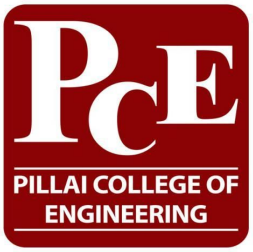
DATA MINING AND BUSINESS INTELLIGENCE

MINI PROJECT

FOREST TYPE DATASET



Pillai's College of Engineering

Dr. K. M. Vasudevan Pillai Campus, Plot No. 10, Sector 16, New Panvel, Navi Mumbai,Maharashtra 410206

Submitted by

Jayesh Thakur 55

Akash Gopinathan 57

Pravin Tambe 53

TABLE OF CONTENTS:

1.INTRODUCTION

1.1 Business Intelligence:

Business intelligence (BI) can be described as "a set of techniques and tools for the acquisition and transformation of raw data into meaningful and useful information for business analysis purposes". The term "data surfacing" is also more often associated with BI functionality. BI technologies are capable of handling large amounts of unstructured data to help identify,develop and otherwise create new strategic business opportunities. The goal of BI is to allow for the easy interpretation of these large volumes of data. Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a competitive market advantage and long-term stability.

BI technologies provide historical, current and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.

BI can be used to support a wide range of business decisions ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions include priorities, goals and directions at the broadest level. In all cases, BI is most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined, external and internal data can provide a more complete picture which, in effect, creates an "intelligence" that cannot be derived by any singular set of data. Among st myriad uses, BI tools empower organizations to gain insight into new markets, assess demand and suitability of products and services for different market segments and gauge the impact of marketing efforts.

1.2 BI Tools:

The following is a list of desirable features BI portals in particular:

- Usable:User should easily find what they need in the BI tool.
- Content Rich:The portal is not just a report printing tool, it should contain more functionality such as advice, help, support information and documentation.
- Clean:The portal should be designed so it is easily understandable and not over-complex as to confuse the users
- Current:The portal should be updated regularly.
- Interactive:The portal should be implemented in a way that makes it easy for the user to use its functionality and encourage them to use the portal. Scalability and customization give the user the means to fit the portal to each user.
- Value Oriented:It is important that the user has the feeling that the DW/BI application is a valuable resource that is worth working on.

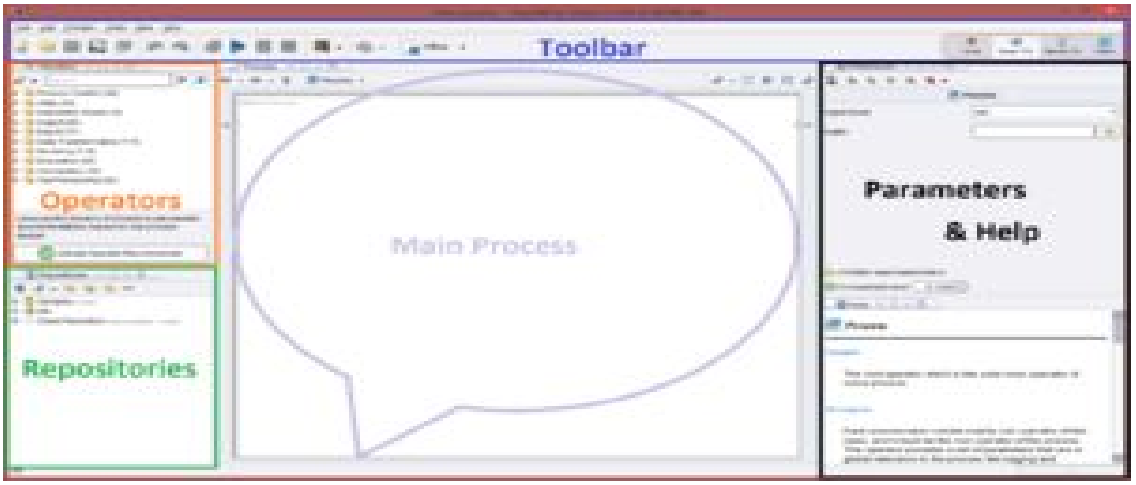
Business intelligence tools are a type of application software designed to retrieve, analyze, transform and report data for business intelligence. The tools generally read data that have been previously stored, often, though not necessarily, in a data warehouse or data mart.

The key general categories of business intelligence tools are:

1. Spreadsheets
2. Reporting and querying software: tools that extract, sort, summarize, and present selected data
3. OLAP: Online analytical processing
4. Digital dashboards
5. Data mining
6. Process Visualization
7. Data warehousing
8. Local information systems

The data mining tool used in this particular dataset mining in Rapidminer.

“RapidMiner provides data mining and machine learning procedures including: data loading and transformation (ETL), data preprocessing and visualization, modelling, evaluation, and deployment. The data mining processes can be made up of arbitrarily nestable operators, described in XML files and created in RapidMiner graphical user interface (GUI). RapidMiner is written in the Java programming language. It also integrates learning schemes and attribute evaluators of the Weka machine learning environment and statistical modelling schemes of the R-Project.”



As you launched RapidMiner Studio (v. 6.1.1000) you will need to install the Text Mining extension. RapidMiner works with extensions that plug into the core system. The Text Mining extension can be found in RapidMiner Marketplace, which can be accessed from Help > Updates and Extensions (Marketplace).
After restarting the software, we can start working with it. First of all create a New Process. You will see now the main window of RapidMiner Studio, and I will briefly describe the main zones of the working space :

- In blue we have the main toolbar
- In orange we can see all the operators that we can use in our processes
- In green we have the repositories
- In purple we have the main process windows, where we will be able to see process results and progression
- In black we have parameters of each element of or process and help

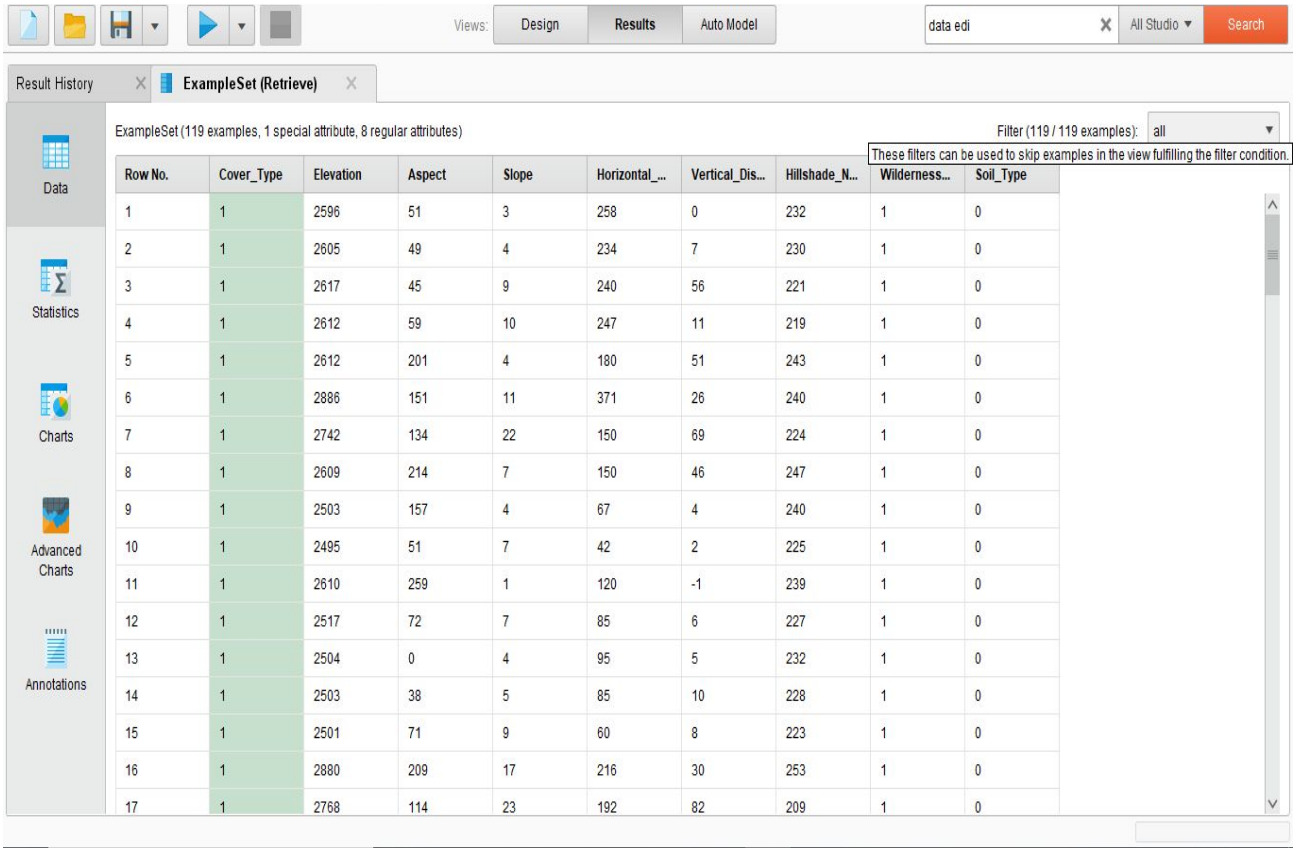
From here, we will first of all find our operator Process Documents from Files and we will drag it into the Process zone, in the center. At this point we have our operator in our process, and we need to set his parameters. Click on our operator in the main process area, and see which parameters you can set on the right side. First parameter is text directories which we will set right away.
Note : On the right side of your toolbar you can see a four-element menu that allows you to switch between Design and Results (also with F8 and F9 keys) that will be very useful. If your results aren't what you were expecting, or you made a mistake when designing your process, you can easily return from the results to the design area.

- In my case, i have a directory on my Desktop which name is "data"
- In /data/, I have crop.csv.
- I will set up my text directories like suggested in the and give both a different name to be able to show results depending on text directory

In next section we will talk about operators, and we will come back to Process Documents from Files parameters to choose which vector we want RapidMiner to create.

1.3 Data set:

- Title: Forest type
- Number of instance: 119
- Number of attributes: 9 (8 Numeric , 1 Nominal)
- Attribute description of forest dataset:
    - Elevation
    - Slope
    - Aspects
    - Horizontal_Distance_To_Hydrology
    - Vertical_Distance_To_Hydrology
    - Hillshade_Noon
    - Wilderness_Area
    - SoilType
    - Cover_Type: Forest cover type {1,2,3,5,6,7}
- Nominal Attribute:
  Cover_Type.
- Numeric Attribute:
  Elevation,Slope,Aspect,Horizontal_Distance_To_Hydrology,
  Vertical_Distance_To_Hydrology,Hillshade_Noon,Wilderness_Area,Soil_Type.



Row No.	Cover_Type	Elevation	Aspect	Slope	Horizontal_...	Vertical_Dis...	Hillshade_N...	Wilderness...	Soil_Type
1	1	2596	51	3	258	0	232	1	0
2	1	2605	49	4	234	7	230	1	0
3	1	2617	45	9	240	56	221	1	0
4	1	2612	59	10	247	11	219	1	0
5	1	2612	201	4	180	51	243	1	0
6	1	2886	151	11	371	26	240	1	0
7	1	2742	134	22	150	69	224	1	0
8	1	2609	214	7	150	46	247	1	0
9	1	2503	157	4	67	4	240	1	0
10	1	2495	51	7	42	2	225	1	0
11	1	2610	259	1	120	-1	239	1	0
12	1	2517	72	7	85	6	227	1	0
13	1	2504	0	4	95	5	232	1	0
14	1	2503	38	5	85	10	228	1	0
15	1	2501	71	9	60	8	223	1	0
16	1	2880	209	17	216	30	253	1	0
17	1	2768	114	23	192	82	209	1	0

Cover_Type is class of the dataset.Below is the snapshot of Dataset:

2. PROBLEM STATEMENT

2.1 Purpose of Dataset:

As the problem of global warming is arising. Government is saving land to grow forest. The following dataset studies the attribute and features of the land and based on the study it predicts which is the perfect cover type forest should be grown on particular land.

2.2 Dataset Analysis:

This dataset performs the test to find t which forest cover type is suitable for particular type of lands.This dataset has various factors based on the values of the particular attributes it selects the cover type of forest trees.
It also considers the soil type,slope,aspect and hillshade at particular time.There are 6 types of forest cover.1,2,3,5,6,7.

3.TECHNIQUE:

The data mining technique that is specifically used for this data set (College) is the classification technique.
Data analysis can be used for extracting models describing the important classes is the classification Data mining technique
Data Classification is a two step process:

1. Learning Step:- The training set is analyzed by using a classification algorithm.
2. Classification Step:- It includes a root node, branches and leaf node

3.1 Classification:

The learning and classification steps indicate that it is fast and simple while also having good accuracy for mining the data.
Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.
A classification task begins with a data set in which the class assignments are known. For example, a classification model that predicts credit risk could be developed based on observed data for many loan applicants over a period of time. In addition to the historical credit rating, the data might track employment history, home ownership or rental, years of residence, number and type of investments, and so on. Classifications are discrete and do not imply order. Continuous, floating-point values would indicate a numerical, rather than a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.
The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multi class targets have more than two values: for example, low, medium, high, or unknown credit rating.
In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown.
Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model.

Scoring a classification model results in class assignments and probabilities for each case. For example, a model that classifies customers as low, medium, or high value would also predict the probability of each classification for each customer.
Classification has many applications in customer segmentation, business modeling, marketing, credit analysis, and biomedical and drug response modeling.

The classification technique consists of the following methods
1. Decision tree method
2. Naive Bayes method
3. Random Forest method
4. Prune Tree method
5. Neural Network

4. ALGORITHM

4.1 Explanation:

Decision tree: Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value.It is one of the predictive modeling approaches used in statistics,data mining and machine learning.Tree models where the target variable can take a finite set of values are called classification tress.In these tree structures,leaves represent class labels and branches represent conjunctions of features that lead to those clas labels.Decision trees where the target variable can take continuous values (typically numbers) are called regression trees.
In decision analysis,a decision tree can be used to visually and explicitly represent decisions and decision making,In data mining, a decision tree describes data but not decisions;rather the resulting classification tree can be an input for decision making

Some techniques,often called ensemble methods,construct more than one decision tree:

1. **Bagging decision trees**,an early ensemble method,builds multiple decision trees by repeatedly re sampling training data with replacement and voting the trees for a consensus prediction.
2. **A Random Forest** classifier uses a number of decision trees,in order to improve the classification rate.
3. **Boosted Trees** can be used for regression-type and classification-type problems.
4. **Rotation Forest** in which every decision tree is trained by first applying principal component analysis(PCA) on a random subset of the input features.

Naive Bayes:- A class' prior may be calculated by assuming equiprobable classes(i.e.,priors=1/(number of classes)),or by calculating an estimate for the class probability from the training set(i.e(prior for a given class)=(number of samples in the class) / (total number of samples)).To estimate the parameters for a feature distribution,one must assume a distribution or generate nonparametric models for the features from the training set.

The assumptions on distribution of features are called the event model of the Naive Bayes classifier .for discrete features like the ones encountered in document classification(include spam filtering),multinomial and Bernoulli distributions are popular,These assumptions lead to two distinct models,which are often confused.

Supervised Learning
In Supervised Learning ,we are given a set of example pairs (x,y), $x \varepsilon X$, $y \varepsilon Y$ and the aim is to find a functions that matches the examples.In other words, we wish to infer the mapping implied by the data;the cost function is related to the mismatch between our mapping and the data and it implicitly contains prior knowledge about the problem domain.

A commonly used cost is the mean-squared error,which tires to minimize the average squared error between the network;s output,f(x),and the target value y over all the example pairs.when one tries to minimize this cost using gradient descent for the class of neural networks called multilayer perceptrons(MLP),one obtains the common and well known backpropagation algorithm for training neural networks.

Unsupervised Learning

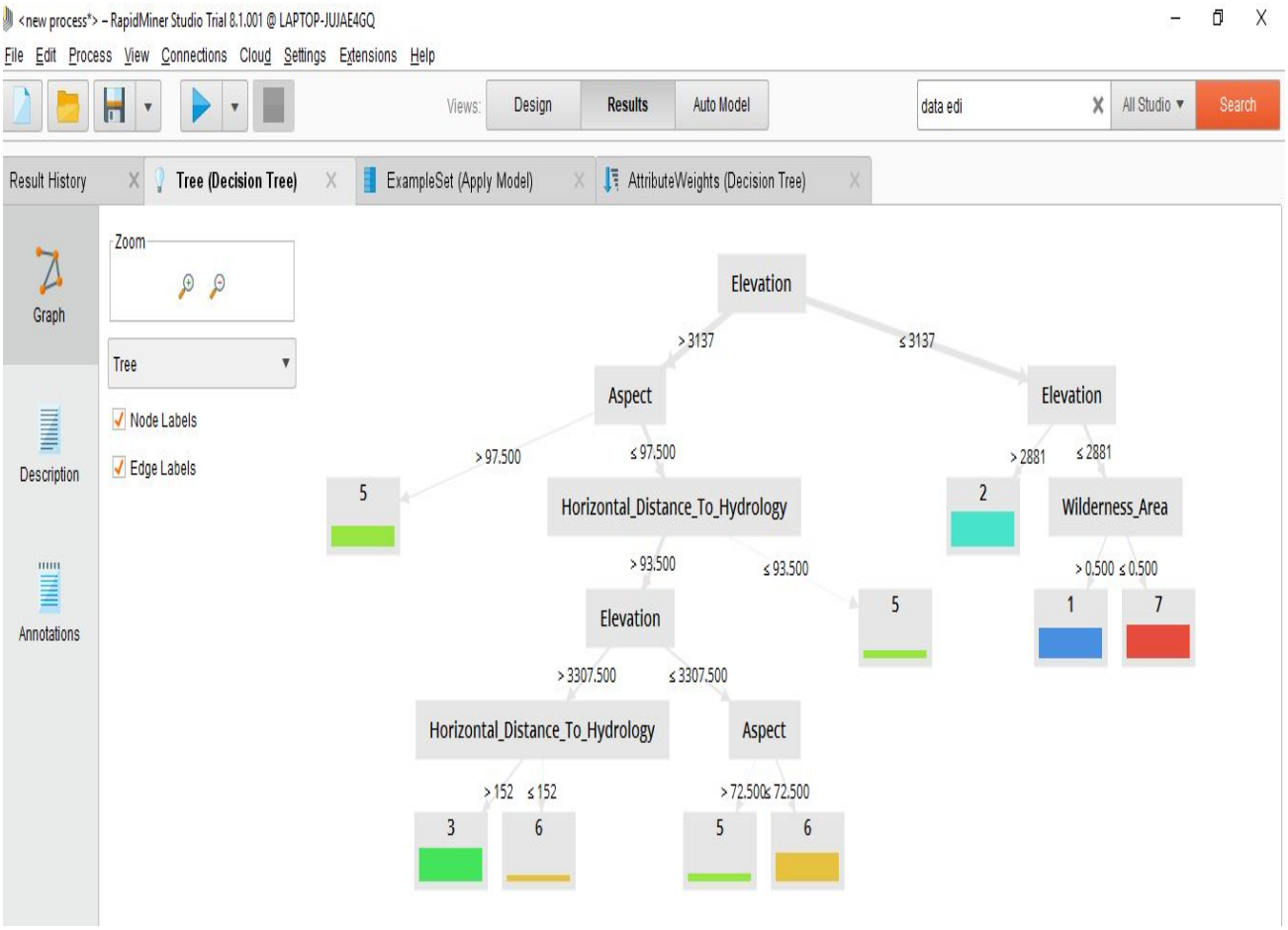
In Unsupervised Learning ,some data x is given and the cost function to be minimized, that can be any function of the data x and the networks output,f.

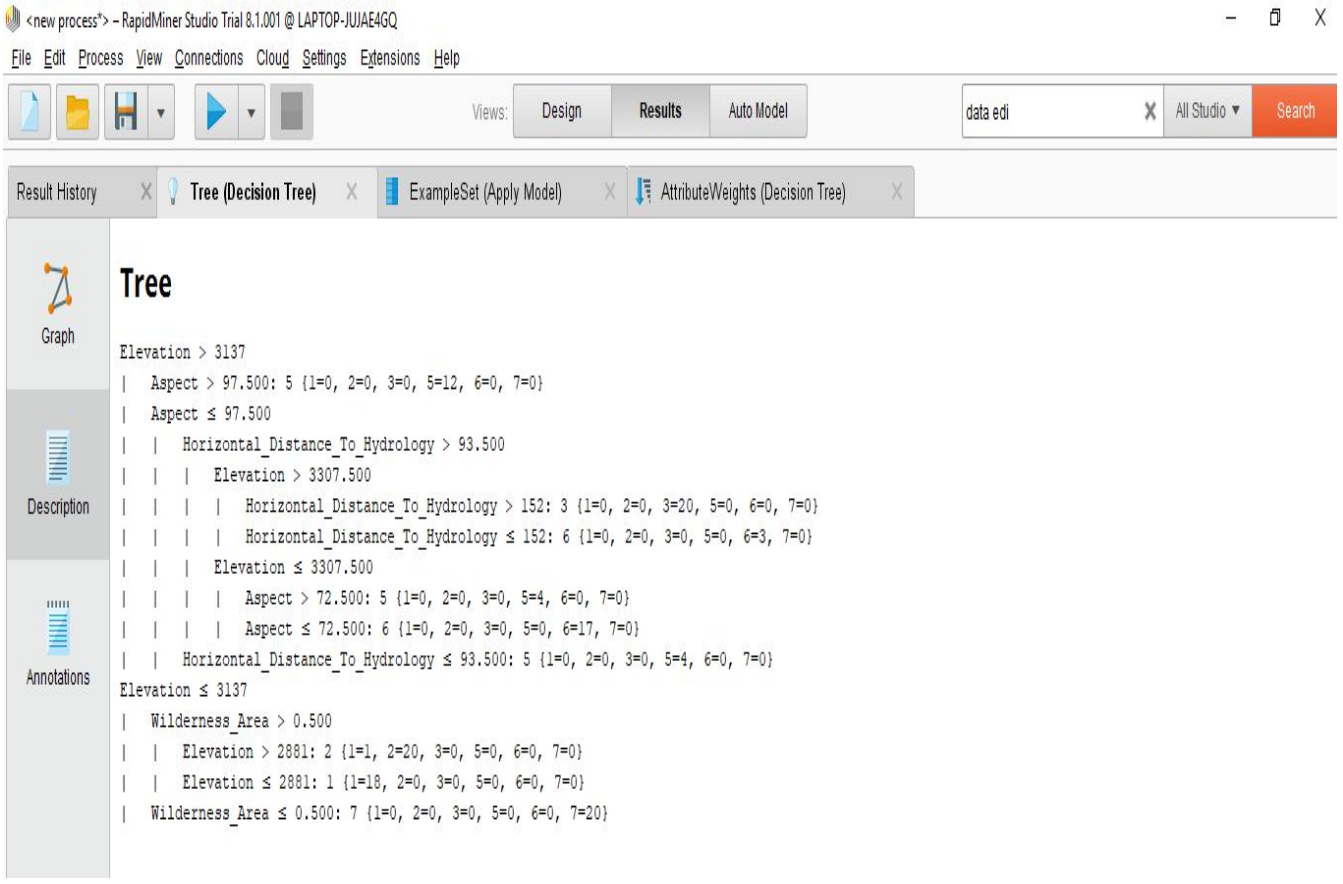
The cost function is dependent on the task(what we are trying to model) and our a priori assumptions(the implicit properties of our model,its parameters and the observed variables).

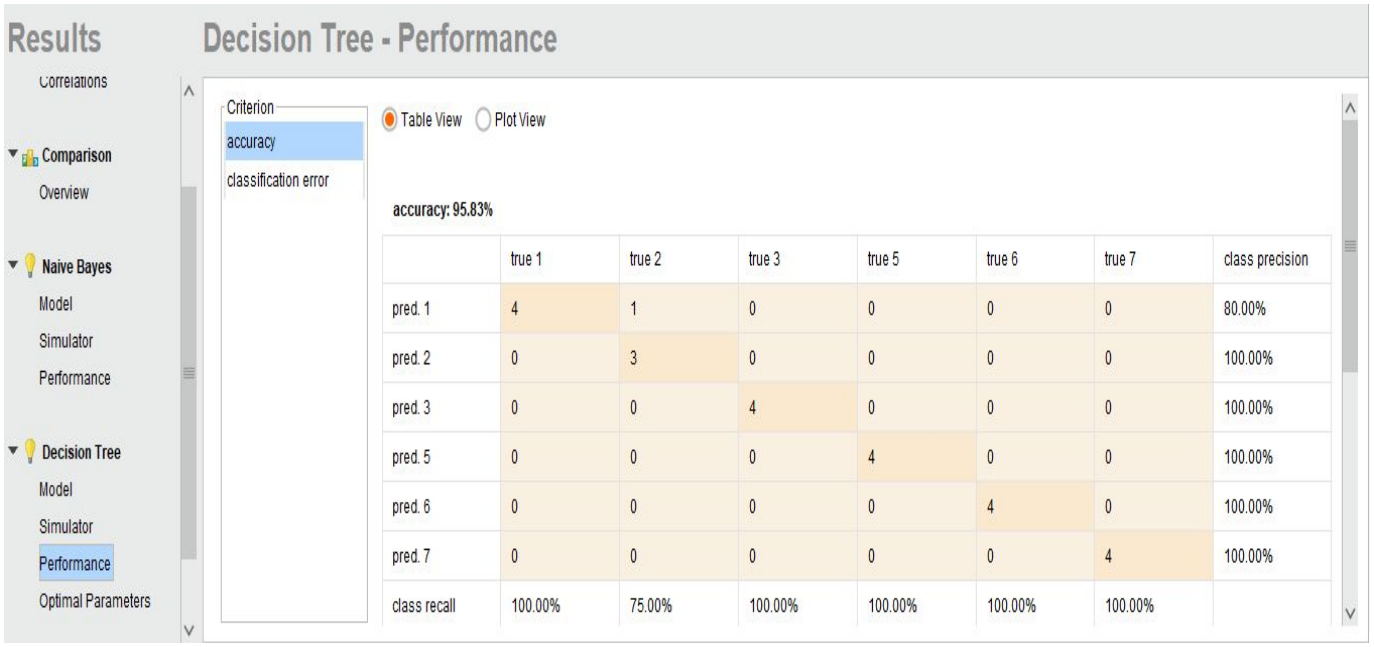
As a trivial example,consider the model f(x)=a where a is a constant and the cost C=E[(x-f(x))^2].Minimizing this cost will give us a value of that is equal to the mean of the data.The cost function can be much more complicated.Its form depends on the application:for example,in compression it could be related to the mutual information between x and whereas in statistical modeling,it could be related to the posterior probability of the model given the data(note that both of those examples those quantities would be maximized rather than minimized).

4.2 Snapshots

DECISION TREE:-


<new process*> – RapidMiner Studio Trial 8.1.001 @ LAPTOP-JUJAE4GQ
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Auto Model
data edi
All Studio
Search
Result History
Tree (Decision Tree)
ExampleSet (Apply Model)
AttributeWeights (Decision Tree)
Graph
Description
Annotations
Zoom
Tree
Node Labels
Edge Labels
Elevation
> 3137
≤ 3137
Aspect
Elevation
> 97.500
≤ 97.500
> 2881
≤ 2881
5
Horizontal_Distance_To_Hydrology
2
Wilderness_Area
> 93.500
≤ 93.500
> 0.500
≤ 0.500
Elevation
5
1
7
> 3307.500
≤ 3307.500
Horizontal_Distance_To_Hydrology
Aspect
> 152
≤ 152
> 72.500
≤ 72.500
3
6
5
6

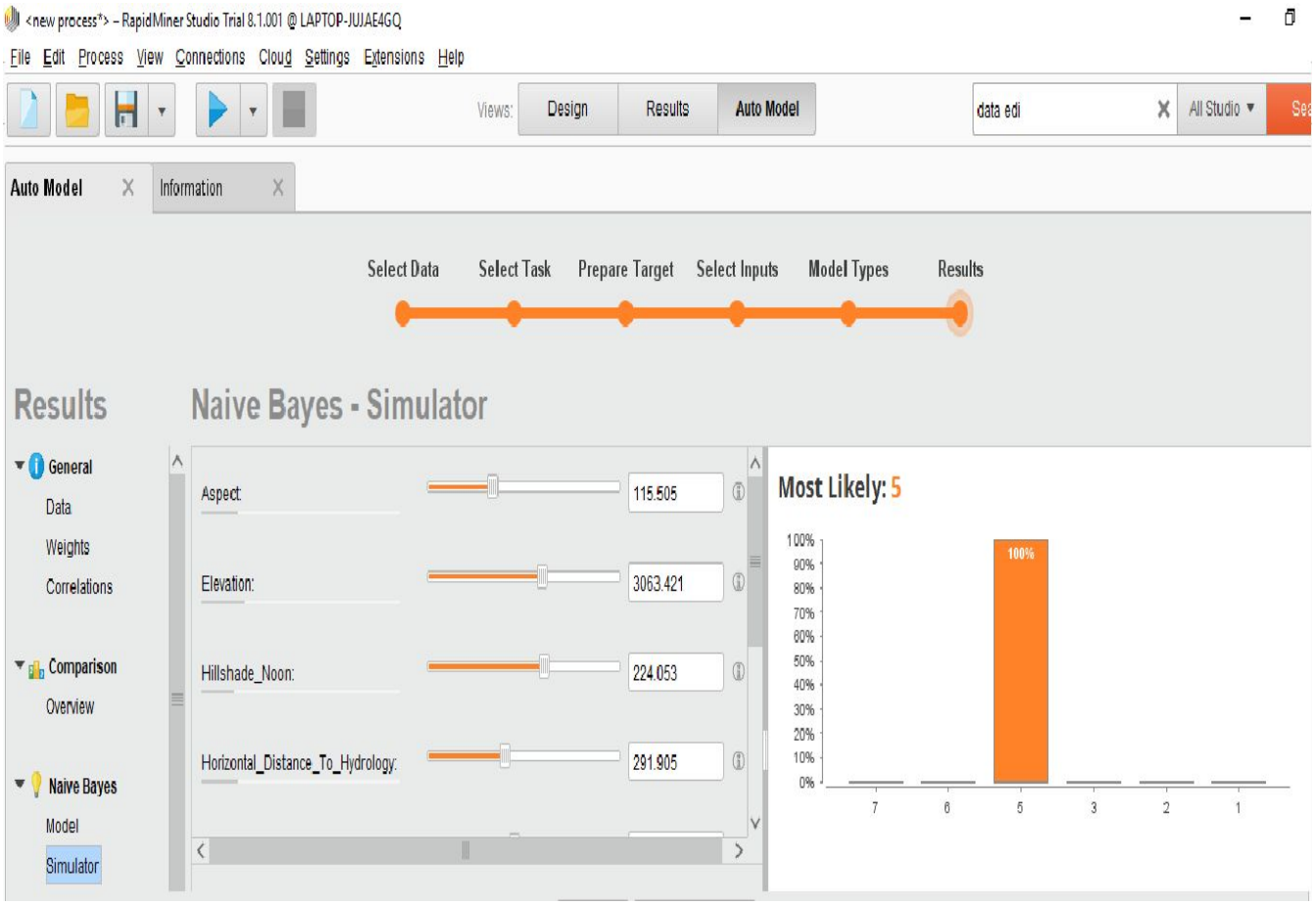

<new process*> – RapidMiner Studio Trial 8.1.001 @ LAPTOP-JUJAE4GQ
File Edit Process View Connections Cloud Settings Extensions Help
Views: Design Results Auto Model
data edi
All Studio
Search
Result History
Tree (Decision Tree)
ExampleSet (Apply Model)
AttributeWeights (Decision Tree)
Graph
Description
Annotations
Tree
Elevation > 3137
| Aspect > 97.500: 5 {1=0, 2=0, 3=0, 5=12, 6=0, 7=0}
| Aspect ≤ 97.500
| | Horizontal_Distance_To_Hydrology > 93.500
| | | Elevation > 3307.500
| | | | Horizontal_Distance_To_Hydrology > 152: 3 {1=0, 2=0, 3=20, 5=0, 6=0, 7=0}
| | | | Horizontal_Distance_To_Hydrology ≤ 152: 6 {1=0, 2=0, 3=0, 5=0, 6=3, 7=0}
| | | Elevation ≤ 3307.500
| | | | Aspect > 72.500: 5 {1=0, 2=0, 3=0, 5=4, 6=0, 7=0}
| | | | Aspect ≤ 72.500: 6 {1=0, 2=0, 3=0, 5=0, 6=17, 7=0}
| | Horizontal_Distance_To_Hydrology ≤ 93.500: 5 {1=0, 2=0, 3=0, 5=4, 6=0, 7=0}
Elevation ≤ 3137
| Wilderness_Area > 0.500
| | Elevation > 2881: 2 {1=1, 2=20, 3=0, 5=0, 6=0, 7=0}
| | Elevation ≤ 2881: 1 {1=18, 2=0, 3=0, 5=0, 6=0, 7=0}
| Wilderness_Area ≤ 0.500: 7 {1=0, 2=0, 3=0, 5=0, 6=0, 7=20}



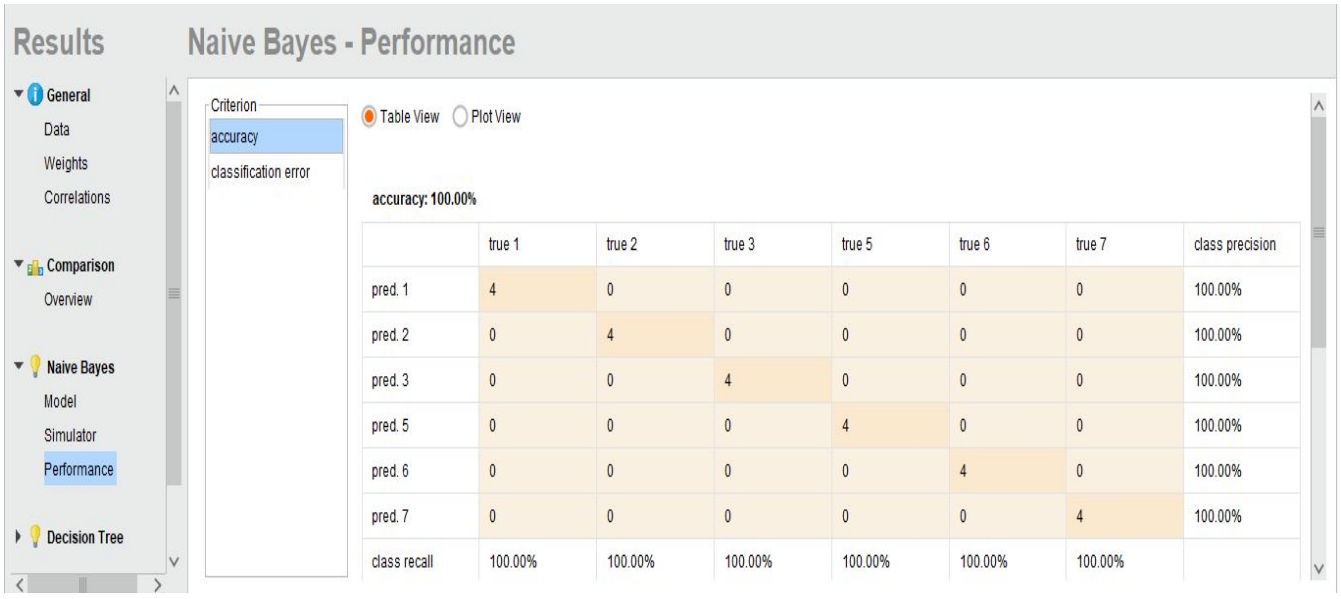
	true 1	true 2	true 3	true 5	true 6	true 7	class precision
pred. 1	4	1	0	0	0	0	80.00%
pred. 2	0	3	0	0	0	0	100.00%
pred. 3	0	0	4	0	0	0	100.00%
pred. 5	0	0	0	4	0	0	100.00%
pred. 6	0	0	0	0	4	0	100.00%
pred. 7	0	0	0	0	0	4	100.00%
class recall	100.00%	75.00%	100.00%	100.00%	100.00%	100.00%	

Naive Bayes Algorithm:

Prediction:



Accuracy of naive bayes algorithm:

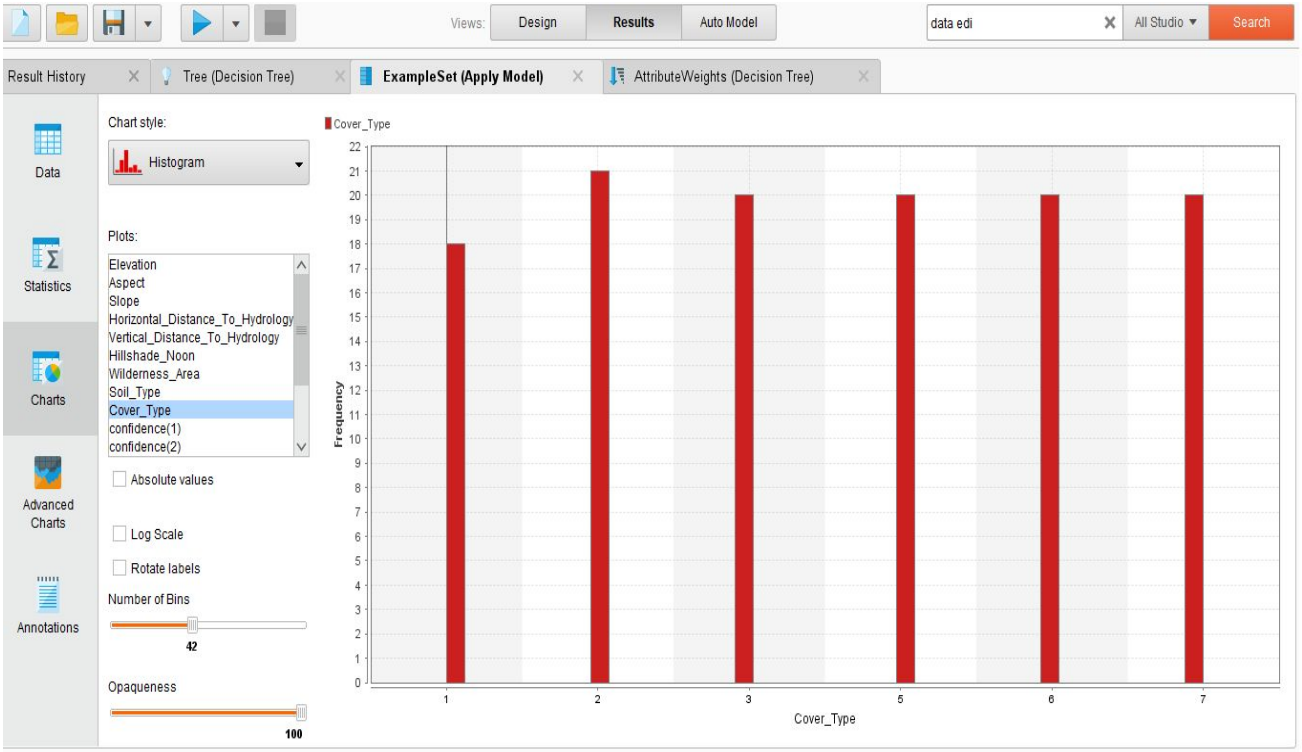


	true 1	true 2	true 3	true 5	true 6	true 7	class precision
pred. 1	4	0	0	0	0	0	100.00%
pred. 2	0	4	0	0	0	0	100.00%
pred. 3	0	0	4	0	0	0	100.00%
pred. 5	0	0	0	4	0	0	100.00%
pred. 6	0	0	0	0	4	0	100.00%
pred. 7	0	0	0	0	0	4	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	

5. INTERPRET AND VISUALIZE

5.1 Explanation :

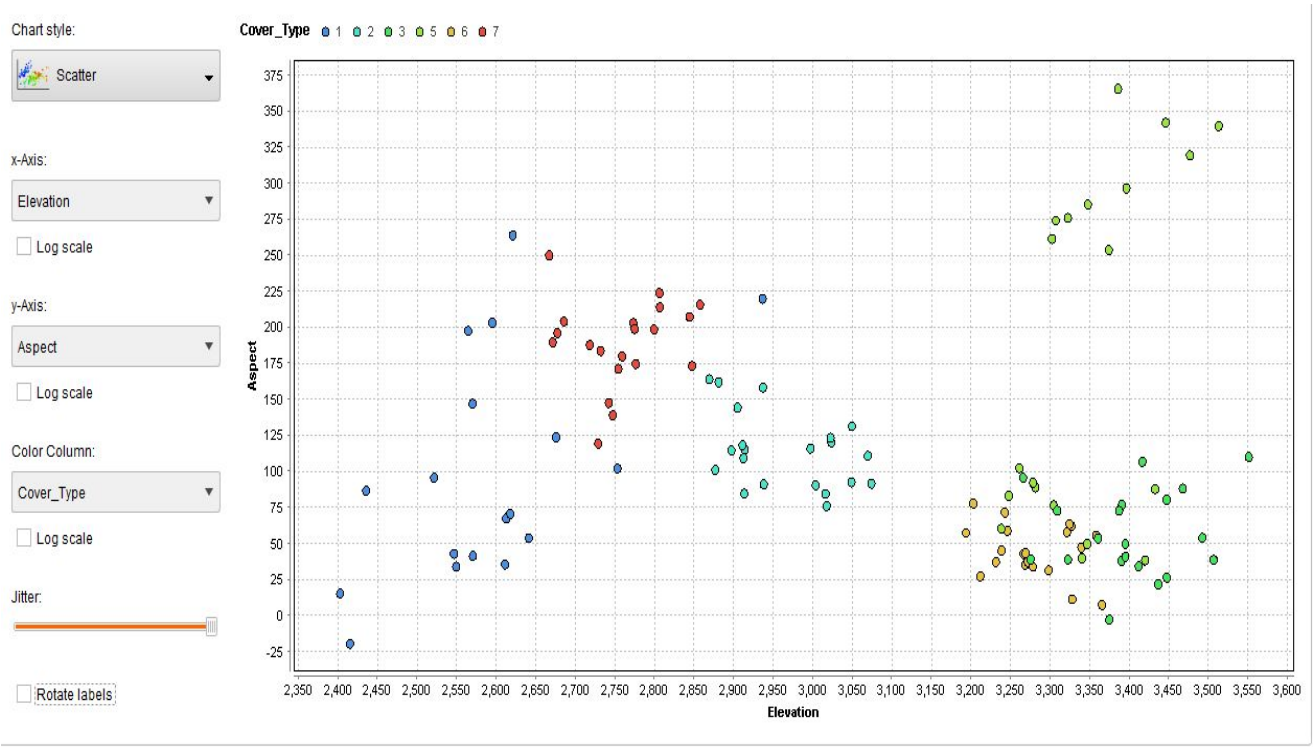
In the following histogram, the forest cover land shown.



Due to this forest cover lands can be deduced, which helps to know about properties of land. Therefore, it becomes possible to help to increase the forest cover land .The particular land can have different type of cover forest that can be planted.So this explains us forest cover that can be introduced on different land.

5.2 Histogram Visualization Technique Snapshots:

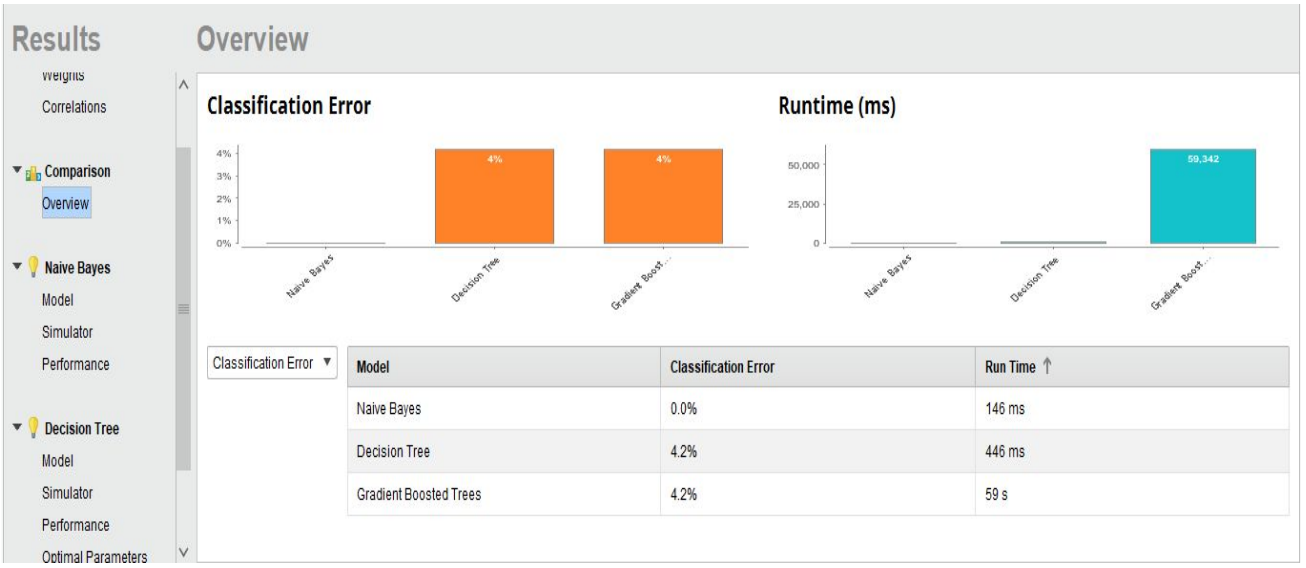
The above Scattered Plot shows the overall forest cover distribution of the lands



6.RESULT

Overall Error report

As we can see from the above report, from the different classification techniques the overall error percentage of the forest data set while using the Naive Bayes techniques has 0% errors compared to the Decision tree technique's error of 4.17%



Model Type	Accuracy	Error
Naive Bayes	100%	0%
Decision Tree	95.8%	4.2%
Gradient Boosted Tree	95.8%	4.2%

Therefore, the Naive Bayes classification technique is the most compatible data mining technique for the forest data set than decision tree and gradient boosted tree.