International Journal of Computing and Artificial Intelligence

E-ISSN: 2707-658X
P-ISSN: 2707-6571
IJCAI 2021; 2(2): 17-22
Received: 10-05-2021
Accepted: 12-06-2021

Akhil R
Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

Boyella Mala Konda Reddy
Assistant Professor, Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

Corresponding Author:
Akhil R
Department of Computer Science, SDHR College, Tirupati, Andhra Pradesh, India

An efficient heart disease detection system utilizing naive bayes classification

Akhil R and Boyella Mala Konda Reddy

DOI: https://doi.org/10.33545/27076571.2021.v2.i2a.33

Abstract

Coronary illness is perhaps the most basic human sicknesses on the planet and influences human existence severely. Heart related sicknesses or cardiovascular diseases (CVDs) are the principal justification countless passing on the planet in the course of the most recent couple of many years and has arisen as the most perilous illness, in India as well as in the entire world. In coronary illness, the heart can't push the necessary measure of blood to different pieces of the body. Precise and on time analysis of coronary illness is significant for cardiovascular breakdown avoidance and treatment. The conclusion of coronary illness through conventional clinical history has been considered as not dependable in numerous angles. In this way, there is a need of solid, precise and achievable framework to analyze such sicknesses on schedule for appropriate therapy. The proposed Naive Bayes characterization framework can undoubtedly recognize and order individuals with coronary illness from sound individuals. The proposed Naive Bayes characterization-based choice emotionally supportive network will help the specialists to determination heart patients proficiently. In this paper we thought about Classification Rule Mining for information revelation and produced the guidelines by applying our created approach on Heart expire data sets. Our proposed model has accomplished 81.48% precision.

Keywords: heart illness, naive bayes, classification, data mining and ml

1. Introduction

We see lately different clinical associations are delivering huge measures of information which are hard to deal with. Clinics have gathered huge amounts of data about patients and their clinical accounts. Information digging is looking for connections and examples that could give valuable information to viable dynamic. Clinical information mining is one of the main points of contention to get valuable clinical information from clinical data sets.

This is the mother justification some connected clinical issues like heart attack, liver disillusionment, kidney frustrations, nerves damages and vision mishap. One of the significant genuine clinical issues is the location of diabetes at its beginning phase.

Heart is the most key organ in human body assuming that organ gets affected, it moreover impacts the other key pieces of the body. Thusly it is crucial for people to go for a coronary sickness investigation [1].

The main organ of the human body is heart. The capacity of the heart is to siphon the blood and circles whole body [3]. The coronary illness (HD) has been considered as one of the complex and life deadliest human sicknesses on the planet. In this sickness, generally the heart can't push the necessary measure of blood to different pieces of the body to satisfy the ordinary functionalities of the body, and because of this, eventually the cardiovascular breakdown happens. As indicated by the World Health Organization (WHO), an expected 17 million individuals bite the dust every year from cardiovascular illness, especially coronary failures and strokes [9].

The side effects of coronary illness incorporate windedness, shortcoming of actual body, swollen feet, and weariness with related signs, for instance, raised jugular venous pressing factor and fringe edema brought about by useful heart or non-cardiac irregularities [8]. The examination methods in beginning phases used to distinguish coronary illness were convoluted, and its subsequent intricacy is one of the significant reasons that influence the norm of life [8].

The coronary illness determination and treatment are extremely intricate, particularly in the agricultural nations, because of the uncommon accessibility of symptomatic mechanical assembly and lack of doctors and others assets which influence legitimate expectation and treatment the coronary illness hazard in patients is vital for decreasing their related dangers of serious heart issues and improving security of heart [9].

1.1 Objective of this examination

The primary target of this paper is the expectation coronary illness utilizing Naive Bayes Classification calculation. The objective is to separate the secret examples by applying information mining methods on the dataset, which are essential to heart illnesses and to foresee the presence of coronary illness in patients where the presence is esteemed on a scale. The forecast of coronary illness requires a gigantic size of information which is excessively perplexing and monstrous to measure and break down by ordinary methods. In this paper we are anticipating the coronary illness event in a patient dependent on some significant attributes which are most appropriate dependent on our informational index that we have gathered.

2. Classification cycle

Preposterous decade there has been an increment in the work done on applying AI calculations to the clinical area. Arrangement is a champion among the most examined issues in AI and data mining [2]. Expecting the consequence of a contamination is a champion among the most interesting and inciting tasks in which to make data mining applications.

Portrayal is the route toward learning the target limit that maps between a great deal of features and predefined class marks. The data for the gathering is a ton of events. Every event is a record of data as (X, Y) where X is the features set and Y is the goal variable.

Grouping of this colossal measure of information is tedious and uses extreme computational exertion, which may not be suitable for some applications. The order of clinical information has become an inexorably difficult issue, because of late advances in clinical mining innovation. Arrangement focuses on to characterizing a theoretical model of a bunch of classes, called classifier, which is worked from a bunch of marked information, the preparation set. The classifier is then used to properly group new information for which the class mark is obscure [5]. Building precise and productive classifiers for Medical information bases is one of the fundamental errands of information mining and AI research. Building successful characterization frameworks is one of the focal assignments of information mining.

The readiness data involves events whose class names are known. The game plan model can be developed ward on the readiness data. The model by then can be surveyed and attempted by using the testing data which contains records with dark class marks.

2.1 Arrangement is a two-phase measure

i) **Model turn of events:** Portraying a ton of destined classes. Each tuple is acknowledged to have a spot with a predefined class, as constrained by the class mark attribute. The course of action of tuples used for model turn of events: getting ready set. The model is addressed as gathering rules, decision trees, or logical formulae.

ii) **Model usage:** For requesting future or dark articles. It checks exactness of the model; the known name of test is differentiated and the described result from the model. Precision rate is the degree of test set models that are adequately requested by the model. Test set is liberated from planning set, for the most part over-fitting will occur.

3. Methodology: Naive Bayes (NB) Classification

The Naive Bayes is an energetic procedure for arrangement of quantifiable farsighted models. NB relies upon the Bayesian speculation [2, 4, 7]. This computation uses class prohibitive independence and has ability to adjust quickly. This portrayal technique assessments the association between every property and the class for every guide to decide a prohibitive probability for the associations between the trademark characteristics and the class. In the midst of setting up, the probability of each class is enrolled by checking how regularly it occurs in the arrangement dataset. This is known as the "prior probability" P(C=c). Despite the prior probability, the computation furthermore enlists the probability for the event x given c with the assumption that the characteristics are independent. This probability transforms into the aftereffect of the probabilities of each single quality. The probabilities would then have the option to be evaluated from the frequencies of the events in the planning set.

3.1 Bayesian Theorem

Given training data X, posterior probability of a hypothesis H, P(H|X), follows the Bayes theorem P(H|X) = $\frac{P(X|H)P(H)}{P(X)}$

Let X be data tuple and H be some hypothesis such that the data tuple X belongs to a specified class C. For classification problems, we want to determine P (H|X), the probability that the hypothesis H holds the given evidence or observed data tuple X.

P (H|X) is the posterior probability of H conditioned on X

P (H) is the prior probability of H

P (X|H) is the posterior probability of X conditioned on H

P(X) is prior probability of X

4. Experimental Results

The assessments have been coordinated by using Python programming language. It is an open-source programming language give stunning utilization of different data examination and Visualization methodologies. It is an earth-shattering library that gives numerous AI gathering estimations, capable mechanical assemblies for data mining and data assessment. The Python Scikit-learn is a pack for data request, backslide, bundling and portrayal. We have considered the Heart Disease information from UCI Machine Learning Repository datasets [10]. This Data set has 270 lines and 13 segments. So, in this information there are two class names i.e., the missing class has 150 and Present class has 120. The property data information is dense in Table-1. The standard dataset is parceled into two sets (70% and 30%), one for getting ready and another set for testing.

Table 1: Provides the attribute information of Heart Disease data

Attribute ID	Attribute Definition
age	Age
sex	Sex
chest	Chest Pain Type
resting_blood_pressure	Resting Blood Pressure
serum_cholestoral	Serum Cholesterol in mg/dl
fasting_blood_sugar	Fasting Blood Sugar
resting_electrocardiographic_results	Resting electrocardiographic result
maximum_heart_rate_achieved	Maximum heart rate achieved
exercise_induced_angina	Exercised-induced angina
oldpeak	Old peak
slope	Slope
number_of_major_vessels	Number of major vessels
thal	Thal
class	Class label: absent, present

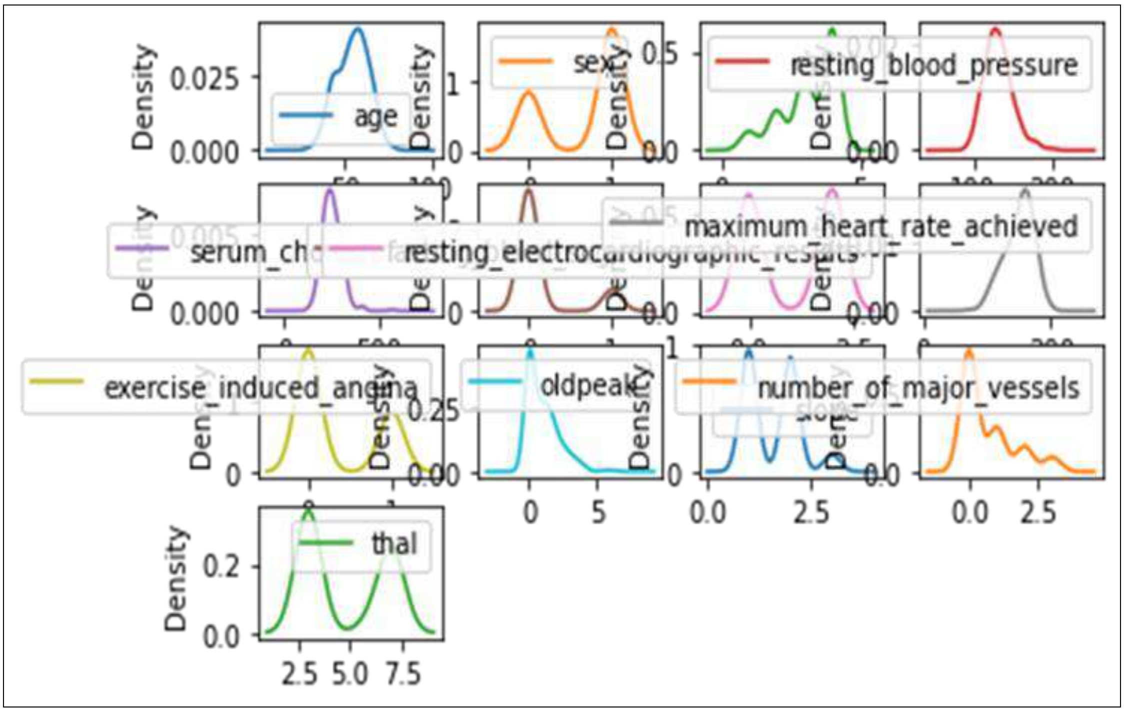


Fig 1: Density plot for data visualization

4.1 Measures for performance evaluation

There exist various measures that can be utilized to assess the exhibition of a classifier, like Accuracy, affectability, explicitness, exactness and review, and so on Every one of these assessment measures have their own constraints and, subsequently, a fitting assessment measure which best suits the issue ought to be chosen. Because of the elements referenced, and to have a solid exhibition assessment, the evaluation ought to be viewed as dependent on the Cross approval.

To limit the predisposition related with the arbitrary testing of the preparation and holdout information tests in looking at the prescient exactness of at least two strategies, analysts will in general utilize k-overlap cross-approval.

Execution of every classifier is measure regarding disarray lattice, affectability, particularity, exactness, review and precision. These measurements are customarily characterized for a parallel order task with positive and negative classes. That is:

Exactness: Accuracy is an action which decides the likelihood that how much outcomes are precisely grouped.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

4.2 Precision: Precision represents how precise the classifier predictions are since it shows the number of true positives that were predicted out of all positive labels assigned to the instances by the classifier. Precision is the proportion of positive predictions that are correct

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

4.3 Recall: Recall is the proportion of positive samples that are correctly predicted positive. It shows the amount of truly predicted positive classes out of the amount of total actual positive classes.

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

Where,

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

Table 2: Confusion Matrix of Prediction cases of classification

		Predicted	
		Positive	**Negative**
Actual Class	Positive	TP	FN
	Negative	FP	TN

Disarray grid is a representation device which is generally used to introduce the exactness of the classifiers in arrangement that help with execution assessment purposes which comprise of the ideas characterized above estimations. This is shown in table-2. It is utilized to show the connections among results and anticipated classes. The degree of viability of the grouping model is determined with the quantity of right and off base arrangement in every conceivable worth of the variable being ordered in the disarray framework.

4.2. Results

To approve the expectation consequences of the Naïve Bayes arrangement and the 10-overlay hybrid approval is utilized. The k-overlap hybrid approval is generally used to diminish the mistake came about because of irregular examining in the examination of the correctness's of various forecast models. The current investigation partitioned the information into 10 folds where 1 overlap was for trying and 9 folds were for preparing for the 10-overlay hybrid approval.

The disarray network of Naïve Bayes order technique is introduced in the table-3. The qualities to quantify the exhibition of the strategies (for example exactness, accuracy, review, and f1-score) are gotten from the disarray grid and appeared in table-4 and same appeared in graphical portrayal in figure-2

Table 3: Confusion Matrix of Heart Disease data classification

Testing Data (81)		
Desired Result	**Output Result**	
	Absent	**Present**
Absent	36	6
Present	9	30

Table 4: Results of Heart Disease Proposed Naïve Bayes Classification

Accuracy	Precision	Recall	f1-score
81.48	82	81	81

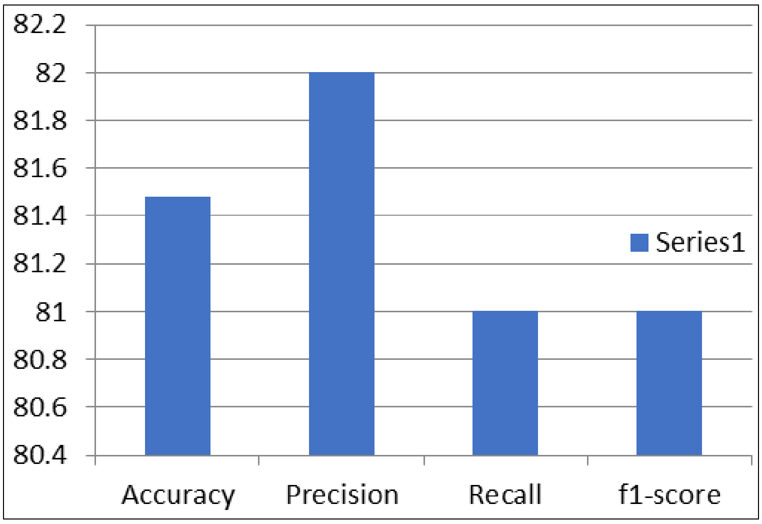
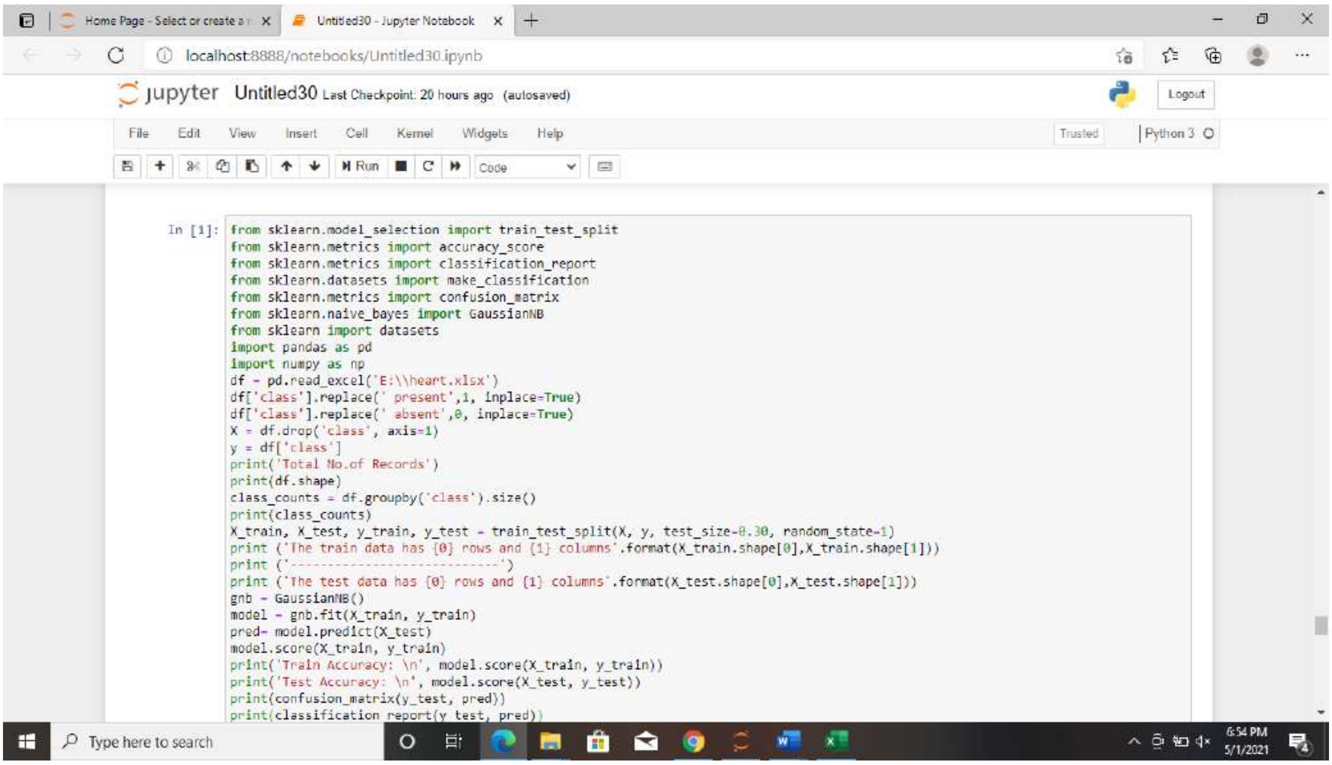


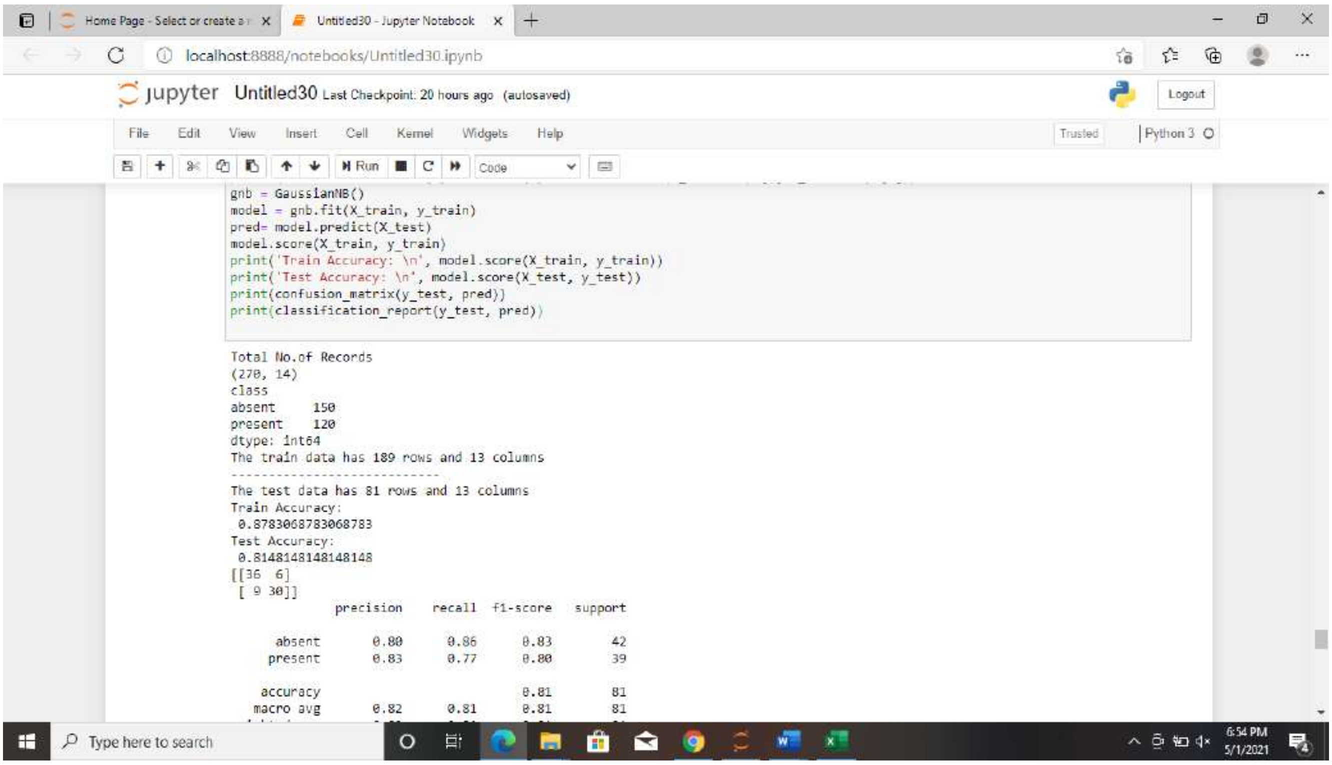
Fig 2: Performance metrics of Heart Disease data

We observe in figure-2 The Naïve Bayes classifier algorithm gives significant improvement in the accuracy. It is accomplished accuracy of 81.48%, precision got 82% and recall is achieved 81%.

4.3 Screen shots


Home Page - Select or create a
Untitled30 - Jupyter Notebook
localhost:8888/notebooks/Untitled30.ipynb
jupyter Untitled30 Last Checkpoint: 20 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Type here to search
6:54 PM
5/1/2021

```
In [1]: from sklearn.model_selection import train_test_split
        from sklearn.metrics import accuracy_score
        from sklearn.metrics import classification_report
        from sklearn.datasets import make_classification
        from sklearn.metrics import confusion_matrix
        from sklearn.naive_bayes import GaussianNB
        from sklearn import datasets
        import pandas as pd
        import numpy as np
        df = pd.read_excel('E:\\heart.xlsx')
        df['class'].replace(' present',1, inplace=True)
        df['class'].replace(' absent',0, inplace=True)
        X = df.drop('class', axis=1)
        y = df['class']
        print('Total No.of Records')
        print(df.shape)
        class_counts = df.groupby('class').size()
        print(class_counts)
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=1)
        print ('The train data has {0} rows and {1} columns'.format(X_train.shape[0],X_train.shape[1]))
        print ('---------------------------')
        print ('The test data has {0} rows and {1} columns'.format(X_test.shape[0],X_test.shape[1]))
        gnb = GaussianNB()
        model = gnb.fit(X_train, y_train)
        pred= model.predict(X_test)
        model.score(X_train, y_train)
        print('Train Accuracy: \n', model.score(X_train, y_train))
        print('Test Accuracy: \n', model.score(X_test, y_test))
        print(confusion_matrix(y_test, pred))
        print(classification_report(y_test, pred))
```


Home Page - Select or create a
Untitled30 - Jupyter Notebook
localhost:8888/notebooks/Untitled30.ipynb
jupyter Untitled30 Last Checkpoint: 20 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Type here to search
6:54 PM
5/1/2021

```
gnb = GaussianNB()
model = gnb.fit(X_train, y_train)
pred= model.predict(X_test)
model.score(X_train, y_train)
print('Train Accuracy: \n', model.score(X_train, y_train))
print('Test Accuracy: \n', model.score(X_test, y_test))
print(confusion_matrix(y_test, pred))
print(classification_report(y_test, pred))
```

```
Total No.of Records
(270, 14)
class
absent     150
present    120
dtype: int64
The train data has 189 rows and 13 columns
---------------------------
The test data has 81 rows and 13 columns
Train Accuracy:
 0.8783068783068783
Test Accuracy:
 0.8148148148148148
[[36  6]
 [ 9 30]]
              precision    recall  f1-score   support

      absent       0.80      0.86      0.83        42
     present       0.83      0.77      0.80        39

    accuracy                           0.81        81
   macro avg       0.82      0.81      0.81        81
```

5. Conclusion

In this paper, Naive Bayes grouping of Data Mining has been talked about that can be utilized for anticipate the precision of Heart illness information. The exactness or expectation pace of Naive Bayes is 81.48%. Choice Support in Heart Disease Prediction System is created utilizing Naive Bayesian Classification method. The framework removes concealed information from a verifiable coronary illness data set. This is the best model to foresee patients with coronary illness. Consequently, proposed Naive Bayes Classifier approach will yield a viable technique for both forecast and location.

6. References

1. Blake CL, Mertz CJ. UCI Machine Learning Databases, http://mlearn.ics.uci.edu/databases/heartdisease/ 2004.
2. Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann 2005.
3. HeonGyu Lee, Ki Yong Noh, Keun Ho Ryu. Mining Biosignal Data: Coronary Artery Disease Diagnosis

using Linear and Nonlinear Features of HRV, LNAI 4819: Emerging Technologies in Knowledge Discovery and Data Mining 2007, 56-66.

4. Ho TJ. Data Mining and Data Warehousing, Prentice Hall 2005.
5. Han J, Kamber M. Data Mining concepts and Techniques, the Morgan Kaufmann series in Data Management Systems, 2nd ed. San Mateo, CA; Morgan Kaufmann 2006.
6. Michael N. Artificial Intelligence – A Guide to Intelligent Systems, 2nd Edition, Addison Wesley 2005.
7. Tan PN, Steinbach M, Kumar V. Introduction to Data Mining, A: Addision-Wesley 2005.
8. Sitar-Taut VA *et al*. Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics 2009.
9. The Atlas of Heart Disease and Stroke, [online]. http://www.who.int/cardiovascular_diseases/resources/atlas/en/
10. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/.