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

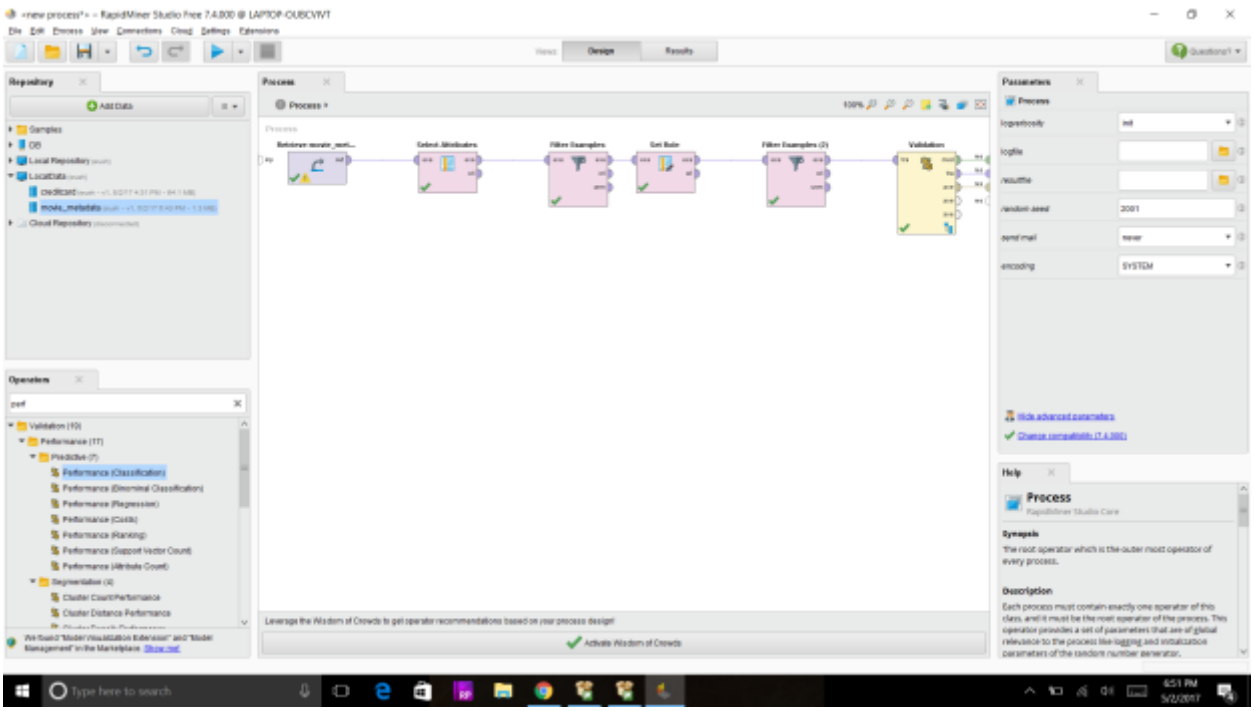
Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision 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

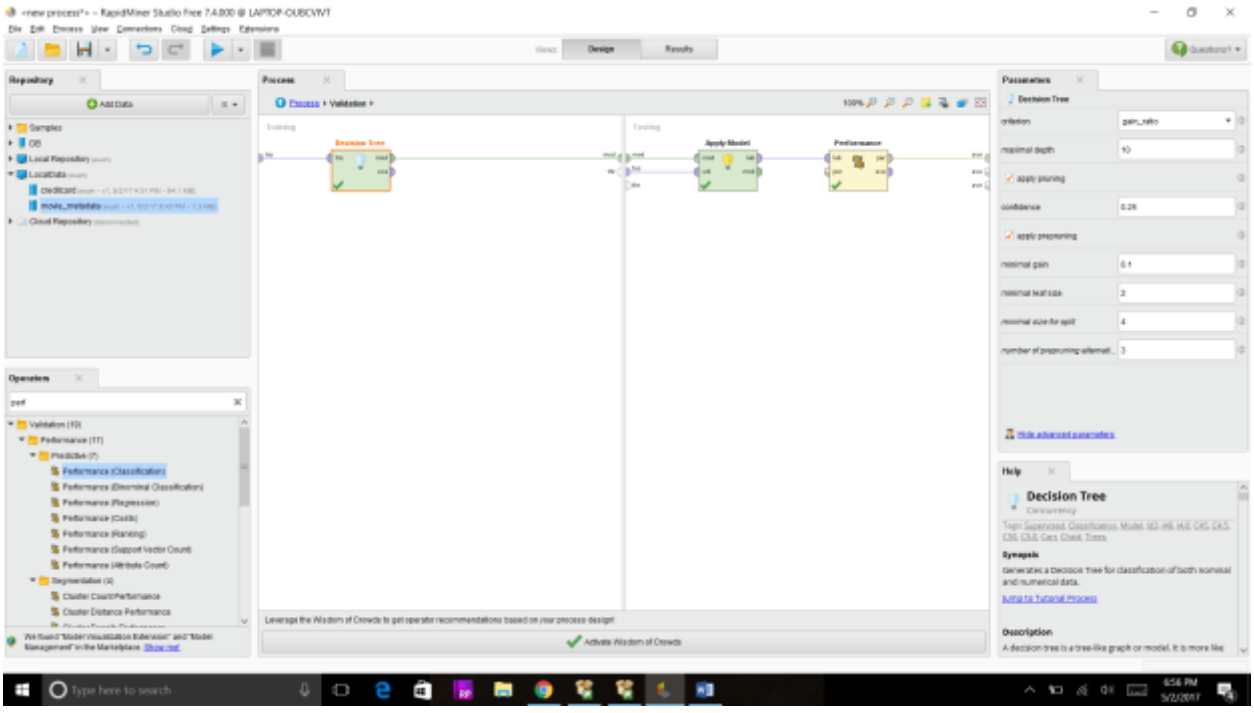
Naive Bayes is a simple technique for constructing classifiers. The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for 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

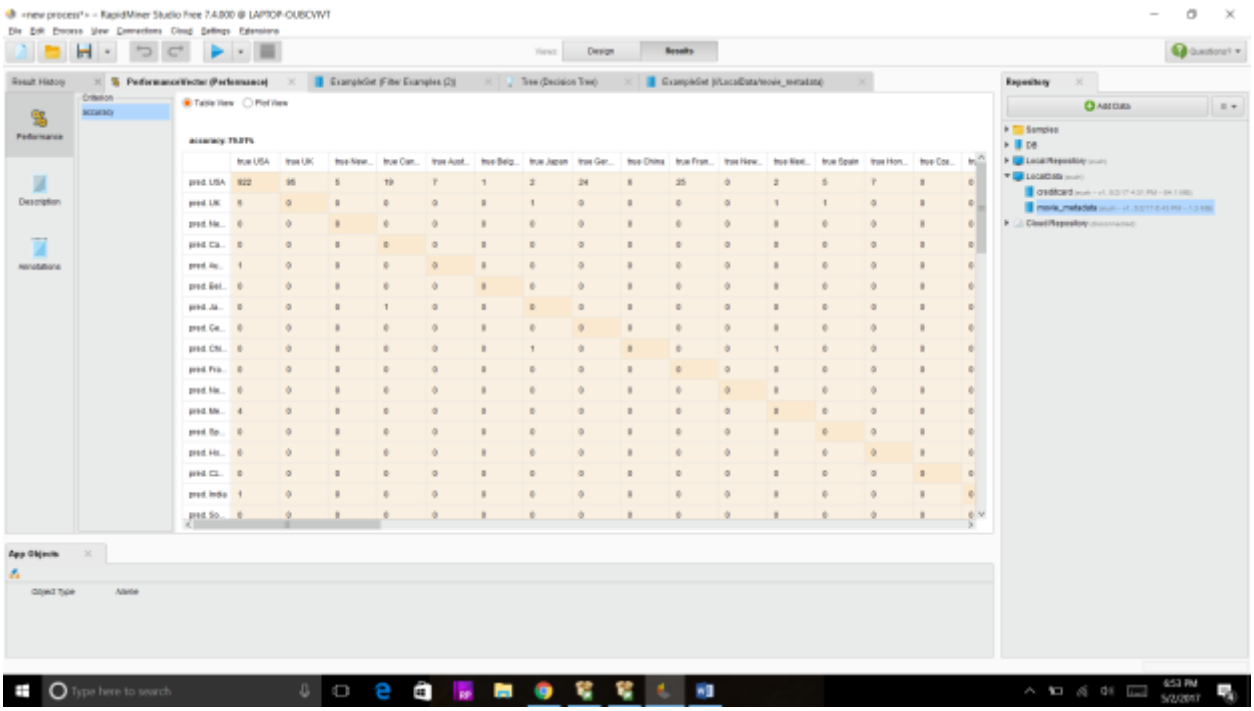
<u>Process</u>



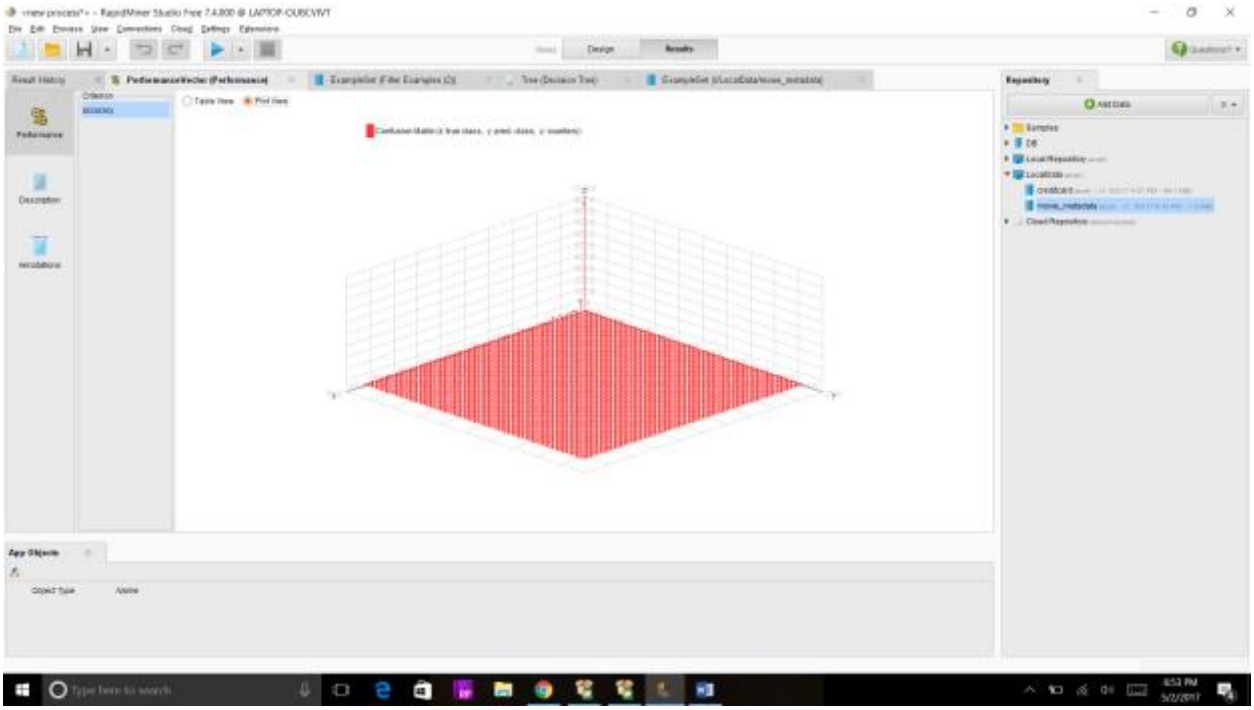
Validation



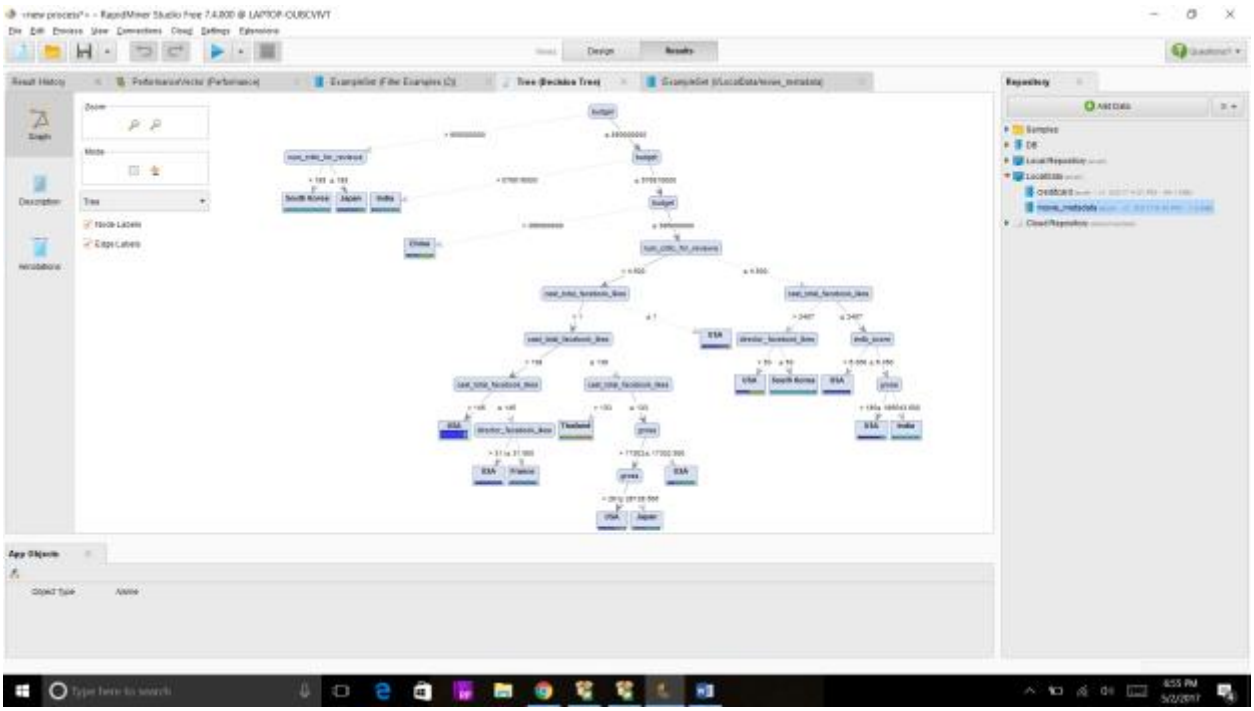
Performance Vector



Confusion Matrix

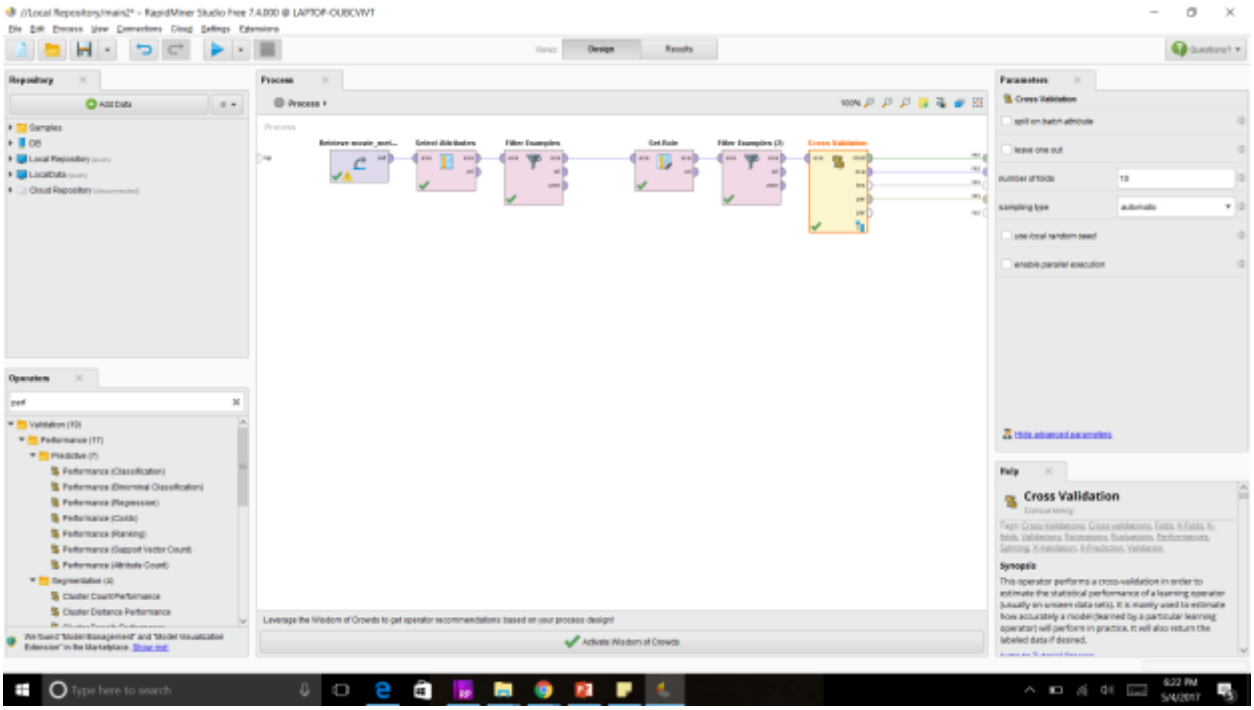


Decision tree

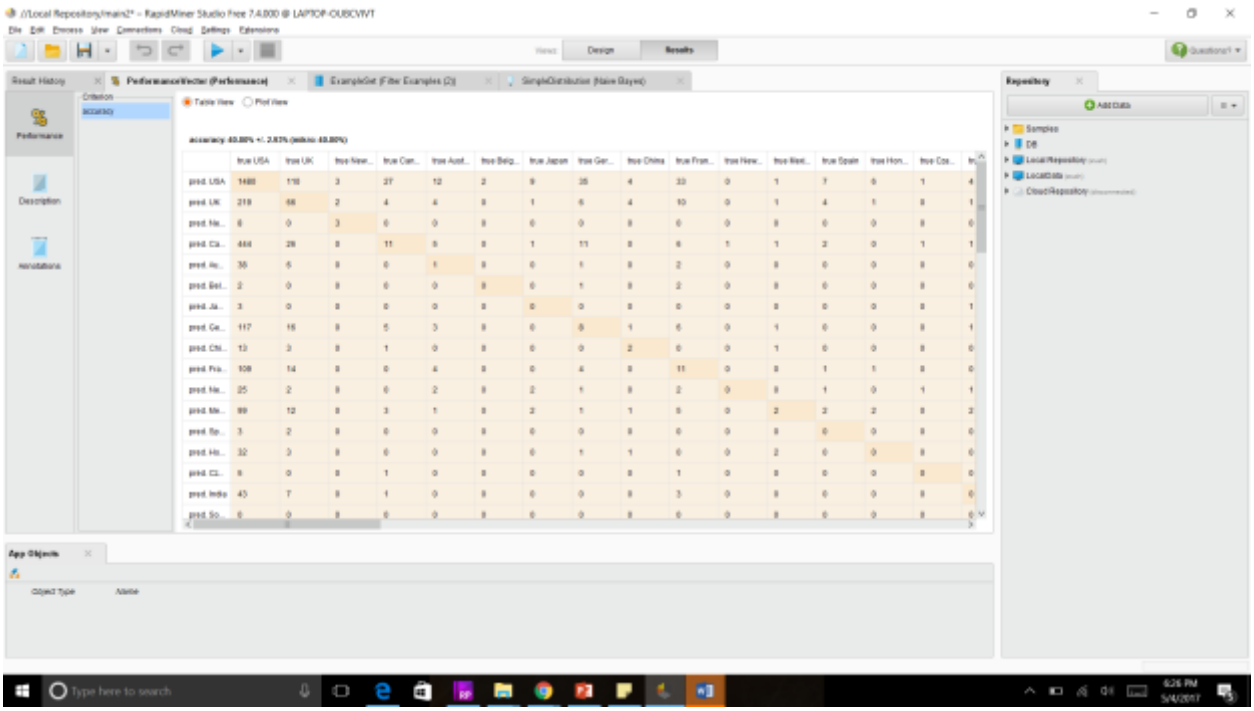


Naïve Bayes

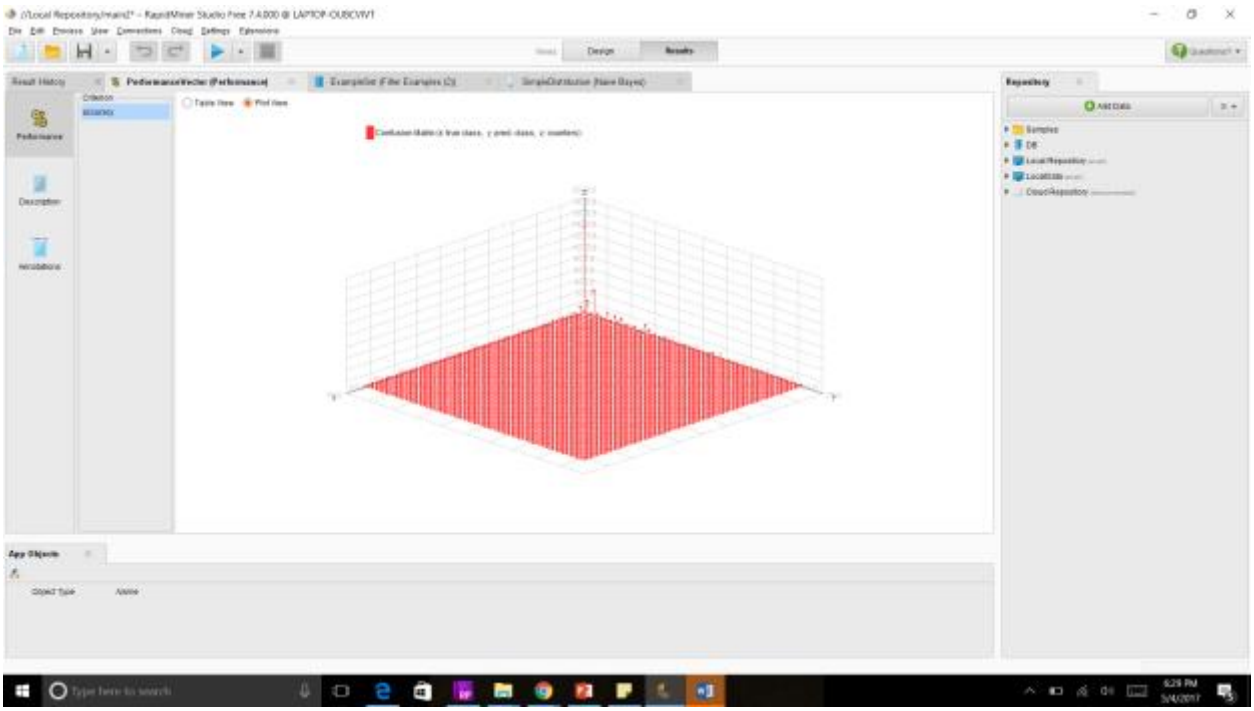
Process



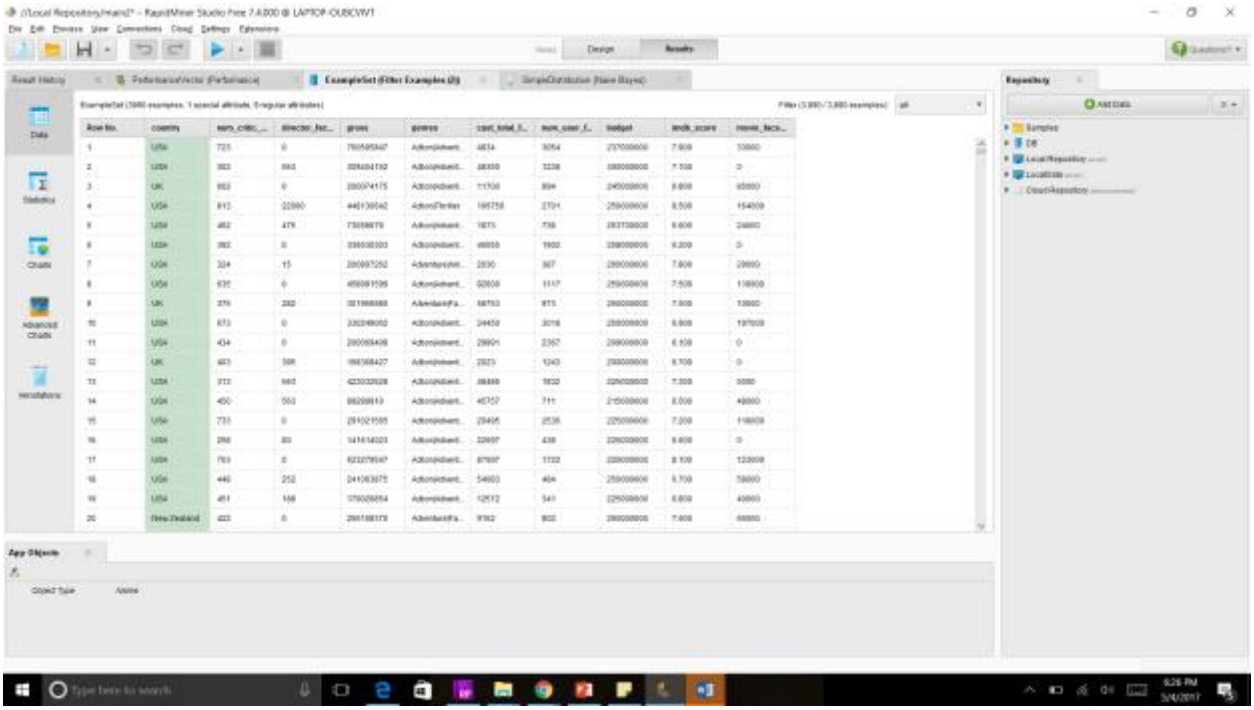
Performance Vector



Confusion Matrix



Example Set



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.