# PROFESSIONAL TRAINING REPORT

**at**

**Sathyabama Institute of Science and Technology**

**(Deemed to be University)**

Submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering

By

## SOMU.K.B

**REG. NO. 39110958**

****

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SCHOOL OF COMPUTING**

**SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,**

**CHENNAI – 600119, TAMILNADU**

**April 2022**

|  |  |  |
| --- | --- | --- |
|  | **SATHYABAMA**  **INSTITUTE OF SCIENCE AND TECHNOLOGY** (DEEMED TO BE UNIVERSITY) **Accredited with Grade “A” by NAAC**  (Established under Section 3 of UGC Act, 1956)  JEPPIAAR NAGAR, RAJIV GANDHI SALAI  CHENNAI– 600119  [**www.sathyabama.ac.in**](http://www.sathyabama.ac.in) |  |

# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonafide work of **SOMU.K.B**

**(Reg. No: 39110958)** who carried out the project entitled **“Cardiovascular Disease Classification Using Data Science Algorithm”** under my supervision from Feb 2022 to April 2022.

## Internal Guide

## Ms. S.Nithya ,M.E,

**Head of the Department**

**Dr.L. Lakshmanan,M.E,.Ph.D,**

**Dr.S. Vigneshwari,M.E,.Ph.D.**



## Submitted for Viva voce Examination held on

**Internal Examiner External Examiner**

**DECLARATION**

I, **SOMU.K. B** hereby declare that the project report entitled “**Cardiovascular Disease Classification Using Data Science Algorithm**” done by me under the guidance of **Ms.S.Nithya,M.E.** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Computer Science and Engineering.

## DATE:

**PLACE:**

**SIGNATURE OF THE CANDIDATE**

**ACKNOWLEDGEMENT**

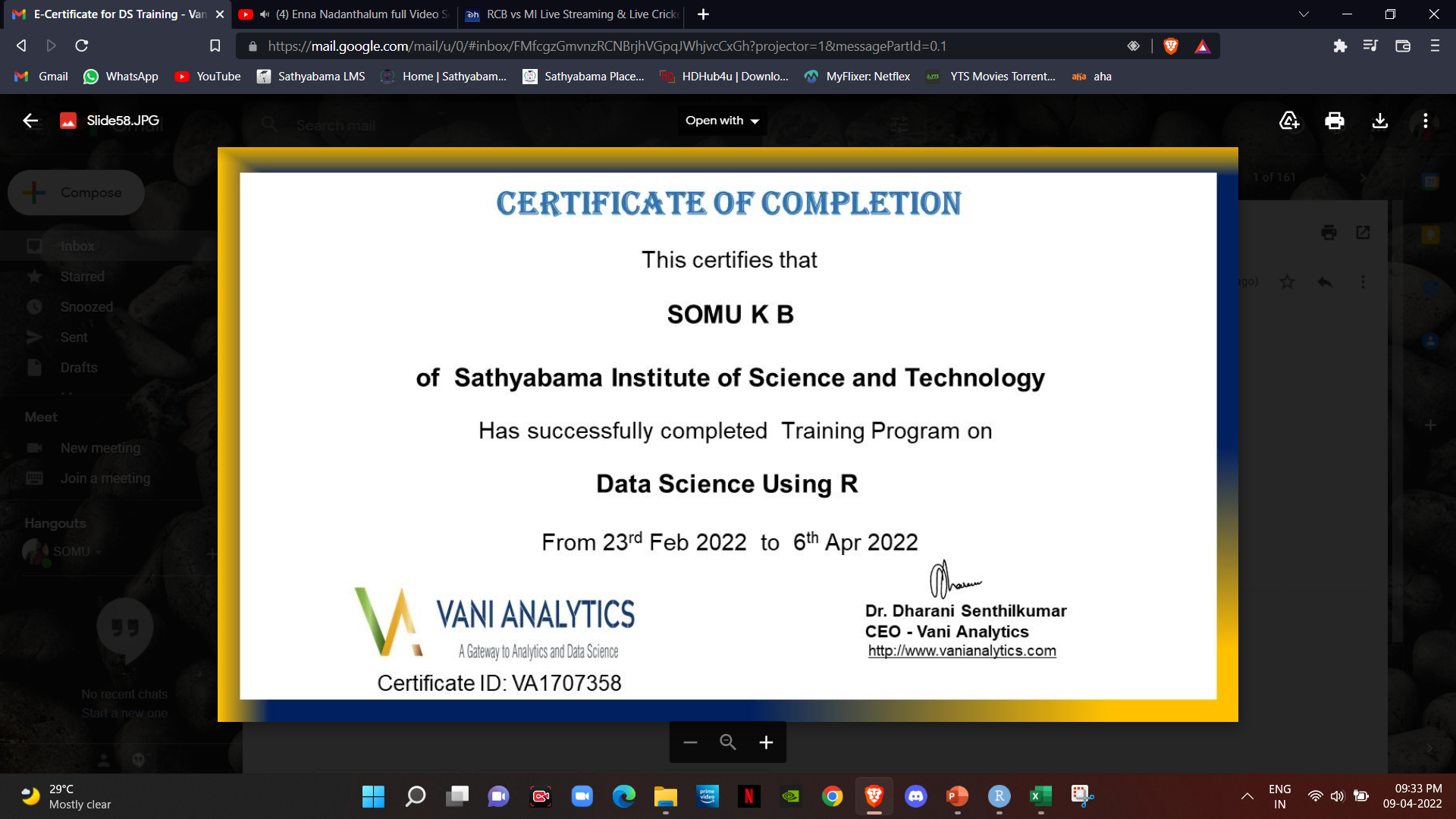
I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D**, **Dean**, School of Computing, **Dr. S. Vigneshwari, M.E., Ph.D. and Dr. L. Lakshmanan, M.E., Ph.D., Heads of the Department** of **Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **MS.S.NITHYA,M.E.** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

**TRAINING CERTIFICATE**

****

**ABSTRACT**

* Heart disease is one of the complex diseases and globally many people suffered from this disease.
* It is even being highlighted as a silent killer which leads to death of a person without obvious symptoms.
* On time and efficient identification of heart disease plays a key role in healthcare, particularly in the field of cardiology. Due to lack of resources in the medical field, the prediction of heart disease occasionally may be a problem.
* Utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients. This issue can be resolved by adopting Data Science techniques.
* Data Science trends have been developed and various research are going on in this field.
* This project intends to adopt Naive Bayes algorithm to design a classification model for the effective identification of Heart disease.
* This Naive Bayes model will predict whether the patient will have the risk heart disease or not.

# 

|  |  |  |
| --- | --- | --- |
| **TABLE OF CONTENTS** | | |
| **CHAPTER** | **TITLE** | **PAGE NO** |
|  | **ABSTRACT** | 6 |
| **1** | **INTRODUCTION** | 8 |
| **2** | **SYSTEM ANALYSIS** | 10 |
|  | 2.1 EXISTING SYSTEM | 10 |
|  | 2.2 PROPOSED SYSTEM | 10 |
| **3** | **SYSTEM SPECIFICATION** | 11 |
|  | 3.1 HARDWARE SPECIFICATION | 11 |
|  | 3.2 SOFTWARE SPECIFICATION | 11 |
| **4** | **SYSTEM IMPLEMENTATION** | 14 |
|  | 4.1 SYSTEM ARCHITECTURE | 14 |
|  | 4.2 MODULES | 15 |
| **5** | **SYSTEM DESIGN** | 20 |
|  | 5.1 DATA FLOW DIAGRAM | 20 |
|  | 5.1.1 LEVEL 0 | 20 |
|  | 5.1.2 LEVEL 1 | 21 |
|  | 5.1.3 LEVEL 2 | 22 |
|  | 5.1.4 LEVEL 3 | 22 |
| **6** | **CONCLUSION** | 23 |
|  | **REFERENCES** | 24 |
|  | **APPENDIX A: SNAP SHOTS** | 25 |
|  | **APPENDIX B: SOURCE CODE** | 28 |

**CHAPTER 1**

**INTRODUCTION**

Heart disease has been the most significant cause of death in the world during the past 10 years. It includes diseases of heart muscles, valves, conduction system, heart disease and others. Myocardial infarction or heart disease is the major one among all other types of heart diseases. Heart diseases are seen in all the classes of people in recent times, in contrast to previous days when it was disease of rich class people. This project, concentrated and focused on the heart disease.

Heart disease is even being highlighted as a silent killer which leads to death of a person without obvious symptoms. Hence continued efforts are being done to predict the possibility of this deadly disease in prior. So that various tools & techniques are regularly being experimented to suit the present day health needs. Data Science techniques can be a boon in this regard.

Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. For this project data have been collected from various sources and by using Naïve Bayes algorithm, datasets are analysed to extract the intelligent pattern of predicting the risk of heart disease. This technique can be very well adapted to the do the prediction of heart disease. As the well-known quote says “Prevention is better than cure”, early prediction and its control can be helpful to prevent & decrease the death rates due to heart disease.

Heart disease is the common term used for Myocardial infarction (MI). It is due to the interruption of blood supply to a part of heart leading to damage to the heart muscle. Chest pain is the common symptom which may pass on to shoulder, neck, back or jaw. It may present in the centre or left side of the chest.

The goal of this project, is to predict whether a patient will have the risk of heart disease or not. Naive Bayes algorithms is used to classify and predict the Heart disease, using factors, such as Age, Gender, Chest Pain, Rest BP, Chol, FBS , Rest ECG etc. This technique can be very well adapted in the hospital to do the prediction of heart disease in prior.

**CHAPTER 2**

**SYSTEM ANALYSIS**

**2.1 EXISTING SYSTEM**

Existing concept deals with Qualitative observations and simple statistical analysis. The qualitative observations deals with the data that can be observed through human senses. They do not involve measurements or number. The simple statistical analysis includes mean, standard deviation, median, finding the size of data, variance etc., the results produced by this techniques are not precise. Using some programming languages and data mining tools, heart dataset is classified. There is an increase in need to manage and understand data.

**2.2 PROPOSED SYSTEM**

Heart disease dataset is collected and analyzed to predict whether a patient will have the risk of heart disease or not after certain period of time. For prediction Data Science techniques of Naive Bayes algorithm have been applied. Data Science is an interdisciplinary field that incorporates computer science, mathematics, statistics and domain knowledge. Naive Bayes Algorithm is applied to the heart disease dataset and efficient prediction accuracy is achieved by delivering the result of whether the patient will have the risk of heart disease or not.

**CHAPTER 3**

**SYSTEM SPECIFICATION**

**3.1 HARDWARE SPECIFICATION**

Processor : Processor Intel CORE i3 and above

Internet Connection : Existing telephone lines, Data card, Fiber net

RAM : 4 GB

**3.2 SOFTWARE SPECIFICATION**

Operating System : Windows, Mac, Linux

Language : R Programming – R-4.1.1

GUI : RStudio

R is a Programming language and environment for Data Manipulation, statistical computing and graphics. Developed at Bell Laboratories by John Chambers and colleagues. Similar to the S language and available across all platforms -Linux, Mac, Windows. It provides variety of Graphical and Statistical techniques – Regression, Clustering, Classification, Association Rule Mining etc.

R is available as Free Software under the terms of the Free Software Foundation’s GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and Mac OS. The source code for the R software environment is written primarily in C, FORTRAN, and R.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

* An effective data handling and storage facility,
* A suite of operators for calculations on arrays, in particular matrices,
* A large, coherent, integrated collection of intermediate tools for data analysis,
* graphical facilities for data analysis and display either on-screen or on hardcopy and
* a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

RStudio is an Integrated Development Environment (IDE) for R. RStudio makes R easier to use. It includes a code editor, compiler or interpreter, debugger and visualization tools as well as tools for plotting, history, syntax-highlighting editor that supports direct code execution and workspace management. RStudio was founded by JJ Allaire, creator of the programming language ColdFusion.

RStudio is available in open source and commercial editions available in two editions: RStudio Desktop, where the program is run locally as a regular desktop application (Windows, macOS, and Linux) and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server. RStudio is written in the C++ programming language. RStudio helps to keep R more organized and it adds more functionality to it.

**CHAPTER 4**

**SYSTEM IMPLEMENTATION**

**4.1 SYSTEM ARCHITECTURE**

Importing Heart Dataset

Data Preprocessing

Naive Bayes Algorithm

Naive Bayes Model

Output Predicted Variable

Predictors Variables

**Fig.4.1 Architecture Diagram**

The fig 4.1 shows the architectural diagram of this project defines the flow of data for classifying the patients for having the heart disease or not. The first step is data importing, Data have to be loaded in to the R environment for analysis. The second step is, preprocessing the collected raw data into understandable format. Standardization and Normalization is the technique which is used to transform the various format of data into the common format and min-max technique is used for normalization of data values.

It is not necessary to hold all the attributes for doing the analysis, we can hold only the attributes which is affecting the analysis. The missing values problem have to be solved by simple statistical techniques. Naive Bayes algorithm is applied on the preprocessed data to create a predictive Naive Bayes model. The algorithm is tested with some predictor variables, the Naive Bayes Model will generate the output of whether a patient will have the risk of heart disease or not.

**4.2 MODULES**

There are four Modules

* Importing Heart dataset
* Data Preprocessing
* Model Generation Using Naive Bayes Algorithm
* Prediction using Naive Bayes Model

**4.2.1 IMPORTING HEART DATASET**

Data is available in any file format like .txt, .csv, .xlsx. spss etc. Data have to be loaded in to R environment for analysis. Once data have been extracted from the file it should be stored in a data frame. Packages necessary for classification algorithm – Naive Bayes have to be installed into the R environment. For Naive Bayes, Naïve Bayes package have to be installed and loaded in to R environment.

**4.2.2 DATA PREPROCESSING**

Data preprocessing is the data mining technique that involves transforming raw data into understandable format. The raw data is highly susceptible to noise, missing values, and inconsistency.

Real world data is often incomplete, inconsistent, and is likely to contain errors. The missing Values problem have to be solved by simple statistical techniques. Data preprocessing is the proven method for resolving such issues. In order to improve the quality of the data consequently, the mining results of raw data is preprocessed so the efficiency process improved. It is not necessary to hold all the attributes for doing the analysis, we can hold only the attributes which is affecting the analysis.

Data standardization is the process by which similar data is collected in various formats is transformed to a common format that enhances the comparison process, allows for collaborative research and large scale analytics. Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

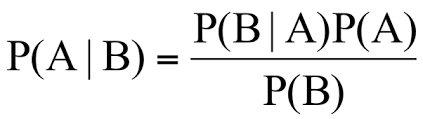
Formula used for Normalization:

*X-Xmin/Xmax-Xmin*

**4.2.3 MODEL GENERATION USING NAIVE BAYES ALGORITHM**

Classification can be performed on structured or unstructured data. Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category/class to which a new data will fall under. Classification can be performed on structured or unstructured data.

Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category/class to which a new data will fall under. Naive Bayes is a classification and prediction algorithm based on Bayes Theorem. Bayes Theorem



P(A|B) - Probability of occurrence of event A given the event B is true - Posterior Probability

P(A) - Probabilities of the occurrence of event A - Class prior Probability

P(B) - Probabilities of the occurrence of event B – Predictor Prior Probability

P(B|A) - Probability of the occurrence of event B given the event A is true - Likelihood

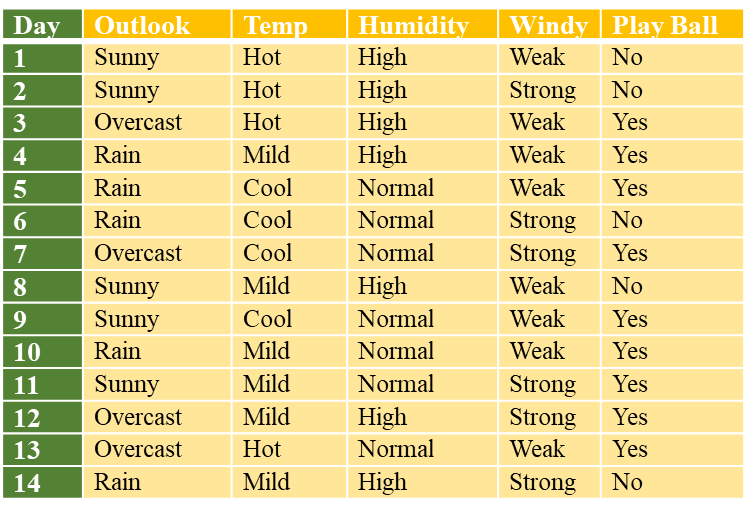
Naive Bayes algorithm based on Bayes’ theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as document classification and spam filtering. This algorithm requires a small amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely fast compared to more sophisticated methods.

***Algorithmic steps for Naive bayes classification***

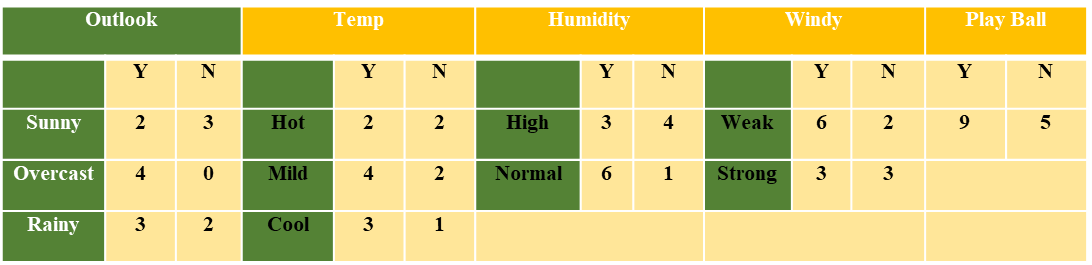
Step 1: Convert the data set into a frequency table

Step 2: Create Likelihood table by finding the probabilities values of each attribute

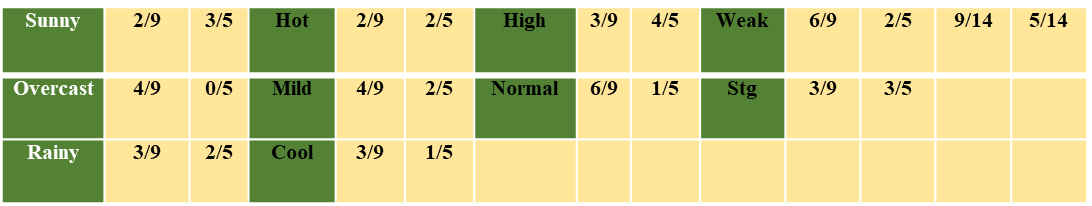
Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.



Frequency Table



Likelihood Table



**Sunny, Cool, High, Strong**

P(Yes |S,C,H,S) = P(S,C,H,S | Yes) \*P(Yes) / P(S,C,H,S)

P(No | S,C,H,S) = P(S,C,H,S | No) \*P(No) / P(S,C,H,S)

P(S,C,H,S) = P(S)\*P(C)\*P(H)\*P(S)

= 5/14 \* 4/14 \*7/14\*6/14 = 0.0216

Likelihood for Yes = 2/9 \* 3/9 \* 3/9\* 3/9 = 0.0082

Likelihood for No = 3/5 \* 1/5 \* 4/5\* 3/5 = 0.0576

Posterior Probability of Yes = (0.0082\* 9/14) / 0.0216 =0.245

Posterior Probability of No = (0.0576 \* 5/14) / 0.0216 =0.952

The posterior probability for the No is higher, so the probability of no is higher.

 Once Naive Bayes model have been created, by using testing data, model can be evaluated and by giving only predictor variables, target variable can be predicted.

**4.2.4 PREDICTION USING NAIVE BAYES MODEL**

By using testing data model can be evaluated. The algorithm is used for prediction by passing predictors variable, the Naive Bayes Model will generate the output of which patient is having the risk of heart disease in future.

**CHAPTER 5**

**SYSTEM DESIGN**

**5.1 DATA FLOW DIAGRAM**

Data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. It differs from the flowchart as it shows the data flow instead of the control flow of the program. A data flow diagram can also be used for the visualization of data processing. The DFD is designed to show how a system is divided into smaller portions and to highlight the flow of data between those parts.

**5.1.1 LEVEL 0**

**R Environment**

**Heart Dataset – Attributes -** AgeGender,ChestPain,RestBP,Chol,Fbs,RestECG, Target etc

**Install Packages**

NB– Naivebayes

**Fig 5.1.1 LEVEL 0**

In the fig 5.1.1, Data is imported in to R environment for analysis. The input dataset is heart dataset, is in csv format have to be loaded into the program to start the analysis. Package necessary for Naive Bayes algorithm have to be installed and loaded in to the program. Naive Bayes - naivebayes.

**5.1.2 LEVEL 1**

Type Casting

Fill in Missing Values

Attributes Selection

Normalization

**Fig 5.1.2 LEVEL 1**

The fig 5.1.2 shows the data preprocessing, which includes attribute selection, standardization and normalization functions. In standardization, raw data is transformed into common, understandable format. In attribute selection, hold only the attributes which is affecting the analysis and It is not necessary to hold all the attributes for doing the analysis. In Normalization, mean of the attribute will be 0 and standard deviation will be 1.

**5.1.3 LEVEL 2**

In fig 5.1.3, preprocessed data is given as an input to Naive Bayes algorithm and Naive Bayes model has been created, which is used for classification and prediction.

Preprocessed

Data

Naive Bayes

Model

Naive Bayes

Algorithm

**Fig 5.1.3 LEVEL 2**

**5.1.4 LEVEL 3**

In fig 5.1.4, Predictors variables is given as an input and the Naive Bayes Model will generate the output of whether the patient is having the risk of a developer will leave the company or he will continue in the company. Thus prediction process is implemented successfully. Using this model classification is also achieved by passing more number of employee datasets, it will classify the employee as two groups. One group is the set of employee who will leave the company and the other group is the set of employees who will remain in the company.

Naive Bayes Model

Output Predicted Variable

Predictors Variables

**Fig 5.1.4 LEVEL 3**

**CHAPTER 6**

**CONCLUSION**

This project specifies, the usage of Data Science algorithm in predicting, whether the patient is having the possibility of the heart disease. The Naive Bayes algorithm is used for predicting the patient’s heart disease using various parameters. This model can potentially help the doctors to do the prediction of heart disease in prior. Thus, this project concludes, Naive Bayes Algorithm performs better and faster when compared to other statistical techniques. Thereby the diagnosis and management of heart disease can be made simpler.

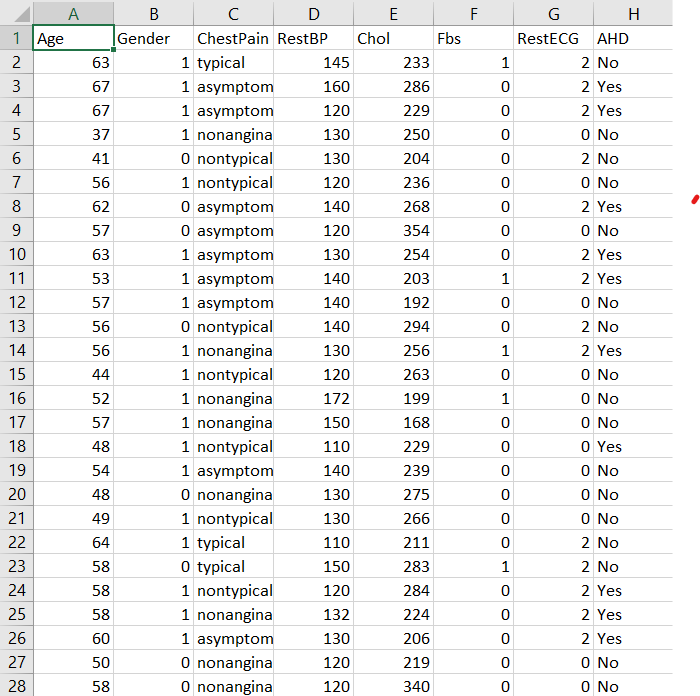
**REFERENCES**

1. Alperen Erdogan, Selda Guney, “Heart Disease Prediction by Using Machine Learning Algorithms", IEEE Signal Processing and Communications Applications Conference, 2020.  
     
   2. Jian Ping Li, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, ” Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare”, IEEE Access,Volume: 8, 2020, pp: 107562 - 107582  
     
   3. Seyedamin Pouriyeh, “A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease ", IEEE International Conference on Computers and Communications , 2017.  
     
   4. Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare, Jian Ping Li; Amin Ul Haq; Salah Ud Din; Jalaluddin Khan;Asif Khan;Abdus Saboor,IEEE Access, 2020, Volume: 8.  
     
   5. An Automated Strategy for Early Risk Identification of Sudden Cardiac Death by Using Machine Learning Approach on Measurable Arrhythmic Risk Markers”,Dakun Lai;Yifei Zhang;Xinshu Zhang;Ye Su;Md Belal Bin Heyat, IEEE Access, Year: 2019, Vol. 7.

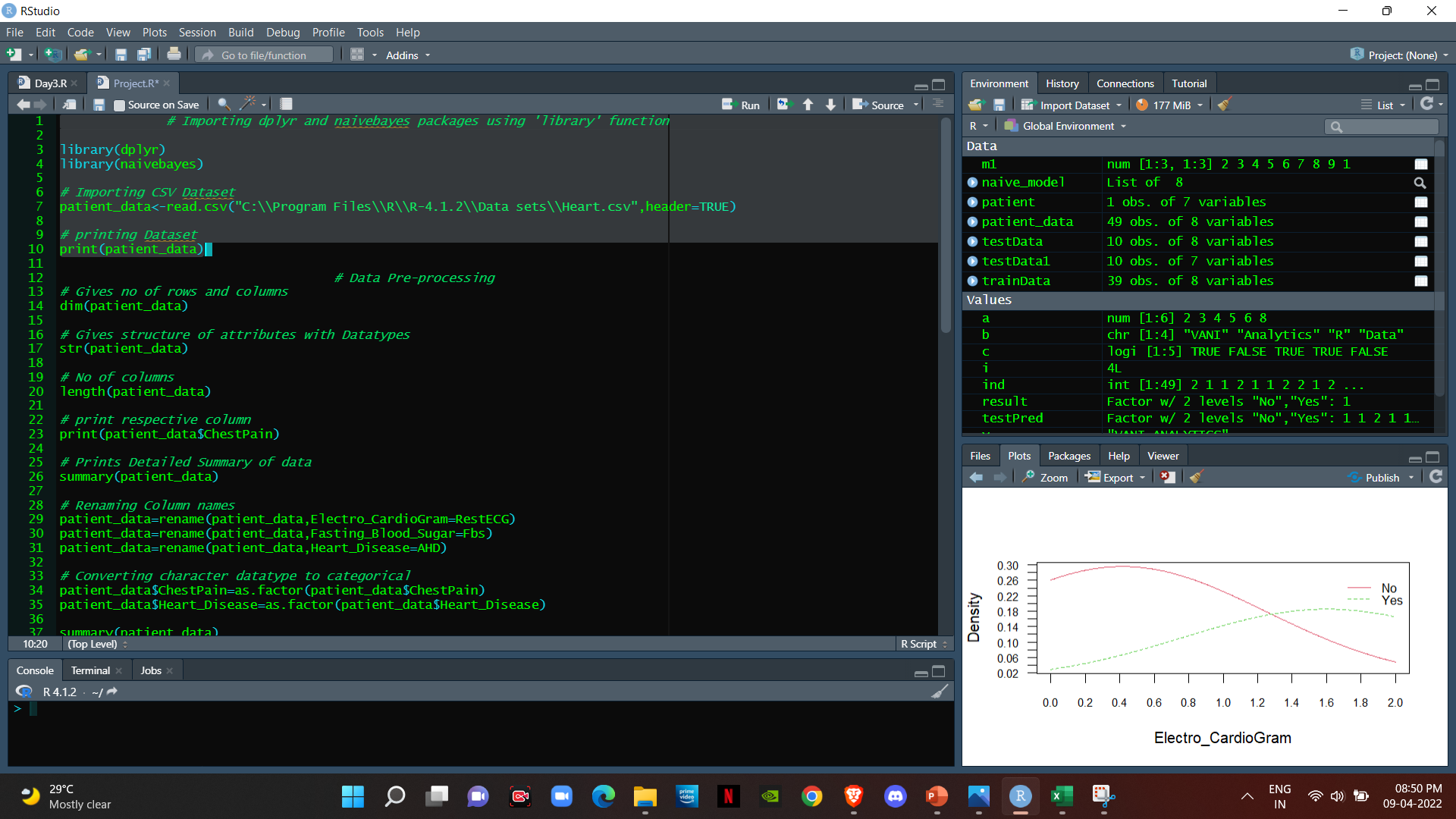
**APPENDIX A**

**SNAP SHOTS**

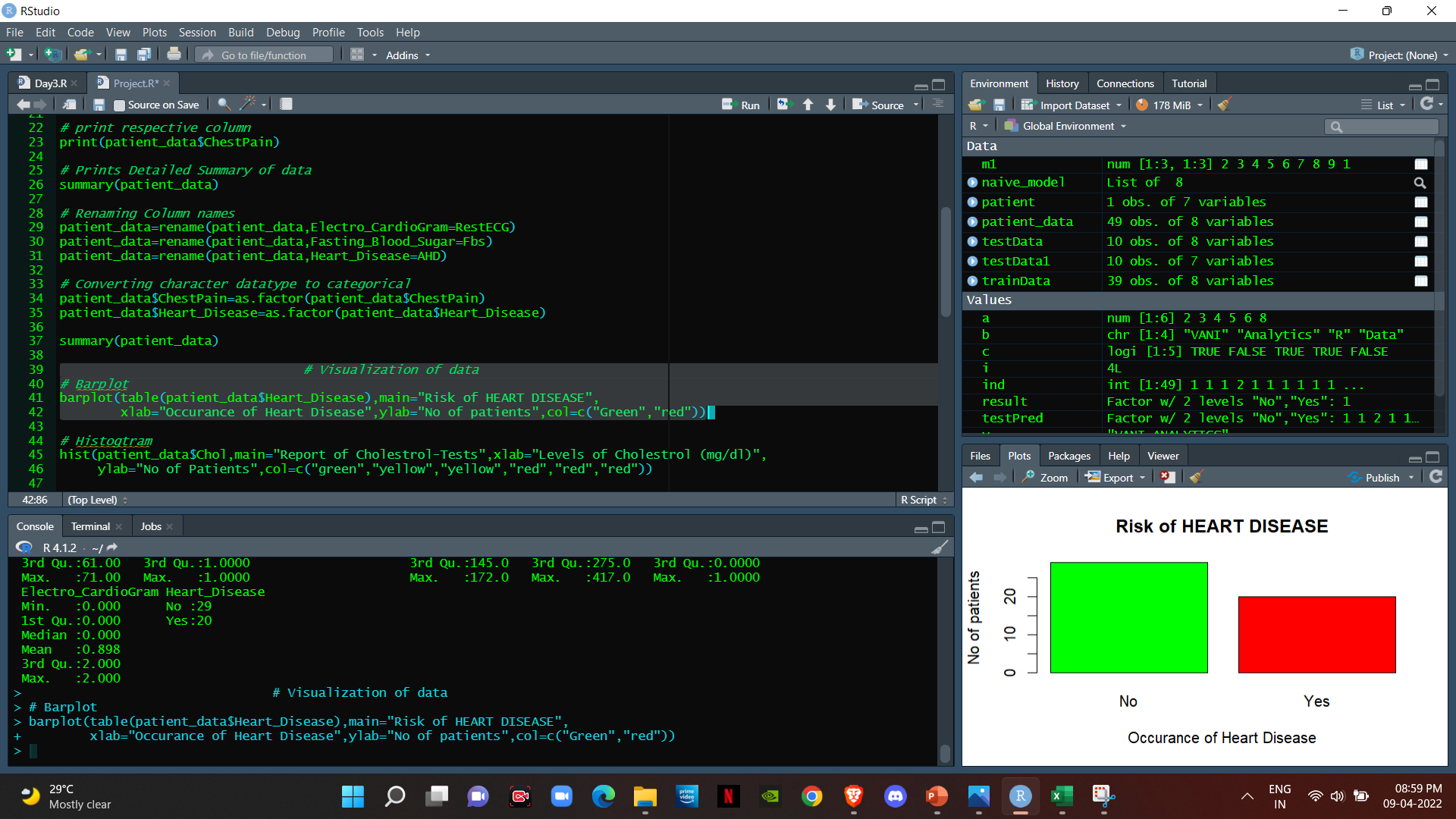
**SAMPLE DATASETS :-**



**DATA IMPORTING :-**

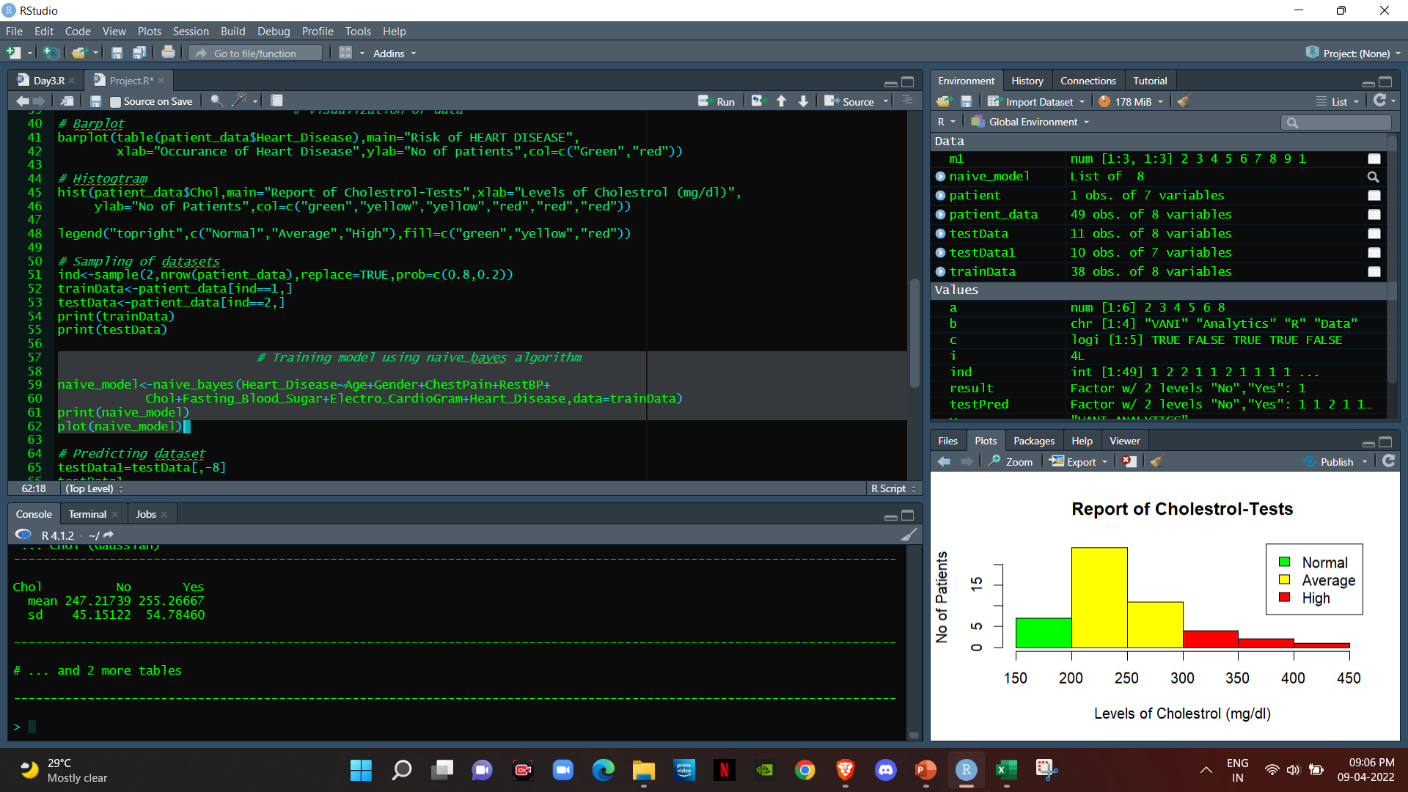
****

**DATA PREPROCESSING :-**

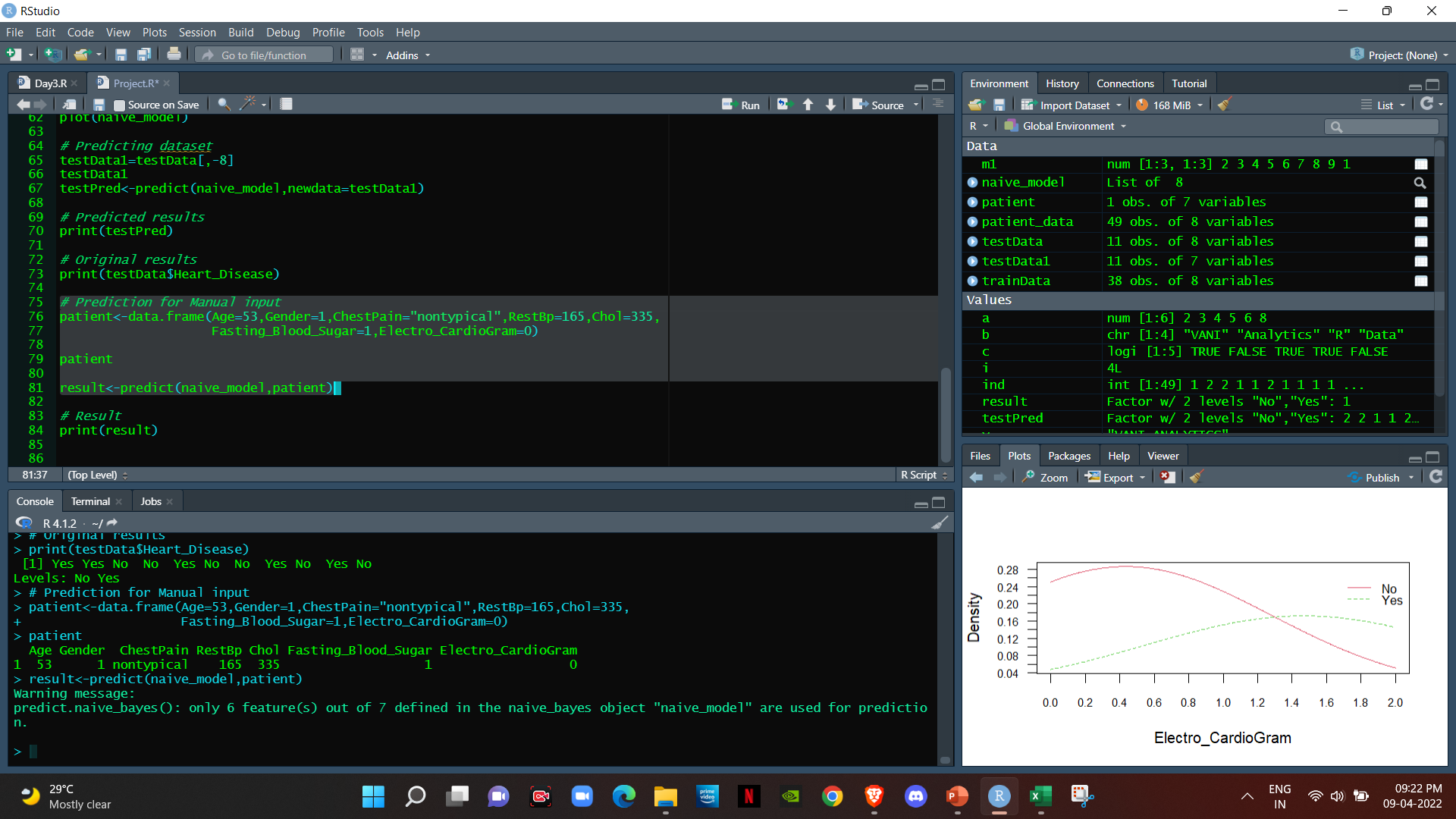
****

**NAIVE BAYES ALGORITHM :-**

**Naive Bayes Classification Model**

****

**Naive Bayes Prediction Model**

****

**APPENDIX B**

**SOURCE CODE**

***# Importing dplyr and naivebayes packages using 'library' function***

**library(dplyr)**

**library(naivebayes)**

***# Importing CSV Dataset***

**patient\_data<-read.csv("C:\\Program Files\\R\\R-4.1.2\\Data sets\\Heart.csv",header=TRUE)**

***# printing Dataset***

**print(patient\_data)**

***# Data Pre-processing***

***# Gives no of rows and columns***

**dim(patient\_data)**

***# Gives structure of attributes with Datatypes***

**str(patient\_data)**

***# No of columns***

**length(patient\_data)**

***# print respective column***

**print(patient\_data$ChestPain)**

***# Prints Detailed Summary of data***

**summary(patient\_data)**

***# Renaming Column names***

**patient\_data=rename(patient\_data,Electro\_CardioGram=RestECG)**

**patient\_data=rename(patient\_data,Fasting\_Blood\_Sugar=Fbs)**

**patient\_data=rename(patient\_data,Heart\_Disease=AHD)**

***# Converting character datatype to categorical***

**patient\_data$ChestPain=as.factor(patient\_data$ChestPain)**

**patient\_data$Heart\_Disease=as.factor(patient\_data$Heart\_Disease)**

**summary(patient\_data)**

***# Visualization of data***

***# Barplot***

**barplot(table(patient\_data$Heart\_Disease),main="Risk of HEART DISEASE",**

**xlab="Occurance of Heart Disease",ylab="No of patients",col=c("Green","red"))**

***# Histogtram***

**hist(patient\_data$Chol,main="Report of Cholestrol-Tests",xlab="Levels of Cholestrol (mg/dl)",**

**ylab="No of Patients",col=c("green","yellow","yellow","red","red","red"))**

**legend("topright",c("Normal","Average","High"),fill=c("green","yellow","red"))**

***# Sampling of datasets***

**ind<-sample(2,nrow(patient\_data),replace=TRUE,prob=c(0.8,0.2))**

**trainData<-patient\_data[ind==1,]**

**testData<-patient\_data[ind==2,]**

**print(trainData)**

**print(testData)**

***# Training model using naive\_bayes algorithm***

**naive\_model<-naive\_bayes(Heart\_Disease~Age+Gender+ChestPain+RestBP+**

**Chol+Fasting\_Blood\_Sugar+Electro\_CardioGram+Heart\_Disease,data=trainData)**

**print(naive\_model)**

**plot(naive\_model)**

***# Predicting dataset***

**testData1=testData[,-8]**

**testData1**

**testPred<-predict(naive\_model,newdata=testData1)**

***# Predicted results***

**print(testPred)**

***# Original results***

**print(testData$Heart\_Disease)**

***# Prediction for Manual input***

**patient<-data.frame(Age=53,Gender=1,ChestPain="nontypical",RestBp=165,Chol=335,**

**Fasting\_Blood\_Sugar=1,Electro\_CardioGram=0)**

**patient**

**result<-predict(naive\_model,patient)**

***# Result***

**print(result)**