

Practical 1

Aim: Cases study of any three data mining applications and make a detailed note on them.

1) Healthcare

In health care areas, due to regulations and due to the availability of computers, a large amount of data is becoming available. On the one hand, practitioners are expected to use all this data in their work but, at the same time, such a large amount of data cannot be processed by humans in a short time to make diagnosis, prognosis and treatment schedules. Using data mining tools in medical and health care applications to develop a tool that can help make timely and accurate decisions.

Health care covers a detailed processes of the diagnosis, treatment and prevention of disease, injury and other physical and mental impairments in humans. The healthcare industry in most countries are evolving at a rapid pace. The healthcare industry can be regarded as place with rich data as they generate massive amounts of data including electronic medical records, administrative reports and other benchmarking finding.

Data mining is able to search for new and valuable information from these large volumes of data. Data mining in healthcare are being used mainly for predicting various diseases as well as in assisting for diagnosis for the doctors in making their clinical decision.

The data mining plays an important role in healthcare industry, especially in predicting various types of diseases. The diagnosis is widely being used in predicting diseases, they are extensively used in medical diagnosing. In healthcare, data mining has proven effective in areas such as predictive medicine, customer relationship management, detection of fraud and abuse, management of healthcare and measuring the effectiveness of certain treatments.

2) Facebook friend suggestion

Facebook does not randomly suggest friends. It has a very clever algorithm that does the job. Facebook is a savvy data miner. If you have not listed your high school information but if any of your family member or friends listed them then they would automatically suggest you members from that high school, Facebook's scrupulous data crawler will make the connection. Whenever you search for a friend, it will show the search result in order of highest number of mutual friends and similarities. For whatsapp contacts, facebook uses your contacts in mobile phone to suggest you the friends which are on facebook with those contact numbers. Facebook uses whole of your contact lists, email ids, or all the data stored on your mobile phone to get you suggest you the friends based on that data. Facebook does some amazingly intelligent stuff to infer who you know by what you do on Facebook. For example, let's say you have three friends: X, Y, and Z.

- On Facebook, you go out and you friend X and Y.
- X and Y happen to both know Z.
- So, because you're friends with X and Y, and because X and Y both know Z, it's possible that Facebook says, "You know what, if these two friends of yours know this person, maybe you do too. So I'll suggest them as a possible friend for you."

The same is the case with facebook pages. It starts suggesting you facebook pages based on the likes of your previous pages as well as your friends liking for that page. Facebook thinks like, "You know, you people all have a lot of common interests. Maybe you should get together; maybe you should know each other. In fact, maybe you do know each other, and I'll suggest them as possible friends." This is not the case always. Facebook Page is where Marketing business comes in the picture. Facebook charges an amount to boost up your page. For some amount, it will boost up your page to an appropriate profiles of your certain business region or area or people of similar interest. That's where it generates it's revenue by selling adds and page suggestions. So everything lies in algorithm which they have developed for the purpose.

3) Crime Analysis

Crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime. The system can predict regions which have high probability for crime occurrence and can visualize crime prone areas.

With the increasing advent of computerized systems, crime data analysts can help the Law enforcement officers to speed up the process of solving crimes. Using the concept of data mining we can extract previously unknown, useful information from an unstructured data.

Here we have an approach between computer science and criminal justice to develop a data mining procedure that can help solve crimes faster. Instead of focusing on causes of crime occurrence like criminal background of offender, political enmity etc. we are focusing mainly on crime factors of each day.

There are steps in doing Crime Analysis:

- 1) Data Collection
- 2) Classification
- 3) Pattern Identification
- 4) Prediction
- 5) Visualization

Practical 2

Aim: Study of data mining tools:

- i) **Rapid Miner**
- ii) **Weka**

Rapid Miner

RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics.

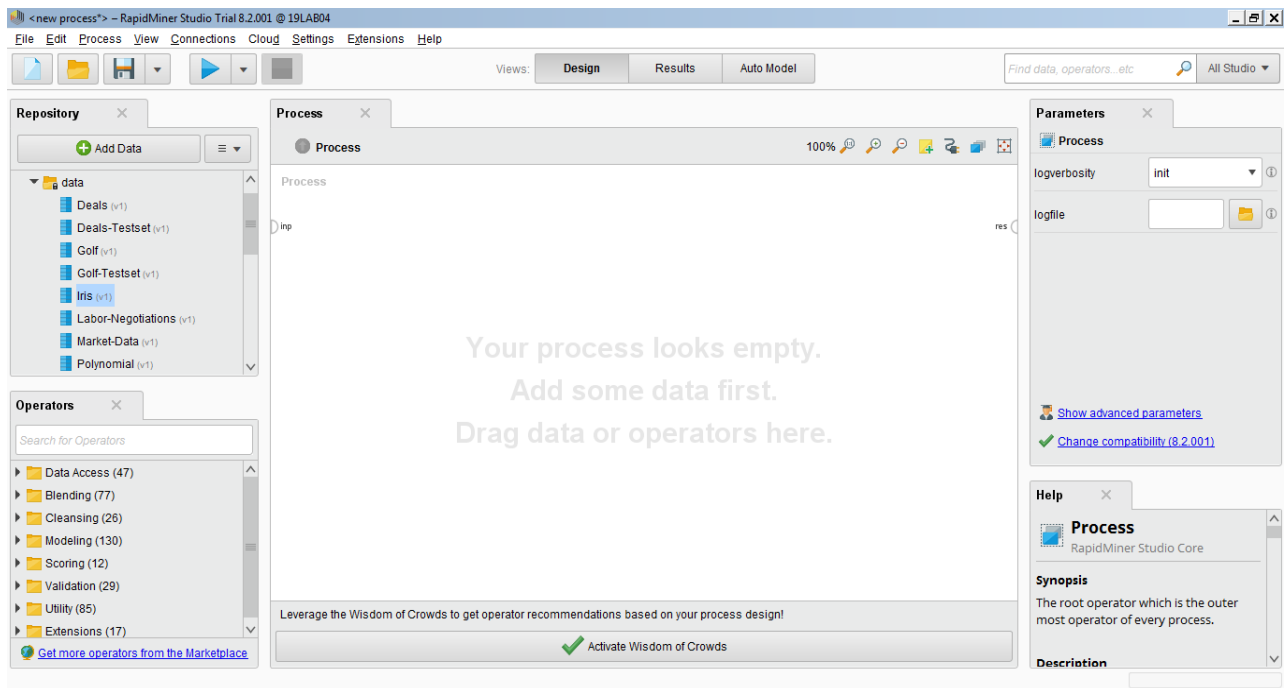


Fig. 1: Main workspace of Rapid Miner

RapidMiner is a software platform for data science teams that unites data prep, machine learning, and predictive model deployment.

RapidMiner uses a client/server model with the server offered as either on-premise, or in public or private cloud infrastructures

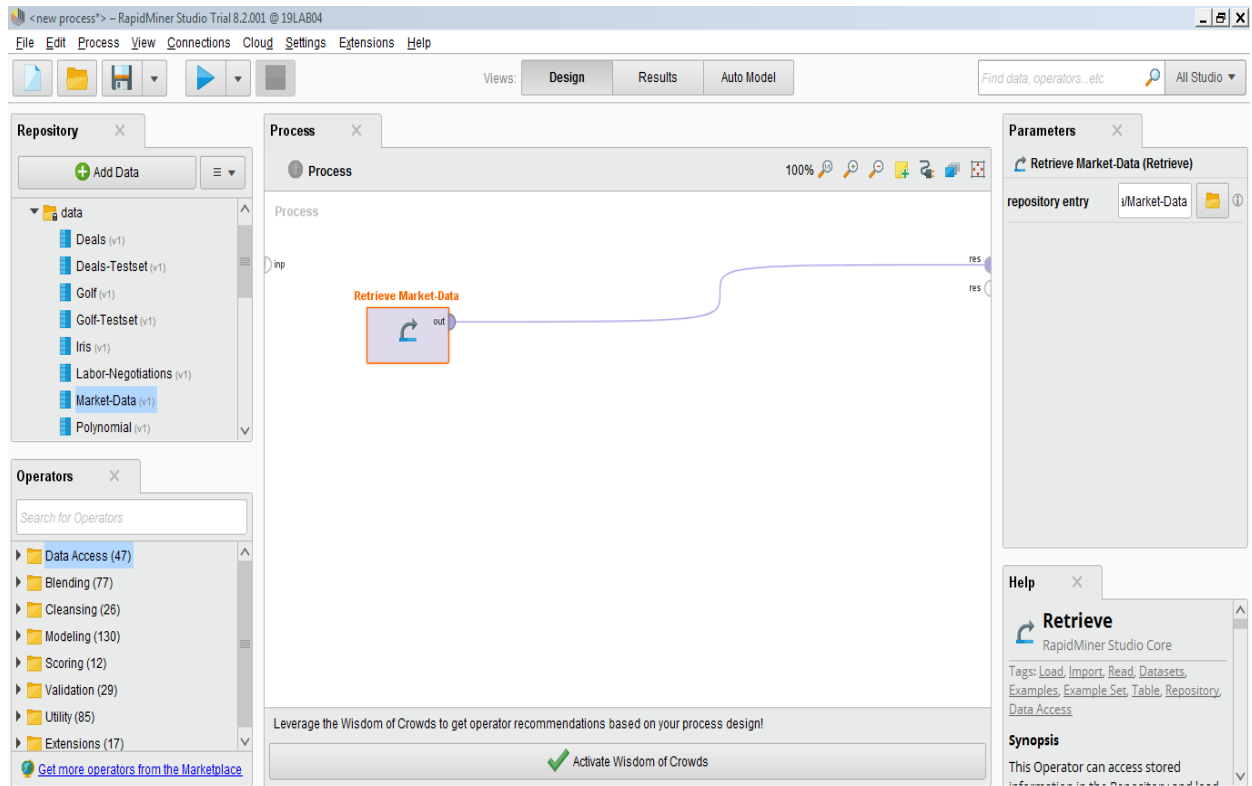


Fig. 2: Data set is added for process in design view

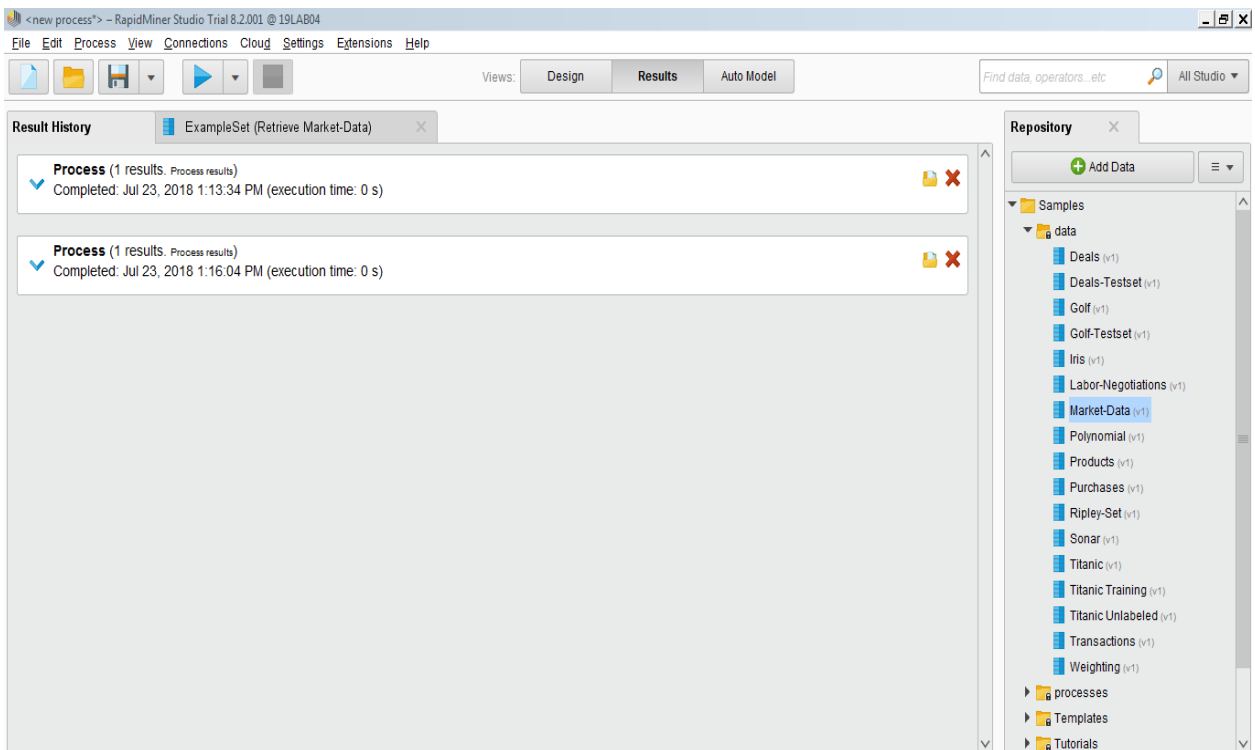
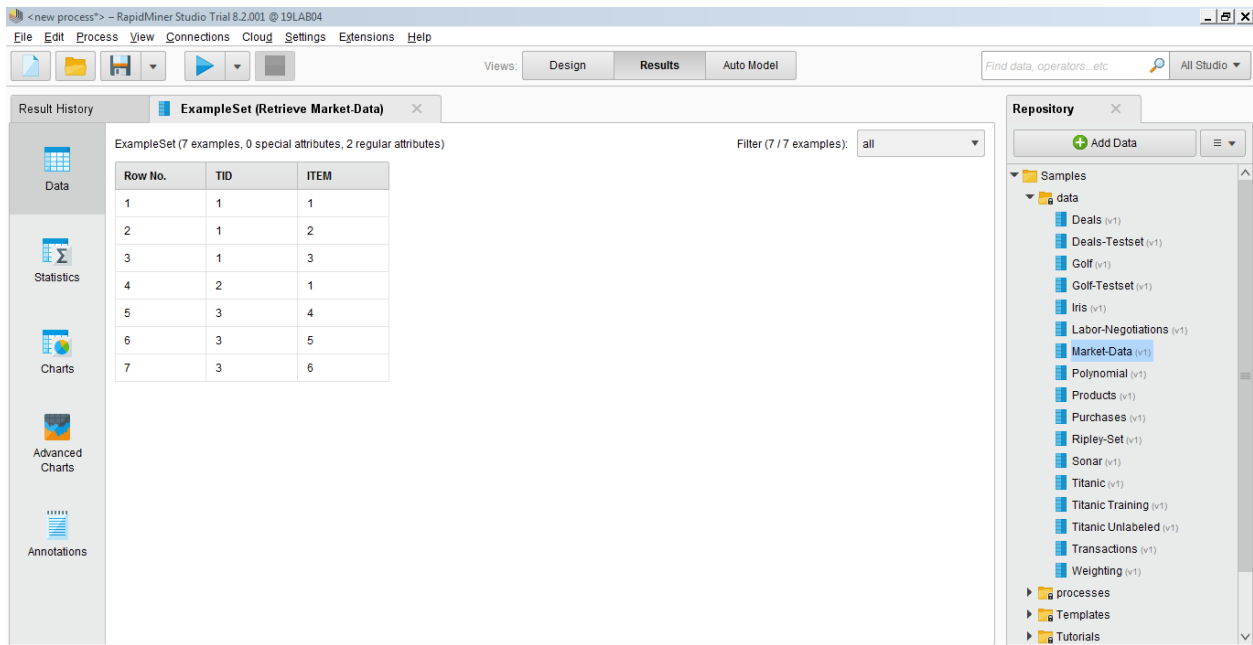


Fig. 3: Result History



Result History: ExampleSet (Retrieve Market-Data)

ExampleSet (7 examples, 0 special attributes, 2 regular attributes)

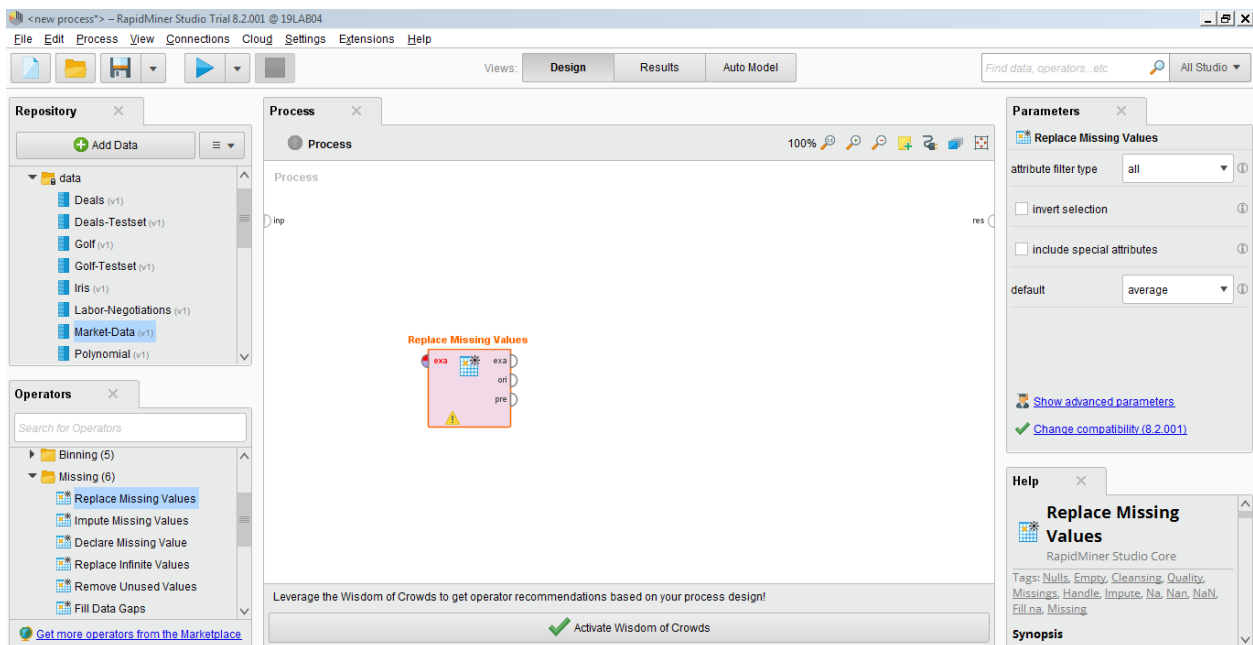
Filter (7 / 7 examples): all

| Row No. | TID | ITEM |
|---------|-----|------|
| 1 | 1 | 1 |
| 2 | 1 | 2 |
| 3 | 1 | 3 |
| 4 | 2 | 1 |
| 5 | 3 | 4 |
| 6 | 3 | 5 |
| 7 | 3 | 6 |

Repository:

- data
 - Deals (v1)
 - Deals-Testset (v1)
 - Golf (v1)
 - Golf-Testset (v1)
 - Iris (v1)
 - Labor-Negotiations (v1)
 - Market-Data (v1)
 - Polynomial (v1)
 - Products (v1)
 - Purchases (v1)
 - Ripley-Set (v1)
 - Sonar (v1)
 - Titanic (v1)
 - Titanic Training (v1)
 - Titanic Unlabeled (v1)
 - Transactions (v1)
 - Weighting (v1)
- processes
- Templates
- Tutorials

Fig. 4: Result dataset



Process: Replace Missing Values

Parameters:

- attribute filter type: all
- invert selection: ☐
- include special attributes: ☐
- default: average

Help: Replace Missing Values

Tags: Nulls, Empty, Cleansing, Quality, Missings, Handle, Impute, Na, Nan, NaN, Fill na, Missing

Synopsis

Fig. 5: Replace missing values

Features of RapidMiner:-

- Graphical user interface.
- Analysis processes design.
- Multiple data management methods.
- Data from file, database, web, and cloud services.
- In-memory, in-database and in-Hadoop analytics.
- Application templates.
- -D graphs, scatter matrices, self-organizing map.
- GUI or batch processing.

Advantages of RapidMiner:-

- Flow based programming allows visualization of pipelines
- Contains modules for statistical analysis, machine learning, etl, etc
- No coding required
- Easy to setup

Disadvantages of RapidMiner:-

- No coding required-Challenging to use for coders. Although it does contain Java/Python modules you must use flow programming interface.
- Commercial-Expensive licenses need to be purchased.
- Unintuitive-Its very easy to get lost in the sea of modules

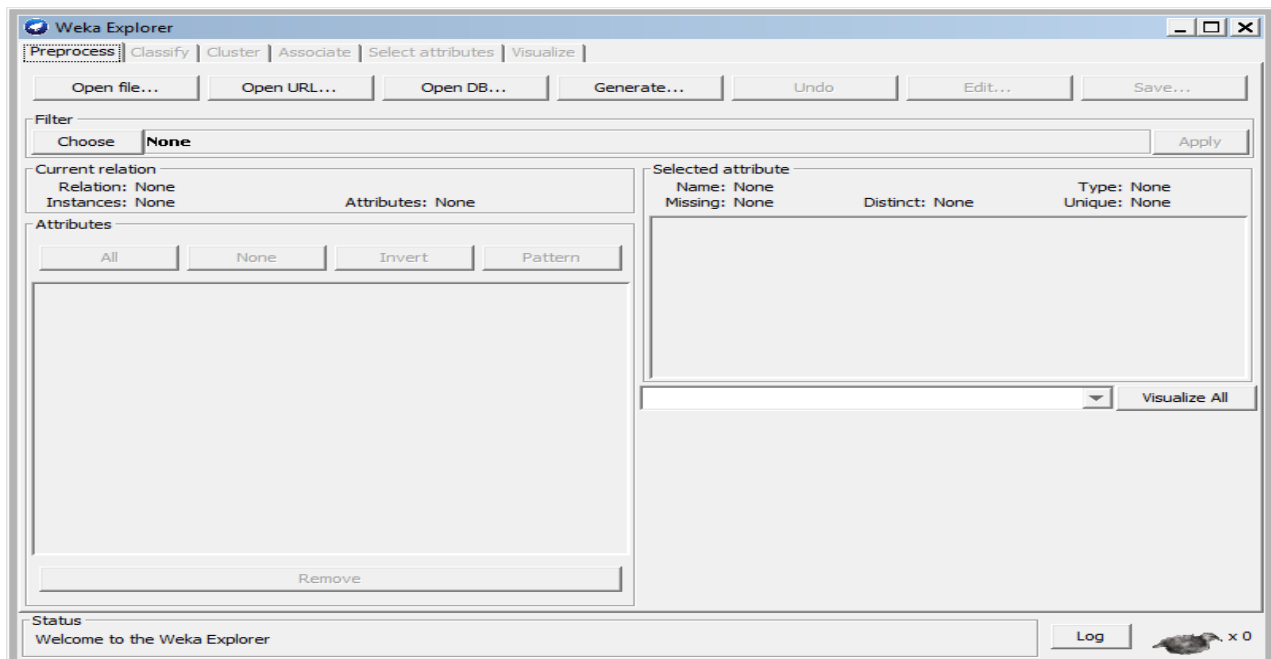
ii) Weka

Waikato Environment for Knowledge Analysis is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License.



Fig. 6: Weka main screen

Weka's main user interface is the *Explorer*.



In weka explorer when we can open or select any file for analysis of data.

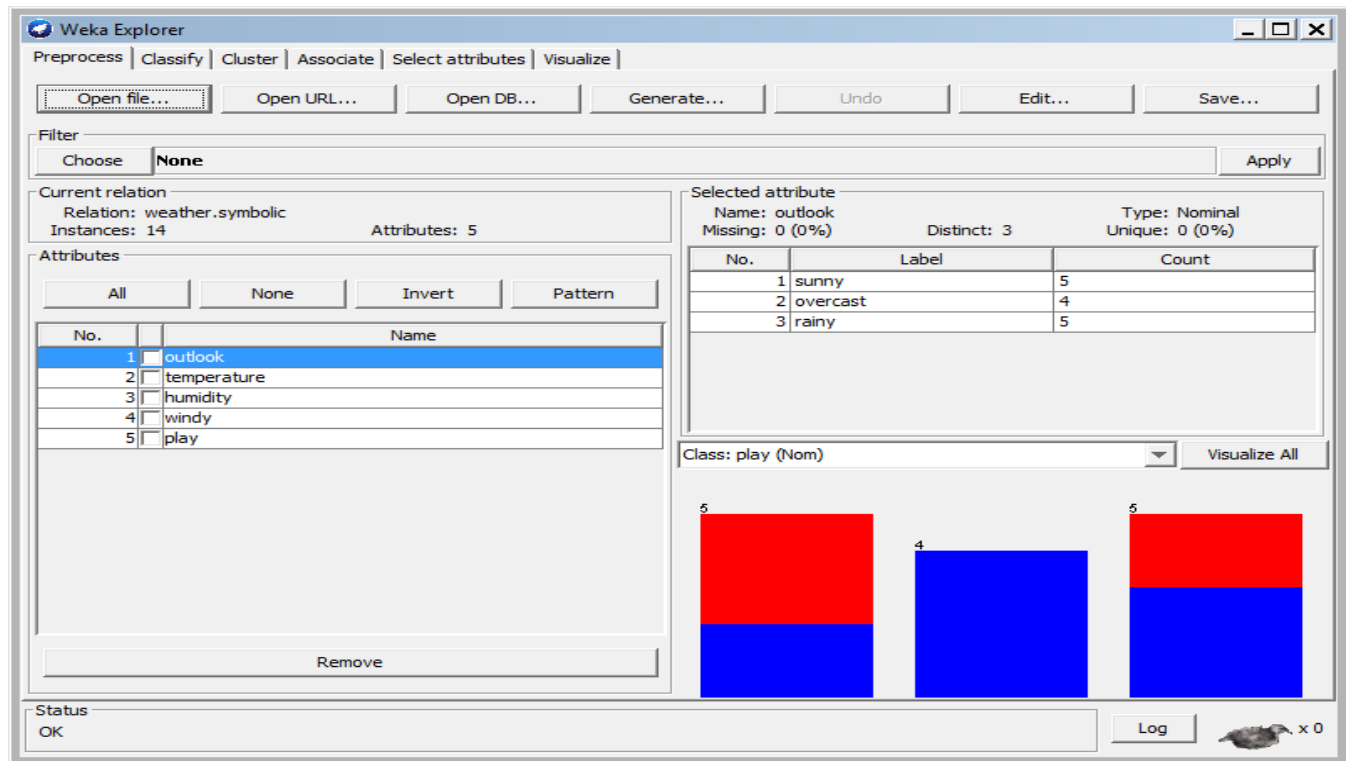


Fig. 7: Import data set from open file option

The screenshot shows the Weka Viewer window displaying the 'weather.symbolic' relation data. The table below shows the data for 14 instances.

| No. | outlook Nominal | temperature Nominal | humidity Nominal | windy Nominal | play Nominal |
|-----|--------------------|------------------------|---------------------|------------------|-----------------|
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 8 | sunny | mild | high | FALSE | no |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |
| 14 | rainy | mild | high | TRUE | no |

Fig. 8: Relation data view

Dataset of weather :

@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}

@attribute temperature {hot, mild, cool}

@attribute humidity {high, normal}

@attribute windy {TRUE, FALSE}

@attribute play {yes, no}

@data

sunny,hot,high,FALSE,no

sunny,hot,high,TRUE,no

overcast,hot,high,FALSE,yes

rainy,mild,high,FALSE,yes

rainy,cool,normal,FALSE,yes

rainy,cool,normal,TRUE,no

overcast,cool,normal,TRUE,yes

sunny,mild,high,FALSE,no

sunny,cool,normal,FALSE,yes

rainy,mild,normal,FALSE,yes

sunny,mild,normal,TRUE,yes

overcast,mild,high,TRUE,yes

overcast,hot,normal,FALSE,yes

rainy,mild,high,TRUE,no

In machine learning and statistics, **classification** is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

The Classify panel enables applying classification and regression algorithms (indiscriminately called classifiers in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, receiver operating characteristic (ROC) curves, etc.,

The Associate panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data

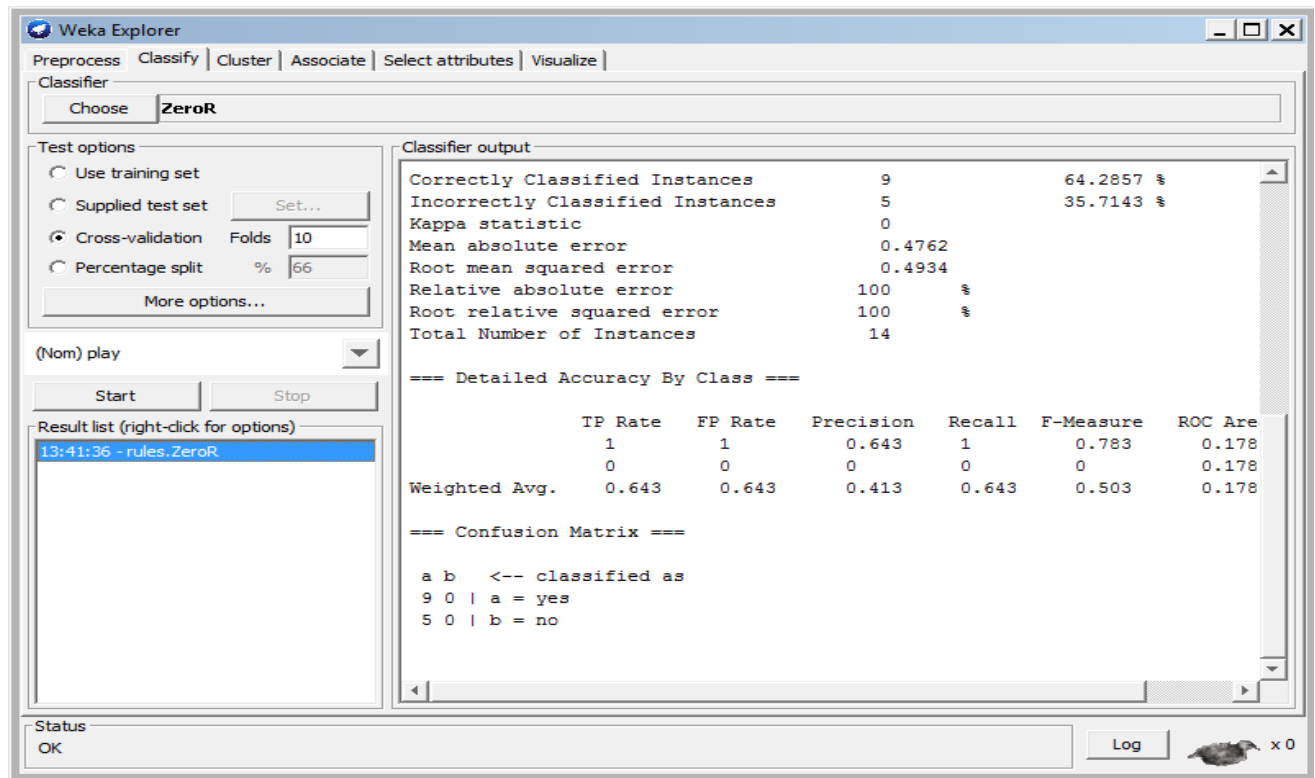


Fig. 9: Classification result view of weather data

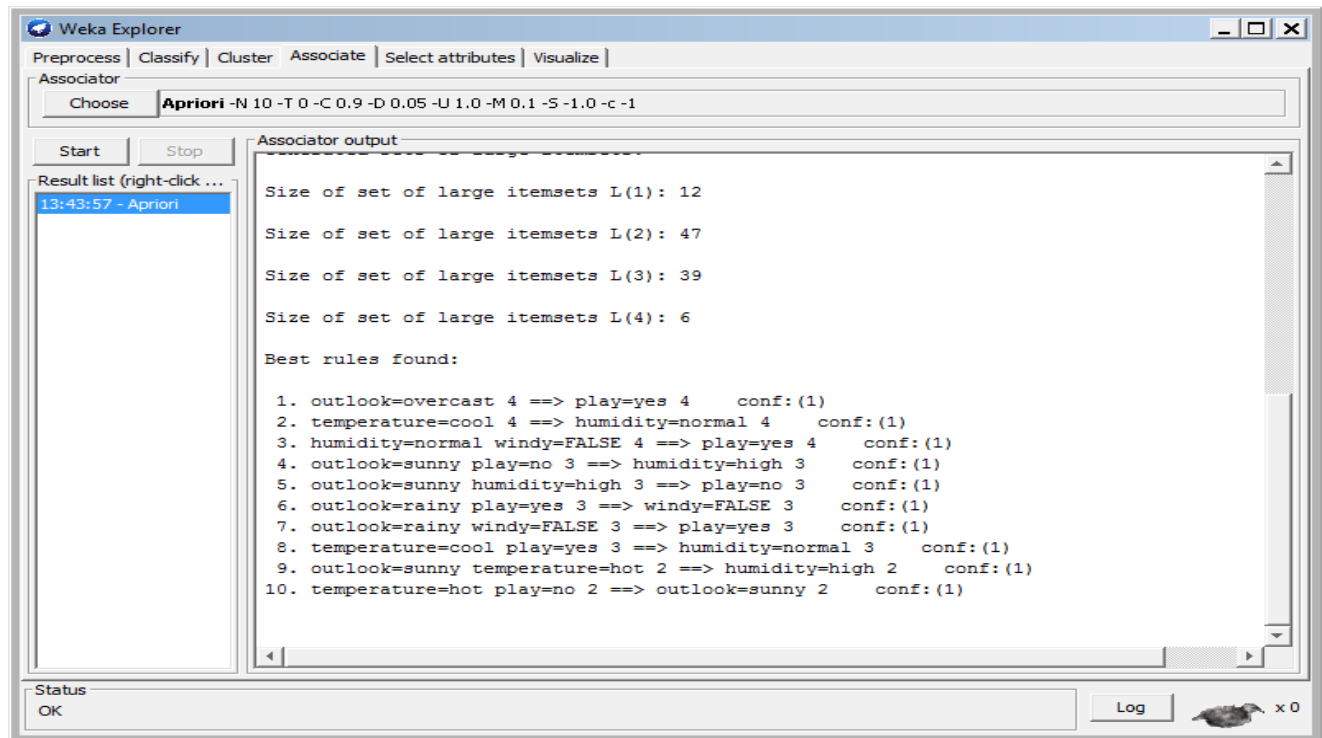


Fig. 10: Association result view of weather data

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

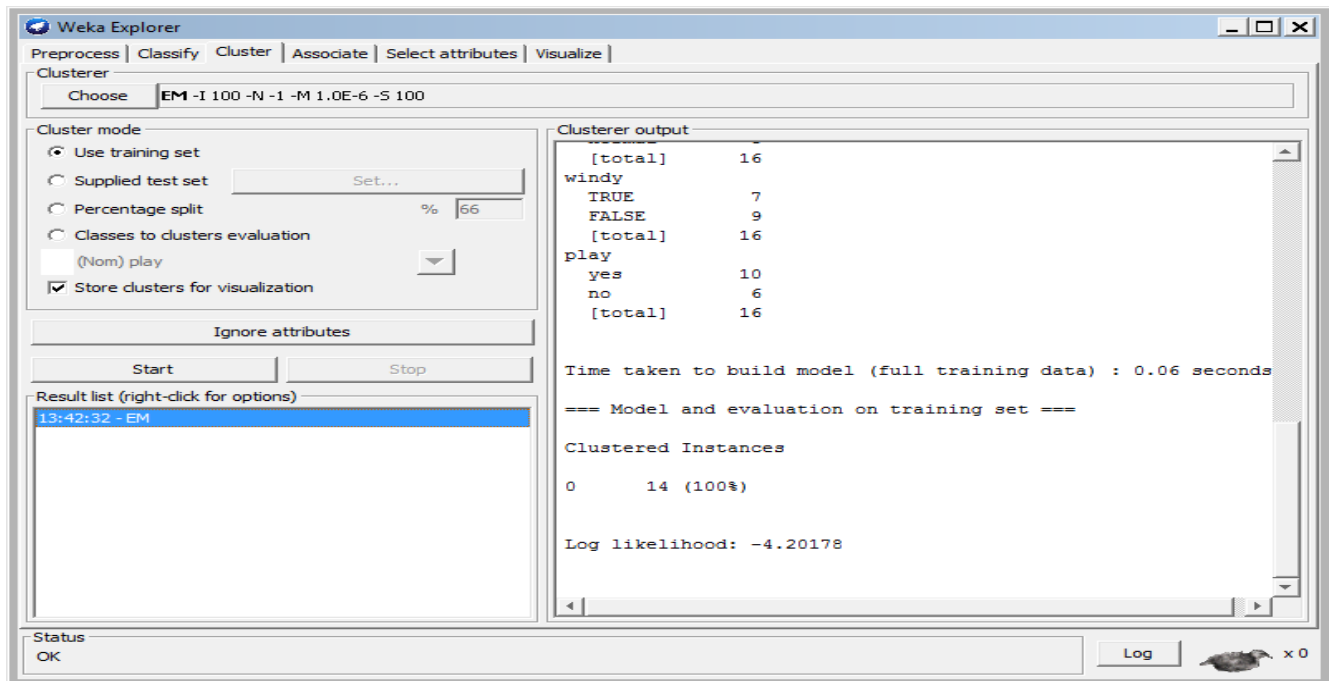
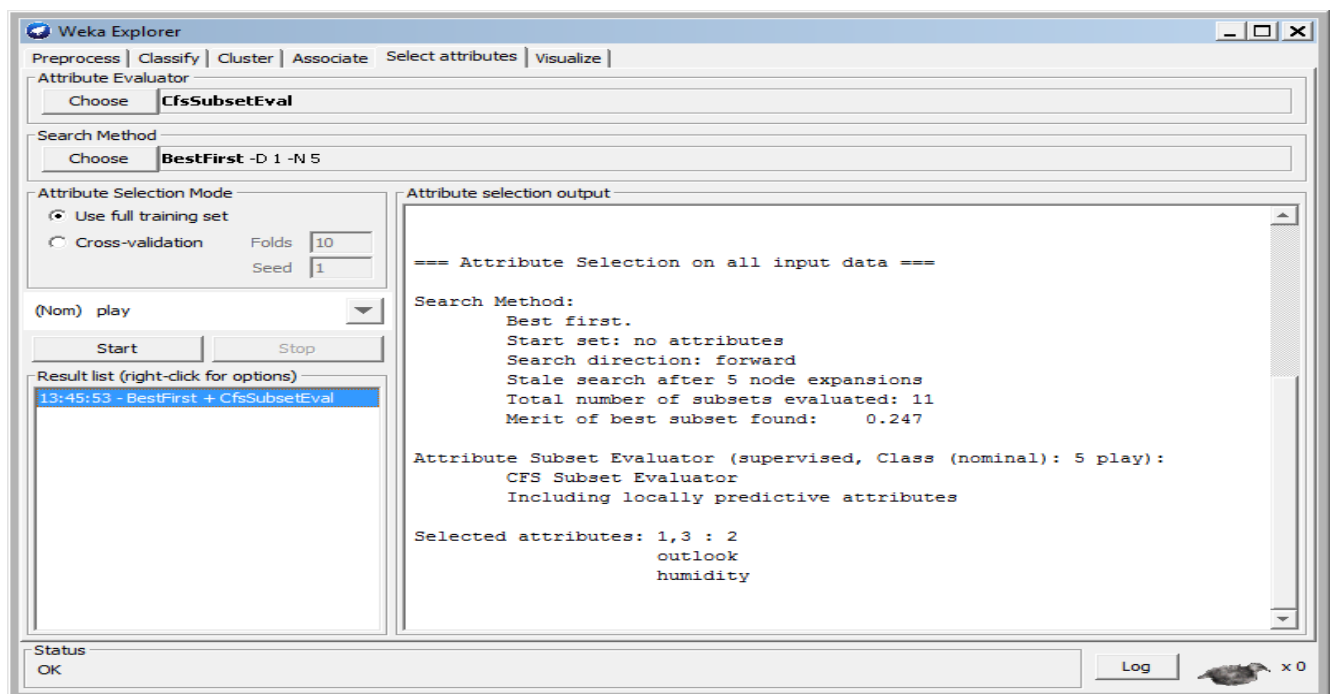


Fig. 11: Cluster result view of weather data



Access to **visualization** from the *Classifier*, *Cluster* and *Attribute Selection* panel is available from a popup menu. Click the right mouse button over an entry in the Result list to bring up the menu. You will be presented with options for viewing or saving the text output and depending on the scheme further options for visualizing errors, clusters, trees etc.

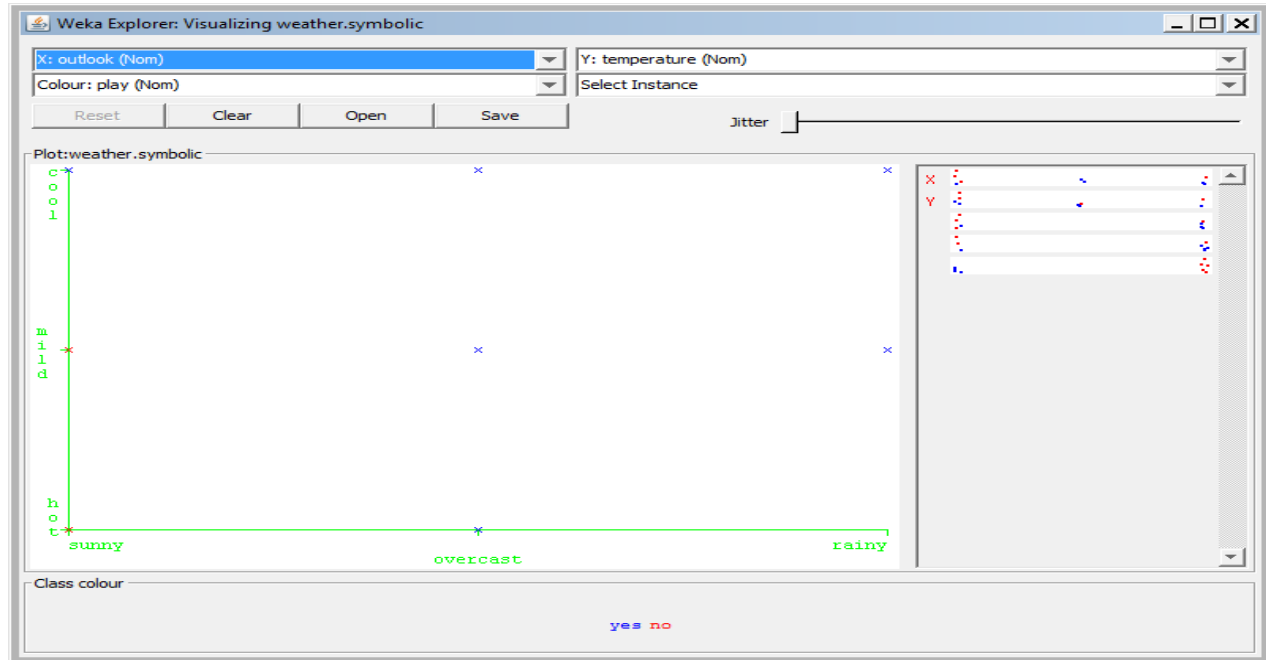


Fig. 12(a): Visualize result view of weather data

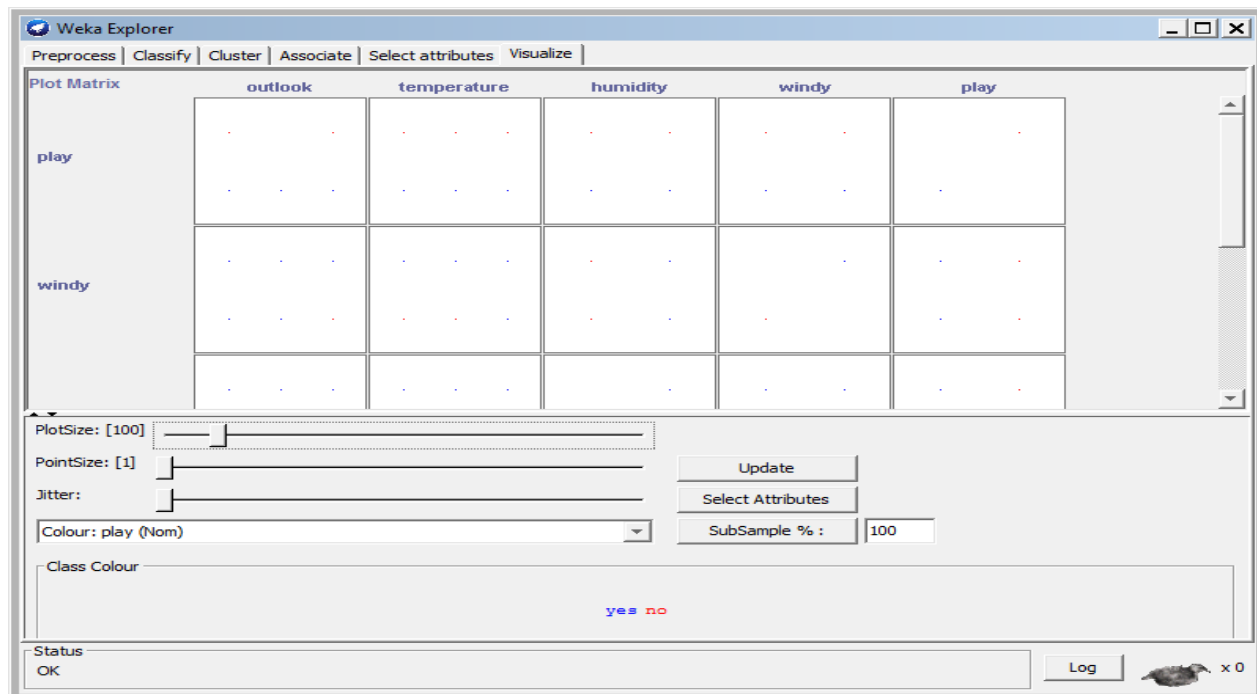


Fig. 12(b): Visualize result view of weather data

Key features for Weka:-

- It provides many different algorithms for data mining and machine learning
- It is open source and freely available
- It is platform-independent
- It is easily useable by people who are not data mining specialists
- It provides flexible facilities for scripting experiments
- It has kept up-to-date, with new algorithms being added as they appear in the research literature.
- Weka provides implementations of state-of-the-art data mining and machine learning algorithms.

Advantages of Weka:-

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.