

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
Technology Degree in Information Technology

By

SAMRITHA.A.R
REG. NO. 39120101



DEPARTMENT OF INFORMATION TECHNOLOGY

SCHOOL OF COMPUTING

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600119, TAMILNADU

NOVEMBER 202



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 **SAMRITHA.A.R (Reg. No: 39120101)** who carried out the project entitled "**Predicting Student Placement Using Naive Bayes Algorithm**" under my supervision from June 2021 to November 2021.

Internal Guide

Dr. R.M. Gomathi, M.Tech., Ph.D.,

Head of the Department

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I, **SAMRITHA. A.R** hereby declare that the project report entitled **Predicting Student Placement Using Naive Bayes Algorithm** done by me under the guidance of **Dr. R.M. Gomathi** is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology 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. S. Vigneshwari, M.E., Ph.D. and Dr. L. Lakshmanan, M.E., Ph.D., Heads of the Department of Information Technology 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 **Dr. R.M. Gomathi, M.Tech., Ph.D.,** 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 Information Technology and Engineering** who were helpful in many ways for the completion of the project.

TRAINING CERTIFICATE

CERTIFICATE OF COMPLETION

This certifies that

SAMRITHA A R 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

One of the biggest challenges that higher learning institutions face today is to improve the placement performance of students. Admission and reputation of institutions mainly depends on placements. Hence all institutions strive to strengthen placement department. All student's dream is to obtain a job offer in their hands before they leave their college. Educational institutes look for more efficient technology that assist better management and support decision making procedures or assist them to set new strategies. Placement predictor system could predict the type of company a student has chances to be placed. It helps the students to have an idea about where they stand and what to be done to obtain a good placement. In this Project, student placement prediction model has been proposed that uses Naive Bayes algorithm to identify the placement chance of students. This model helps the placement cell within an organization to identify the prospective students and pay attention to and improve their technical as well as interpersonal skills. With this model, students can put in more hardwork for getting placed in to the companies that belong to higher hierarchies.

TABLE OF CONTENTS		
CHAPTER	TITLE	PAGE NO
	ABSTRACT	
	LIST OF FIGURES	
	LIST OF TABLES	
1	INTRODUCTION	1
2	SYSTEM ANALYSIS	3
	2.1 EXISTING SYSTEM	3
	2.2 PROPOSED SYSTEM	3
3	SYSTEM SPECIFICATION	4
	3.1 HARDWARE SPECIFICATION	4
	3.2 SOFTWARE SPECIFICATION	4
4	SYSTEM IMPLEMENTATION	6
	4.1 SYSTEM ARCHITECTURE	6
	4.2 MODULES	7
	4.2.1 IMPORTING PLACEMENT DATASET	7
	4.2.2 DATA PREPROCESSING	7
	4.2.3 MODEL GENERATION USING NAIVEBAYES ALGORITHM	8
	4.2.4 PREDICTION USING NAÏVE BAYES MODEL	11
5	SYSTEM DESIGN	12
	5.1 DATA FLOW DIAGRAM	12
	5.1.1 LEVEL 0	12
	5.1.2 LEVEL 1	13
	5.1.3 LEVEL 2	13
	5.1.4 LEVEL 3	14

6	CONCLUSION	15
	REFERENCES	16
	APPENDIX A: SNAP SHOTS	17
	APPENDIX B: SOURCE CODE	22

LIST OF FIGURES		
FIGURE NO	FIGURE NAME	PAGE NO
4.1	ARCHITECTURE DIAGRAM	6
5.1.1	LEVEL 0	12
5.1.2	LEVEL 1	13
5.1.3	LEVEL 2	14
5.1.4	LEVEL 3	14

LIST OF TABLES		
TABLE NO	TABLE NAME	PAGE NO
4.2.3.1	SAMPLE DATASET	10
4.2.3.2	FREQUENCY TABLE	10
4.2.3.3	LIKELIHOOD TABLE	11

CHAPTER 1

INTRODUCTION

Majority of students in higher education join a course for securing a good job. Therefore, taking a wise career decision regarding the placement after completing a particular course is crucial in a student's life. It is a well-known fact all around the world that admission of students in an educational institution depends on the placements. Placement is one of the factors considered for determining the quality of the institution. Hence every institution strives hard to provide better placements to their students.

According to statistics 1.5 million engineers are graduating each year in India. The demand and need for qualified graduates in the field of IT industry is rising day by day. But most of the students are unaware about the need of the IT industry. The number of student graduates who satisfies the requirements and quality of the company is very low. The placement cell and professors of an institute should take proper steps in order to produce a set of students suitable for each company requirements.

Campus placement performs an essential position in each academic group in supporting college students to achieve their goals. Managing placement and education information in any large corporation is tough because the scholar quantity is high; in such a situation differentiation and classification on different categories will become tedious. All college students dream to get employment to provide by their hands earlier than they depart their college.

In this project, Naive Bayes algorithms is used to predict the student placement using factors, such as gender, ssc_p, ssc_b, hsc_p, hsc_b, hsc_s, degree_p, degree_t, workex, etest_p, specialisation, mba_p, status etc. A placement probability predictor allows college students to possess a thought regarding wherever

they stand and what to be performed to obtain a decent placement. The result acquired from this can assist the students to better understand their weak regions to work upon. Working in those regions will permit college students to obtain a better number of placements in a college institution.

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 these techniques are not precise. Using some programming languages and data mining tools, student placement dataset is classified. There is an increase in need to manage and understand data.

2.2 PROPOSED SYSTEM

Student placement dataset is collected and analyzed to predict whether the students get placed or not. Placement opportunities are abundant in market due to many emerging technologies. Skill development of student should be done in placement point of view. 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 placement data, efficient prediction accuracy is achieved by delivering the result of the placement chance of students

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

CHAPTER 4

SYSTEM IMPLEMENTATION

4.1 SYSTEM ARCHITECTURE

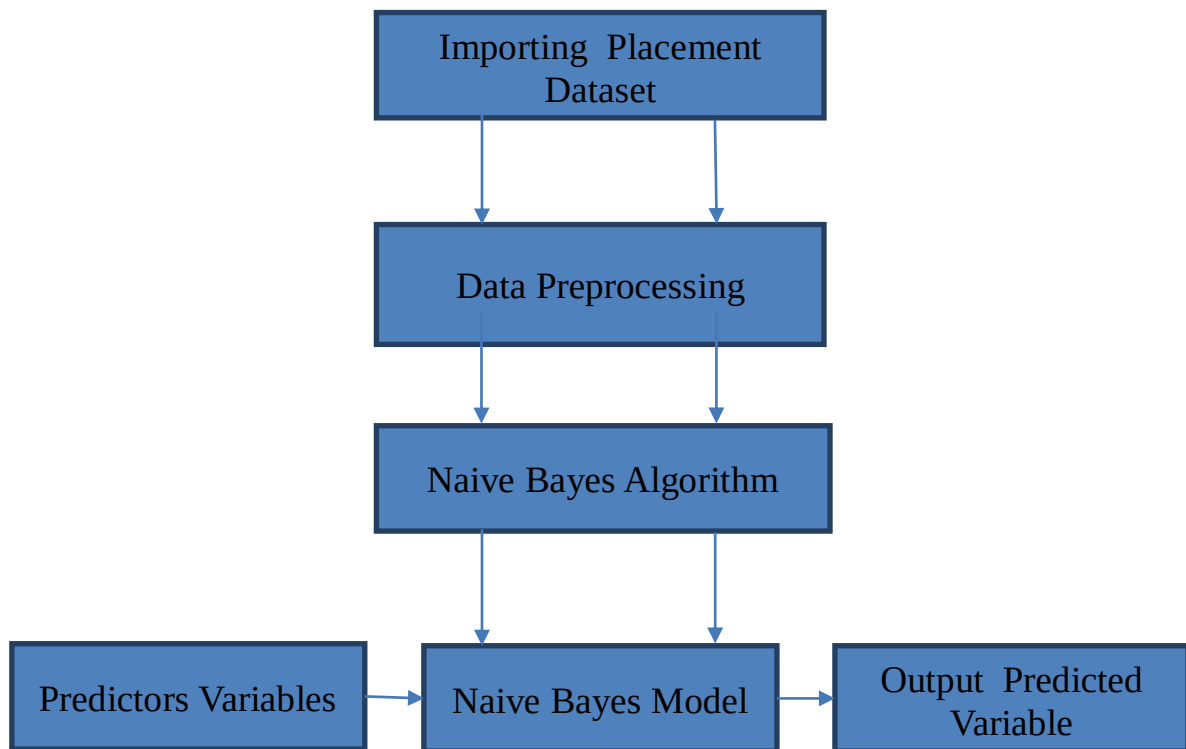


Fig.4.1 Architecture Diagram

The fig 4.1 shows the architectural diagram of this project defines the flow of data for predicting the student placement. 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 formatof 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 student will get the placement in the company or not.

4.2 MODULES

There are four Modules

- Importing Placement dataset
- Data Preprocessing
- Model Generation Using Naive Bayes Algorithm
- Prediction using Naive Bayes Model

4.2.1 IMPORTING PLACEMENT 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 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:

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

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

Table 4.2.3.1 Sample Dataset

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								

Table 4.2.3.2 Frequency Table

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								

Table 4.2.3.3 Likelihood Table

Sunny, Cool, High, Strong

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

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

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

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

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

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

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

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

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 predict the probability of a student getting placed in a company.

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

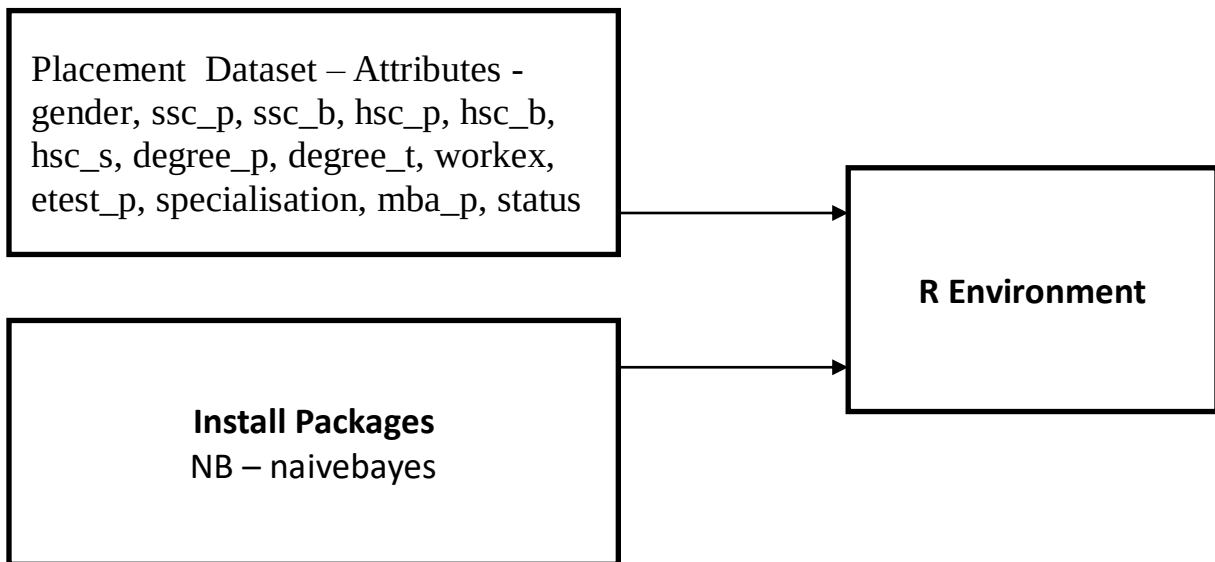


Fig 5.1.1 LEVEL0

In the fig 5.1.1, Data is imported in to R environment for analysis. The input dataset is placement 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

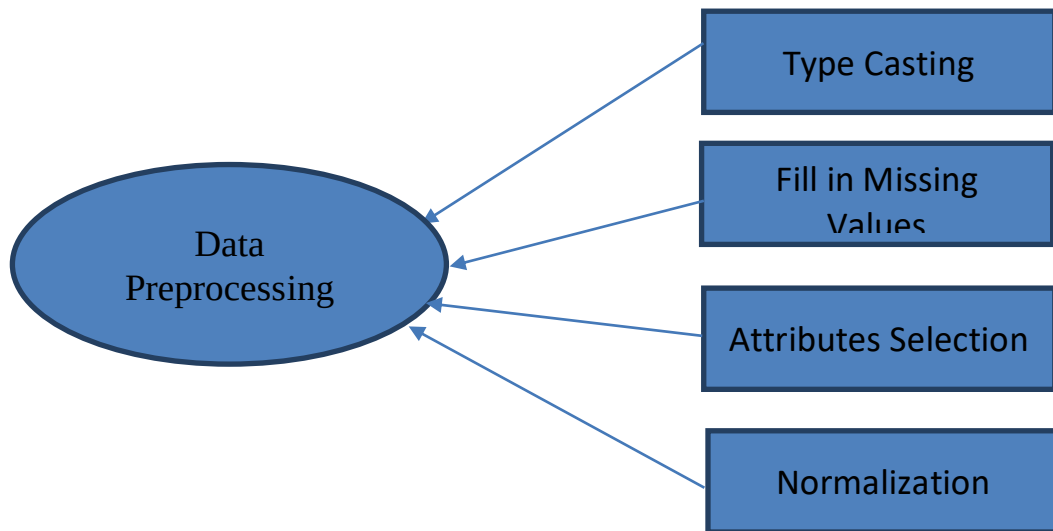


Fig 5.1.2 LEVEL1

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.

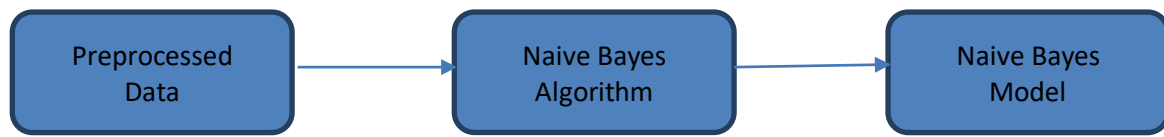


Fig 5.1.3 LEVEL2

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 a student will get the placement in the company or not. Thus, prediction process is implemented successfully. Using this model classification is also achieved by passing more number of placement datasets, it will classify the students as two groups. One group is the set of students who will get the placement and the other group is the set of students who will not get the placement in the company.

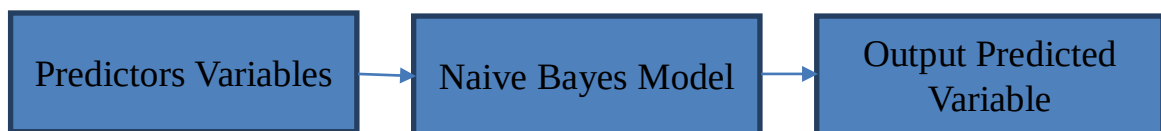


Fig 5.1.4 LEVEL3

CHAPTER 6

CONCLUSION

This project presents a recommendation system that uses Data Science algorithm for classifying and predicting, the student placement. The Naive Bayes Data Science algorithm is used for classifying and predicting the placement of students in a company. The students in pre-final and final years of their course can also use this system to know their individual placement status that they are most likely to achieve. This application endows the targeted group of students to boost their placement probability. Thus, this project concludes, Naive Bayes Algorithm performs better and faster when compared to other statistical techniques. This project can be further extended using some more data science algorithm and deep learning algorithms to improve the performance of prediction system.

REFERENCES

1. TadiAravind, " A Comparative Study on Machine Learning Algorithms for Predicting the Placement Information of Under Graduate Students", IEEE International conference on I-SMAC,2019
2. Kanika Thakur, "Ensemble method to predict impact of student intelligent quotient and academic achievement on placement", IEEE International Conference on Intelligent Engineering and Management, 2021
3. Jinny Rhee; Sheri Sheppard, " Supporting student career development of undergraduate engineering", IEEE Frontiers in Education Conference (FIE), 2017.
4. Jaime Ramírez Castillo; AldabbaghGhadah, "Towards improved student placement and preparation methods on information technologies post-secondary education", IEEE Federated Conference on Computer Science and Information Systems, 2013.
5. Liya Claire Joy, " A Review on Student Placement Chance Prediction" IEEE International Conference on Advanced Computing & Communication Systems, 2019.

APPENDIX A

SNAP SHOTS

SAMPLE DATASETS

gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisat	mba_p	status	salary
M	67	Others	91	Others	Commerce	58	Sci&Tech	No	55	Mkt&HR	58.8	Placed	270000
M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Placed	200000
M	65	Central	68	Central	Arts	64	Comm&M	No	75	Mkt&Fin	57.8	Placed	250000
M	56	Central	52	Central	Science	52	Sci&Tech	No	66	Mkt&HR	59.43	Not Placed	
M	85.8	Central	73.6	Central	Commerce	73.3	Comm&M	No	96.8	Mkt&Fin	55.5	Placed	425000
M	55	Others	49.8	Others	Science	67.25	Sci&Tech	Yes	55	Mkt&Fin	51.58	Not Placed	
F	46	Others	49.2	Others	Commerce	79	Comm&M	No	74.28	Mkt&Fin	53.29	Not Placed	
M	82	Central	64	Central	Science	66	Sci&Tech	Yes	67	Mkt&Fin	62.14	Placed	252000
M	73	Central	79	Central	Commerce	72	Comm&M	No	91.34	Mkt&Fin	61.29	Placed	231000
M	58	Central	70	Central	Commerce	61	Comm&M	No	54	Mkt&Fin	52.21	Not Placed	
M	58	Central	61	Central	Commerce	60	Comm&M	Yes	62	Mkt&HR	60.85	Placed	260000
M	69.6	Central	68.4	Central	Commerce	78.3	Comm&M	Yes	60	Mkt&Fin	63.7	Placed	250000
F	47	Central	55	Others	Science	65	Comm&M	No	62	Mkt&HR	65.04	Not Placed	
F	77	Central	87	Central	Commerce	59	Comm&M	No	68	Mkt&Fin	68.63	Placed	218000
M	62	Central	47	Central	Commerce	50	Comm&M	No	76	Mkt&HR	54.96	Not Placed	
F	65	Central	75	Central	Commerce	69	Comm&M	Yes	72	Mkt&Fin	64.66	Placed	200000
M	63	Central	66.2	Central	Commerce	65.6	Comm&M	Yes	60	Mkt&Fin	62.54	Placed	300000
F	55	Central	67	Central	Commerce	64	Comm&M	No	60	Mkt&Fin	67.28	Not Placed	
F	63	Central	66	Central	Commerce	64	Comm&M	No	68	Mkt&HR	64.08	Not Placed	
M	60	Others	67	Others	Arts	70	Comm&M	Yes	50.48	Mkt&Fin	77.89	Placed	236000
M	62	Others	65	Others	Commerce	66	Comm&M	No	50	Mkt&HR	56.7	Placed	265000
F	79	Others	76	Others	Commerce	85	Comm&M	No	95	Mkt&Fin	69.06	Placed	393000
F	69.8	Others	60.8	Others	Science	72.23	Sci&Tech	No	55.53	Mkt&HR	68.81	Placed	360000
F	77.4	Others	60	Others	Science	64.74	Sci&Tech	Yes	92	Mkt&Fin	63.62	Placed	300000
M	76.5	Others	97.7	Others	Science	78.86	Sci&Tech	No	97.4	Mkt&Fin	74.01	Placed	360000
F	52.58	Others	54.6	Central	Commerce	50.2	Comm&M	Yes	76	Mkt&Fin	65.33	Not Placed	
M	71	Others	79	Others	Commerce	66	Comm&M	Yes	94	Mkt&Fin	57.55	Placed	240000
M	63	Others	67	Others	Commerce	66	Comm&M	No	68	Mkt&HR	57.69	Placed	265000
M	76.76	Others	76.5	Others	Commerce	67.5	Comm&M	Yes	73.35	Mkt&Fin	64.15	Placed	350000
M	62	Central	67	Central	Commerce	58	Comm&M	No	77	Mkt&Fin	51.29	Not Placed	

DATA IMPORTING

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project (None)

Environment History Connections Tutorial

Files Plots Packages Help Viewer

Install Update

Name Description Version

System Library

- ☒ base The R Base Package 4.1.1
- ☐ boot Bootstrap Functions (Originally by Angelo Canty for S) 1.3-28
- ☐ class Functions for Classification 7.3-19
- ☐ cluster "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al. 2.1.2
- ☐ codetools Code Analysis Tools for R 0.2-18
- ☐ compiler The R Compiler Package 4.1.1
- ☒ datasets The R Datasets Package 4.1.1
- ☐ foreign Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ... 0.8-81
- ☒ graphics The R Graphics Package 4.1.1
- ☒ grDevices The R Graphics Devices and Support for Colours and Fonts 4.1.1
- ☐ grid The Grid Graphics Package 4.1.1
- ☐ KernSmooth Functions for Kernel Smoothing *Simon J. Wand & Jones (1995)* 2.23-20

```
1 library(naivebayes)
2
3
4 placement<- read.csv("c:\\Program Files\\R\\R-4.1.1\\datasets\\Placement.csv", header = TRUE)
5 print(placement)
6
7
8 dim(placement)
9 length(placement)
10 placement[6,]
11 placement[,c(10,11,12,13)]
12 placement[1:10,]
13 placement[,9]
14 str(placement)
15
```

65:1 (Top Level)

R Script

Console Terminal Jobs

R 4.1.1 - C:/Users/admin/Desktop/PT-1/

```
type "q()" to quit R.
> library(naivebayes)
naivebayes 0.9.7 loaded
> placement<- read.csv("c:\\Program Files\\R\\R-4.1.1\\datasets\\Placement.csv", header = TRUE)
> print(placement)
```

	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	workex	etest_p
1	M	67.00	others	91.00	others	Commerce	58.00	Sci&Tech	No 55.00
2	M	79.33	Central	78.33	others	Science	77.48	Sci&Tech	Yes 86.50
3	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No 75.00
4	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No 66.00
5	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No 96.80
6	M	55.00	others	49.80	others	Science	67.25	Sci&Tech	Yes 55.00
7	F	46.00	others	49.20	others	Commerce	79.00	Comm&Mgmt	No 74.28
8	M	82.00	Central	64.00	Central	Science	66.00	Sci&Tech	Yes 67.00
9	M	73.00	Central	79.00	Central	Commerce	72.00	Comm&Mgmt	No 91.34
10	M	58.00	Central	70.00	Central	Commerce	61.00	Comm&Mgmt	No 54.00
11	M	58.00	Central	61.00	Central	Commerce	60.00	Comm&Mgmt	Yes 62.00
12	M	69.60	Central	68.40	Central	Commerce	78.30	Comm&Mgmt	Yes 60.00
13	F	47.00	Central	55.00	others	Science	65.00	Comm&Mgmt	No 62.00
14	F	77.00	Central	87.00	Central	Commerce	59.00	Comm&Mgmt	No 68.00

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project (None)

Environment History Connections Tutorial

Files Plots Packages Help Viewer

Install Update

Name Description Version

System Library

- ☒ base The R Base Package 4.1.1
- ☐ boot Bootstrap Functions (Originally by Angelo Canty for S) 1.3-28
- ☐ class Functions for Classification 7.3-19
- ☐ cluster "Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al. 2.1.2
- ☐ codetools Code Analysis Tools for R 0.2-18
- ☐ compiler The R Compiler Package 4.1.1
- ☒ datasets The R Datasets Package 4.1.1
- ☐ foreign Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ... 0.8-81
- ☒ graphics The R Graphics Package 4.1.1
- ☒ grDevices The R Graphics Devices and Support for Colours and Fonts 4.1.1
- ☐ grid The Grid Graphics Package 4.1.1
- ☐ KernSmooth Functions for Kernel Smoothing *Simon J. Wand & Jones (1995)* 2.23-20

```
1 library(naivebayes)
2
3
4 placement<- read.csv("c:\\Program Files\\R\\R-4.1.1\\datasets\\Placement.csv", header = TRUE)
5 print(placement)
6
7
8 dim(placement)
9 length(placement)
10 placement[6,]
11 placement[,c(10,11,12,13)]
12 placement[1:10,]
13 placement[,9]
14 str(placement)
15
```

65:1 (Top Level)

R Script

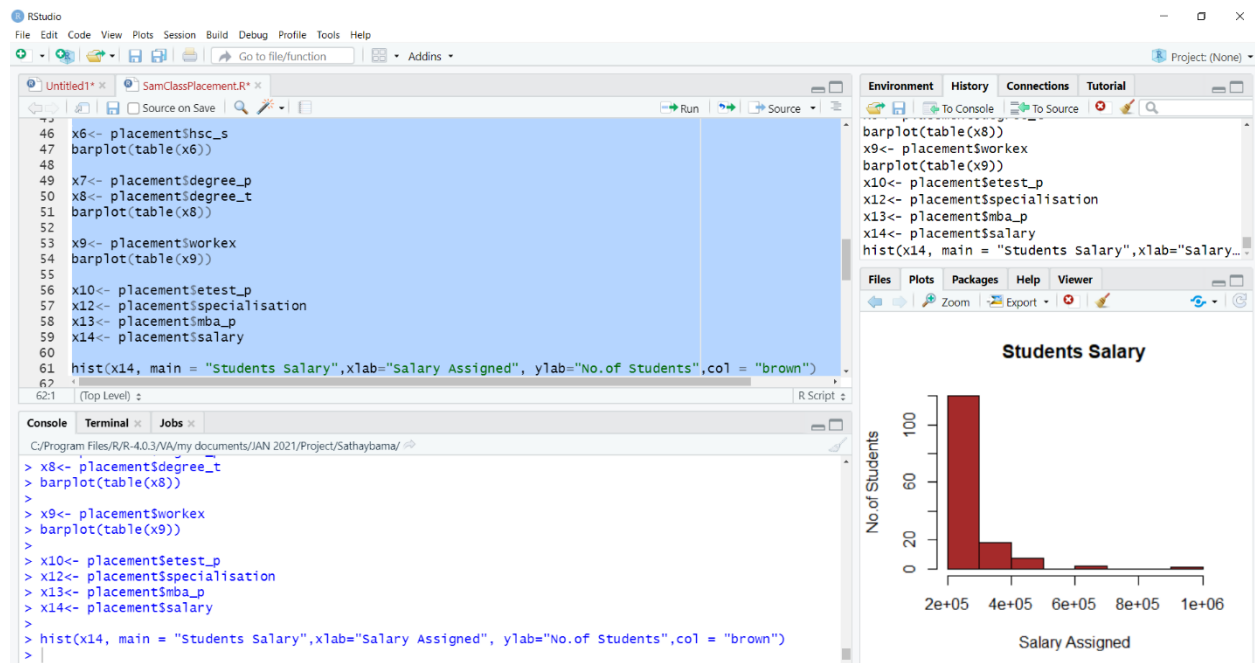
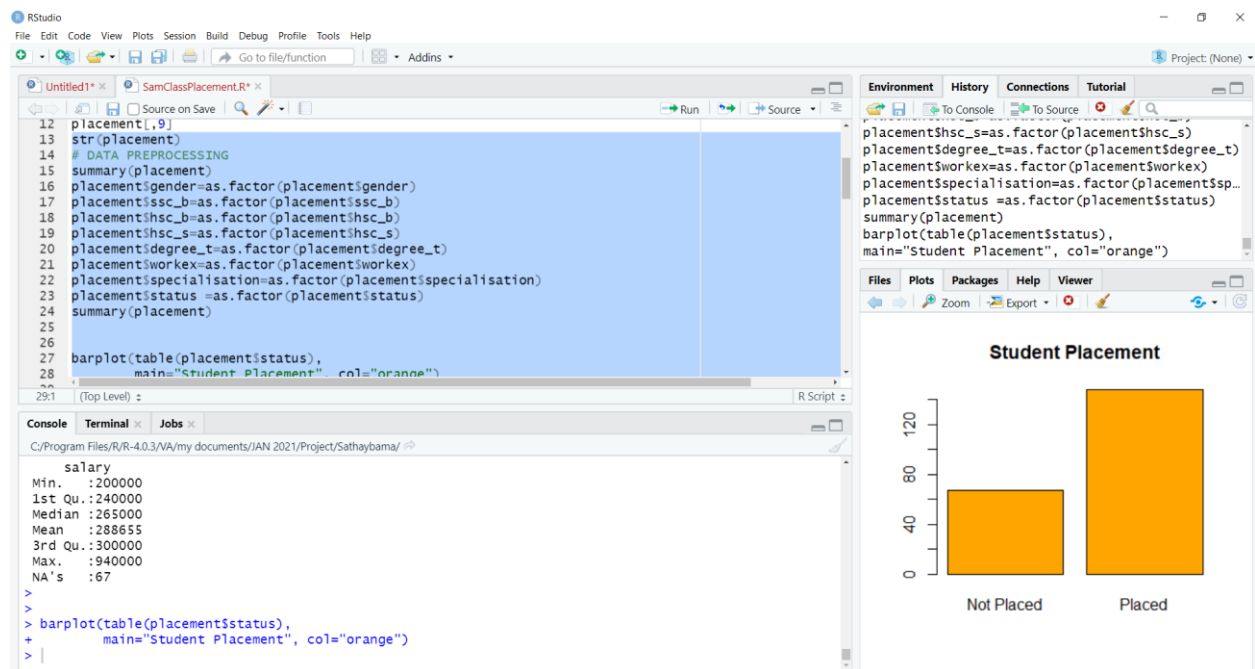
Console Terminal Jobs

R 4.1.1 - C:/Users/admin/Desktop/PT-1/

```
type "q()" to quit R.
> library(naivebayes)
naivebayes 0.9.7 loaded
> placement<- read.csv("c:\\Program Files\\R\\R-4.1.1\\datasets\\Placement.csv", header = TRUE)
> print(placement)
```

	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	workex	etest_p
14	F	77.00	Central	87.00	Central	Commerce	59.00	Comm&Mgmt	No 68.00
15	M	62.00	Central	47.00	Central	Commerce	50.00	Comm&Mgmt	No 76.00
16	F	65.00	Central	75.00	Central	Commerce	69.00	Comm&Mgmt	Yes 72.00
17	M	63.00	Central	66.20	Central	Commerce	65.60	Comm&Mgmt	Yes 60.00
18	F	55.00	Central	67.00	Central	Commerce	64.00	Comm&Mgmt	No 60.00
19	F	63.00	Central	66.00	Central	Commerce	64.00	Comm&Mgmt	No 68.00
20	M	60.00	others	67.00	others	Arts	70.00	Comm&Mgmt	Yes 94.00
21	M	62.00	others	65.00	others	Commerce	66.00	Comm&Mgmt	No 50.00
22	F	79.00	others	76.00	others	Commerce	85.00	Comm&Mgmt	No 95.00
23	F	69.80	others	60.80	others	Science	72.23	Sci&Tech	No 55.53
24	F	77.40	others	60.00	others	Science	64.74	Sci&Tech	Yes 92.00
25	M	76.50	others	97.70	others	Science	78.86	Sci&Tech	No 97.40
26	F	52.58	others	54.60	Central	Commerce	50.20	Comm&Mgmt	Yes 76.00
27	M	71.00	others	79.00	others	Commerce	66.00	Comm&Mgmt	Yes 94.00
28	M	63.00	others	67.00	others	Commerce	66.00	Comm&Mgmt	No 68.00
29	M	76.76	others	76.50	others	Commerce	67.50	Comm&Mgmt	Yes 73.35
30	M	62.00	Central	67.00	Central	Commerce	58.00	Comm&Mgmt	No 77.00
31	F	64.00	Central	73.50	Central	Commerce	73.00	Comm&Mgmt	No 52.00
32	F	67.00	Central	53.00	Central	Science	65.00	Sci&Tech	No 64.00
33	F	61.00	Central	81.00	Central	Commerce	66.40	Comm&Mgmt	No 50.89

DATA PREPROCESSING



NAIVE BAYES ALGORITHM

Naive Bayes Classification Model

RStudio interface showing the Naive Bayes Classification Model workflow.

```
67 ind <- sample(2, nrow(placement), replace=TRUE, prob=c(0.8, 0.2))
68 trainData <- placement[ind==1,]
69 testData <- placement[ind==2,]
70
71 print(trainData)
72 print(testData)
73
74 naive_model <- naive_bayes(status ~ gender+ssc_p+ssc_b+hsc_p+hsc_b+hsc_s+degree_p+degree_t+
75   workex+etest_p+specialisation+mba_p, data=trainData, type="C-classifi
76   plot(naive_model)]
77
78 print(naive_model)
```

Console output:

```
A priori probabilities:
Not Placed    Placed
0.3371429    0.6628571

Tables:
::: gender (Bernoulli)
```

Environment pane shows installed packages:

Name	Description	Versi...
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-
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

NAIVE BAYES ALGORITHM

NB Model Inference – MBA_P Vs PLACEMENT

RStudio interface showing the Naive Bayes Model Inference workflow.

```
72 print(trainData)
73 print(testData)
74
75 naive_model <- naive_bayes(status ~ gender+ssc_p+ssc_b+hsc_p+hsc_b+hsc_s+degree_p+degree_t+
76   workex+etest_p+specialisation+mba_p, data=trainData, type="C-classifi
77   print(naive_model)
78 plot(naive_model)]
79
80 #predict on test data
81 testData1 = testData[,-c(13,14)]
82
83 testPred <- predict(naive_model, newdata = testData1)
```

Console output:

```
hsc_b    Not Placed    Placed
Central 0.4237288 0.4310345
Others 0.5762712 0.5689655

# ... and 7 more tables

> plot(naive_model)
>
```

Environment pane shows installed packages:

Name	Description	Versi...
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-
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

Plot of Density vs mba_p:

mba_p	Density (Not Placed)	Density (Placed)
52	0.005	0.005
56	0.015	0.015
60	0.025	0.035
64	0.020	0.045
68	0.010	0.030
72	0.005	0.015
76	0.000	0.005

NB Model Inference – Placement Vs Specialization

The RStudio interface shows a script for training a Naive Bayes model and plotting a confusion matrix. The script defines training and test data, fits a model, and predicts on new data. The Environment pane shows the model object. The Plots pane displays a confusion matrix plot with 'specialisation' on the y-axis and 'placement' on the x-axis. The plot shows a 2x2 matrix with red and yellow cells.

```

72 print(trainData)
73 print(testData)
74
75
76 naive_model1 <- naive_bayes(status ~ gender+ssc_p+ssc_b+hsc_p+hsc_b+hsc_s+degree_p+degree_t+
77   workex+etest_p+specialisation+mba_p, data=trainData, type="C-classifi
78 print(naive_model1)
79 plot(naive_model1)
80
81 #predict on test data
82 testData1 = testData[,-c(13,14)]
83 testData1
84 testPred <- predict(naive_model1, newdata = testData1)
85
86

```

Environment:

```

naive_model1 <- naive_bayes(status ~ gender+ssc_...
workex+etest_p+specialisation+mba_p, data=train...
print(naive_model1)
plot(naive_model1)
naive_model1 <- naive_bayes(status ~ gender+ssc_...
workex+etest_p+specialisation+mba_p, data=train...
print(naive_model1)
plot(naive_model1)

```

Plots:

Console:

```

hsc_b    Not Placed    Placed
Central  0.4237288    0.4310345
Others   0.5762712    0.5689655

```

... and 7 more tables

```

> plot(naive_model1)
>

```

Naive Bayes Prediction Model

The RStudio interface shows a script for predicting using a Naive Bayes model. The script defines test data, fits a model, and predicts on new data. The Environment pane shows the model object. The Plots pane displays a confusion matrix plot with 'specialisation' on the y-axis and 'placement' on the x-axis. The plot shows a 2x2 matrix with red and yellow cells.

```

84 testPred <- predict(naive_model1, newdata = testData1)
85 print(testPred)
86 print(testData1$status)
87
88
89 a<-data.frame(gender= "M",ssc_p =76.50, ssc_b = "Others", hsc_p =97.70, hsc_b = "Others", hsc_s =
90   degree_p = 78.86, degree_t = "Sci&Tech", workex= "No",etest_p= 97.40, specialisation
91   mba_p =74.01)
92
93 result <- predict(naive_model1,a)
94 print(result)
95
96

```

Environment:

```

testPred <- predict(naive_model1, newdata = test...
print(testPred)
print(testData1$status)
a<-data.frame(gender= "M",ssc_p =76.50, ssc_b = ...
degree_p = 78.86, degree_t = "Sci&Tech", workex=...
mba_p =74.01)
result <- predict(naive_model1,a)
print(result)

```

Plots:

Console:

```

[1] Placed Placed Placed Not Placed Placed Placed Placed Not Placed
[25] Placed Placed Placed Placed Not Placed Not Placed Placed Not Placed
[33] Not Placed Placed Placed Placed Placed Placed Placed Placed
Levels: Not Placed Placed
>
> a<-data.frame(gender= "M",ssc_p =76.50, ssc_b = "Others", hsc_p =97.70, hsc_b = "Others", hsc_s = "Sci
+   degree_p = 78.86, degree_t = "Sci&Tech", workex= "No",etest_p= 97.40, specialisation = "M
+   mba_p =74.01)
+
> result <- predict(naive_model1,a)
> print(result)
[1] Placed
Levels: Not Placed Placed
>

```

APPENDIX B

SAMPLE CODING

```
library(naivebayes)
```

```
#DATA IMPORTING
```

```
placement<- read.csv("C:\\Program Files\\R\\R-4.1.1\\datasets\\Placement.csv",
```

```
  header = TRUE)
```

```
print(placement)
```

```
dim(placement)
```

```
length(placement)
```

```
placement[6,]
```

```
placement[,c(10,11,12,13)]
```

```
placement[1:10,]
```

```
placement[,9]
```

```
str(placement)
```

```
# DATA PREPROCESSING
```

```
summary(placement)
```

```
placement$gender=as.factor(placement$gender)
```

```
placement$ssc_b=as.factor(placement$ssc_b)
```

```
placement$hsc_b=as.factor(placement$hsc_b)
```

```
placement$hsc_s=as.factor(placement$hsc_s)
```

```
placement$degree_t=as.factor(placement$degree_t)
```

```
placement$workex=as.factor(placement$workex)
```

```
placement$specialisation=as.factor(placement$specialisation)
```

```
placement$status =as.factor(placement$status)
```

```
summary(placement)
```

```
barplot(table(placement$status),
```

```
main="Student Placement", col="orange")
```

```
y<-placement$status
```

```
print(y)
```

```
x1<- placement$gender
```

```
barplot(table(placement$gender),
```

```
main="Students Gender", col="Yellow")
```

```
x2<- placement$ssc_p
```

```
hist(x2)
```

```
x3<- placement$ssc_b
```

```
barplot(table(x3))
```

```
x4<- placement$hsc_p
```

```
hist(x4, main = "SSC Pass Percentage",xlab="Pass Percentage",
```

```
ylab="No.ofStudents",col = "red")
```

```
x5<- placement$hsc_b
```

```
barplot(table(x5))
```

```
x6<- placement$hsc_s
```

```
barplot(table(x6))
```

```
x7<- placement$degree_p
```

```
x8<- placement$degree_t
```

```
barplot(table(x8))
```

```
x9<- placement$workex
```

```
barplot(table(x9))
```

```
x10<- placement$etest_p
```

```
x12<- placement$specialisation
```

```
x13<- placement$mba_p
```

```
x14<- placement$salary
```

```
hist(x14, main = "Students Salary",xlab="Salary Assigned",
```

```
ylab="No.ofStudents",col = "brown")
```

```
new_data<- cbind(x3,x4,x5,x6,x7,x8,y)
print(placement)
```

```
# Sampling of Datasets
```

```
ind<- sample(2, nrow(placement), replace=TRUE, prob=c(0.8, 0.2))
trainData<- placement[ind==1,]
testData<- placement[ind==2,]
print(trainData)
print(testData)
```

```
#Naive Bayes Model Generation
```

```
naive_model<- naive_bayes(status ~
  gender+ssc_p+ssc_b+hsc_p+hsc_b+hsc_s+degree_p+degree_t+
  workex+etest_p+specialisation+mba_p, data=trainData, type="C-classification")
```

```
print(naive_model)
plot(naive_model)
```

```
#predict on test data
```

```
testData1 = testData[,-c(13,14)]
testData1
testPred<- predict(naive_model, newdata = testData1)
print(testPred)
print(testData$status)
```

```
#Naive Bayes Prediction Model
```

```
a<-data.frame(gender= "M",ssc_p =76.50, ssc_b = "Others", hsc_p =97.70,
  hsc_b = "Others", hsc_s ="Science",
  degree_p = 78.86, degree_t ="Sci&Tech", workex= "No",etest_p= 97.40,
  specialisation ="Mkt&Fin",
```

```
mba_p =74.01)
```

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

```
print(result)
```