Heart Failure Prediction using Gaussian Naïve Bayes Algorithm

Dr. S Suriya[1], N H Madhumitha[2]

[1]Associate Professor, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamil Nadu, India.

[2]PG Scholar, Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamil Nadu, India.

E-mail: suriyas84@gmail.com[1], ss.cse@psgtech.ac.in[1], 22mz02@psgtech.ac.in[2]

Abstract

Heart failure affects a minimum of 26 million individuals and its occurrence has been increasing day-by-day. Heart failure occurs when the heart cannot pump enough blood to meet the body's needs. This is caused due to various reasons such as coronary heart disease, heart valve malfunctioning, diabetes, anaemia etc. So, it is important to predict heart failure in its early stage to reduce the mortality rate. Cardiovascular disease is the major contributing factor for the prediction of heart failure. This research uses Gaussian Naïve Bayes technique which comes under supervised learning algorithm to predict heart failure. Gaussian Naïve Bayes algorithm treats each feature to be independent and each feature has equal importance in predicting heart failure. For the purpose of predicting heart failure, the mean and standard deviation of each column in a dataset is used. Different datasets are used for the implementation of heart failure prediction, and performance metrics are identified for each dataset. The findings of this study suggest that the Gaussian Naive Bayes algorithm can be a useful algorithm in predicting heart failure, and it has the potential to improve patient outcomes by facilitating early detection of heart failure. There are various metrics used for the performance analysis of the system such as, accuracy, recall, precision, confusion matrix and ROC curve. Finally, this study concludes by presenting the implementation of heart failure using Gaussian Naïve Bayes.

Journal of Information Technology and Digital World, June 2023, Volume 5, Issue 2, Pages 125-142

DOI: https://doi.org/10.36548/jitdw.2023.2.004
Received: 18-04-2023. received in revised form: 18-05-2023, accepted: 02-06-2023, published: 12-06-2023

Keywords: Heart Failure (HF), Supervised learning algorithms, Gaussian Naïve Bayes (GNB), Performance metrics, Receiver Operating Characteristic (ROC) Curve

1. Introduction

Heart Failure (HF) happens when it is incapable pumping the blood required by the body. Millions of people are affected by it, making it as a major global reason for mortality. This can occur due to a variety of factors, such as damage to the heart muscle, narrowed or blocked blood vessels or other underlying health conditions. The symptoms of HF can vary depending on the severity of the condition and the underlying causes. It includes shortness of breath, tiredness, swollen legs and feet, and rapid and irregular heartbeat. HF can lead to complications such as fluid build-up in the lungs or other organs, and increased risk of heart attack or stroke. The diagnosis and management of HF are complex due to the variety of causes and clinical presentations. Early identification and treatment of HF are crucial to prevent adverse outcomes such as hospitalization and death.

Machine learning enables the computer to learn from data and past experience without being explicitly programmed. It makes the computer to think like human; how humans take decision in a particular situation. Machine learning is of three types: supervised, unsupervised and reinforcement learning [2]. The type of machine learning depends upon the dataset used to train the system and the learning environment. Machine learning is used to identify hidden patterns in the dataset. It consists of thousands of data points, so it takes less time to train the system. Thus, machine learning domain is chosen to predict HF.

The accuracy of heart failure diagnosis can be improved by using machine learning techniques. Gaussian Naive Bayes (GNB) is a simple but powerful machine learning algorithm that has been widely used in various applications, including medical diagnosis. GNB is a probabilistic model that is based on Bayes' theorem. Bayes theorem has two main assumptions they are, each feature has equal contribution in predicting the disease and the second one is each feature is independent. GNB follow a Gaussian distribution. GNB has been used to predict HF outcomes based on clinical data. It is a probabilistic model that assumes that the features are independent and normally distributed. GNB has been applied to various medical domains, including cancer diagnosis and cardiovascular risk prediction.

GNB is a type of algorithm that comes under supervised machine learning algorithm, as it requires a labelled dataset to train the model. The labelled dataset consists of patient records with HF status (positive or negative) and their corresponding features. The GNB algorithm calculates the probability of HF using Bayes' theorem, which states that the probability of HF given some observed evidence (the features) is proportional to the product of the prior probability of the HF and the likelihood of the evidence given for the HF. In other words, GNB estimates HF probability based on the characteristics of the patient.

The prior probability of HF is the overall prevalence of the disease in the population. This can be estimated from the labelled dataset. The likelihood of the evidence given HF is calculated using the Gaussian distribution, which assumes that the features are normally distributed. The training dataset is used estimating the mean as well as the standard deviation.

2. Literature Survey

Research [1] described the major advancements in machine learning over the past decade, including the rise of deep learning, the increasing availability of large datasets, and the emergence of new applications in areas such as voice recognition, Natural Language Processing (NLP), computer vision. The development of a recommendation system is also an example.

In [2], an overview about different types of algorithms that comes under machine learning was given. The author also elaborated major machine learning algorithms. The inference from this study is if there is lesser amount of data to train the system, then it is better to go with machine learning.

A description of Machine Learning Algorithms was given in paper [3]. According to the authors, these algorithms are evaluated on metrics including accuracy, recall and F1 score. The work, in general, gives an important comparison between a number of widely used machine learning algorithms. Based on the information set and the number of cases and variables used with diabetes data from Waikato Environment for Knowledge Analysis (WEKA), the author determines the most efficient classification algorithm. The author concludes that SVM and ANN have also a similar accuracy. Both have good accuracy in comparison with other algorithms.

Research [4] implemented disease prediction system and the results show that Support Vector Machine outperforms Naïve Bayes Classifier in disease detection, while joint implementation outperforms both classifiers individually. This study focuses on concurrently implementing supervised learning algorithm in disease detection. Here SVM and Naive Bayes were used to detect disease. Thus, the author concludes hybrid system has higher accuracy than SVM and Naive Bayes classifier implemented separately.

In [5], the working of various supervised algorithms was explained. This study mentions advantages and disadvantages of each algorithm and compares each algorithm with other algorithms. This research clearly states the working of each machine learning algorithm. It gives a clear overview of working of the algorithms.

The authors of [6] have clearly stated the working of each supervised algorithm. The author compares the advantage and disadvantage of supervised algorithms and also compares the performance of supervised algorithm in disease prediction. The author concludes that in heart disease prediction, SVM, NB show high accuracy, in diabetes and Parkinson disease prediction SVM shows high accuracy and for breast cancer prediction, ANN shows high accuracy.

In [7], MARKER-HF algorithm that uses machine learning to predict the risk of cardiac failure in patients with cardiovascular disease, was used. The algorithm uses a combination of clinical, demographic, and laboratory data to predict the risk of heart failure. However, generalization problems were there in using demographic data.

The dataset collected from the UCI database has been used for the study [8]. In developing and comparing models, the authors used a broad range of machine learning supervised algorithms in order to assess their performance and concluded that SVM performed well.

Research [9] highlighted the potential of machine learning to improve the efficiency of heart failure diagnosis and classification, and to predict patient outcomes such as mortality and hospital readmission. Overall, the author provided a review of the current state of machine learning in the clinical setting for heart failure diagnosis, classification, and prediction.

The inference from [10] is that the survival rate for the patients was predicted. The results showed that the Random Forests method obtained a high degree of accuracy when predicting survival in patients with heart failure, compared to all other methods. Regarding future developments, the authors think of applying machine learning approach to other cardiovascular heart disease dataset and other diseases such as cancer.

The study [11] used electronic health record data for predicting heart failure in cancer patients once after completing the cancer diagnoses. The study showed that it is possible to predict cardiac failure in cancer patients using machine learning techniques. LR and RFs performed well in comparison with the four machine learning models.

The author of [12] studied about the survival and quality of life in patients with heart failure. The study had limitations because the mortality follow-up duration was only three years. In recent years, the brain natriuretic peptide has become a good indicator of the severity of HF. The next step is to identify patients with a lower quality of life, given that Quality of Life is an indicator of mortality.

The authors of [13] performed a systematic review of studies that developed or validated risk prediction models for patients with heart failure. Several electronic databases were searched for relevant articles published between 1980 and 2013. The authors evaluated system performance taking into consideration the discrimination and calibration.

In [14], in order to analyze the data and develop a model which can predict patient condition and determine any possible complications, the approaches of machine learning were utilized. This is a dedicated, purposed dashboard for treating patients with heart failure that includes key management and other essential functions associated with the disease.

Research [15] used data from a heart failure registry to develop a machine learning algorithm that could predict patients' risk of hospitalization or death within the next six months. Several algorithms were compared based on the precision, accuracy and sensitivity, The authors found that the random forest algorithm had good performance, with an AUC-ROC of 0.75.

In [8], the performance was evaluated, and the results showed that the SVM algorithm outperformed other algorithms with an accuracy of 84.85%. MARKER HF algorithm was used in [9].

3. Proposed Methodology

In this study, Gaussian Naïve Bayes has been chosen over SVM and other popular algorithms as it is easy to implement with less parameters to achieve good accuracy levels. The prediction of heart failure using Gaussian Naïve Bayes (group of algorithms that are based on the bayes theorem) is proposed in this work. It is used for solving classification problems. The fundamental assumption of Naïve Bayes assumption is that each feature is independent and has equal contribution in predicting the outcome.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{1}$$

Gaussian Naïve Bayes supports continuous values associated with each feature and are assumed to be distributed according to a Gaussian distribution. Normal distribution is also referred to as the Gaussian distribution. It is assumed that the probability of these features will be Gaussian. Hence the conditional probability is given by Equation (2).

$$P(x \mid \text{class}) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{(x-\text{mean})^2}{2\sigma^2}\right) \tag{2}$$

3.1 Steps Involved in Gaussian Naïve Bayes

Step-1: Gather the data required. Split the dataset for training as well as testing.

Step-2: Compute the prior probability.

Step-3: Compute mean and standard deviation for each class.

Step-4: Calculate class conditional probability. Use the mean and standard deviation from step 3.

Step-5: Calculate posterior probability.

Step-6: Based on calculated posterior value, make predictions.

Step-7: The model is ready.

3.2 Manual Tracing with Respect to Small Sample of the Dataset

Before implementing the algorithm, a small portion of the dataset shown in Table1 is used for manual tracing purpose.

ISSN: 2582-418X

Table 1. Manual Tracing of Sample Dataset

AGE	SEX	CHESTPAIN	RESTING ECG	EXERCISE INDUCED ANGINA (EXANG)	TARGET
50	0	1	1	0	1
60	1	2	0	0	0
60	0	2	1	0	1
50	1	0	0	1	0
60	1	0	0	1	1

3.3 Prior Probability for the Dataset

Target

Yes: 2/5

No: 3/5

AGE:	**Yes**	**No**	**p(yes)**	**p(no)**	**SEX:**	**Yes**	**No**	**p(yes)**	**p(no)**
50	1	1	1/2	1/3	0	2	0	2/2	0/3
60	2	1	2/2	1/3	1	1	2	1/2	2/3

CP:	**Yes**	**No**	**p(yes)**	**p(no)**	**EXANG:**	**Yes**	**No**	**p(yes)**	**p(no)**
0	1	1	1/2	1/3	0	2	1	2/2	1/3
1	1	0	1/2	0/3	1	1	1	1/2	1/3
2	1	1	1/2	1/3					

RESTING ECG:	**Yes**	**No**	**p(yes)**	**p(no)**
0	1	2	1/2	2/3
1	2	0	2/2	0/3

Mean And Standard Deviation for Each Class: For each class, the mean and standard deviation of each feature in the training set are calculated. For example, the mean age of patients who have experienced heart failure might be 65 years with a standard deviation 5, while the mean age of

patients who have not experienced heart failure might be 55 years with a standard deviation of 10.

Class Conditional Probability: The mean and standard deviation values calculated from step 3 are used to calculate class conditional probability of each feature in a class. For example, the class-conditional probability of a 60-year-old patient experiencing heart failure might be calculated as follows:

$$P(age = 60|HF) = \left(\frac{1}{\sqrt{(2\pi 5^2)}}\right) * e^{\left(-\frac{(60-65)^2}{2*5^2}\right)} = 0.12$$

$$P(age = 60|NoHF) = \left(\frac{1}{\sqrt{(2\pi 10^2)}}\right) * e^{\left(-\frac{(60-55)^2}{2*10^2}\right)} = 0.04$$

Posterior Probability: Bayes' theorem is used to calculate the posterior probability for each class, which is the probability that an instance belongs to a given class given its feature values. For example, the posterior probability of a 60-year-old patient experiencing heart failure is as follows

$$P(HF|age = 60) = \frac{P(age = 60|HF) * P(HF)}{P(age = 60|HF)P(HF) + P(age = 60|NoHF) * P(NoHF)} = \frac{0.12 * 0.2}{0.12 * 0.2 + 0.04 * 0.08} = 0.429$$

Make Predictions: The posterior probabilities calculated in step 5 are used to make predictions for the instances in the test set. For example, if the posterior probability of a patient in the test set having heart failure is greater than 0.5, then it is predicted that the patient has a heart failure.

Below are the steps involved in manual tracking of gaussian Naïve Bayes algorithm with respect to small sample of the dataset.

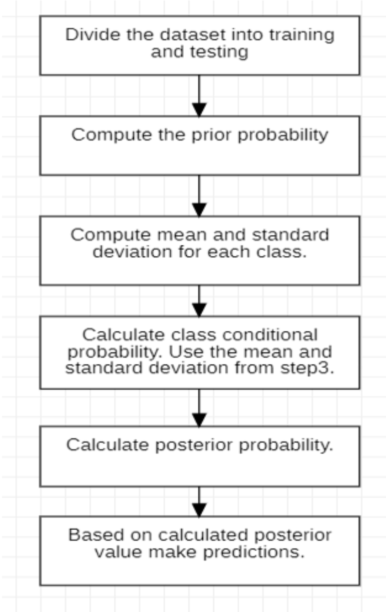


Figure 1. Flow of the proposed design

3.4 Dataset Description

Integrating different datasets, which had been available separately but never integrated together, was used to create this database. For research purposes, this dataset consists of five heart datasets comprising of more than 11 common characteristics and is the largest data base for cardiovascular disease to date. [16]

The description of the dataset is as follows;

- ✓ Age: in years.
- ✓ Sex: gender of the patient (0: Male, 1: Female)
- ✓ ChestPainType: specifies the type of the chest pain such as typical angina (0), atypical angina (1), non-anginal Pain (2), and asymptomatic (3)]
- ✓ RestingBP: blood pressure at rest
- ✓ Blood cholesterol: serum cholesterol [mm/dl]
- ✓ Fasting blood sugar (Fasting BloodSugar): 1 if Fasting BloodSugar is greater than normal, 0 otherwise.
- ✓ Resting ECG: resting electrocardiogram result [Normal: Normal, ST: having ST-T wave abnormality]

- ✓ MaxHR: maximum heart rate achieved [between 60 and 202]
- ✓ ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
- ✓ Oldpeak: ST [Numeric value measured in depression]
- ✓ ST_Slope: the slope of the peak exercise [Up: upsloping, Flat: flat, Down: downsloping]
- ✓ HeartFailure: output [1: heart disease, 0: Normal]

The dataset for Heart Failure Prediction is shown in Table2.

Table 2. Dataset for Heart Failure

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
42	1	0	140	226	0	1	178	0	0	2	0	2	1
61	1	2	150	243	1	1	137	1	1	1	0	2	1

4. Results and Discussion

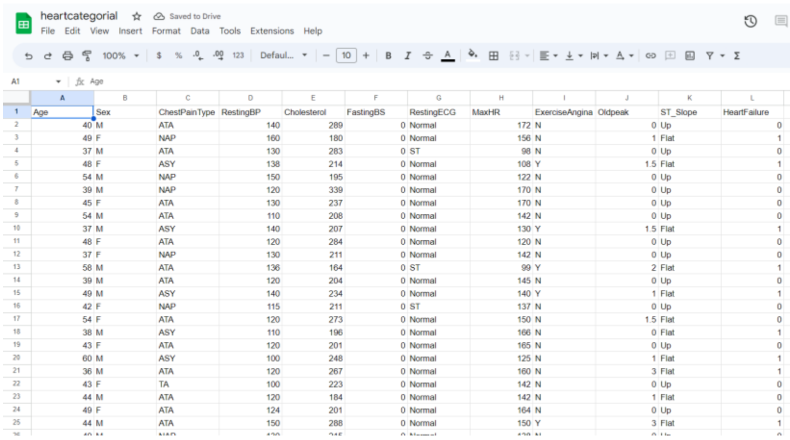
A) Python

Python was the platform used and NumPy, Pandas, Scikit learn and Matplot were the libraries used in data pre-processing, model fitting, making predictions and plotting the results. The results achieved for various heart failure dataset are shown below. Since it is categorial dataset One Hot Encoding technique is used to convert categorial data into numerical one.

Dataset 1:

A small portion of dataset 1 is shown in Table 3.

Table 3. Dataset 1 for Heart Failure Prediction



Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartFailure
40	M	ATA	140	289	0	Normal	172	N	0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
37	M	ATA	130	283	0	ST	98	N	0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0	Up	0
45	F	ATA	130	237	0	Normal	170	N	0	Up	0
54	M	ATA	110	208	0	Normal	142	N	0	Up	0
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
48	F	ATA	120	284	0	Normal	120	N	0	Up	0
37	F	NAP	130	211	0	Normal	142	N	0	Up	0
58	M	ATA	136	164	0	ST	99	Y	2	Flat	1
39	M	ATA	120	204	0	Normal	145	N	0	Up	0
49	M	ASY	140	234	0	Normal	140	Y	1	Flat	1
42	F	NAP	115	211	0	ST	137	N	0	Up	0
54	F	ATA	120	273	0	Normal	150	N	1.5	Flat	0
38	M	ASY	110	196	0	Normal	166	N	0	Flat	1
43	F	ATA	120	201	0	Normal	165	N	0	Up	0
60	M	ASY	100	248	0	Normal	125	N	1	Flat	1
36	M	ATA	120	267	0	Normal	160	N	3	Flat	1
43	F	TA	100	223	0	Normal	142	N	0	Up	0
44	M	ATA	120	184	0	Normal	142	N	1	Flat	0
49	F	ATA	124	201	0	Normal	164	N	0	Up	0
44	M	ATA	150	288	0	Normal	150	Y	3	Flat	1

Dimensions: 919 x 12

Results:

Performance metrics, confusion matrix and ROC curve for dataset 1 are shown in Figures 2, 3 and 4 respectively. The result indicates that the accuracy of the system is 0.86.

```
Training Accuracy: 0.8583106267029973
Training Confusion Matrix:
 [[279  54]
 [ 50 351]]
Training Precision: 0.8666666666666667
Training Recall: 0.8753117206982544
Testing Accuracy: 0.8641304347826086
Testing Confusion Matrix:
 [[70  7]
 [18 89]]
Testing Precision: 0.9270833333333334
Testing Recall: 0.8317757009345794
```

Figure 2. Performance Metrics for Dataset 1

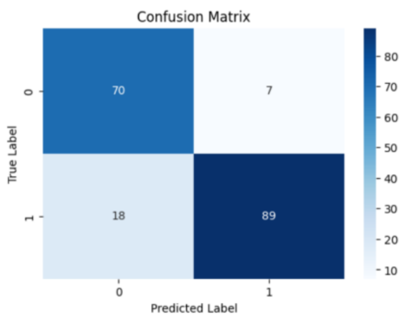


Figure 3. Confusion Matrix for Dataset 1

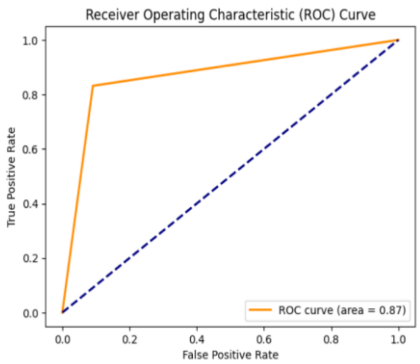


Figure 4. ROC Curve for Dataset 1

Dataset 2

A small portion dataset 2 is shown in table 4. The dataset consists of features such as age, anaemia, creatinine phosphate, ejection fraction, high blood pressure, platelets, serum creatinine, gender, and smoking. These features play a major role in predicting heart failure. It consists of 300 records with 13 features.

Table 4. Dataset 2 for Heart Failure Prediction

heart_failure_clinical_records_dataset (2)

age	anaemia	creatinine_phos	diabetes	ejection_fraction	high_blood_pres	platelets	serum_creatinin	serum_sodium	sex	smoking	time	HEART_FAILURE
75	0	582	0	20	1	265000	1.9	130	1	0	4	1
55	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
65	0	146	0	20	0	162000	1.3	129	1	1	7	1
50	1	111	0	20	0	210000	1.9	137	1	0	7	1
65	1	160	1	20	0	327000	2.7	116	0	0	8	1
90	1	47	0	40	1	204000	2.1	132	1	1	8	1
75	1	246	0	15	0	127000	1.2	137	1	0	10	1
60	1	315	1	60	0	454000	1.1	131	1	1	10	1
65	0	157	0	65	0	263358.03	1.5	138	0	0	10	1
80	1	123	0	35	1	388000	9.4	133	1	1	10	1
75	1	81	0	38	1	368000	4	131	1	1	10	1
62	0	231	0	25	1	253000	0.9	140	1	1	10	1
45	1	981	0	30	0	136000	1.1	137	1	0	11	1
50	1	168	0	38	1	276000	1.1	137	1	0	11	1
49	1	80	0	30	1	427000	1	138	0	0	12	0
82	1	379	0	50	0	47000	1.3	136	1	0	13	1
87	1	149	0	38	0	262000	0.9	140	1	0	14	1
45	0	582	0	14	0	166000	0.8	127	1	0	14	1
70	1	125	0	25	1	237000	1	140	0	0	15	1
48	1	582	1	55	0	87000	1.9	121	0	0	15	1
65	1	52	0	25	1	276000	1.3	137	0	0	16	0
65	1	128	1	30	1	297000	1.6	136	0	0	20	1
68	1	220	0	35	1	289000	0.9	140	1	1	20	1
53	0	63	1	60	0	368000	0.8	135	1	0	22	0

Dimensions: 300 x 13

Results:

Performance metrics, confusion matrix and ROC curve for dataset 2 are shown in Figures 5, 6 and 7 respectively. The result indicates that the accuracy of the system is 0.7.

```
Training Accuracy: 0.8117154811715481
Training Confusion Matrix:
 [[154  14]
 [ 31  40]]
Training Precision: 0.7407407407407407
Training Recall: 0.5633802816901409
Testing Accuracy: 0.7
Testing Confusion Matrix:
 [[33  2]
 [16  9]]
Testing Precision: 0.8181818181818182
Testing Recall: 0.36
```

Figure 5. Performance Metrics for Dataset 2

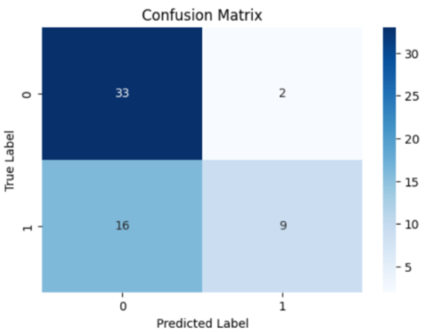


Figure 6. Confusion Matrix for Dataset 2

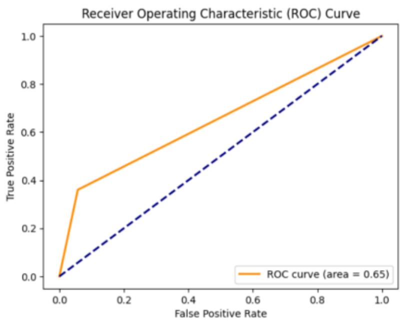


Figure 7. ROC Curve for Dataset 2

Dataset 3

A small portion of dataset 3 is shown in table 5. It consists of the same attributes present in dataset 1. The number of records present in this dataset is 304, and it has 14 features.

Table 5. Dataset 3 for Heart Failure Prediction

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
42	1	0	140	226	0	1	178	0	0	2	0	2	1
61	1	2	150	243	1	1	137	1	1	1	0	2	1

Dimensions: 304 x 14

Results:

Performance metrics, confusion matrix and ROC curve for dataset 3 are shown in Figures 8, 9 and 10 respectively. The result indicates that the accuracy of the system is 0.87. The ROC curve in figure 10 is similar to that in figure 4. This indicates that both dataset 1 and dataset 3 have similar accuracy rate. However, the precision score for dataset 1 is lesser than that of dataset 3.

```
Training Accuracy: 0.8181818181818182
Training Confusion Matrix:
 [[ 84  25]
 [ 19 114]]
Training Precision: 0.8201438848920863
Training Recall: 0.8571428571428571
Testing Accuracy: 0.8688524590163934
Testing Confusion Matrix:
 [[26  3]
 [ 5 27]]
Testing Precision: 0.9
Testing Recall: 0.84375
```

Figure 8. Performance Metrics for Dataset 3

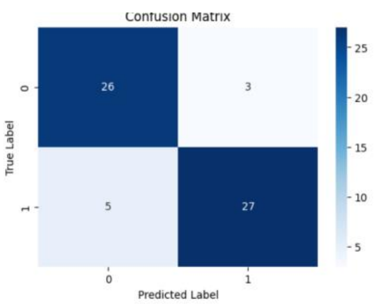


Figure 9. Confusion Matrix for Dataset 3

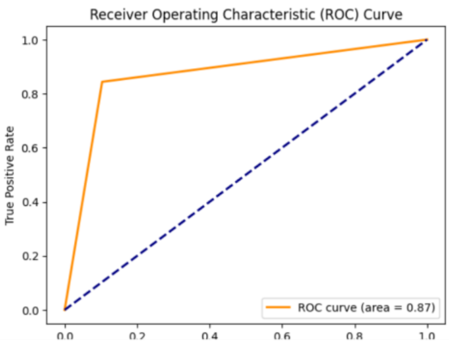
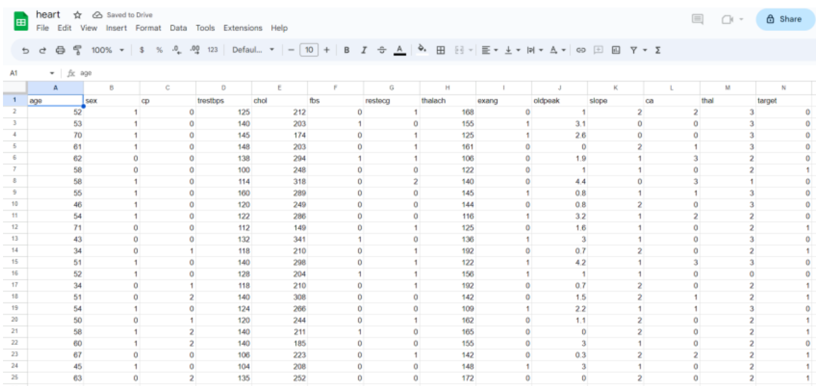


Figure 10. ROC Curve for Dataset 3

Dataset 4

A small portion of dataset 4 is shown in table 6. The dataset consists of features that are similar to that of dataset 1. The dataset 4 consists of 1026 records with 14 features.

Table 6. Dataset 4 for Heart Failure Prediction



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
52	1	0	128	204	1	1	156	1	1	1	0	0	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1
60	1	2	140	185	0	0	155	0	3	1	0	2	0
67	0	0	106	223	0	1	142	0	0.3	2	2	2	1
45	1	0	104	208	0	0	148	1	3	1	0	2	1
63	0	2	135	252	0	0	172	0	0	2	0	2	1

Dimensions: 1026 x 14

Results:

Performance metrics, confusion matrix and ROC curve for dataset 4 are shown in Figures 11, 12 and 13 respectively. The accuracy of the system is 0.8.

```
Training Accuracy: 0.8390243902439024
Training Confusion Matrix:
 [[320  77]
 [ 55 368]]
Training Precision: 0.8269662921348314
Training Recall: 0.8699763593380615
Testing Accuracy: 0.8
Testing Confusion Matrix:
 [[72 30]
 [11 92]]
Testing Precision: 0.7540983606557377
Testing Recall: 0.8932038834951457
```

Figure 11. Performance Metrics for Dataset 4

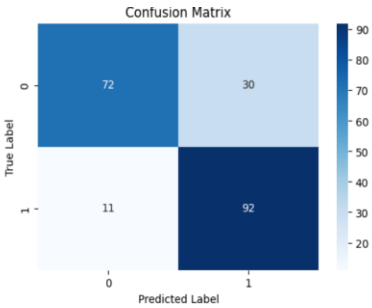


Figure 12. Confusion Matrix for Dataset 4

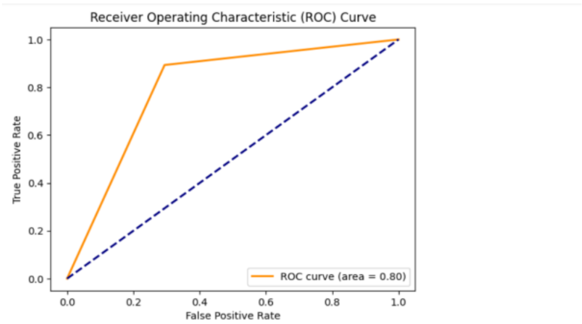


Figure 13. ROC Curve for Dataset 4

Performance Comparison with Previous Studies:

The proposed GNB algorithm for heart failure prediction is compared with the existing implementations for heart failure prediction in table 7.

Table 7. Performance Comparison with Existing Systems

Sl.NO	EXISTING SYSTEM	ACCURACY	PROPOSED SYSTEM	ACCURACY
1.	SVM	84.85	GAUSSIAN NAÏVE BAYES	87
2.	LOGISTIC REGRESSION	82.56		
3.	DECISION TREE	82.22		
4.	RANDOM FOREST	84.17		

B. Discussion

From the above results it is observed that, dataset 1 and dataset 3 have achieved good accuracy rate, precision, and recall scores. The main reason behind this is the quality of the dataset which helps the algorithm to predict accurately during testing stage. It is observed that the training and testing accuracy is moreover same. This indicates that the model doesn’t memorize. The accuracy for dataset 1 and dataset 3 is 87%.

5. Conclusion and Future Works

It is observed that Gaussian Naïve Bayes gets a better accuracy with clinical dataset for the prediction of heart disease. Thus, Gaussian Naïve Bayes algorithm is effective in predicting the diseases. The future works include, investigating which features have the most significant impact on heart failure prediction and identify new features that may improve model accuracy and exploring methods for optimizing the model's hyperparameters to improve its accuracy and generalizability. Techniques such as cross-validation and grid search could be used to determine the optimal values for the smoothing parameter and other GNB-specific parameters.

References

[1] Jordan, M.I. and Mitchell, T.M.,” Machine learning: Trends, perspectives, and prospects”, *Science*, Vol. No *349*, Issue No.6245, pp.255-260, year 2015.

[2] Mahesh, B., "Machine learning algorithms-a review.” *International Journal of Science and Research (IJSR).[Internet]*, Vol. No *9*, pp.381-386, year 2020.

[3] Osisanwo, F.Y., Akinsola, J.E.T., Awodele, O., Hinmikaiye, J.O., Olakanmi, O. and Akinjobi, J., “Supervised machine learning algorithms: classification and comparison”. *International Journal of Computer Trends and Technology (IJCTT)*, Vol. No *48, Issue No* 3, pp.128-138, year 2017.

[4] Tafa, Zhilbert, “Concurrent implementation of supervised learning algorithms in disease detection.” Journal of Advances in Information Technology Vol. No 7, Issue No 2, year 2016.

[5] Josephine, P.K., Prakash, V.S. and Divya, K.S., “Supervised Learning Algorithms: A Comparison”. Kristu Jayanti Journal of Computational Sciences (KJCS), pp.01-12, year 2021.

[6] Uddin, S., Khan, A., Hossain, M.E. and Moni, M.A.," Comparing different supervised machine learning algorithms for disease prediction". BMC medical informatics and decision making,Vol. No 19, Issue No 1, pp.1-16, year 2019.

[7] Adler, E.D., Voors, A.A., Klein, L., Macheret, F., Braun, O.O., Urey, M.A., Zhu, W., Sama, I., Tadel, M., Campagnari, C. and Greenberg, B., "Improving risk prediction in heart failure using machine learning. *European journal of heart failure"*, Vol. No *22, Issue No* 1, pp.139-147, year 2020.

[8] Alotaibi, Fahd Saleh," Implementation of machine learning model to predict heart failure disease". *International Journal of Advanced Computer Science and Applications*, Vol. No *10, Issue No* 6, year 2019.

[9] Olsen, C.R., Mentz, R.J., Anstrom, K.J., Page, D. and Patel, P.A.," Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure". *American Heart Journal*, Vol. No*229*, pp.1-17, year 2020.

[10] Chicco, D. and Jurman, G., "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone." BMC medical informatics and decision making,Vol No. 20,Issue No. 1, pp.1-16, year 2020.

[11] Yang, X., Gong, Y., Waheed, N., March, K., Bian, J., Hogan, W.R. and Wu, Y.," Identifying cancer patients at risk for heart failure using machine learning methods." American Medical Informatics Association, Vol. 2019, pp. 933, year 2019.

[12] Hoekstra, T., Jaarsma, T., van Veldhuisen, D.J., Hillege, H.L., Sanderman, R. and Lesman-Leegte, I., "Quality of life and survival in patients with heart failure". *European journal of heart failure*, Vol. No *15, Issue No* 1, pp.94-102, year 2013.

[13] Rahimi, K., Bennett, D., Conrad, N., Williams, T.M., Basu, J., Dwight, J., Woodward, M., Patel, A., McMurray, J. and MacMahon, S.," Risk prediction in patients with heart failure: a systematic review and analysis." *JACC: Heart Failure*, Vol. No *2, Issue No* 5, pp.440-446, year 2014.

[14] Guidi, G., Pettenati, M.C., Melillo, P. and Iadanza, E., “A machine learning system to improve heart failure patient assistance”. IEEE journal of biomedical and health informatics, Vol. No 18, Issue No 6, pp.1750-1756, year 2014.

[15] Jing, L., Ulloa Cerna, A.E., Good, C.W., Sauers, N.M., Schneider, G., Hartzel, D.N., Leader, J.B., Kirchner, H.L., Hu, Y., Riviello, D.M. and Stough, J.V.,” A machine learning approach to management of heart failure populations”. Heart Failure, Vol. No 8, Issue No 7, pp.578-587, year 2020.

[16] https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction