

PROFESSIONAL TRAINING REPORT

at

**Sathyabama Institute of Science and
Technology (Deemed to be University)**

Submitted in partial fulfillment of requirements for the award
of Bachelor of Technology Degree in Information Technology

By-

NANDANA. S. NAIR

Reg. No: 39120074



**DEPARTMENT OF INFORMATION TECHNOLOGY
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



DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the Bonafide work of **NANDANA. S. NAIR (Reg. No: 39120074)** who carried out the project entitled "**Movie Success Rate Classification Using Naive Bayes Algorithm**" under my supervision from February 2022 to April 2022.

Internal Guide

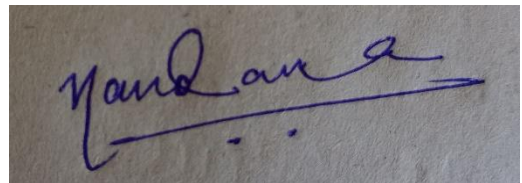
Dr. P. Jeyanthi

DECLARATION

I, **NANDANA. S. NAIR** hereby declare that the project report entitled **Movie Success Rate Classification Using Naive Bayes Algorithm** done by me under the guidance of **Dr. P. Jeyanthi** is submitted in partial fulfillment of requirements for the award of Bachelor of Technology Degree in Information Technology.

DATE: 21/04/22

PLACE: Chennai



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. R. Subhashini M.E., Ph.D., Head of the Department of Information Technology** 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. Jeyanthi M.E., Ph.D.** 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

NANDANA. S. NAIR

of Sathyabama Institute of Science and Technology

Has successfully completed Training Program on

Data Science Using R

From 23rd Feb 2022 to 6th Apr 2022



Certificate ID: VA1707318



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

ABSTRACT

Films have always been the most popular medium of entertainment at all times. The film industry faces a challenge that is related to profit and loss. A risk always exists related to the movie income. The risk is whether, a film will be able to earn the targeted revenue or not according to the huge budget it invests. The number of movie production rate is growing day by day worldwide and the movie makers invest highly in the movie industry. In such a scenario, it is very important to evaluate the movie status before its release time. The prediction of the success of a movie is very essential to the film industry. In this Project, Naive bayes algorithm is used to create Movie Prediction Model, which will predict the success of movie which is essential to the film industry. The accurate prediction of whether a movie will be hit or flop, prior to its arrival at the box office will benefit the production houses.

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	13
	4.1 SYSTEM ARCHITECTURE	13
	4.2 MODULES	14
5	SYSTEM DESIGN	18
	5.1 DATA FLOW DIAGRAM	18
	5.1.1 LEVEL 0	18
	5.1.2 LEVEL 1	19
	5.1.3 LEVEL 2	20
	5.1.4 LEVEL 3	20
6	CONCLUSION	21
	REFERENCES	22
	APPENDIX A: SNAP SHOTS	23
	APPENDIX B: SOURCE CODE	28

CHAPTER 1

INTRODUCTION

Movies continue to be a major source of entertainment in any country. A movie making a billion- dollar industry. Film making involves good story, screen-writing, casting, direction, sound recording and many other activities. Movie industry produces hundreds of movies every year of different genres such as animation, war, comedy, thriller horror etc. Movie has a distinct way to inspire each one of us. People can learn a lot from this medium.

A comedy helps to forgot the sorrows, a sci-fi helps to think big, a biopic helps to achieve the dreams and so on. A great movie is a crowd puller and the entire family enjoy together that translates into revenue generation from purchase of tickets, snacks etc. hence movie industry contributes significantly to any country's economy. When a movie is a blockbuster hit, profits are huge, but when the movie flops the losses are huge. Both have direct positive and negative impact. Movie industry worldwide produces a large number of movies per year. However very few movies are a success and are ranked high.

Due to rapid digitization and emergence of social media the movie industry is growing by leaps and bounds. The average number of movies

produced per year is greater than 1000. So, to make the movie profitable, it becomes a matter of concern that the movie succeeds. As movie industry is growing too fast day by day, there are now a considerable amount of data available on the internet, which makes it an exciting field for data analysis.

Data Science techniques can be a boon in this regard. Using Data Science, many kinds of knowledge can be discovered such as association rules, classifications, clustering and prediction. For this project data have been collected from various sources. The main objective of this project, is to predict the movie success in prior.

By Using, Naive Bayes Algorithm, datasets are analysed to extract the intelligent pattern of predicting the success or failure of the movies using the factors such as whether the his studies, using factors, such Marketing expense, Production expense, Multiplex coverage, Budget, Movie_length, Lead_Actor_Rating, Lead_Actress_rating, Director_rating, Producer_rating, Critic_rating, Trailer_views, Time_taken, Twitter_hastags, Avg_age_actors, Num_multiplex, Collection, Start_Tech_Oscar.

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, Movie dataset is classified. There is an increase in need to manage and understand data.

2.2 PROPOSED SYSTEM

Movie dataset is collected and analyzed to predict whether a movie will be hit or flop. For prediction Data Science techniques of Naive Bayes classification 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 Movie dataset and efficient prediction accuracy is achieved by delivering the result of whether the movie will be a hit or flop after analyzing different parameters. These factors can be considered by the movie maker to decide their financial roadmap and also evaluate their comfort zone on taking risk.

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-4.1.3
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 it is 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

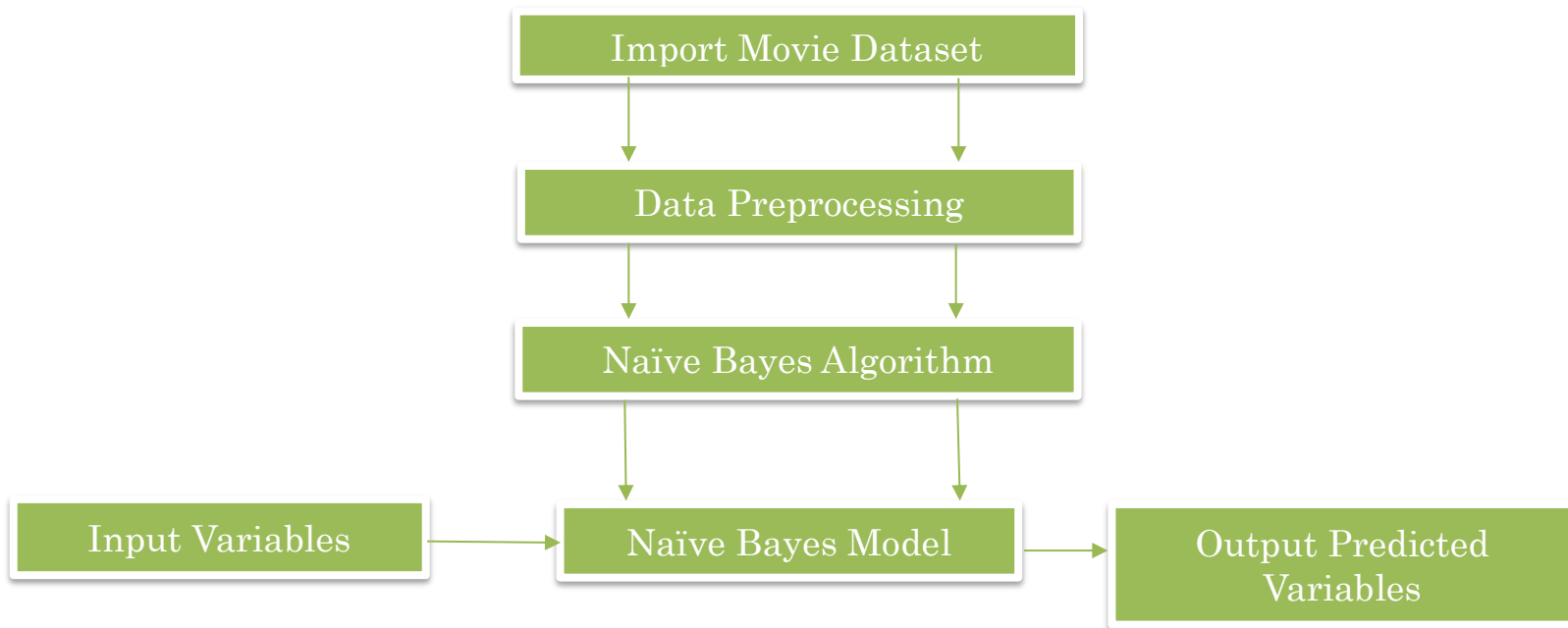


Fig.4.1 Architecture Diagram

The fig 4.1 shows the architectural diagram of this project defines the flow of data for predicting the Movie success. The first step is data importing. Data has 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 are the techniques which are used to transform the various forms of data into a 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 highly affect 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 Movie will be hit or flop.

4.2 MODULES

There are four Modules

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

4.2.1 IMPORTING MOVIE 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 is 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 collected in various formats is transformed to a common format that enhances the comparison process, allows for collaborative research and largescale 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.

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 predictor variables, the Naive Bayes Model will generate the output of whether the movie will be flop or hit.

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

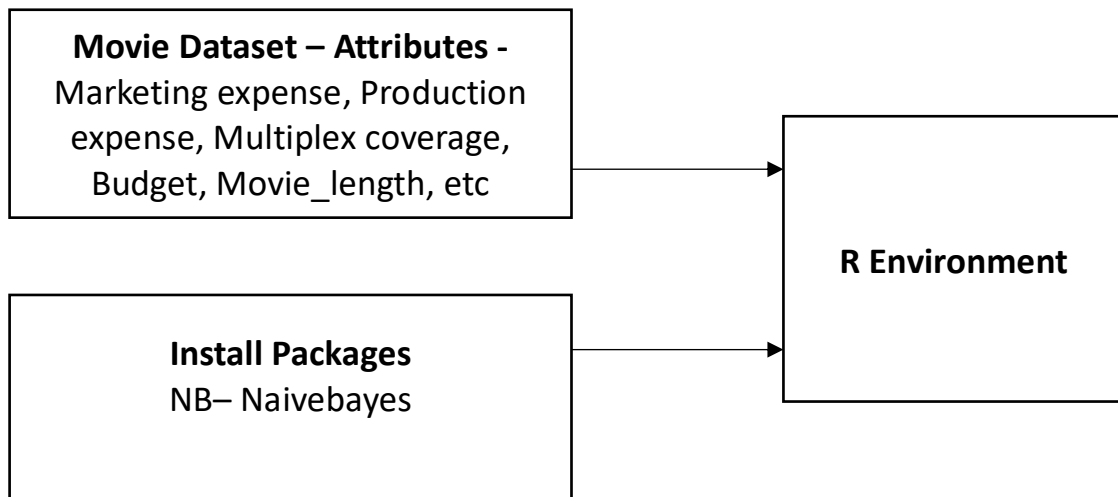


Fig 5.1.1 LEVEL 0

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

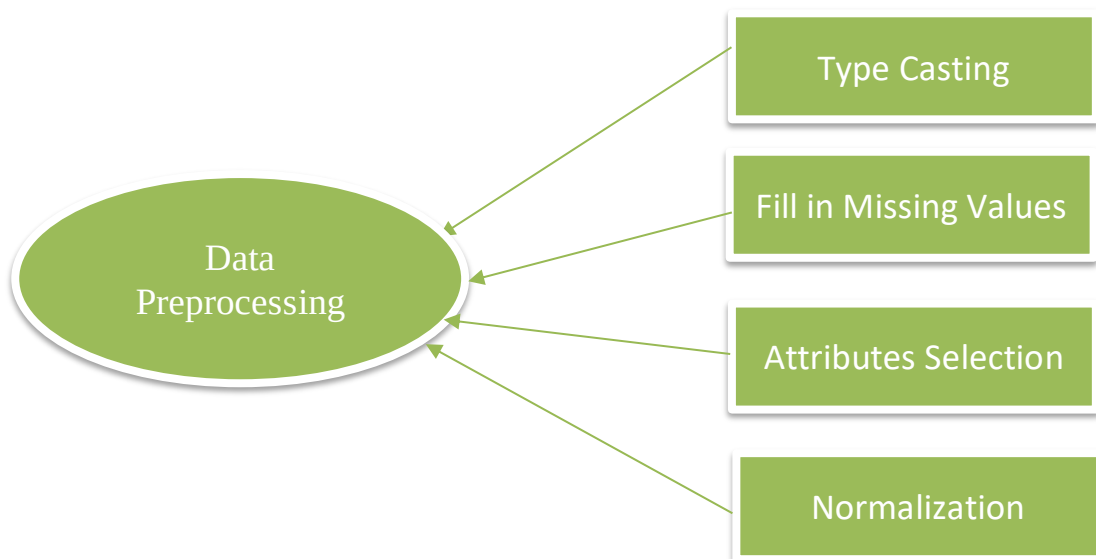


Fig 5.1.2 LEVEL 1

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.



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 movie will be hit or flop. Thus, prediction process is implemented and generated the movie result successfully.



Fig 5.1.4 LEVEL 3

CHAPTER 6

CONCLUSION

The film industry has grown into a multi-billionaire industry in terms of entertainment. The success of the film industry depends on the criteria that how much profit a movie would make which gives the tag of a 'hit' or a 'flop'. Predicting the success is guided by various factors like genre, date of release, actors, net gross and many more. There are many approaches used for prediction, in this project Naive Bayes is used. By using Naive bayes algorithm, Prediction model have been created, which helps to predict the movie success. Understanding the stakes involved with a movie release that can affect its success or a failure, before-hand can be a great step towards the expansion of the film industry business.

REFERENCES

1. Narayana Darapaneni, Christopher Bellarmine, Anwesh Reddy P, "Movie Success Prediction Using ML ", IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2020.
2. Ankit; M. Lakshmi, K. Aditya Shastry, Aditya Sandilya, Rahul S, "A comparative analysis of Machine Learning approaches for Movie Success Prediction", Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2020.
3. Anand Bhawe; Himanshu Kulkarni; Vinay Biramane, "Role of different factors in predicting movie success ", International Conference on Pervasive Computing (ICPC), 2015.
4. R. Dhir and A. Raj, "Movie Success Prediction using Machine Learning Algorithms and their Comparison", First International Conference on Secure Cyber Computing and Communication (ICSCCC), pp. 385-390, 2018
5. Nahid Quader; Md. Osman Gani; Dipankar Chaki; Md. Haider Ali "A machine learning approach to predict movie box-office success", 20th International Conference of Computer and Information Technology (ICCIT), 2018.

APPENDIX A

SNAP SHOTS

SAMPLE DATASETS

Marketing	Production	Multiplex	Budget	Movie_len	Lead_Actr	Lead_Actr	Director_r
20.1264	59.62	0.462	36524.13	138.7	7.825	8.095	7.91
20.5462	69.14	0.531	35668.66	152.4	7.505	7.65	7.44
20.5458	69.14	0.531	39912.68	134.6	7.485	7.57	7.495
20.6474	59.36	0.542	38873.89	119.3	6.895	7.035	6.92
21.381	59.36	0.542	39701.59	127.7	6.92	7.07	6.815
20.597	59.36	0.542	35718.65	132.2	6.89	7.1	6.885
21.7658	70.74	0.476	33396.66	140.1	7.065	7.265	7.15
22.891	70.74	0.476	34285.46	169.6	6.98	7.075	6.875
24.2248	70.74	0.476	31280.21	173.5	6.91	7.075	6.85
23.4008	70.74	0.476	33352.22	159.4	6.665	6.725	6.575
24.4978	70.74	0.476	35424.24	167.8	6.675	6.845	6.725
22.3494	70.74	0.476	33380	156.4	6.865	7.035	6.745
21.8756	70.74	0.476	32713.4	112.5	7.12	7.43	7.21
32.5952	71.28	0.462	33046.7	135.3	7.64	7.705	7.535
32.7592	71.28	0.462	33863.28	158	7.7	7.9	7.76
32.5478	71.28	0.462	32407.87	130	7.7	7.825	7.64
41.0786	71.28	0.462	32968.93	102.8	7.67	7.805	7.74
35.684	71.28	0.462	33274.45	155.2	7.72	7.925	7.82
36.0542	71.28	0.462	30308.08	110.1	8.1	8.24	8.07
34.516	71.28	0.462	31813.49	143	8.01	8.175	8
45.0358	71.28	0.462	30941.35	171.6	8.035	8.205	7.955
37.0408	71.28	0.462	33135.58	162.7	7.945	8.14	7.83
44.6494	71.28	0.462	34118.81	165.2	7.91	8.01	7.845
39.7686	71.28	0.462	32291.22	173.5	7.825	8.01	7.935
35.0052	71.28	0.462	32907.82	167.6	7.655	7.965	7.64
36.8108	71.28	0.462	31102.45	159.2	7.655	7.835	7.64

DATA IMPORTING

RStudio interface showing the initial steps of data importing. The script editor contains the following code:

```
1 library(naivebayes)
2
3
4 Movie_data <- read.csv("c:\\Program Files\\R\\R-4.1.2\\Datasets\\movie.csv", header =
5 print(Movie_data)
6
7 head(Movie_data)
8 dim(Movie_data)
9 length(Movie_data)
10
```

The console output shows the first 10 rows of the loaded data:

	Marketing.expense	Production.expense	Multiplex.coverage	Budget	Movie_length
1	20.1264	59.62	0.462	36524.12	138.7
2	20.5462	69.14	0.531	35668.65	152.4
3	20.5458	69.14	0.531	39912.68	134.6
4	20.6474	59.36	0.542	38873.89	119.3
5	21.3810	59.36	0.542	39701.58	127.7
6	20.5970	59.36	0.542	35718.65	132.2
7	21.7658	70.74	0.476	33396.66	140.1
8	22.8910	70.74	0.476	34285.46	169.6
9	24.2248	70.74	0.476	31280.21	173.5
10	23.4008	70.74	0.476	33352.22	159.4

RStudio interface showing further data manipulation. The script editor contains the following code:

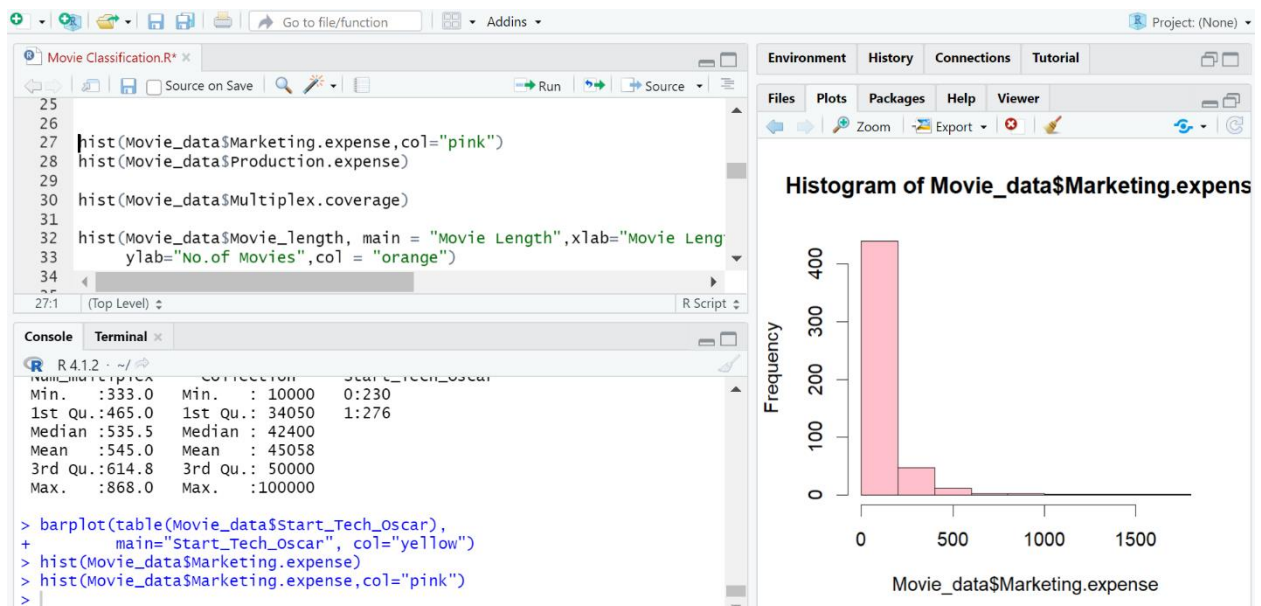
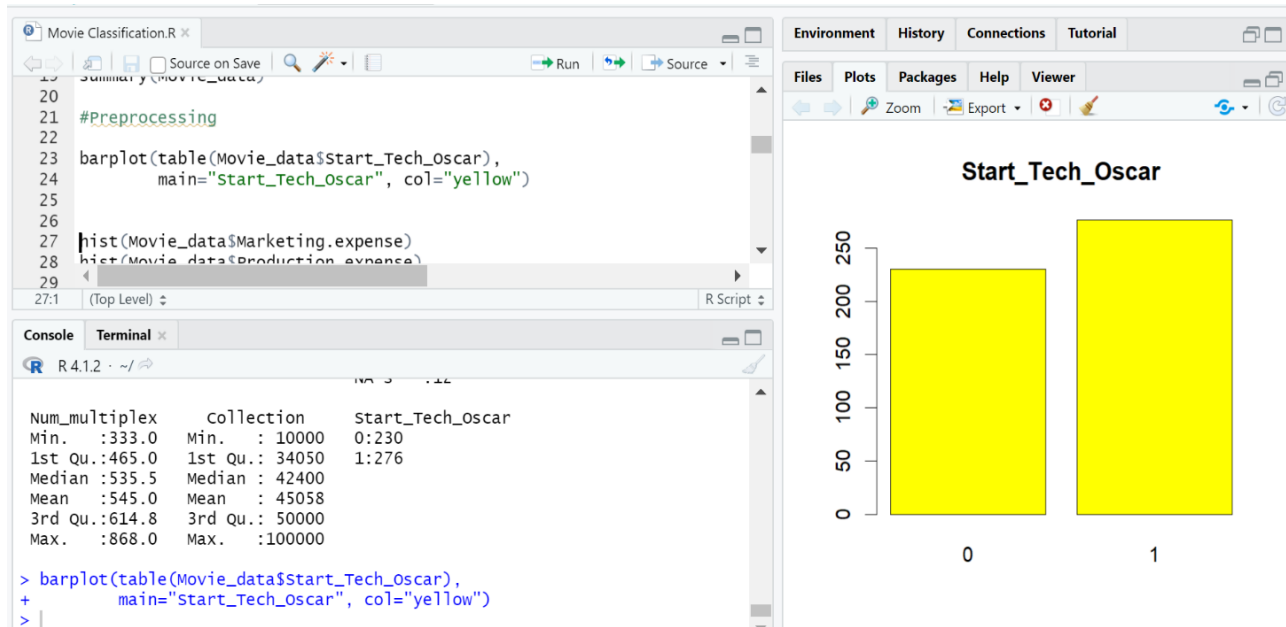
```
10 length(Movie_data)
11 Movie_data[10:20,12:15]
12 Movie_data[,11]
13 Movie_data[25,]
14
15 print(Movie_data$Movie_length)
16
17 summary(Movie_data)
18 Movie_data$Start_Tech_Oscar=as.factor(Movie_data$Start_Tech_Oscar)
19
20 summary(Movie_data)
21
22
```

The console output shows the summary of the data and the new factor variable:

Lead_Actress_rating	Director_rating	Producer_rating	Critic_rating	Trailer_views	Time_taken	Twitter_hashtags
Min. : 4.035	Min. : 3.840	Min. : 4.030	Min. : 6.600	Min. : 212912	Min. : 0.0	Min. : 201.2
1st Qu.: 7.504	1st Qu.: 7.296	1st Qu.: 7.508	1st Qu.: 7.200	1st Qu.: 409128	1st Qu.: 132.3	1st Qu.: 223.8
Median : 8.495	Median : 8.312	Median : 8.465	Median : 7.960	Median : 462460	Median : 160.0	Median : 254.4
Mean : 8.186	Mean : 8.020	Mean : 8.191	Mean : 7.811	Mean : 449861	Mean : 157.4	Mean : 260.8
3rd Qu.: 9.030	3rd Qu.: 8.884	3rd Qu.: 9.030	3rd Qu.: 8.260	3rd Qu.: 500248	3rd Qu.: 181.9	3rd Qu.: 283.4
Max. : 9.540	Max. : 9.425	Max. : 9.635	Max. : 9.400	Max. : 567784	Max. : 217.5	Max. : 2022.4

Avg_age_actors	Num_multiplex	Collection	Start_Tech_Oscar
Min. : 3.00	Min. : 333.0	Min. : 10000	Min. : 0.0000
1st Qu.: 28.00	1st Qu.: 465.0	1st Qu.: 34050	1st Qu.: 0.0000
Median : 39.00	Median : 535.5	Median : 42400	Median : 1.0000
Mean : 39.18	Mean : 545.0	Mean : 45058	Mean : 0.5455
3rd Qu.: 50.00	3rd Qu.: 614.8	3rd Qu.: 50000	3rd Qu.: 1.0000
Max. : 60.00	Max. : 868.0	Max. : 100000	Max. : 1.0000

DATA PREPROCESSING



NAIVE BAYES ALGORITHM

Naive Bayes Classification Model

The screenshot shows the RStudio interface with a script editor, console, and environment pane. The script editor contains the following code:

```
50  
51 #model generation  
52 naive_model <- naive_bayes(Start_Tech_Oscar ~., data=trainData, type="c-classification")  
53 print(naive_model)  
54 plot(naive_model)  
55  
56
```

The console output shows the results of the model generation:

```
R 4.1.2 ~/  
===== Naive Bayes =====  
  
call:  
naive_bayes.formula(formula = Start_Tech_Oscar ~., data = trainData,  
  type = "c-classification")  
  
-----  
  
Laplace smoothing: 0  
  
-----  
  
A priori probabilities:  
  
      0      1  
0.4415011 0.5584989
```

The environment pane shows the system library with the following packages installed:

Package	Description	Version
base	The R Base Package	4.1.2
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	2.1.2

The screenshot shows the RStudio interface with a script editor, console, and environment pane. The script editor contains the following code:

```
50  
51 #model generation  
52 naive_model <- naive_bayes(Start_Tech_Oscar ~., data=trainData, type="c-classification")  
53 print(naive_model)  
54 plot(naive_model)  
55  
56 #predict on test data  
57 testData1 = testData[, -c(17)]  
58 testData1  
59
```

The console output shows the results of the prediction and plotting:

```
R 4.1.2 ~/  
mean 142.02450 142.16364  
sd 27.48918 28.47097  
  
-----  
  
# ... and 11 more tables  
  
-----  
  
> plot(naive_model)  
>
```

The environment pane shows the system library with the following packages installed:

Package	Description	Version
base	The R Base Package	4.1.2
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	2.1.2

A plot titled "Collection" is shown in the viewer pane. The x-axis is labeled "Collection" and ranges from 1e+04 to 9e+04. The y-axis is labeled "Density" and ranges from 0.0e+00 to 1.3e-05. The plot shows two overlapping density curves: a solid red line and a dashed green line. The red curve is slightly higher and shifted to the right compared to the green curve.

Naive Bayes Prediction Model

The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains R code for predicting the 'Start_Tech_Oscar' variable. The code includes loading a dataset, creating a new data frame 'a' with specific values, and using the 'predict' function on a Naive Bayes model.
- Console:** Shows the output of the prediction, which is a single value '0'.
- Environment Pane:** Shows the loaded packages, including 'base', 'boot', 'class', and 'cluster'.

```
# Predicting Technical Achievement award
a <- data.frame(Marketing.expense= 35.0052, Production.expense=71.28, Multiplex.
  Lead_.Actor_Rating= 7.655, Lead_Actress_rating=7.965, Director.
  Critic_rating=7.8, Trailer_views=414145, Time_taken= 164.24, T
  Avg_age_actors=27, Num_multiplex= 502, Collection=31200)
a
result <- predict(naive_model,a)
print(result)
```

Console Output:

```
> result <- predict(naive_model,a)
> print(result)
[1] 0
Levels: 0 1
>
```

Environment Pane:

Name	Description	Ver...
base	The R Base Package	4.1.2
boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-28
class	Functions for Classification	7.3-19
cluster	"Finding Groups in Data"	2.1.2

APPENDIX B

SOURCE CODE

```
library(naivebayes)

Movie_data <- read.csv("C:\\Program Files\\R\\R-3.3.2\\datasets\\movie.csv",
  header = TRUE)
print(Movie_data)

head(Movie_data)
dim(Movie_data)
length(Movie_data)
Movie_data[10:20,12:15]
Movie_data[,11]
Movie_data[25,]

print(Movie_data$Movie_length)

summary(Movie_data)
Movie_data$Start_Tech_Oscar=as.factor(Movie_data$Start_Tech_Oscar)

summary(Movie_data)

#Preprocessing

barplot(table(Movie_data$Start_Tech_Oscar),
  main="Start_Tech_Oscar", col="yellow")
```

```
hist(Movie_data$Marketing.expense)
```

```
hist(Movie_data$Production.expense)
```

```
hist(Movie_data$Multiplex.coverage)
```

```
hist(Movie_data$Movie_length, main = "Movie Length",xlab="Movie Length in  
Minutes",  
ylab="No.of Movies",col = "orange")
```

```
new_data<- Movie_data[,c(2,3,4,5)]
```

```
print(new_data)
```

```
#Sampling of Datasets
```

```
set.seed(1)
```

```
ind <- sample(2, nrow(Movie_data), replace=TRUE, prob=c(0.9, 0.1))
```

```
trainData <- Movie_data[ind==1,]
```

```
testData <- Movie_data[ind==2,]
```

```
print(trainData)
```

```
print(testData)
```

```
naive_model <- naive_bayes(Start_Tech_Oscar ~., data=trainData, type="C-  
classification")
```

```
print(naive_model)
```

```
plot(naive_model)
```

```
#predict on test data
```

```
testData1 = testData[,-c(17)]
testData1
testPred <- predict(naive_model, newdata = testData1)
print(testPred)
print(testData$Start_Tech_Oscar)
```

#Predicting Technical Achievement award

```
a<-data.frame(Marketing.expense= 35.0052,Production.expense=71.28,
  Multiplex.coverage=0.462,Budget= 32907.82, Movie_length=167.6,
  Lead_.Actor_Rating= 7.655, Lead_Actress_rating=7.965,
  Director_rating=7.64, Producer_rating=7.94,
  Critic_rating=7.8, Trailer_views=414145, Time_taken= 164.24,
  Twitter_hastags=262.496,
  Avg_age_actors=27, Num_multiplex= 502, Collection=31200)
a
result <- predict(naive_model,a)
print(result)
```