaussian Naïve Bayes Classifier Results
```

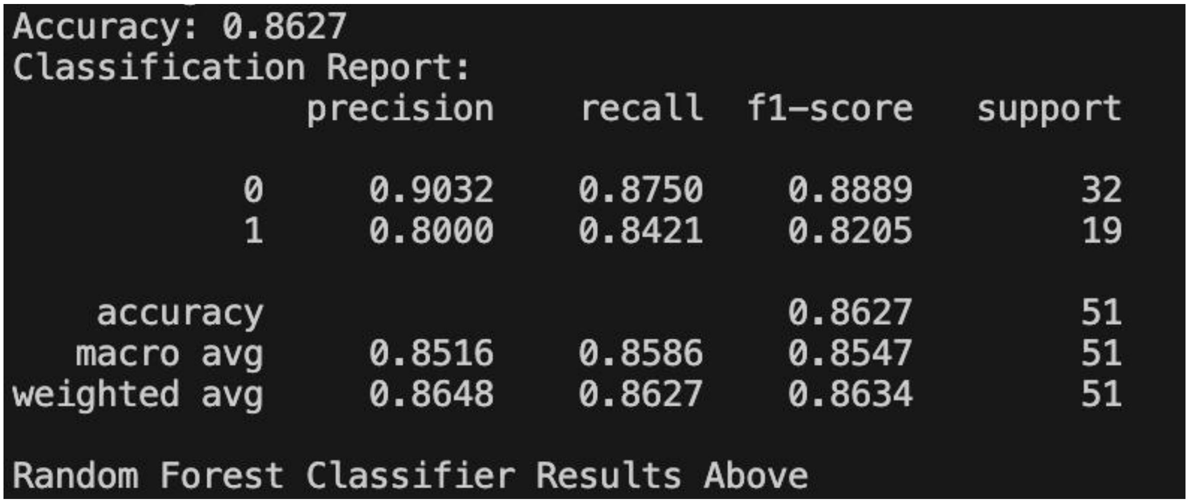


```
warnings.warn(
Accuracy: 0.9020
Classification Report:
              precision    recall  f1-score   support

           0     0.9091    0.9375    0.9231        32
           1     0.8889    0.8421    0.8649        19

    accuracy                         0.9020        51
   macro avg     0.8990    0.8898    0.8940        51
weighted avg     0.9016    0.9020    0.9014        51

Support Vector Machine Results Above
```



```
Accuracy: 0.8627
Classification Report:
              precision    recall  f1-score   support

           0     0.9032    0.8750    0.8889        32
           1     0.8000    0.8421    0.8205        19

    accuracy                         0.8627        51
   macro avg     0.8516    0.8586    0.8547        51
weighted avg     0.8648    0.8627    0.8634        51

Random Forest Classifier Results Above
```