



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 Information Technology

By: M.Prithiviraj
Reg.No:39120085



DEPARTMENT

TECHNOLOGY

SCHOOL OF COMPUTING

**SATHYABAMA INSTITUTE OF SCIENCE AND
TECHNOLOGY**

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,

CHENNAI – 600119, TAMILNADU

NOVEMBER 2021

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

DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Prithiviraj.M(Reg. No: 39120085)** who carried out the project entitled “**Heart attack prediction using Data science algorithm**” under my supervision from June 2021 to November 2021.

Internal Guide

Dr.P.Ajitha

Submitted for Viva voce Examination held on_____

Internal Examiner

External Examiner

DECLARATION

I, **M.Prithiviraj** hereby declare that the project report entitled **Heart attack prediction using Data science algorithm** done by me under the guidance of

DR.P.Ajitha is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in Information Technology

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.P.Ajitha** 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 **Dr.P.Ajitha** for her 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 Information Technology** who were helpful in many ways for the completion of the project.

TRAINING CERTIFICATE

CERTIFICATE OF COMPLETION

This certifies that

PRITHIVIRAJ M of Sathyabama Institute of Science and Technology

Has successfully completed Training Program on

Data Science Using R

From 9th Sep 2021 to 30th Oct 2021



Dr. Dharani Senthilkumar
CEO - Vani Analytics
<http://www.vanianalytics.com>

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 and predictive model for the effective identification of Heartdisease. 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	10-11

	2.1 EXISTING SYSTEM	10
	2.2 PROPOSED SYSTEM	10-11
3	SYSTEM SPECIFICATION	11-12
	3.1 HARDWARE SPECIFICATION	11
	3.2 SOFTWARE SPECIFICATION	11-12
4	SYSTEM IMPLEMENTATION	13-19
	4.1 SYSTEM ARCHITECTURE	13
	4.2 MODULES	14-19
5	SYSTEM DESIGN	19-22
	5.1 DATA FLOW DIAGRAM	19
	5.1.1 LEVEL 0	19-20
	5.1.2 LEVEL 1	20
	5.1.3 LEVEL 2	21
	5.1.4 LEVEL 3	22
6	CONCLUSION	22
	REFERENCES	23-24
	APPENDIX A: SNAP SHOTS	24-28
	APPENDIX B: SOURCE CODE	29-31

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 Heartdisease, using factors, such as Age, Gender, ChestPain, RestBP, Chol,Fbs, RestECG 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.

2.2PROPOSED 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

3.2SOFTWARESPECIFICATION

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

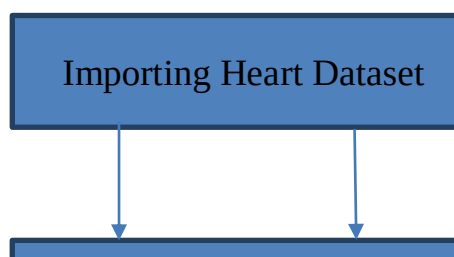


Fig.4.1 Architecture Diagram

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.

4.2MODULES

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, naivebayes package have to be installed and loaded in to R environment.

4.2.2 DATAPREPROCESSING

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:

$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

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

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

$P(A|B)$ - 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.

Day	Outlook	Temp	Humidity	Windy	Play Ball
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Frequency Table

Outlook			Temp			Humidity			Windy			Play Ball	
	Y	N		Y	N		Y	N		Y	N	Y	N
Sunny	2	3	Hot	2	2	High	3	4	Weak	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	Strong	3	3		
Rainy	3	2	Cool	3	1								

Likelihood Table

Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	Weak	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	Stg	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Sunny, Cool, High, Strong

$$P(\text{Yes} | S, C, H, S) = P(S, C, H, S | \text{Yes}) * P(\text{Yes}) / P(S, C, H, S)$$

$$P(\text{No} | S, C, H, S) = P(S, C, H, S | \text{No}) * P(\text{No}) / P(S, C, H, S)$$

$$\begin{aligned} P(S, C, H, S) &= P(S) * P(C) * P(H) * P(S) \\ &= 5/14 * 4/14 * 7/14 * 6/14 = 0.0216 \end{aligned}$$

$$\text{Likelihood for Yes} = 2/9 * 3/9 * 3/9 * 3/9 = 0.0082$$

$$\text{Likelihood for No} = 3/5 * 1/5 * 4/5 * 3/5 = 0.0576$$

$$\text{Posterior Probability of Yes} = (0.0082 * 9/14) / 0.0216 = 0.245$$

$$\text{Posterior Probability of No} = (0.0576 * 5/14) / 0.0216 = 0.952$$

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.

5.1.1LEVEL 0

Heart Dataset – Attributes -
Age,Gender,ChestPain,RestBP,Chol
,Fbs,RestECG, Target etc

R Environment

Install Packages
NB– Naivebayes

Fig 5.1.1 LEVEL0

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 LEVEL1

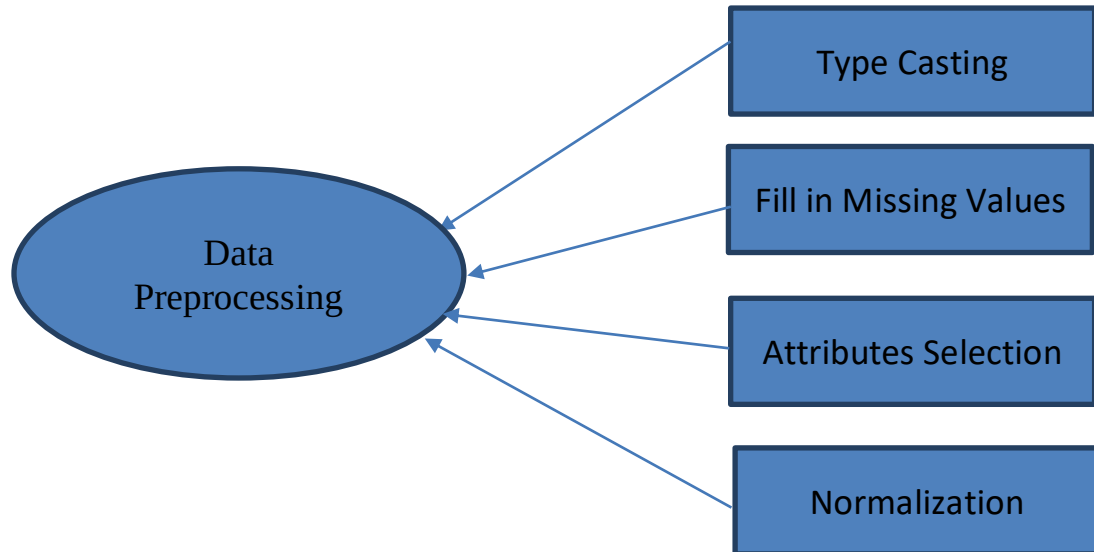


Fig 5.1.2 LEVEL1

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 LEVEL2

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.

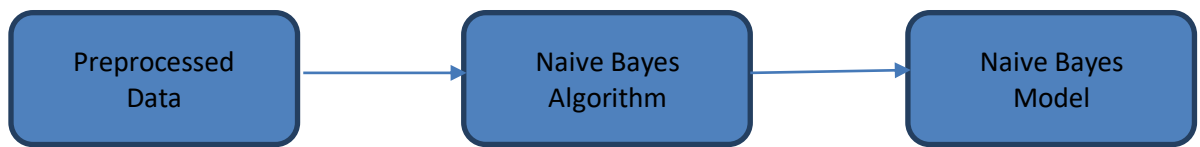


Fig 5.1.3 LEVEL2

5.1.4 LEVEL3

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.



Fig 5.1.4 LEVEL3

CHAPTER 6

CONCLUSION

This project specifies, the usage of Data Science algorithm in classifying and predicting, whether the patient is having the possibility of the heart disease. The Naive Bayes algorithm is used for classifying and 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

RStudio interface showing the execution of R code for loading and summarizing the heart_data dataset.

```

1 library(naivebayes)
2
3
4 #DATA IMPORTING
5 heart_data<- read.csv("C:\\Program Files\\R\\R-3.3.2\\datasets\\Heart.csv", header = TRUE)
6 print(heart_data)
7
8 dim(heart_data)
9 length(heart_data)
10 heart<- heart_data[1:5,]
11 heart
12 heart_data[,5]
13 print(heart_data$Age)
14
15 summary(heart_data)
16
17

```

Console output:

```

C:/Program Files/R/R-4.0.3/VA/my documents/JAN 2021/Internship/
> library(naivebayes)
>
> #DATA IMPORTING
> heart_data<- read.csv("C:\\Program Files\\R\\R-3.3.2\\datasets\\Heart.csv", header = TRUE)
> print(heart_data)
  Age Gender ChestPain RestBP Chol Fbs RestECG AHD
1  63      1    typical   145  233    1      2   No
2  67      1 asymptomatic 160  286    0      2   Yes
3  67      1 asymptomatic 120  229    0      2   Yes
4  37      1 nonanginal  130  250    0      0   No
5  41      0 nontypical  130  204    0      2   No
6  56      1 nontypical  120  236    0      0   No
7  62      0 asymptomatic 140  268    0      2   Yes
8  57      0 asymptomatic 120  354    0      0   No

```

Environment pane shows the loaded objects:

- Global Environment
 - heart_data: 49 obs. of 8 variables
 - naive_model: List of 8
 - new_data: chr [1:49, 1:7] "63" "67" "67" "..."
 - testData: 12 obs. of 8 variables
 - trainData: 37 obs. of 8 variables

User Library pane shows installed packages:

- abind: Combine Multidimensional Arrays (1.4-5)
- ape: Analyses of Phylogenetics and Evolution (5.4-1)
- arules: Mining Association Rules and Frequent Itemsets (1.6-6)
- arulesViz: Visualizing Association Rules and Frequent Itemsets (1.3-3)
- assertthat: Easy Pre and Post Assertions (0.2.1)
- backports: Reimplementations of Functions Introduced Since R-3.0.0 (1.2.0)
- base64enc: Tools for base64 encoding (0.1-3)
- BH: Boost C++ Header Files (1.72.0-3)
- bit: A Class for Vectors of 1-Bit Booleans (1.1-15.2)
- bit64: A S3 Class for Vectors of 64bit Integers (0.9-7)
- bitops: Bitwise Operations (1.0-6)

Age	Gender	ChestPain	RestBP	Chol	Fbs	RestEC	
						G	AHD
63	1	typical	145	233	1	2	No
67	1	asymptomatic	160	286	0	2	Yes
67	1	asymptomatic	120	229	0	2	Yes
37	1	nonanginal	130	250	0	0	No
41	0	nontypical	130	204	0	2	No
56	1	nontypical	120	236	0	0	No
62	0	asymptomatic	140	268	0	2	Yes
57	0	asymptomatic	120	354	0	0	No
63	1	asymptomatic	130	254	0	2	Yes
53	1	asymptomatic	140	203	1	2	Yes
57	1	asymptomatic	140	192	0	0	No
56	0	nontypical	140	294	0	2	No
56	1	nonanginal	130	256	1	2	Yes
44	1	nontypical	120	263	0	0	No
52	1	nonanginal	172	199	1	0	No
57	1	nonanginal	150	168	0	0	No
48	1	nontypical	110	229	0	0	Yes

		asymptomatic					
54	1	c	140	239	0	0	No
48	0	nonanginal	130	275	0	0	No
49	1	nontypical	130	266	0	0	No
64	1	typical	110	211	0	2	No
58	0	typical	150	283	1	2	No
58	1	nontypical	120	284	0	2	Yes
58	1	nonanginal	132	224	0	2	Yes
		asymptomatic					
60	1	c	130	206	0	2	Yes
50	0	nonanginal	120	219	0	0	No
58	0	nonanginal	120	340	0	0	No
66	0	typical	150	226	0	0	No
		asymptomatic					
43	1	c	150	247	0	0	No
		asymptomatic					
40	1	c	110	167	0	2	Yes

Age	Gender	ChestPain	RestBP	Chol	Fbs	RestECG	AHD
63	1	typical	145	233	1	2	No
67	1	asymptomatic	160	286	0	2	Yes
67	1	asymptomatic	120	229	0	2	Yes
37	1	nonanginal	130	250	0	0	No
41	0	nontypical	130	204	0	2	No
56	1	nontypical	120	236	0	0	No
62	0	asymptomatic	140	268	0	2	Yes
57	0	asymptomatic	120	354	0	0	No
63	1	asymptomatic	130	254	0	2	Yes
53	1	asymptomatic	140	203	1	2	Yes
57	1	asymptomatic	140	192	0	0	No
56	0	nontypical	140	294	0	2	No
56	1	nonanginal	130	256	1	2	Yes
44	1	nontypical	120	263	0	0	No
52	1	nonanginal	172	199	1	0	No
57	1	nonanginal	150	168	0	0	No
48	1	nontypical	110	229	0	0	Yes
54	1	asymptomatic	140	239	0	0	No
48	0	nonanginal	130	275	0	0	No
49	1	nontypical	130	266	0	0	No
64	1	typical	110	211	0	2	No
58	0	typical	150	283	1	2	No
58	1	nontypical	120	284	0	2	Yes
58	1	nonanginal	132	224	0	2	Yes
60	1	asymptomatic	130	206	0	2	Yes
50	0	nonanginal	120	219	0	0	No

58	0	nonanginal	120	340	0	0	No
66	0	typical	150	226	0	0	No
43	1	asymptomatic	150	247	0	0	No
40	1	asymptomatic	110	167	0	2	Yes
Age	Gender	ChestPain	RestBP	Chol	Fbs	RestECG	AHD
63	1	typical	145	233	1	2	No
67	1	asymptomatic	160	286	0	2	Yes
67	1	asymptomatic	120	229	0	2	Yes
37	1	nonanginal	130	250	0	0	No
41	0	nontypical	130	204	0	2	No
56	1	nontypical	120	236	0	0	No
62	0	asymptomatic	140	268	0	2	Yes
57	0	asymptomatic	120	354	0	0	No
63	1	asymptomatic	130	254	0	2	Yes
53	1	asymptomatic	140	203	1	2	Yes
57	1	asymptomatic	140	192	0	0	No
56	0	nontypical	140	294	0	2	No
56	1	nonanginal	130	256	1	2	Yes
44	1	nontypical	120	263	0	0	No
52	1	nonanginal	172	199	1	0	No
57	1	nonanginal	150	168	0	0	No
48	1	nontypical	110	229	0	0	Yes
54	1	asymptomatic	140	239	0	0	No
48	0	nonanginal	130	275	0	0	No
49	1	nontypical	130	266	0	0	No
64	1	typical	110	211	0	2	No
58	0	typical	150	283	1	2	No
58	1	nontypical	120	284	0	2	Yes
58	1	nonanginal	132	224	0	2	Yes
60	1	asymptomatic	130	206	0	2	Yes
50	0	nonanginal	120	219	0	0	No
58	0	nonanginal	120	340	0	0	No
66	0	typical	150	226	0	0	No
43	1	asymptomatic	150	247	0	0	No
40	1	asymptomatic	110	167	0	2	Yes

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

Environment History Connections Tutorial

Global Environment

Data

Object	Class	Size
heart	data.frame	5 obs. of 8 variables
heart_data	data.frame	49 obs. of 8 variables
naive_model	list	List of 8
new_data	chr	chr [1:49, 1:7] "63" "67" "67" ...
testData	data.frame	12 obs. of 8 variables

Files Plots Packages Help Viewer

Install Update

User Library

Package	Description	Version
abind	Combine Multidimensional Arrays	1.4-5
ape	Analyses of Phylogenetics and Evolution	5.4-1
arules	Mining Association Rules and Frequent Itemsets	1.6-6
arulesViz	Visualizing Association Rules and Frequent Itemsets	1.3-3
assertthat	Easy Pre and Post Assertions	0.2.1
backports	Reimplementations of Functions Introduced Since R-3.0.0	1.2.0
base64enc	Tools for base64 encoding	0.1-3
BH	Boost C++ Header Files	1.72.0-3
bit	A Class for Vectors of 1-Bit Booleans	1.1-15.2
bit64	A S3 Class for Vectors of 64bit Integers	0.9-7
bitops	Bitwise Operations	1.0-6

```

4 #DATA IMPORTING
5 heart_data<- read.csv("C:\\Program Files\\R\\R-3.3.2\\datasets\\Heart.csv", header = TRUE)
6 print(heart_data)
7
8 dim(heart_data)
9 length(heart_data)
10 heart<- heart_data[1:5,]
11 heart
12 heart_data[,5]
13 print(heart_data$Age)
14
15 summary(heart_data)
16
17 # Preprocessing
18 barplot(table(heart_data$AHD),
19          main="Heart Attack", col="blue")
20

```

Console

```

> heart<- heart_data[1:5,]
> heart
  Age Gender ChestPain RestBP Chol Fbs RestECG AHD
1  63      1    typical   145  233      1      2   No
2  67      1 asymptomatic  160  286      0      2   Yes
3  67      1 asymptomatic  120  229      0      2   Yes
4  37      1 nonanginal   130  250      0      0   No
5  41      0 nontypical   130  204      0      2   No
> heart_data[,5]
[1] 233 286 229 250 204
[23] 284 224 206 219 340
[45] 330 230 175 243 417

```

DATA PREPROCESSING

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

Environment History Connections Tutorial

Global Environment

Data

Object	Class	Size
heart	data.frame	5 obs. of 8 variables
heart_data	data.frame	49 obs. of 8 variables
naive_model	list	List of 8
new_data	chr	chr [1:49, 1:7] "63" "67" "67" ...
testData	data.frame	12 obs. of 8 variables

Files Plots Packages Help Viewer

Zoom Export

Heart Attack

Category	Frequency
No	233
Yes	16

```

11 heart
12 heart_data[,5]
13 print(heart_data$Age)
14
15 summary(heart_data)
16
17 # Preprocessing
18 barplot(table(heart_data$AHD),
19          main="Heart Attack", col="blue")
20
21
22 levels(heart_data$AHD) = c("No disease", "Disease")
23 levels(heart_data$Gender) = c("female", "male", "")
24 mosaicplot(heart_data$Gender ~ heart_data$AHD,
25            main="Heart Attack by Gender", shade=FALSE, col=TRUE,
26            xlab="Gender", ylab="Heart disease")
27

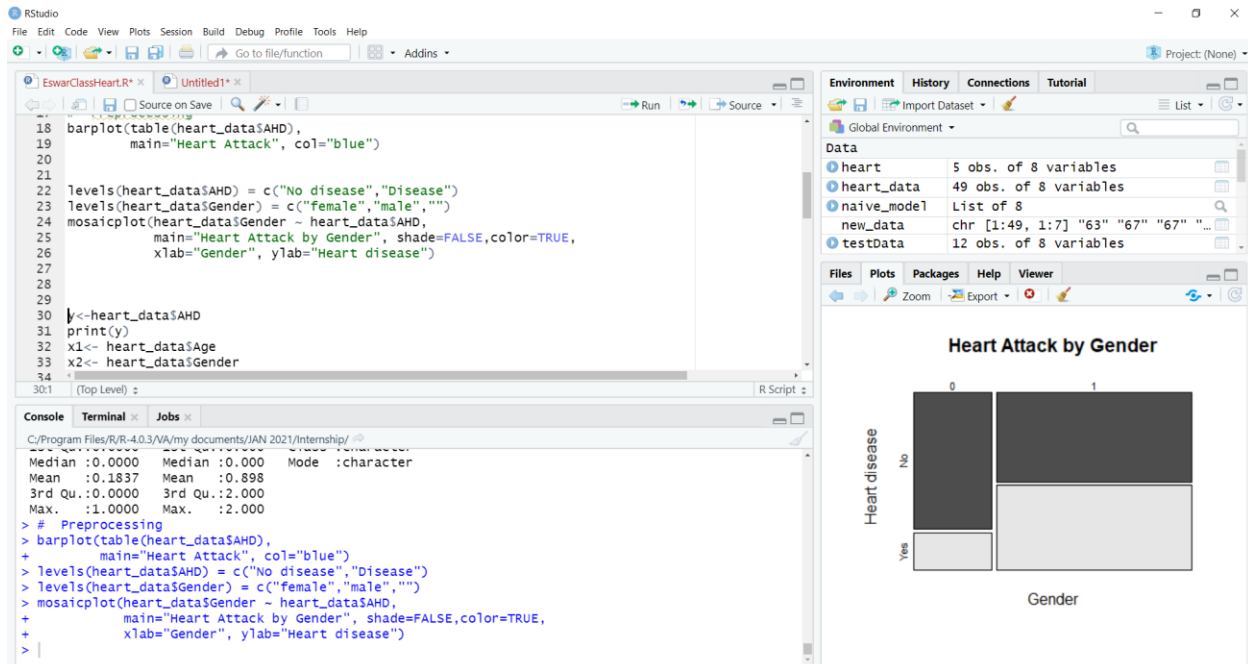
```

Console

```

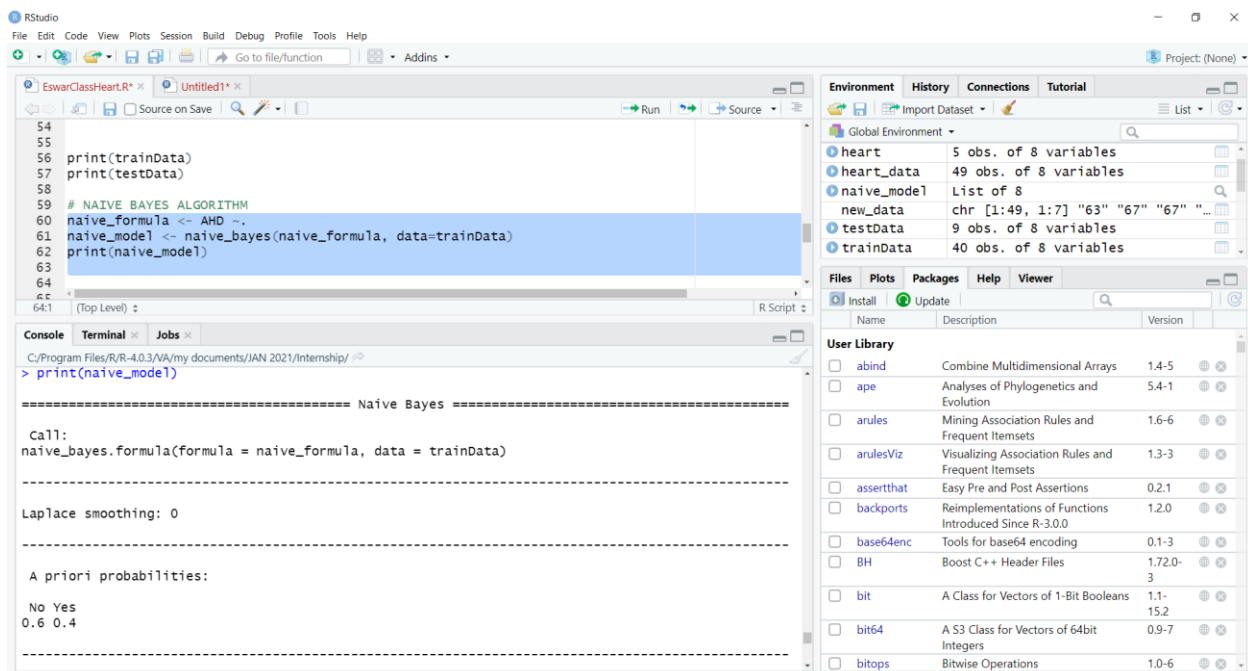
3rd Qu.: 61.00 3rd Qu.: 1.0000 3rd Qu.: 145.0 3rd Qu.: 275.0
Max.: 71.00 Max.: 1.0000 Max.: 172.0 Max.: 417.0
Fbs RestECG AHD
Min.: 0.0000 Min.: 0.000 Length: 49
1st Qu.: 0.0000 1st Qu.: 0.000 Class: character
Median: 0.0000 Median: 0.000 Mode: character
Mean: 0.1837 Mean: 0.898
3rd Qu.: 0.0000 3rd Qu.: 2.000
Max.: 1.0000 Max.: 2.000
> # Preprocessing
> barplot(table(heart_data$AHD),
+          main="Heart Attack", col="blue")
>

```



NAIVE BAYES ALGORITHM

Naive Bayes Classification Model



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

Environment History Connections Tutorial

Global Environment

trainData 40 obs. of 8 variables

Values

ind	naive_formula	AHD	result	testPred
int [1:49]	1 1 1 1 1 1 1 1 1 1 ...	~ .	Factor w/ 2 levels "No", "Yes": 2	Factor w/ 2 levels "No", "Yes": 1 ...

Files Plots Packages Help Viewer

Install Update

User Library

Name	Description	Version
abind	Combine Multidimensional Arrays	1.4-5
ape	Analyses of Phylogenetics and Evolution	5.4-1
arules	Mining Association Rules and Frequent Itemsets	1.6-6
arulesViz	Visualizing Association Rules and Frequent Itemsets	1.3-3
assertthat	Easy Pre and Post Assertions	0.2.1
backports	Reimplementations of Functions Introduced Since R-3.0.0	1.2.0
base64enc	Tools for base64 encoding	0.1-3
BH	Boost C++ Header Files	1.72.0-3
bit	A Class for Vectors of 1-Bit Booleans	1.1-15.2
bit64	A S3 Class for Vectors of 64bit Integers	0.9-7
bitops	Bitwise Operations	1.0-6

```

64 #Evaluation on test data
65 testData1 = testData[, -c(8)]
66
67
68 testPred <- predict(naive_model, newdata = testData1)
69 print(testPred)
70
71 # Naive model Prediction
72
73 new_patient<-data.frame(Age = 50,Gender=1,ChestPain = "asymptomatic",RestBP= 140,Chol= 268,Fbs=
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

```

Console Terminal Jobs

C:/Program Files/R/R-4.0.3/NA/my documents/JAN 2021/Internship/

```

Chol      No      Yes
mean 255.29167 264.06250
sd   55.83787  44.28313

# ... and 2 more tables

> #Evaluation on test data
> testData1 = testData[, -c(8)]
>
> testPred <- predict(naive_model, newdata = testData1)
> print(testPred)
[1] No Yes No No No No No Yes No
Levels: No Yes
>

```

Naive Bayes Prediction Model

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

Environment History Connections Tutorial

Global Environment

trainData 40 obs. of 8 variables

Values

ind	naive_formula	AHD	result	testPred
int [1:49]	1 1 1 1 1 1 1 1 1 1 ...	~ .	Factor w/ 2 levels "No", "Yes": 2	Factor w/ 2 levels "No", "Yes": 1 ...

Files Plots Packages Help Viewer

Install Update

User Library

Name	Description	Version
abind	Combine Multidimensional Arrays	1.4-5
ape	Analyses of Phylogenetics and Evolution	5.4-1
arules	Mining Association Rules and Frequent Itemsets	1.6-6
arulesViz	Visualizing Association Rules and Frequent Itemsets	1.3-3
assertthat	Easy Pre and Post Assertions	0.2.1
backports	Reimplementations of Functions Introduced Since R-3.0.0	1.2.0
base64enc	Tools for base64 encoding	0.1-3
BH	Boost C++ Header Files	1.72.0-3
bit	A Class for Vectors of 1-Bit Booleans	1.1-15.2
bit64	A S3 Class for Vectors of 64bit Integers	0.9-7
bitops	Bitwise Operations	1.0-6

```

69 print(testPred)
70
71 # Naive model Prediction
72
73 new_patient<-data.frame(Age = 50,Gender=1,ChestPain = "asymptomatic",RestBP= 140,Chol= 268,Fbs=
74
75
76 result <- predict(naive_model,new_patient)
77 print(result)
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

```

Console Terminal Jobs

C:/Program Files/R/R-4.0.3/NA/my documents/JAN 2021/Internship/

```

> #Evaluation on test data
> testData1 = testData[, -c(8)]
>
> testPred <- predict(naive_model, newdata = testData1)
> print(testPred)
[1] No Yes No No No No No Yes No
Levels: No Yes
> # Naive model Prediction
>
> new_patient<-data.frame(Age = 50,Gender=1,ChestPain = "asymptomatic",RestBP= 140,Chol= 268,Fbs=
0,RestECG= 2)
>
> result <- predict(naive_model,new_patient)
> print(result)
[1] Yes
Levels: No Yes
>

```

APPENDIX B

SAMPLE CODING

```
library(naivebayes)

#DATA IMPORTING
heart_data<- read.csv("C:\\Program Files\\R\\R-3.3.2\\datasets\\Heart.csv",
  header = TRUE)
print(heart_data)
dim(heart_data)
length(heart_data)
heart<- heart_data[1:5,]
heart
heart_data[,5]
print(heart_data$Age)
summary(heart_data)

# Preprocessing
barplot(table(heart_data$AHD),
  main="Heart Disease", col="blue")

levels(heart_data$AHD) = c("No disease","Disease")
levels(heart_data$Gender) = c("female","male","")
mosaicplot(heart_data$Gender ~ heart_data$AHD,
  main="Heart Disease by Gender", shade=FALSE,color=TRUE,
  xlab="Gender", ylab="Heart disease")
```

```
y<-heart_data$AHD
print(y)
x1<- heart_data$Age
x2<- heart_data$Gender
hist(x1)
x3<- heart_data$ChestPain
x4<- heart_data$RestBP
hist(x4, main = "Blood Preasure",xlab="BP Value", ylab="No.of Patients",col
     = "yellow")
x5<- heart_data$Chol
hist(x5)
x6<- heart_data$Fbs
print(x6)
x7<-heart_data$RestECG
print(x7)
new_data <- cbind(x1,x2,x3,x5,x6,x7,y)
print(heart_data)

# Sampling the datasets
ind <- sample(2, nrow(heart_data), replace=TRUE, prob=c(0.8, 0.2))
trainData <- heart_data[ind==1,]
testData <- heart_data[ind==2,]
print(trainData)
print(testData)
```

```
# NAIVE BAYES ALGORITHM
```

```
naive_formula <- AHD ~.
```

```
naive_model <- naive_bayes(naive_formula, data=trainData)
```

```
print(naive_model)
```

```
#Evaluation on test data
```

```
testData1 = testData[, -c(8)]
```

```
testPred <- predict(naive_model, newdata = testData1)
```

```
print(testPred)
```

```
# Naive model Prediction
```

```
new_patient<-data.frame(Age = 50,Gender=1,ChestPain =
```

```
  "asymptomatic",RestBP= 140,Chol= 268,Fbs= 0,RestECG= 2)
```

```
result <- predict(naive_model,new_patient)
```

```
print(result)
```