

DATA ANALYTICS

DATA MINING PROJECT

ABHISHEK CHITRE

DARSHIL SHAH

HARSHA MIRANI

JASHNA KAPADIA

PRIYANKA RANA

Contents

Description:.....	3
Algorithms.....	3
Decision Tree	3
Naïve Bayes.....	3
Naïve Bayes.....	5
Decision Tree	10

Description

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.

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.

Algorithms

Decision Tree

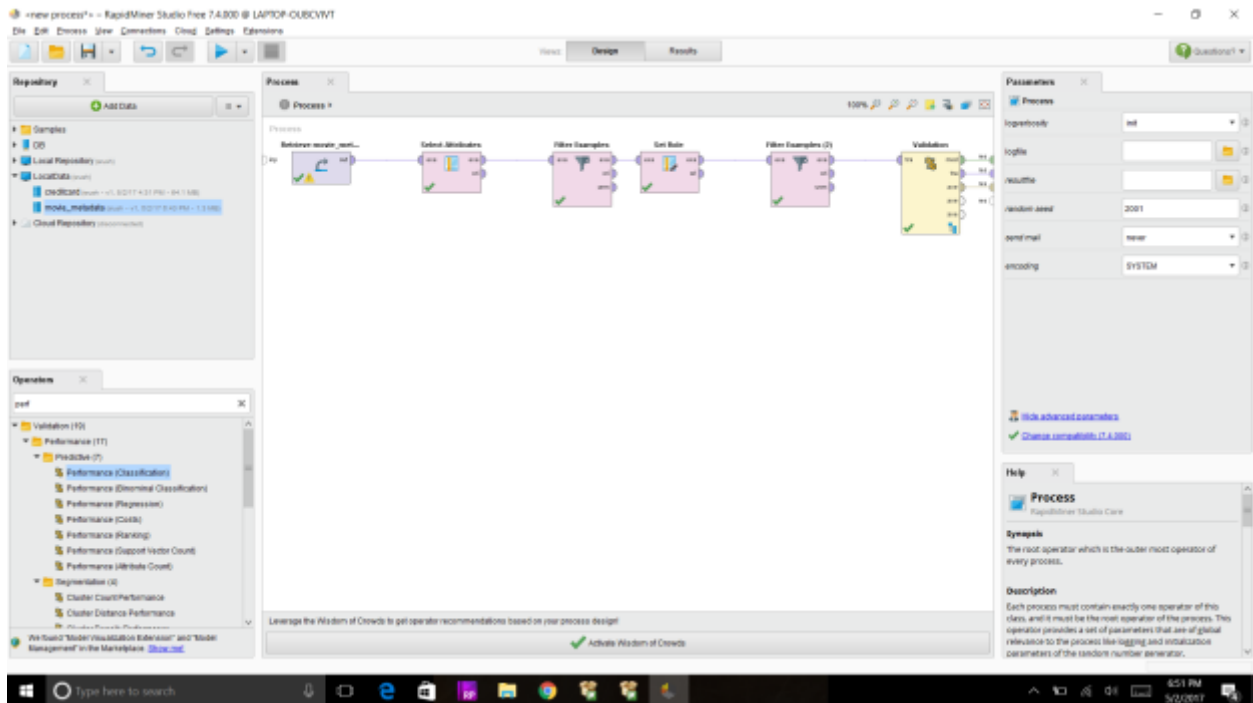
Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision tree is composed by several IF-THEN in cascade. Decision trees are an excellent tool because it provides a highly effective structure which you can lay out options and investigate the possible outcomes of choosing those options. We used decision tree for having a good prediction, find correlation between features and as we saw before for pre-processing the data set.

Naïve Bayes

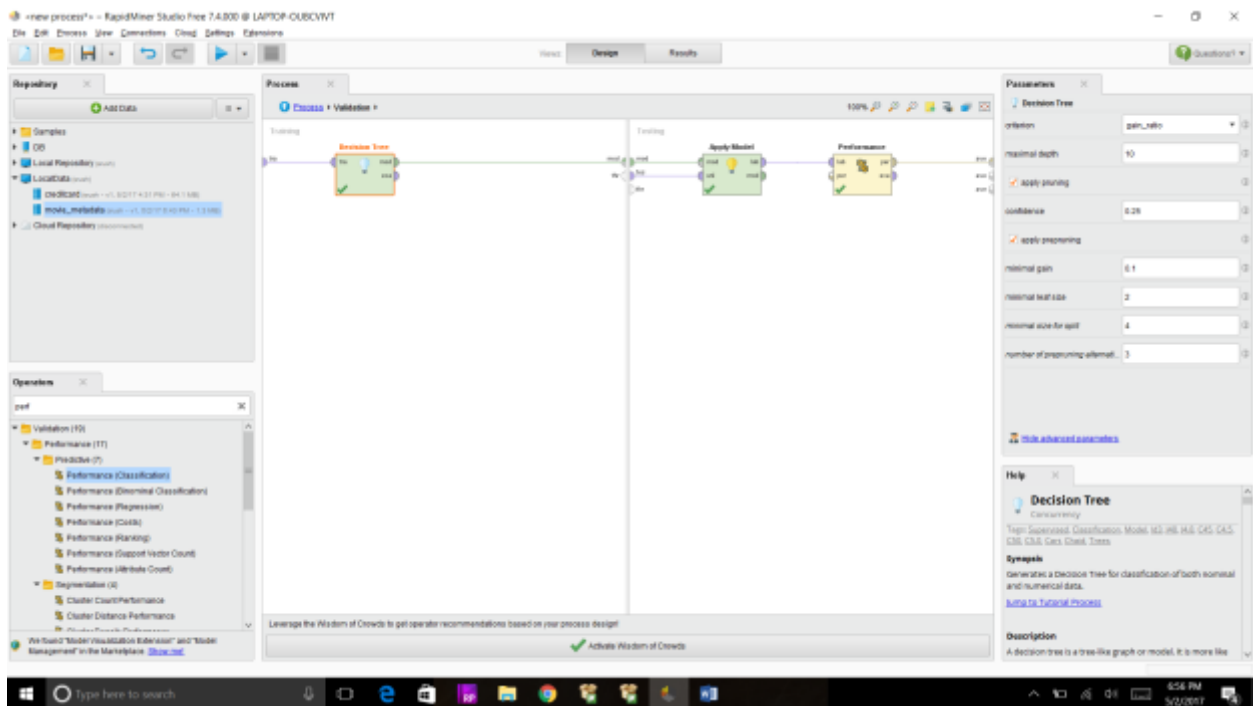
Naive Bayes is a simple technique for constructing classifiers. The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Decision Tree

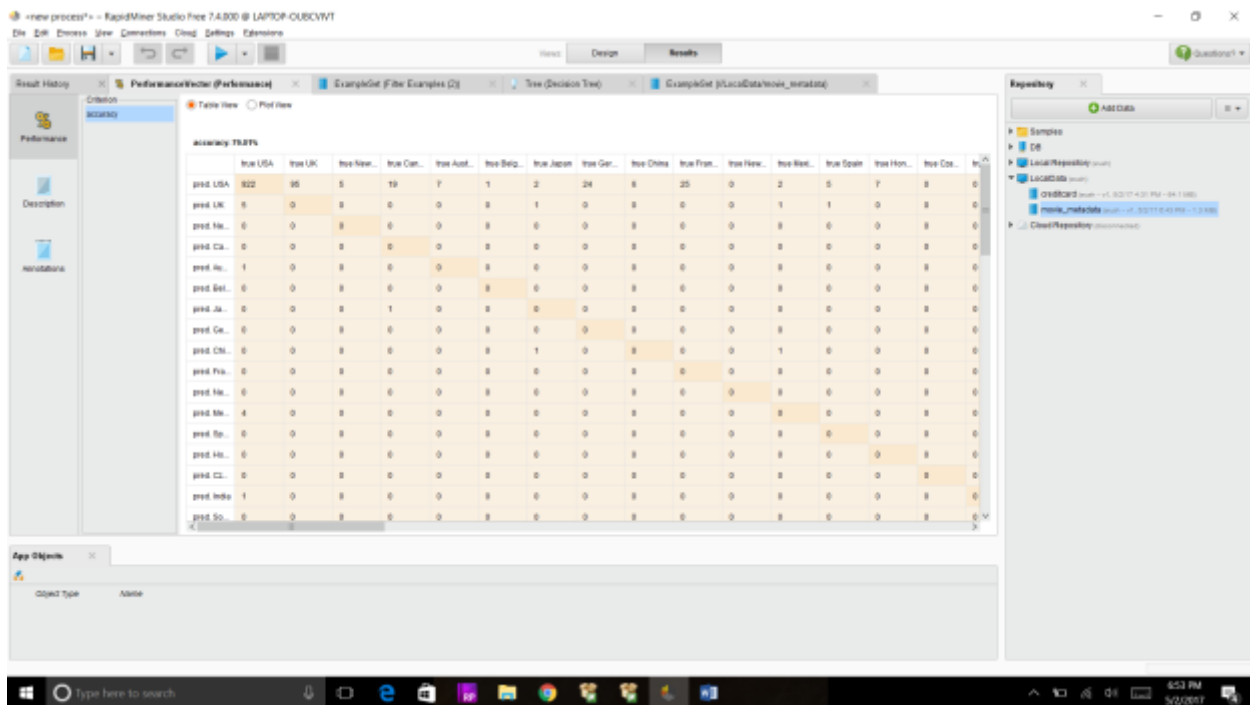
Process



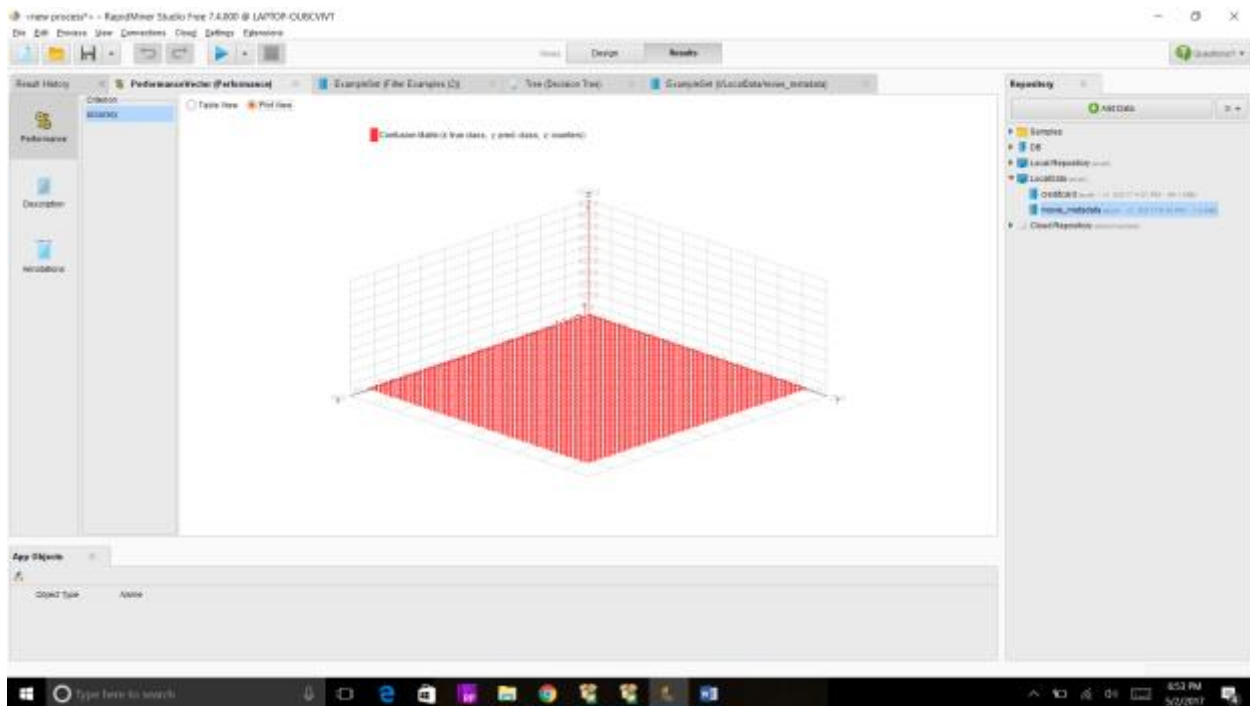
Validation



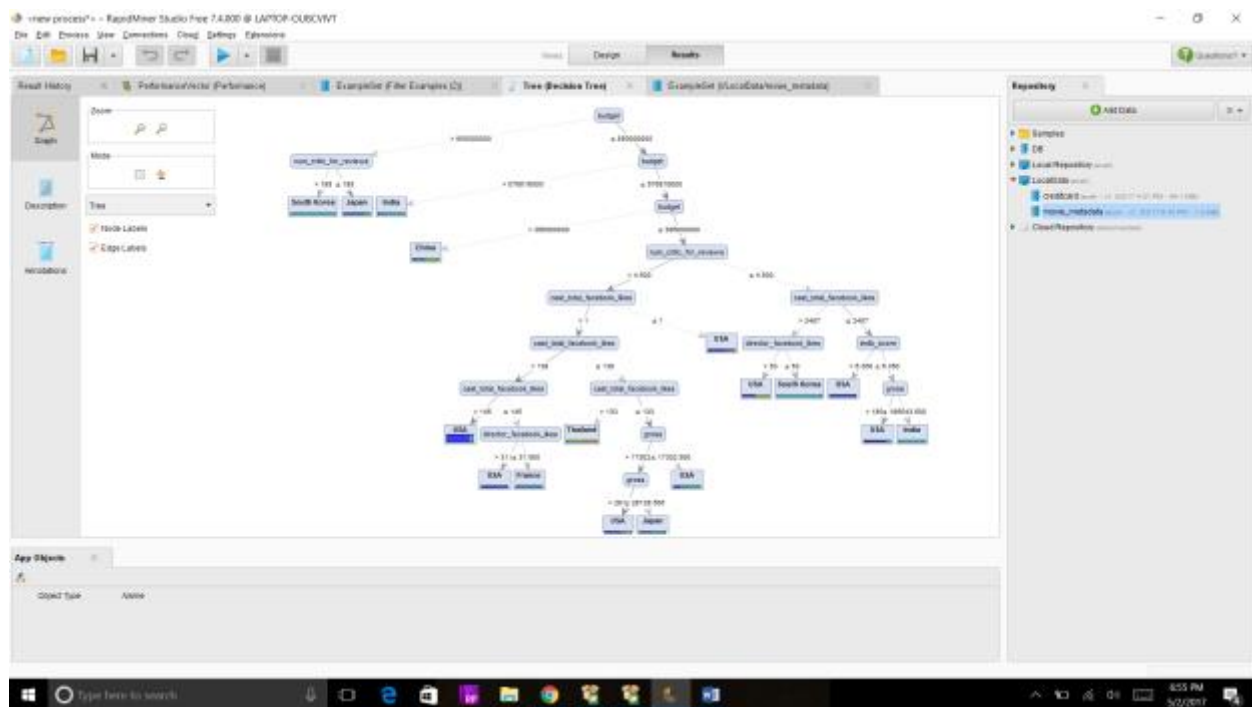
Performance Vector



Confusion Matrix

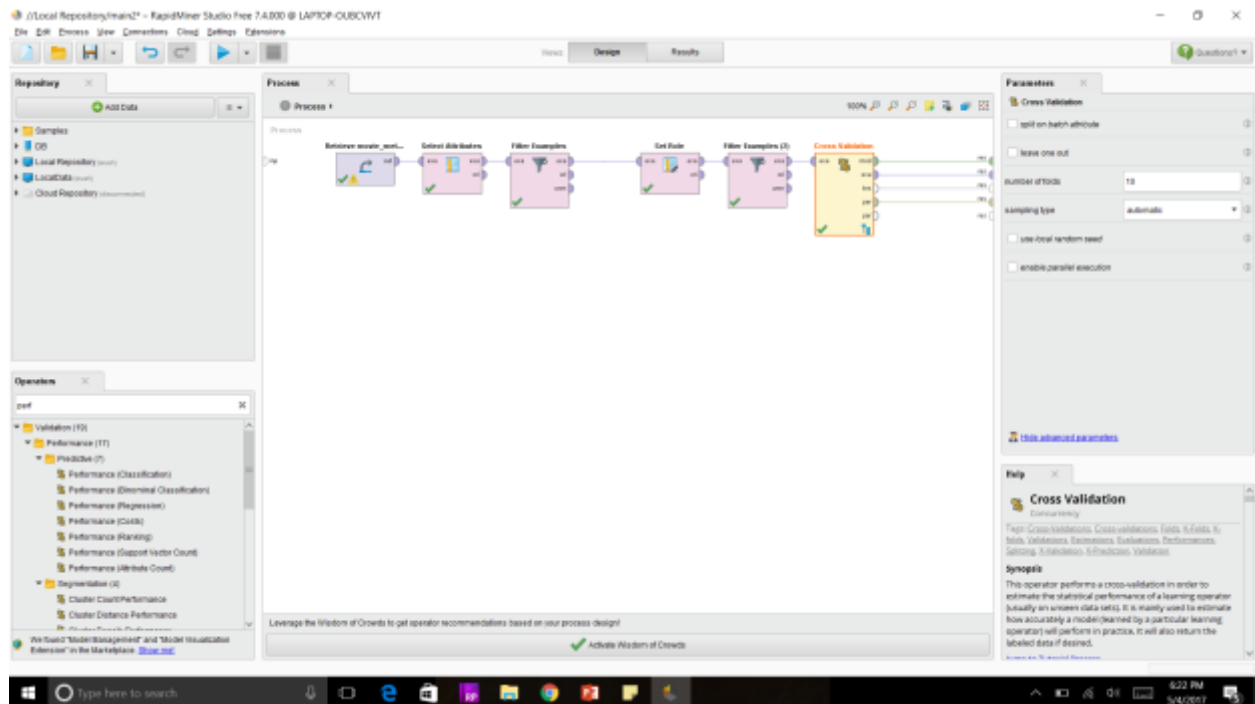


Decision tree

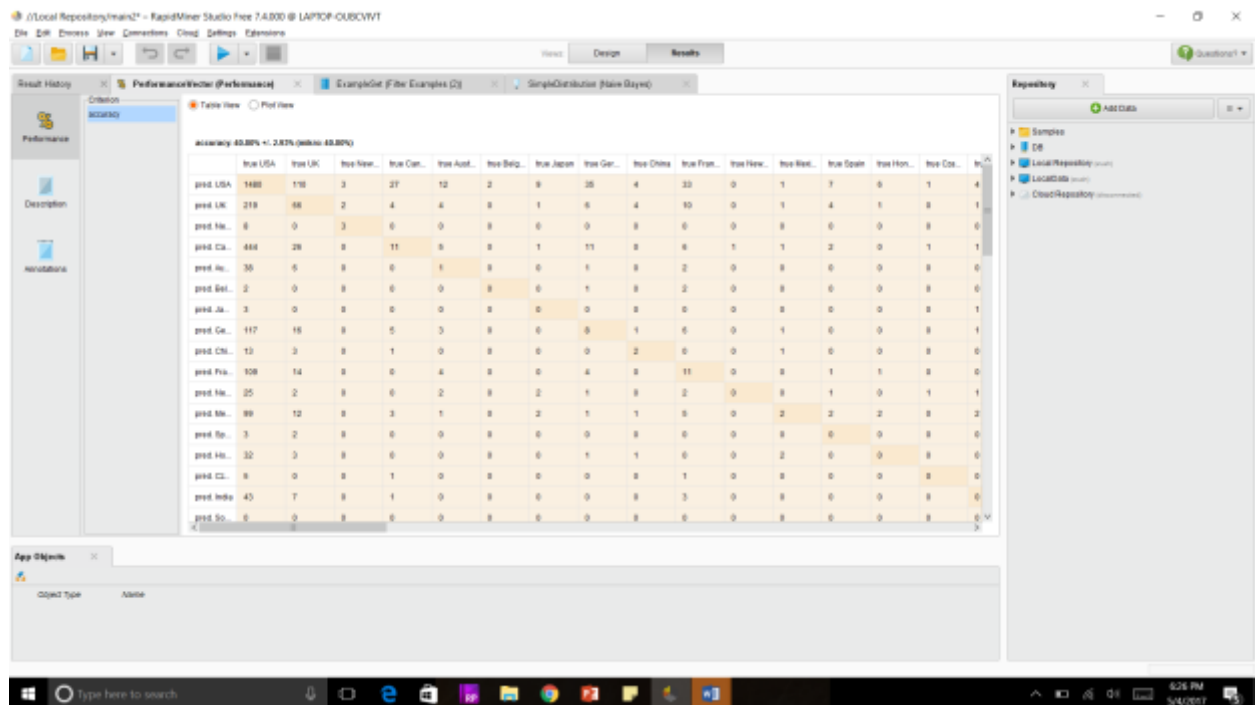


Naïve Bayes

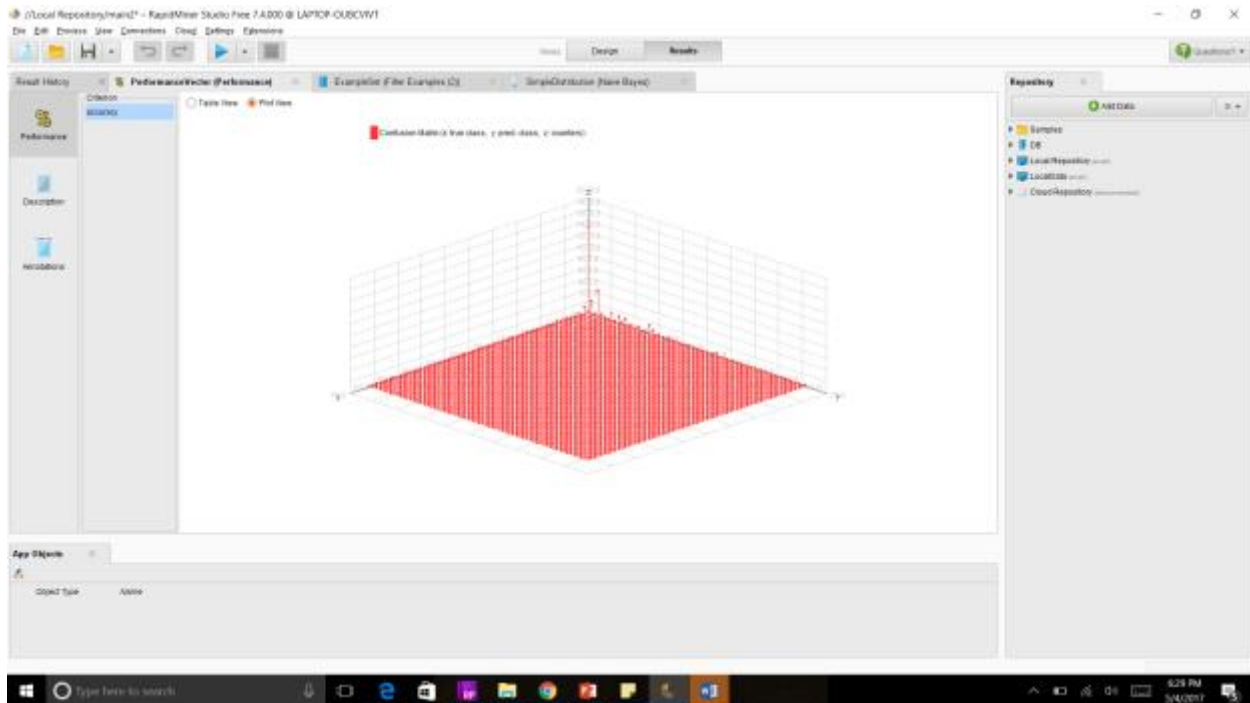
Process



Performance Vector



Confusion Matrix



Example Set

Local Repository - RapidMiner Studio Free 7.0.00 @ LAP-OP-UBCVR1

File Edit Process View Connections Cloud Settings Extensions

Icons Design Results

Results

Result History

Performance (vector Performance)

ExampleSet (Other Examples (3))

Single Distribution (Flow Bayes)

Data

Statistics

Charts

Advanced Charts

Visualizations

ExampleSet (300 examples, 1 potential attribute, 5 regular attributes)

File (3,000 / 3,000 examples) All

Row No.	country	year	director	genre	start_year	rank	user	budget	score	rank
1	USA	1959	0	Adventure	1959	3054	21000000	7.800	10000	
2	USA	1963	0	Adventure	1963	1238	18000000	7.100	0	
3	UK	1963	0	Adventure	1963	894	24000000	8.800	60000	
4	USA	1915	22000	Adventure	1915	2701	25000000	8.500	15000	
5	USA	1952	275	Adventure	1952	718	10170000	8.600	24000	
6	USA	1962	0	Adventure	1962	1902	13000000	8.200	0	
7	USA	1934	15	Adventure	1934	367	28000000	7.800	28000	
8	USA	1937	0	Adventure	1937	1117	25000000	7.500	18000	
9	UK	1936	282	Adventure	1936	875	28000000	7.800	12000	
10	USA	1933	0	Adventure	1933	3018	18000000	8.800	19000	
11	USA	1934	0	Adventure	1934	2367	29000000	8.100	0	
12	UK	1936	180	Adventure	1936	1243	28000000	8.700	0	
13	USA	1933	180	Adventure	1933	1823	22000000	7.500	3000	
14	USA	1950	100	Adventure	1950	4572	21000000	8.500	4000	
15	USA	1933	0	Adventure	1933	2538	22000000	7.200	18000	
16	USA	1936	80	Adventure	1936	2297	22000000	8.800	0	
17	USA	1936	0	Adventure	1936	1722	28000000	8.100	12000	
18	USA	1936	250	Adventure	1936	484	25000000	8.700	3000	
19	USA	1936	180	Adventure	1936	1212	22000000	8.800	4000	
20	USA	1936	0	Adventure	1936	802	28000000	7.800	6000	

Repository

ARCSDB

Examples

DB

Local Repository

Local Database

Cloud Repository

App Objects

Object Type Name

Windows Taskbar

10:26 PM 10/07/2017

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.