Experiment No.3

Title: Execution of classification algorithm using Rapidminer.

Batch:B4 Roll No.:16010420117 Experiment No.:3

Aim: Execution of data mining algorithm using RapidMiner.

Resources needed: Any RDBMS, Java

Theory:

Rapidminer is a collection of open source of many data mining and machine learning algorithms, including,

– pre-processing on data

– Classification

– clustering

– Association rule extraction

A dataset is a collection of examples, each one of class Instance. Each Instance consists of a number of attributes, any of which can be nominal (= one of a predefined list of values), numeric (= a real or integer number) or a string (= an arbitrary long list of characters, enclosed in "double quotes"). The external representation of an Instances class is an ARFF file, which consists of a header describing the attribute types and the data as comma- separated list.

Rapidminer Main Features:

Main features are as follows:

• 49 data pre-processing tools

• 76 classification/regression algorithms

• 8 clustering algorithms

• 15 attribute/subset evaluators + 10 search algorithms for feature selection.

• 3 algorithms for finding association rules

• 3 graphical user interfaces

The Explorer (exploratory data analysis)

Used for pre-processing, attribute selection, learning, visualization

The Experimenter (experimental environment)

Used for testing and evaluating machine learning algorithms

The Knowledge Flow (new process model inspired interface)

Used for visual design of KDD process

Procedure / Approach /Algorithm / Activity Diagram:

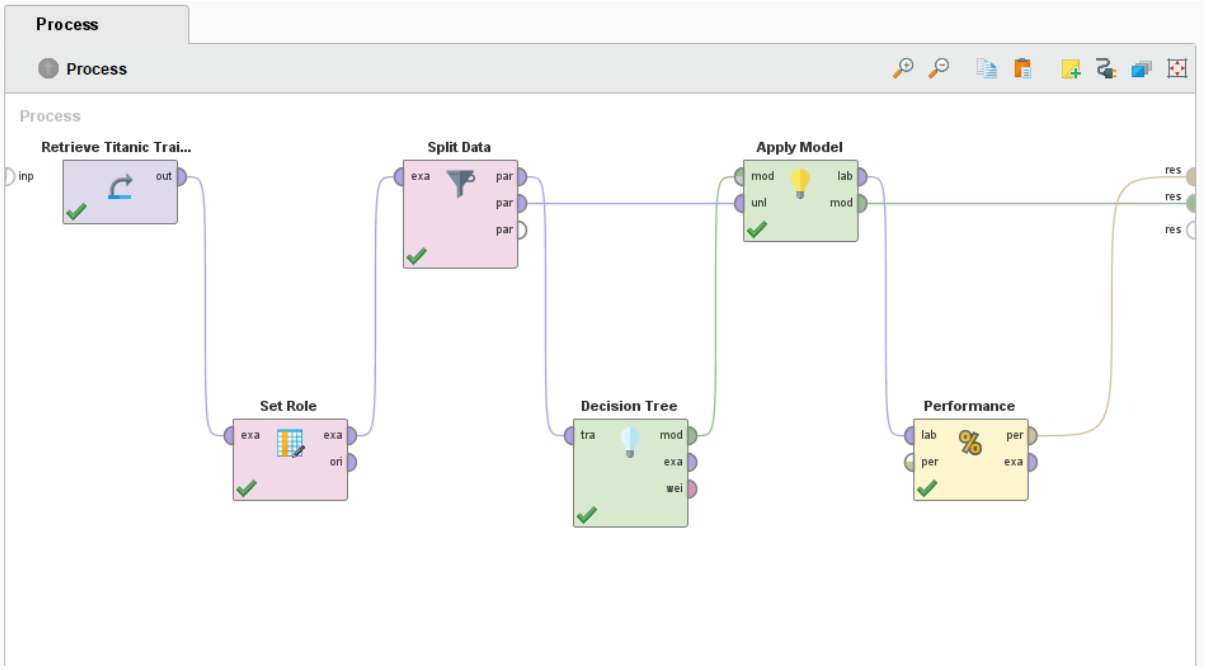
1. Execute any two data mining classification algorithm using Rapidminer tool
2. Analyze the results produced by Rapidminer

Results: (Program printout with output / Document printout as per the format)

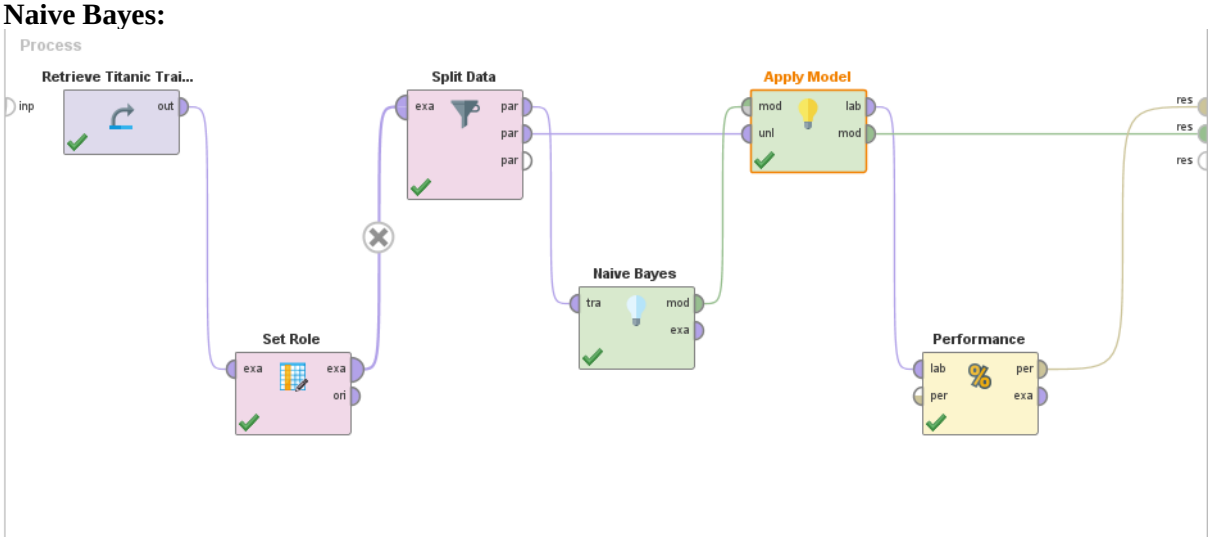
Naïve Bayes algo on titanic dataset

Program:

Decision Tree:



Naive Bayes:



Output:
Decision Tree:

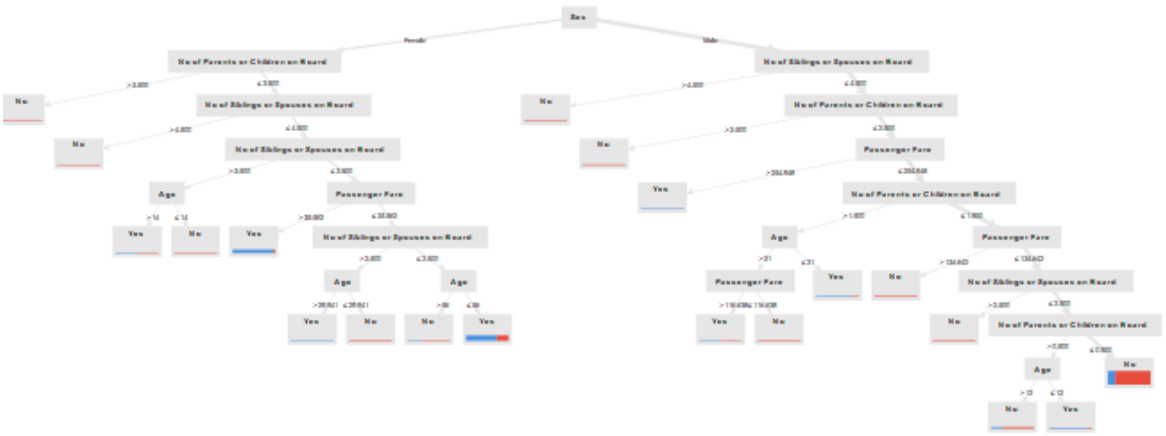


Table View Plot View

accuracy: 79.27%

	true Yes	true No	class precision
pred. Yes	77	29	72.64%
pred. No	28	141	83.43%
class recall	73.33%	82.94%	

precision: 83.43% (positive class: No)

	true Yes	true No	class precision
pred. Yes	77	29	72.64%
pred. No	28	141	83.43%
class recall	73.33%	82.94%	

Naive Bayes:

SimpleDistribution

```
Distribution model for label attribute Survived


Class Yes (0.381)
6 distributions

Class No (0.619)
6 distributions
```

Table View Plot View

accuracy: 80.36%

	true Yes	true No	class precision
pred. Yes	76	25	75.25%
pred. No	29	145	83.33%
class recall	72.38%	85.29%	

Table View Plot View

precision: 83.33% (positive class: No)

	true Yes	true No	class precision
pred. Yes	76	25	75.25%
pred. No	29	145	83.33%
class recall	72.38%	85.29%	

Analysis: The dataset chosen for model training is Titanic Dataset.When this is trained with Naive Bayes model , it shows higher accuracy than Decision Tree.

Post Lab Question- Answers (If Any):

Q: List any five open sources / freeware tools available for data mining.

Ans:5 best open source data mining tools:.

1. Apache Mahout: A well-liked distributed linear algebra framework is Apache Mahout. It is a mathematical representation of the Scala DSL, which enables statisticians and data scientists to quickly build their methods. Multiple backends, including Apache Spark, are supported by the tool. Mahout enables apps to analyse huge datasets more quickly.

2. ELKI: Environment for Developing KDD-Applications Supported by Index-Structures is referred to as ELKI. It is a Java-based open-source data mining programme. The platform is made for studying algorithms. The ELKI platform aims to conduct algorithmic research with an emphasis on unsupervised cluster analysis techniques. Data index structures like the R*-tree are offered by ELKI. The platform provides a vast array of extremely parameterizable algorithms.

3. KNIME: The Java-based framework was created using Eclipse. It is a multilingual environment for developing applications. KNIME is a free platform for data reporting, analytics, and integration. It offers a user-friendly interface and 2000 nodes from which you can pick to build visual workflows.

4. Orange: A variety of data visualisation, exploration, preprocessing, and modelling strategies work best with the open-source, component-based data mining software for machine learning and data visualisation. Orange offers interactive data visualisation with basic data analysis capabilities.

5. Rattle: Data summaries are presented statistically and visually using this open-source GUI for data mining. Rattle enables data transformation for quick modelling. Additionally, it creates data-driven supervised and unsupervised ML models. The finest aspect of Rattle is that it makes extensive use of R's statistical capabilities to offer data mining functionality.

CO: Comprehend basics of ML

Conclusion: In this experiment, I successfully understood and implemented classification algorithms on Titanic Dataset using Rapid Miner

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition