IS 665001 Data Analytics for Information Systems

Data Mining Project

Project Report

On

Decision Tree

Naïve Bayes

Data Used: IMDB 5000 MOVIE DATASET

Tool Used: Rapid Miner

By,

Team 4

Nikitha Srinivas

Praloy Choudhury

Rajil Nambiar

Saurabh Moyal

Shivani Ratnaparkhi

Srushti Sushilkumar

Introduction:

What is Predictive Data Mining?

Predictive modeling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation or it may be a complex neural network, mapped out by sophisticated software. As additional data becomes available, the statistical analysis model is validated or revised.

Problem Statement:

We have taken an IMDB movies dataset and mined the data for Gross value using 2 algorithms viz., Decision Tree and Naïve Bayes Algorithm. Then we have compared the data in the 2 algorithms and concluded which one is better.

Which tool did we use for mining the Data?

We have used Rapid Miner for the mining purpose.

RapidMiner is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including data preparation, results visualization, validation and optimization. RapidMiner is developed on an open core model.

According to Bloor Research, RapidMiner provides 99% of an advanced analytical solution through template-based frameworks that speed delivery and reduce errors by nearly eliminating the need to write code. RapidMiner provides data mining and machine learning procedures including: data loading and transformation (Extract, transform, load (ETL)), data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment.

About the Data:

The dataset is obtained from kaggle.com and it helps to understand the immensity of the movie before it is released. It has 28 variables consisting of 5043 movies spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses.

Implementation

1. Using Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

Model Construction

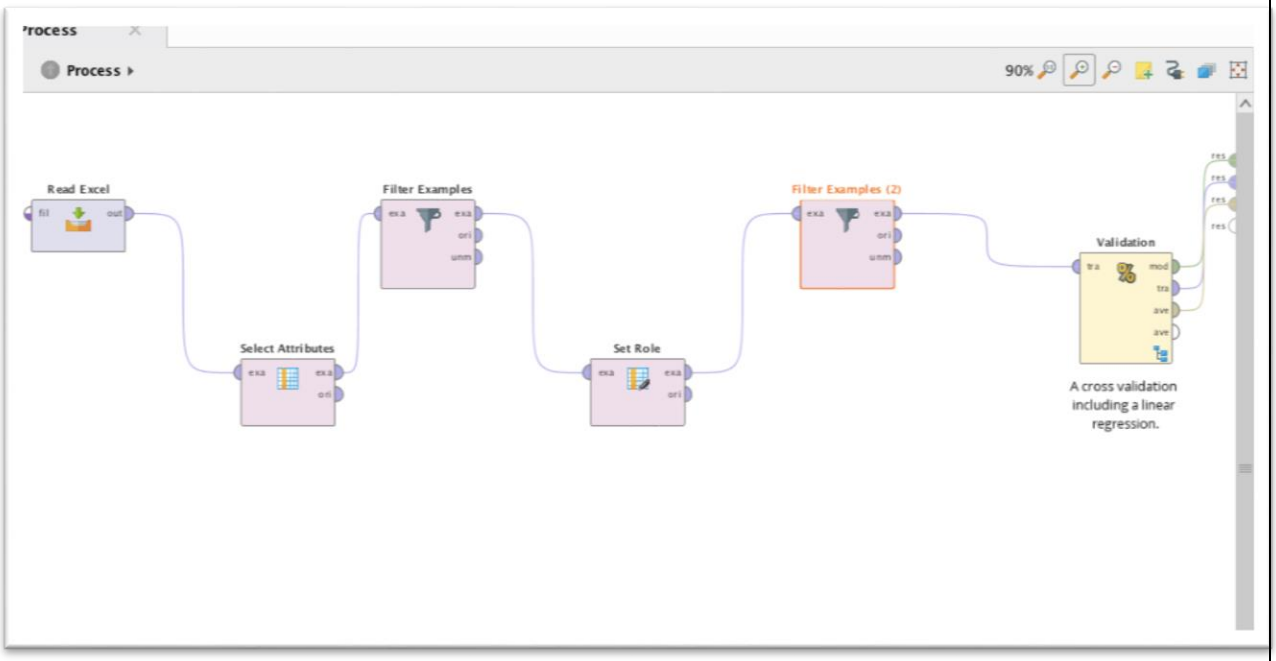


Fig 1: Decision Tree Model Design

Steps:

1) Using “Read Excel operator”, take the data input
2) Select dependent and independent variables based on problem statement using “Select Attributes”.
3) Using “Filter Examples” operator, clean the data
4) Using “Set Role” operator, select labels for the decision tree
5) Using “Validation” operator, validate the parameters used for decision tree.

Results:

PerformanceVector

```
PerformanceVector:
accuracy: 78.60% +/- 2.33% (mikro: 78.60%)
ConfusionMatrix:
True:   USA     UK      New Zealand     Canada  Australia       Belgium Japan   Germany China   France  New Line        Mexico  Spain   Hong Kong       C
USA:    2988    320     11      61      40      2       12      81      13      101     1       8       22      13      3       10      0       6       1
UK:     4       2       0       0       0       0       0       0       0       1       0       0       0       0       0       0       0       0       0
New Zealand:    0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Canada: 1       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Australia:      1       0       0       0       0       0       0       1       0       0       0       0       0       0       0       0       0       0
Belgium:        0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Japan:  3       0       0       1       0       0       3       0       1       0       0       0       0       0       0       0       2       0
Germany:        2       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
China:  0       0       0       0       0       0       0       0       0       1       0       0       0       0       0       1       0       0       0
France: 4       1       0       0       0       0       0       0       0       1       0       0       0       0       0       0       0       0       0
New Line:       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Mexico: 0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Spain:  0       0       0       0       0       0       0       0       1       0       0       0       0       0       0       0       0       0       0
Hong Kong:      0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Czech Republic: 0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
India:  2       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Soviet Union:   0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
South Korea:    0       0       0       0       0       0       0       1       0       0       0       0       0       0       0       0       0       0
Peru:   0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Italy:  1       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Russia: 0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Aruba:  0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Denmark:        0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Libya:  0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Ireland:        0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
South Africa:   1       0       0       0       0       0       1       0       0       0       0       0       0       0       0       0       0       0       0
Iceland:        0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Switzerland:    0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Romania:        0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
West Germany:   0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Chile:  0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Netherlands:    0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Hungary:        0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Panama: 0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Greece: 0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Sweden: 0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Norway: 1       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Taiwan: 1       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Official site:  0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
```

Fig 2 : Results: Performance vector

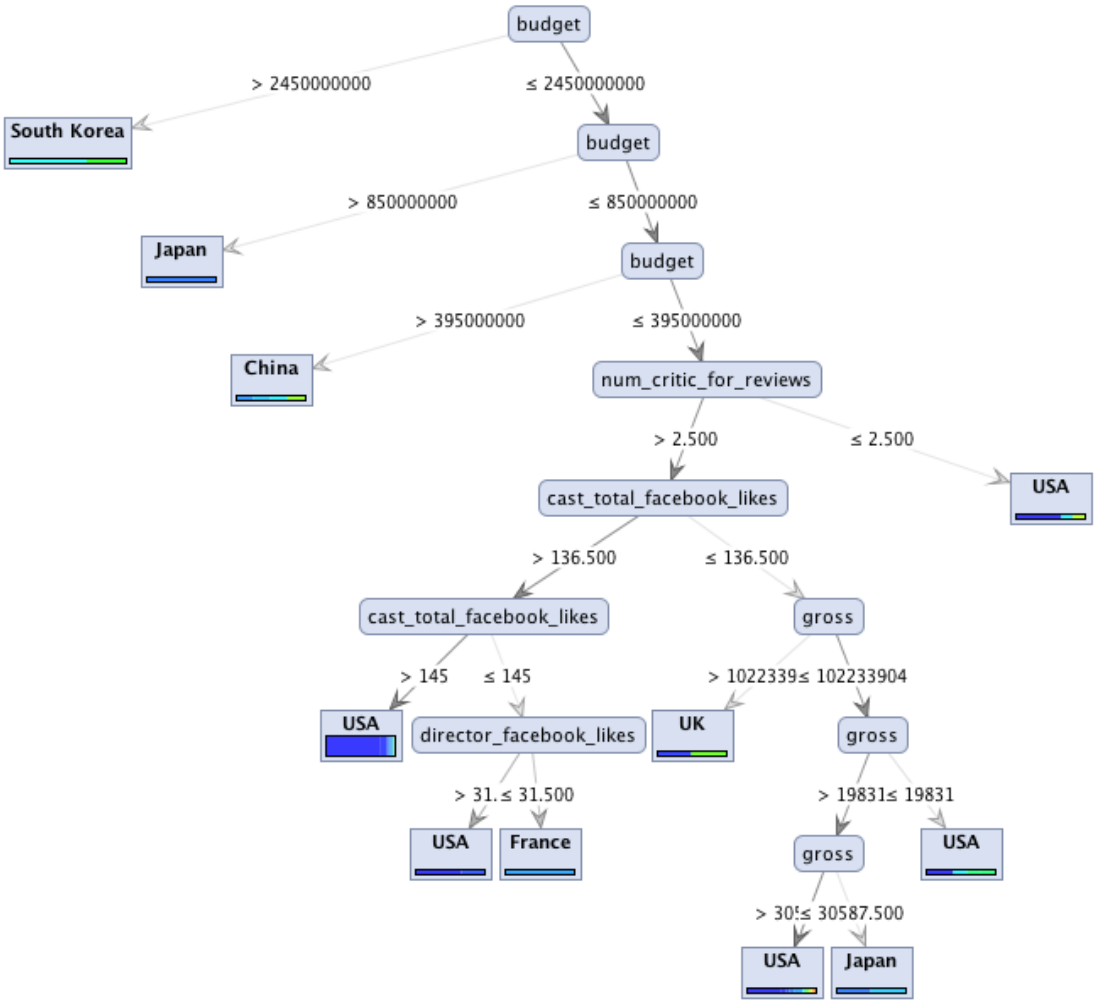


Fig 3 : Results: Tree Representation of the data

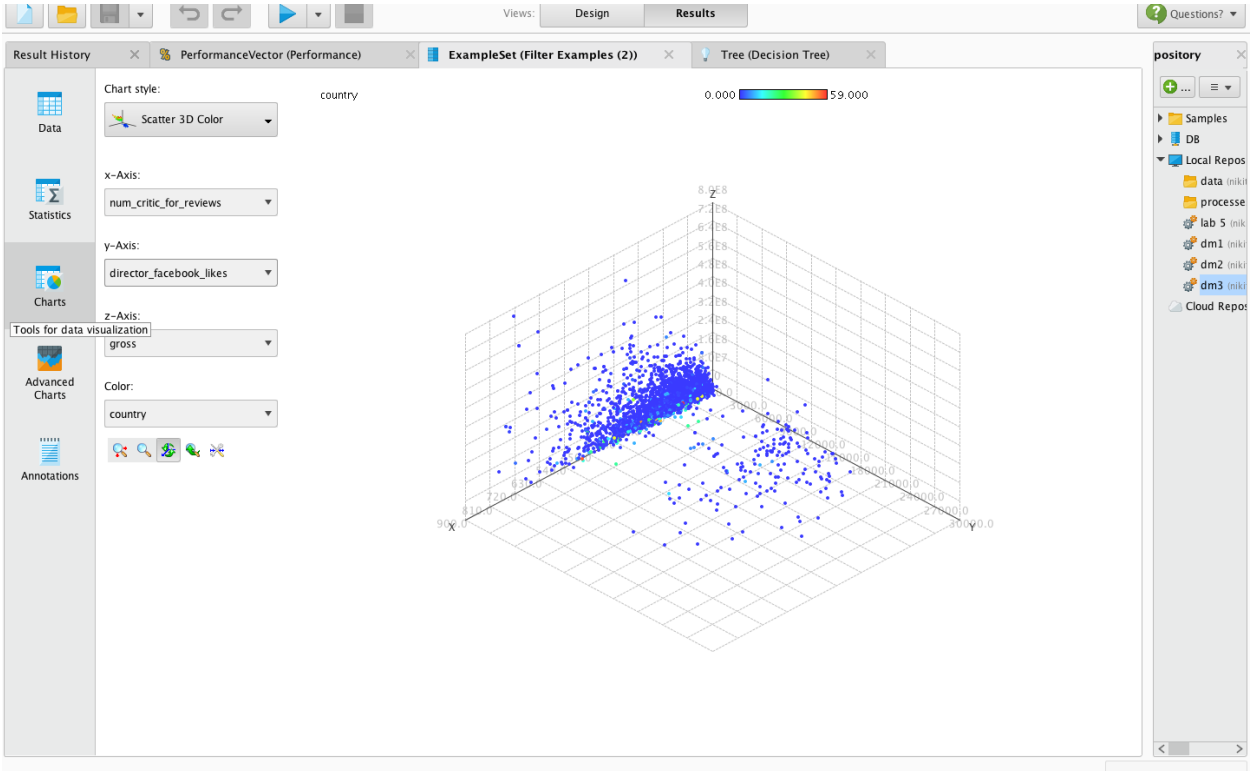


Fig 4: Results: 3 D plotting of the data.

2. Using Naïve Bayes Algorithm

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

Model Construction:

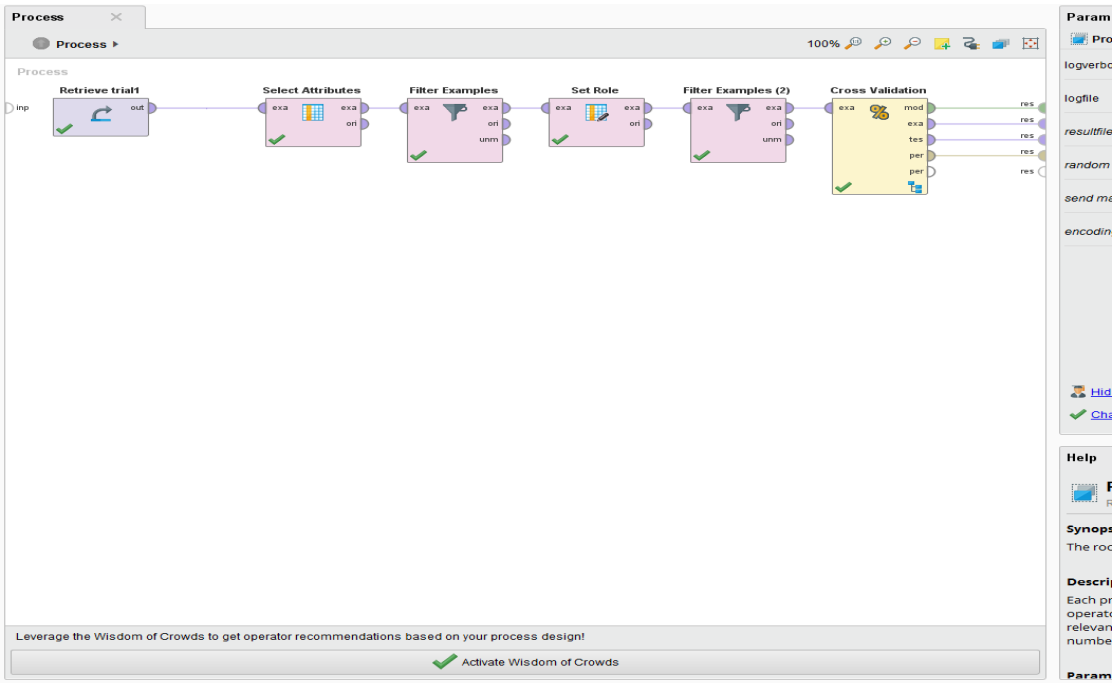


Fig 5 : Design Model using Naïve Bayes algorithm

Steps:

1) Using “Retrieve” operator, import the dataset
2) Using “Select attributes” operator, find dependent and independent variables.
3) Using “Filter Examples”, clean the data.
4) Using “Cross Validation” operator, implement Naïve Bayes algorithm.

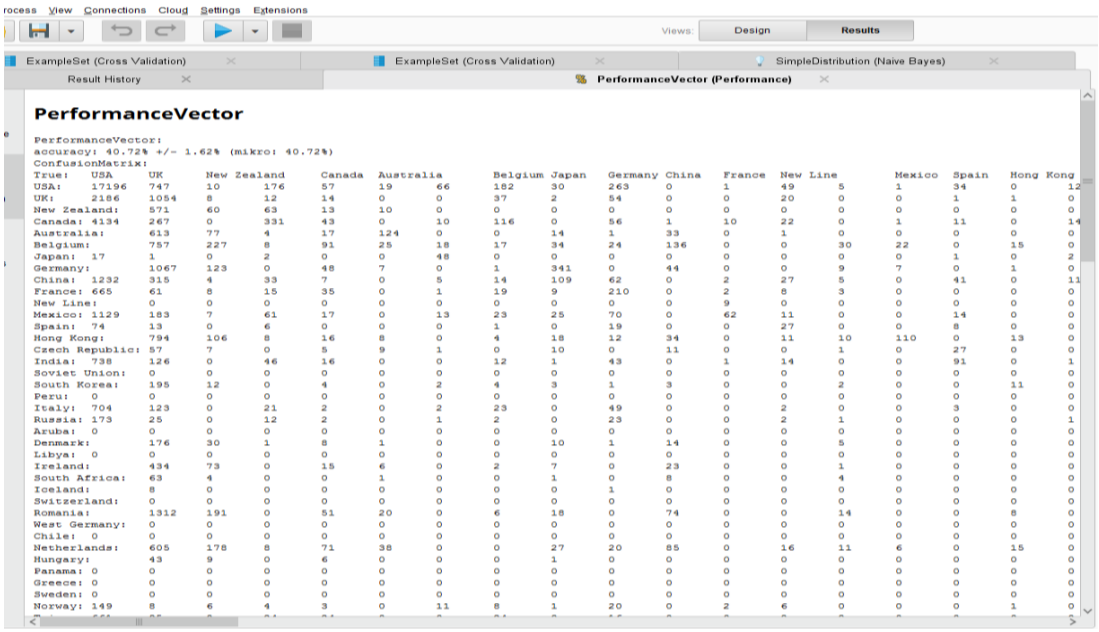
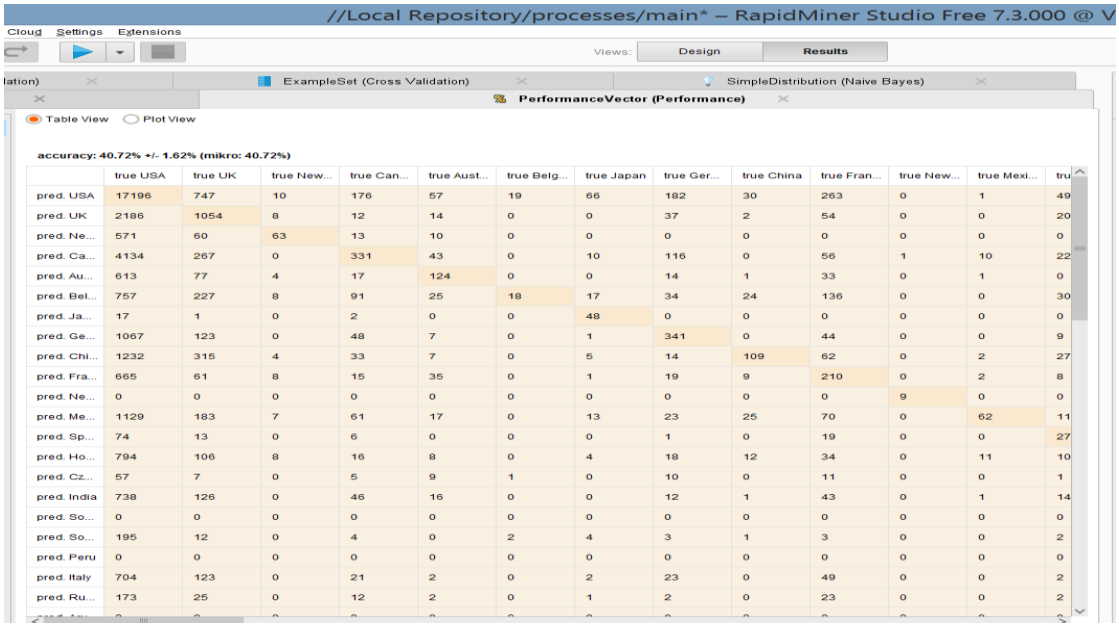


Fig 6: Result: Performance Result of Naïve Bayes



	true USA	true UK	true New...	true Can...	true Aust...	true Belg...	true Japan	true Ger...	true China	true Fran...	true New...	true Mexi...	tru
pred. USA	17196	747	10	176	57	19	66	182	30	263	0	1	49
pred. UK	2186	1054	8	12	14	0	0	37	2	54	0	0	20
pred. Ne...	571	60	63	13	10	0	0	0	0	0	0	0	0
pred. Ca...	4134	267	0	331	43	0	10	116	0	56	1	10	22
pred. Au...	613	77	4	17	124	0	0	14	1	33	0	1	0
pred. Bel...	757	227	8	91	25	18	17	34	24	136	0	0	30
pred. Ja...	17	1	0	2	0	0	48	0	0	0	0	0	0
pred. Ge...	1067	123	0	48	7	0	1	341	0	44	0	0	9
pred. Chi...	1232	315	4	33	7	0	5	14	109	62	0	2	27
pred. Fra...	665	61	8	15	35	0	1	19	9	210	0	2	8
pred. Ne...	0	0	0	0	0	0	0	0	0	0	9	0	0
pred. Me...	1129	183	7	61	17	0	13	23	25	70	0	62	11
pred. Sp...	74	13	0	6	0	0	0	1	0	19	0	0	27
pred. Ho...	794	106	8	16	8	0	4	18	12	34	0	11	10
pred. Cz...	57	7	0	5	9	1	0	10	0	11	0	0	1
pred. India	738	126	0	46	16	0	0	12	1	43	0	1	14
pred. So...	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. So...	195	12	0	4	0	2	4	3	1	3	0	0	2
pred. Peru	0	0	0	0	0	0	0	0	0	0	0	0	0
pred. Italy	704	123	0	21	2	0	2	23	0	49	0	0	2
pred. Ru...	173	25	0	12	2	0	1	2	0	23	0	0	2

Fig 7 : Result: Naïve Bayes

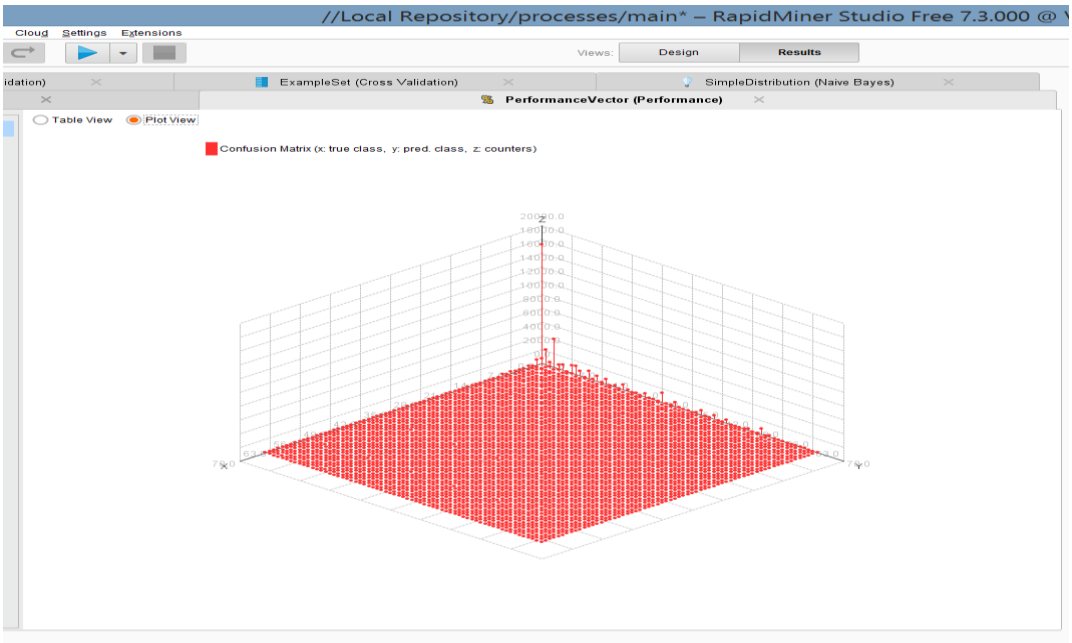


Fig 8 : 3 D plotting of Naïve Bayes

Observation:

Accuracy of Decision tree is 78.06 % and that of Naïve Bayes is 48.72 which is way less than prior one. Hence, we can clearly observe that Decision tree is more accurate in predicting the results than Naïve Bayes Algorithm in this case.

Conclusion:

By observing above scenario, we can conclude that, in our case, Decision tree is more effective in data mining. The reasons are as follows:

a. Decision trees are more flexible than Naïve Bayes
b. Naïve Bayes algorithm is effective when output parameters are binary whereas Decision tree works with multiple output parameters.
c. In our case, the output parameters were more than 2 I.e. Gross, Genre, Country, Budget etc. Hence the complexity increased, and Naïve Bayes accuracy declined.

References:

http://searchdatamanagement.techtarget.com/definition/predictive-modeling

https://en.wikipedia.org/wiki/RapidMiner

https://en.wikipedia.org/wiki/Decision_tree

http://stackoverflow.com/questions/10317885/decision-tree-vs-naive-bayes-classifier

https://en.wikipedia.org/wiki/Naive_Bayes_classifier