International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217399

Prediction of Heart Disease and Diabetes Using Naive Bayes Algorithm

Ninad Marathe[1], Sushopti Gawade[2], Adarsh Kanekar[1]
[1]Information Technology, Pillai College of Engineering, Panvel, Maharashtra, India
[2]Computer Engineering, Pillai College of Engineering, Panvel, Maharashtra, India

Article Info
Volume 7, Issue 3
Page Number: 447-453
Publication Issue :
May-June-2021
Article History
Accepted : 25 May 2021
Published : 31 May 2021

ABSTRACT

Based on the test report values, diagnose a potential problem. The patient's report can be entered as feedback by the doctors (Sugar level, Age, Blood pressure, etc.). Through evaluating the available data collection, we can predict whether the patient has heart disease or diabetes using the method. Apart from that, we use Rstudio's R shiny addon for Web UI design. As a coding language, we use the R programming language. The Rstudio IDE was used. The datasets were obtained from the University of California at Irvine's repository.

Keywords : R Shiny, Naive Bayes, Sqldf, Cognitive Approach.

I. INTRODUCTION

The test report input from the user interface is captured and stored in a table. We've built a table with all of the records from the current data collection. To construct a model, we used the Naive Bayes algorithm. We are predicting whether the patient will develop heart disease or diabetes based on the model and test results. Age, chest paint form, resting blood pressure, cholesterol, fasting blood sugar, maximum heart rate, the slope of peak exercise, and other factors will be considered in the patient's study. The person's report is available in R shiny and can also be viewed in a browser.

Algorithm Details

Naive Bayes:

To obtain results, Naive Bayes is a simple supervised machine learning algorithm that employs the Bayes' theorem and strong independence assumptions between features. That is, the algorithm simply assumes that each input variable is unrelated to the others. Making such a naive statement regarding real-world data is truly naive It is widely used in text-sharing operations that require a large database of training. Naive Bayes Classifier is a simple and effective programming algorithm that helps to build fast learning models capable of making fast predictions. Possible separator, which means that we make predictions based on the inclination of the object. Spam filtering, emotional analysis, and article classification are all examples of the Naive Bayes Algorithm.

The Naive Bayes algorithm is made up of two names: Naive and Bayes, which can be defined as Naive: It is called Nave because it believes that the appearance of one element is not related to the appearance of others. For example, when the fruit is identified by its color, shape, and taste, the red, round, and sugary fruit is identified as an apple. As a result, each function contributes to the identification of an apple without the dependence of the other. It is called Bayes because it is based on the concept of Bayes Theorem.

Bayes Theorem: The Bayes' theorem, also known as the Bayes' Rule or Bayes' rule, is a way of calculating the possibilities of thinking based on prior knowledge. Conditional opportunities determine this. The Bayes theorem formula is as follows:

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

P(A|B) is the product of P(A|B) and P(A|B). Probability in the future: On the observed case B, the probability of hypothesis A.

P(B|A) = P(B|A) = P(B|A) = Probability of Likelihood: Provided that the likelihood of a hypothesis is valid, the probability of the proof.

Prior Probability (P(A)) is the probability of a hypothesis before seeing the proof.

P(B) stands for Probability of Evidence Marginal Probability.

II. Literature Review

2.1 Heart disease prognosis using the Naive Bayes Classifier.

The term “heart disease” refers to cardiovascular problems. Cardiovascular disease is one of the most common flying diseases in the world. High blood pressure, body inertia, hypertension, high blood pressure, cholesterol, family history of heart disease, and smoking all contribute to heart disease. The main purpose of the Phase Algorithm is to guess the target phase by analyzing the training database [1]. Data mining techniques have been developed to conduct convincing medical data analysis to assist clinicians in strengthening their treatment conclusions. Data mining techniques have played a major role in the study of heart disease. The main purpose of the Phase Algorithm is to predict a focused phase using a training database [1]. With convincing medical data research, data mining techniques have been developed to assist clinicians in improving their care conclusions. In heart disease research, data mining methods have been very helpful.

.

2.2 Heart disease prediction program based on the hidden naïve Bayes divider

Heart disease is the leading cause of death in the world. It takes a long time to diagnose heart disease. A system is needed to support wise decisions in predicting disease. Using data analysis methods, patients are often classified as normal or have heart disease. Hidden Naive Bayes is a mining model that alleviates the freedom of conditional thinking of the Naive Bayes model. According to our proposed model, the Hidden Naive Bayes (HNB) model can be used to diagnose heart disease (forecasting). According to our test findings on cardiovascular outcomes, HNB reports 100% accuracy and surpasses the Naive Bayes.

2.3 Prediction of Heart Disease using Classification Algorithms

Heart disease is still the leading cause of death worldwide, including in South Africa, and early detection could help avoid attacks [1]. Medical practitioners produce data that contains a wealth of secret knowledge that isn't being properly used for prediction [1]. For this reason, the research

transforms the unused data into a dataset that can be used in simulations using a variety of data mining techniques. People are killed as a result of signs that were overlooked. Health care providers must be able to predCharacteristics that make a heart attack more likely Smoking, a lack of physical activity, high blood pressure, high cholesterol, an unhealthy diet, harmful alcohol intake, and high blood sugar levels are all examples of assaults. [2][3][4]. Coronary heart disease, cerebrovascular disease (Stroke), hypertensive heart disease, congenital heart disease, and peripheral artery disease are all conditions that affect the heart. Cardiovascular disorder (CVD) includes rheumatic heart disease and inflammatory heart disease [3].

2.4 Cognitive approach on how to Understand Heart Disease Predict using Machine Learning

Predicting disease prevention and control patterns is a challenging and prominent challenge in the medical field. In this paper, we propose a machine learning framework to predict the likelihood of heart disease using various algorithms. The framework was developed using five algorithms Random Forest, Naïve Bayes, Support Vector Machine, Hoeffding Decision Tree, and Logistic Model Tree (LMT). The Cleveland database is used for training and model testing. The database is processed, followed by a feature selection to select the most prominent features. The database is followed and used for draft training. The results are combined and show that the random forest provides high accuracy.

2.5 Comparison of Diabetes Classifiers Predictability Classifiers

Diabetes mellitus, also known as diabetes mellitus (DM), is a group of metabolic disorders marked by high blood glucose levels. Excessive urination, chronic thirst, and high appetite are all symptoms of high blood sugar [1]. Diabetes, if not treated immediately, can lead to serious health problems. Ketoacidosis of diabetes, hyperosmolar hyperglycemic status, or death are all possible side effects. This can lead to long-term problems such as heart disease, stroke, kidney failure, foot ulcers, and eye problems, among others [2]. Diabetes begins when the pancreas fails to produce enough insulin or when synthetic insulin is not used properly by cells and tissues. There are three types of diabetes mellitus [3]: Type 1 diabetes is found in the insulin-producing pancreas, a condition is known as "insulin-subordinate diabetes Mell Diabetes-1 diabetes requires foreign insulin to compensate for pancreas" decreased production of Insulin. Type 2 diabetes mellitus is characterized by an insufficiency of insulin when cells of the body respond to insulin in a different way than they normally do. This can eventually lead to insulin resistance. non-insulin subordinate diabetes mellitus. "People with high BMI or an active lifestyle are more likely to develop this type of diabetes. The third major risk factor during pregnancy is diabetes - glucose levels in the average person range from 70 to 99 milliliters per deciliter. one by one.

III. Methodology

The aim of this study is to find the most effective DM technique with the highest precision among the various classification techniques that will be used. Furthermore, determining the influence of the train/test data ratio on prediction accuracy.

System Architecture

Figure 1 depicts the machine architecture. This section goes through each block in detail.

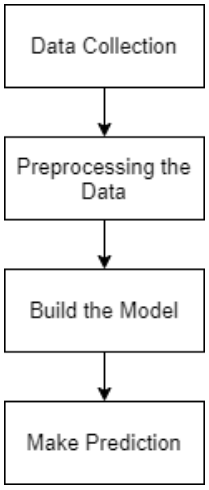


Fig. 1 Proposed System Architecture

A. Data Collection and Understanding Process: The study makes use of a real dataset. We used the Heart Disease dataset, which has 2000 records and ten fields with categorical and numerical data. Each area in the disease data set represents a feature of that individual person, and each record represents single feature information.

B. Data Preparation and Pre-processing: After the data collection process is completed, the data preparation process begins. It's important to clean up this data so that it can be used in models and produce better results. In this phase, we performed tasks like cleaning, filling the missing data, and removing unwanted data.

C. Feature Selection: One of the most important principles in R programming is feature selection. The predictor variable has a lot of columns, whereas it is a process of selecting necessary useful variables in a dataset to improve the results of a program and make it more accurate. As a result, the correlation coefficient is measured to determine which ones are important, and these are then used to develop training methods.

D. Test and Train Dataset:

Separating data into test and training datasets is a crucial part of testing data mining models since it reduces the impact of data inconsistency and allows for a clearer understanding of the model's characteristics. The test data set contains all of the data needed for data prediction, while the training data set contains all of the data that isn't needed. To investigate Prediction estimation, we divided the dataset into variable ratios.

Using various feature selection algorithms, this paper aims to identify the most significant variables that may have a positive impact on the accuracy of the features of music performance prediction models.

IV. Modeling and Experiments

The test report input from UI is captured and stored in a table. We've built a table with all of the records from the current data collection. To construct a model, we used the Naive Bayes algorithm. We are predicting whether the patient will develop heart disease or diabetes based on the model and test results.

Sr No	Parameters	Description
1	Age	Range of Age
2	Sex	Gender of person
3	thal	Value of thal
4	Resting Bp	Range of Resting Bp
5	Cholesterol	Range of Cholesterol
6	Fasting Blood Sugar	Value of Blood Sugar
7	Electrocardiographic	Value of Electrocardiographic

8	Maximum Heart rate	Range of Heart Rate
9	Oldpeak	Value of Oldpeak
10	The Slope of the peak exercise	Value of peak exercise

V. Requirement Analysis

A. Software

Windows 7 and higher can be used as operating systems. R is the programming language used. Along with that, we are using Rstudio for programming.

B. Hardware

The key memory requirement is 8 GB or more so that the entire program can run at the same time. This would eliminate the need to swap the system's memory contents. The hard disc drive is needed in order to permanently store the software on the storage. The processor must process the data on the device quickly. To communicate with the device when on the go, the user requires a computer or laptop.

VI. Implementation and Result Analysis

The test report input from the user interface is captured and stored in a table. We've built a table with all of the records from the current data collection. To construct a model, we used the Naive Bayes algorithm. We are predicting whether the patient will develop heart disease or diabetes based on the model and test results. Age, chest paint form, resting blood pressure, cholesterol, fasting blood sugar, maximum heart rate, the slope of peak exercise, and other factors will be considered in the patient's study. The person's report is available in R shiny and can also be viewed in a browser.

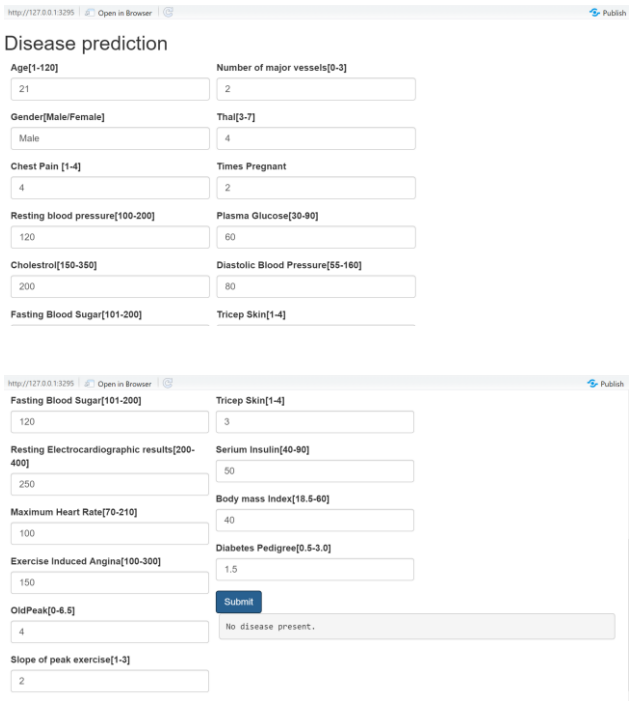


Fig 1: o/p- No disease found

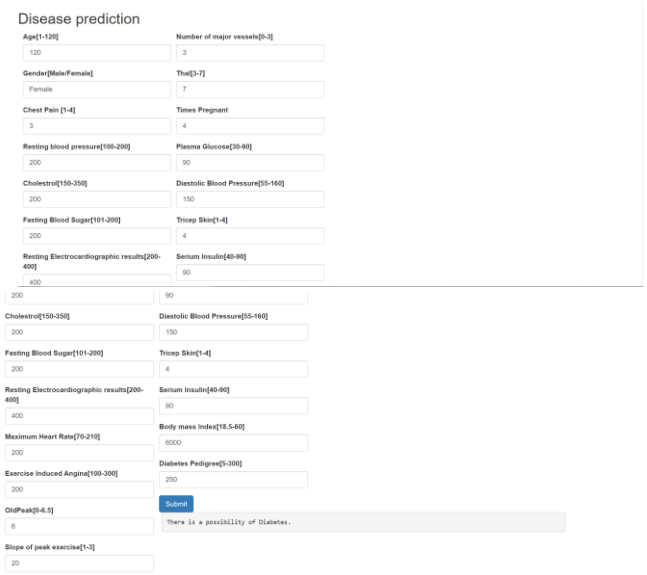
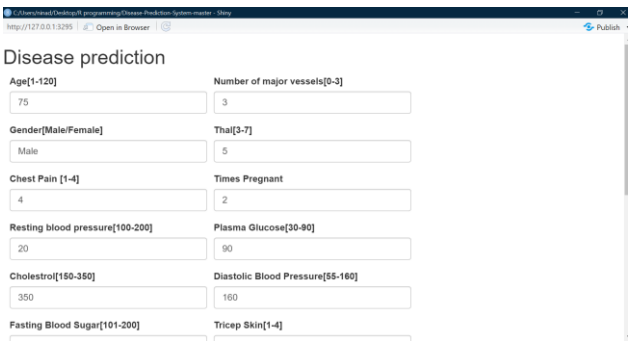


Fig 2: o/p- There is a possibility of Diabetes



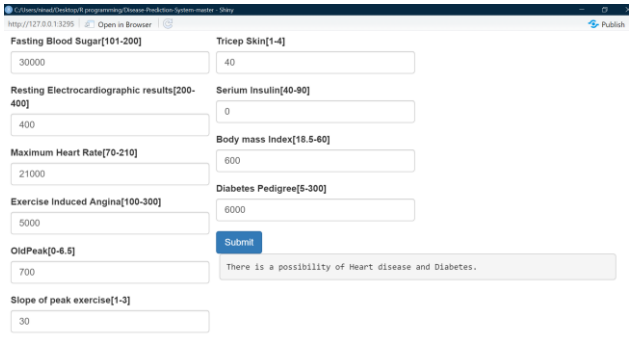


Fig 3: o/p- There is a possibility of Heart Disease and Diabetes

Code :

1 Server.R

This is the server logic for a Shiny web application.

```
library(shiny)
library(sqldf)
library(e1071)

shinyServer(function(input, output) {

  observeEvent(input$submit,{
    output$text <- renderText({

      age <- as.numeric(as.character(input$age))
      sex <- as.factor(input$sex)
      chestpaintype <- as.factor(input$chestpaintype)
      restingbp <- as.numeric(as.character(input$restingbp))
      cholestrol <- as.numeric(as.character(input$cholestrol))
      fastingbloodsugar <- as.factor(input$fastingbloodsugar)
      electrocardiographic <- as.factor(input$electrocardiographic)
      maxheartrate<- as.numeric(as.character(input$maxheartrate))
      exerciseangina<- as.factor(input$exerciseangina)
      oldpeak <- as.numeric(as.character(input$oldpeak))
      slopeofpeakexercise <- as.factor(input$slopeofpeakexercise)
      ca <- as.factor(input$ca)
      thal <- as.factor(input$thal)
      num <- as.factor(0)

      #Diabetes Data Input
      num <- as.factor(0)

      #Diabetes Data Input

      pregnant <- as.numeric(as.character(input$pregnant))
      plasma <- as.numeric(as.character(input$plasma))
      bp <- as.numeric(as.character(input$bp))
      tricep <- as.numeric(as.character(input$tricep))
      insulin <- as.numeric(as.character(input$insulin))
      bmi <- as.numeric(as.character(input$bmi))
      pedigree<- as.numeric(as.character(input$pedigree))
      #dbage <- as.numeric(as.character(input$dbage))
      class <-  as.factor(0)



      DiabetesTestData <- data.frame("times.pregnant" = pregnant, "plasma.glucose" = plasma, "diastolic.bp" =
      write.csv(DiabetesTestData, file="DiabetesTestData.csv", row.names = FALSE)
      diabetesCsvData <- read.csv("DiabetesTestData.csv")

      #Code for HeartDisease
      HeartTestData <- data.frame("Age" = age, "Sex" = sex, "chesp.pain.type" = chestpaintype, "resting.bp" =
      write.csv(HeartTestData, file="hearttestData.csv", row.names = FALSE)
      csvTestData <- read.csv("hearttestData.csv")
      db<-dbConnect(SQLite(), dbname="diseasedb")
      sqldf("attach 'diseasedb' as new")

      dbWriteTable(conn = db, name = "Heart", value = csvTestData , row.names= FALSE, header = FALSE, append =

      dbWriteTable(conn = db, name = "DiabetesData", value = diabetesCsvData , row.names= FALSE, header = FALS
      HeartDiseaseData <- dbReadTable(db,"Heart")

      DiabetesData <- dbReadTable(db,"DiabetesData")


      HeartDiseaseData$num[HeartDiseaseData$num > 0] <- 1

      namesFactor <- c(2:3,6:7,9,11:14)
      HeartDiseaseData[,namesFactor] <- lapply(HeartDiseaseData[,namesFactor], as.factor)
      namesNumeric <- c(1,4:5,8,10)
      HeartDiseaseData[,namesNumeric] <- lapply(HeartDiseaseData[,namesNumeric], as.numeric)
      HeartTestData <- tail(HeartDiseaseData,1)
      Heartmodel <- naiveBayes(num~., data = HeartDiseaseData)
      Heartresult <- predict(Heartmodel,HeartTestData)


      #Diabetes
      numericvalues <- c(1:8)
      DiabetesData[,numericvalues] <- lapply(DiabetesData[,numericvalues], as.numeric)
      factorValues <- c(9)
      DiabetesData[,factorValues] <- as.factor(as.character(DiabetesData[,factorValues]))
```

```
      DiabetesData[,factorValues] <- as.factor(as.character(DiabetesData[,factorValues]))

      DiabetesTestData <- tail(DiabetesData,1)

      DiabetesModel <- naiveBayes(class~., data = DiabetesData)
      DiabetesResult <- predict(DiabetesModel, DiabetesTestData)

      if(Heartresult ==1 && DiabetesResult== 1)
      {
        displayMessage <- "There is a possibility of Heart disease and Diabetes."
      }
      else if(Heartresult == 0 && DiabetesResult == 1)
      {
        displayMessage <- "There is a possibility of Diabetes."
      }
      else if(Heartresult == 1 && DiabetesResult == 0)
      {
        displayMessage <- "There is a possibility of Heart disease."
      }
      else
      {
        displayMessage<- "No disease present."
      }
      paste(displayMessage)
    })
})
```

2 ui.R

This is the user-interface definition of a Shiny web application.

```
# This is the user-interface definition of a Shiny web application.
# You can find out more about building applications with Shiny here:

library(shiny)

shinyUI(bootstrapPage(

  # Application title
  titlePanel("Disease prediction"),
  div(style="float:left; margin-left: 10px",
    id = "form",
    textInput("age", "Age[1-120]", ""),
    textInput("sex", "Gender[Male/Female]", ""),
    textInput("chestpaintype", "Chest Pain [1-4]", ""),
    textInput("restingbp", "Resting blood pressure[100-200]", ""),
    textInput("cholestrol", "Cholestrol[150-350]", ""),
    textInput("fastingbloodsugar", "Fasting Blood Sugar[101-200]", ""),
    textInput("electrocardiographic", "Resting Electrocardiographic results[200-400]", ""),
    textInput("maxheartrate", "Maximum Heart Rate[70-210]", ""),
    textInput("exerciseangina", "Exercise Induced Angina[100-300]", ""),
    textInput("oldpeak", "OldPeak[0-6.5]", ""),
    textInput("slopeofpeakexercise", "Slope of peak exercise[1-3]", "")


  ),
    textInput("oldpeak", "OldPeak[0-6.5]", ""),
    textInput("slopeofpeakexercise", "Slope of peak exercise[1-3]", "")


  ),

  div(style="float:left; margin-left: 10px",
      textInput("ca", "Number of major vessels[0-3]", ""),
      textInput("thal", "Thal[3-7]", ""),
      textInput("pregnant", "Times Pregnant", ""),
    textInput("plasma", "Plasma Glucose[30-90]", ""),
    textInput("bp", "Diastolic Blood Pressure[55-160]", ""),
    textInput("tricep", "Tricep Skin[1-4]", ""),
    textInput("insulin", "Serium Insulin[40-90]", ""),
    textInput("bmi", "Body mass Index[18.5-60]", ""),
    textInput("pedigree", "Diabetes Pedigree[100-5000]", ""),
    actionButton("submit","Submit", class="btn-primary")
  ),

  # Show a plot of the generated distribution
  mainPanel(
    verbatimTextOutput("text")
  )
)
)
```

VII. Conclusion

Our mini project's goal is to build an R project that will aid in the analysis of the patient's well-being based on the reports. It can aid in the early detection of heart disease and diabetes. This project diagnoses the possible health-related issue based on the test report values. In the proposed work naive Bayes algorithm is used to classify the data set because naive Bayes provides accurate results. With these results, heart diseases among people are predicted. Thus the heart disease prediction system successfully diagnoses the medical data and predicts heart diseases and diabetes.

VIII. REFERENCES

[1]. Rosamond W, Flegal K, Furie K, et al. Heart disease and stroke statistics—2008 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee.Circulation.2008;117(4):e25–146.

[2]. M. A. Jabbar and S. Samreen, "Heart disease prediction system based on hidden naïve Bayes classifier," 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), 2016, pp. 1-5, doi: 10.1109/CIMCA.2016.8053261.

[3]. [1] Ritika Chandha and Shubhankar Mayank, "Prediction of heart disease using data mining techniques," Springer, 2016

[4]. B. Alić, L. Gurbeta and A. Badnjević, "Machine learning techniques for classification of diabetes and cardiovascular diseases," 2017 6th Mediterranean Conference on Embedded Computing (MECO), 2017, pp. 1-4, doi: 10.1109/MECO.2017.7977152.

[5]. Sujata Joshi and Mydhili k.Nair, "Prediction of Heart Disease Using Classification Based Data Mining Techniques," Springer, vol. 2, 2015.

[6]. K.Gomathi and Dr. Shanmugapriya, "Predictive Heart Disease Used to Classify Mining Data," International Journal for Research in Applied Science & Engineering Technology (JRASET), vol. 4, no. II, February 2016

[7]. Rishabha Saxena, Aakriti Johri, Vikas Deep and Purushottam Sharma, "Heart Disease Prediction System Using CHC-TSS Evolutionary,KNN,and Decision Tree Classification Algorithm," Springer, 2019.

[8]. Purushottam, Pro.(Dr)Kanak Saxena and Richa Sharma,"Efficient Heart Disease PredictionSystem," Elsevier, 2016

[9]. S. Fathima and N. Hundewale, "Comparison of classification techniques-SVM and naives bayes to predict the Arboviral disease-Dengue," 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBME), 2011, pp.

[10]. M. Gandhi and S. N. Singh, "Predictions in heart disease using techniques of data mining," 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015, pp. 520-525, doi: 10.1109/ABLAZE.2015.7154917.

[11]. J. Singh, A. Kamra and H. Singh, "Prediction of heart diseases using associative classification," 2016 5th International Conference on Wireless Networks and Embedded Systems (WECON), 2016,pp.1-7,doi:10.1109/WECON.2016.799348.

Cite this article as :

Ninad Marathe, Sushopti Gawade, Adarsh Kanekar, "Prediction of Heart Disease and Diabetes Using Naive Bayes Algorithm ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 3, pp. 447-453, May-June 2021. Available at

doi : https://doi.org/10.32628/CSEIT217399

Journal URL : https://ijsrcseit.com/CSEIT217399