DEV SANSKRITI VISHWAVIDYALAYA

SESSION 2018-21

Practical File
of

Data Mining and Warehousing

Submitted To:
Mr. Naveen Pandey
Assistant Professor

Submitted By:
Pranav mishra
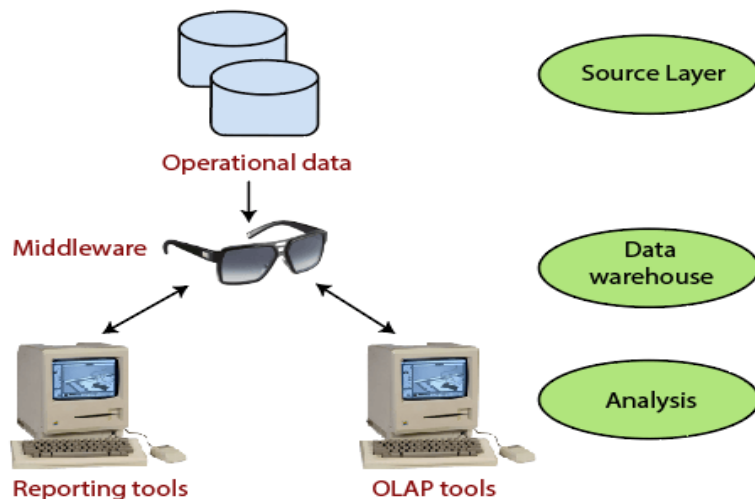BCA (6th Sem.)

**Department of Computer Science,
DSVV, Haridwar**

Index

Signature…………………

1. Draw and explain DATA WAREHOUSE ARCHITECTURE

- ## Single-Tier Architecture

Single-Tier architecture is not periodically used in practice. Its purpose is to minimize the amount of data stored to reach this goal; it removes data redundancies.
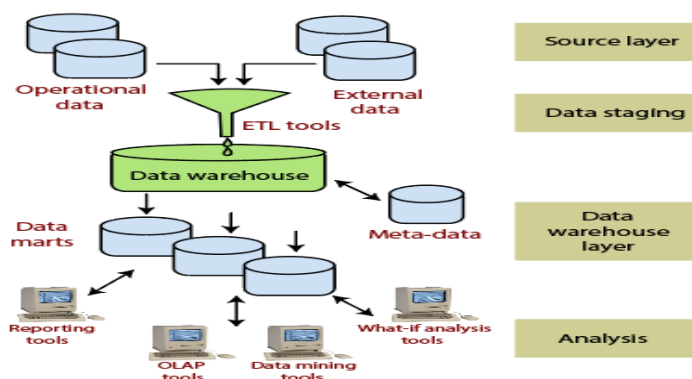

Single-Tier Data Warehouse Architecture

The vulnerability of this architecture lies in its failure to meet the requirement for separation between analytical and transactional processing.

- ## Two-Tier Architecture

The requirement for separation plays an essential role in defining the two-tier architecture for a data warehouse system.

Two-Tier Data Warehouse Architecture

Although it is typically called two-layer architecture to highlight a separation between physically available sources and data warehouses.

- Three-Tier Architecture

The three-tier architecture consists of the source layer (containing multiple source system), the reconciled layer and the data warehouse layer (containing both data warehouses and data marts). The reconciled layer sits between the source data and data warehouse.

This architecture is especially useful for the extensive, enterprise-wide systems.
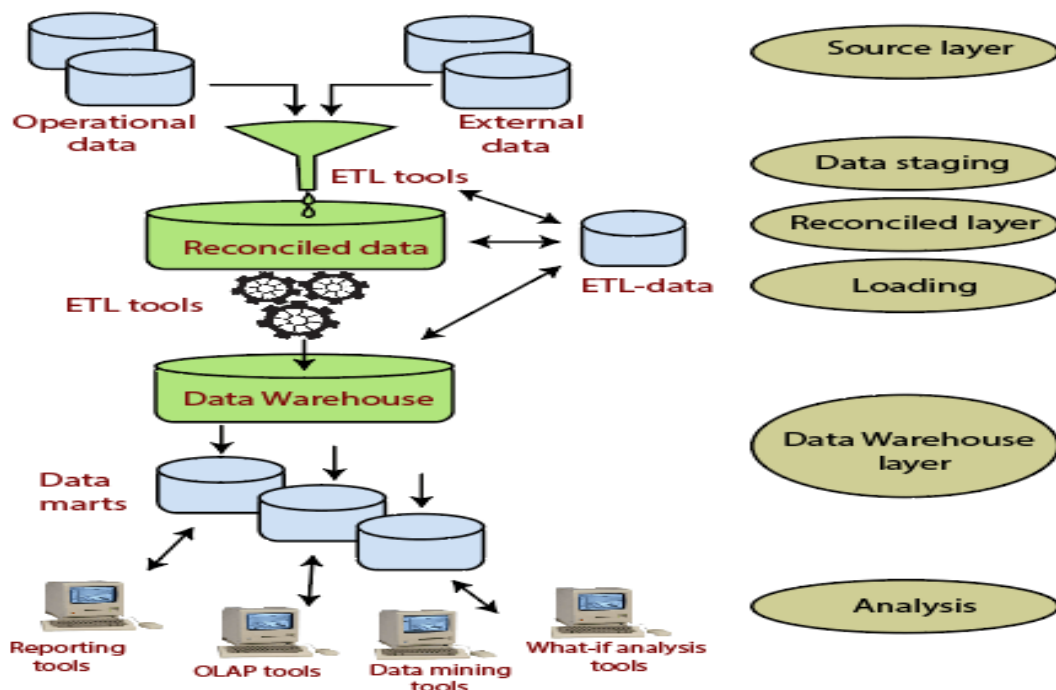
Three-Tier Architecture for a data warehouse system

2. Enlist and explain Data mining techniques

Data Mining Techniques

Data mining includes the utilization of refined data analysis tools to find previously unknown, valid patterns and relationships in huge data sets. These tools can incorporate statistical models, machine learning techniques, and mathematical algorithms, such as neural networks or decision trees. Thus, data mining incorporates analysis and prediction.

Classification:

This technique is used to obtain important and relevant information about data and metadata. This data mining technique helps to classify data in different classes.

Clustering:

Clustering is a division of information into groups of connected objects. Describing the data by a few clusters mainly loses certain confine details, but accomplishes improvement. It models data by its clusters. Data modelling puts clustering from a historical point of view rooted in statistics, mathematics, and numerical analysis.

Regression:

Regression analysis is the data mining process is used to identify and analyse the relationship between variables because of the presence of the other factor. It is used to define the probability of the specific variable. Regression, primarily a form of planning and modelling.\

Association Rules:

This data mining technique helps to discover a link between two or more items. It finds a hidden pattern in the data set.

Association rules are if-then statements that support to show the probability of interactions between data items within large data sets in different types of databases.

Outer detection:

This type of data mining technique relates to the observation of data items in the data set, which do not match an expected pattern or expected behaviour. This technique may be used in various domains like intrusion, detection, fraud detection, etc. It is also known as Outlier Analysis or Outlier mining.

Sequential Patterns:

The sequential pattern is a data mining technique specialized for **evaluating sequential data** to discover sequential patterns. It comprises of finding interesting sub sequences in a set of sequences, where the stake of a sequence can be measured in terms of different criteria like length, occurrence frequency, etc.

Prediction:

Prediction used a combination of other data mining techniques such as trends, clustering, classification, etc. It analyses past events or instances in the right sequence to predict a future event.

3. Case studies on Current trends and application of Data warehouse & Data mining Techniques

Data mining concepts are still evolving and here are the latest trends that we get to see in this field −

- Application Exploration.
- Scalable and interactive data mining methods.
- Integration of data mining with database systems, data warehouse systems and web database systems.
- Standardization of data mining query language.
- Visual data mining.
- New methods for mining complex types of data.
- Biological data mining.
- Data mining and software engineering.
- Web mining.
- Distributed data mining.
- Real time data mining.
- Multi database data mining.
- Privacy protection and information security in data mining.

Applications of Data Warehouse

The most important industries in which data warehouses are used -

- **Supply Chain** - Data warehouses are commonly used for the distribution and marketing of supermarket chains. It also helps monitor products, consumer shopping trends, and promotions and also helps to establish pricing strategies.
- **Investment** - The warehouses are mainly used to analyse patterns of knowledge, consumer trends, and to monitor market movements.
- **Public Sector** - Data warehouses are used in the public sector for collecting intelligence. It allows government agencies to retain and evaluate tax records, health policy records, for every person.
- **Health** - Data warehouse is also used by the healthcare industry to strategize and forecast performance, produce condition reports for patients, exchange data with tie-in insurance providers, medical assistance programs, etc.

- **Banking** - Identify the possible risk of default and collections are handled and monitored. The performance analysis of each commodity, service, interchange, and exchange rates can be easily calculated. Monitor the productivity of accounts and user details. It facilitates bankers with suggestions about consumer relationships and profitability.
- **Government** - Tax records, health policy records, and their respective providers are kept and evaluated. It can forecast the patterns and trends of criminal activities. This supports the hunt for terrorist profiles. It is possible to conduct threat evaluation and fraud detection.

Applications of Data Mining Techniques

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics.
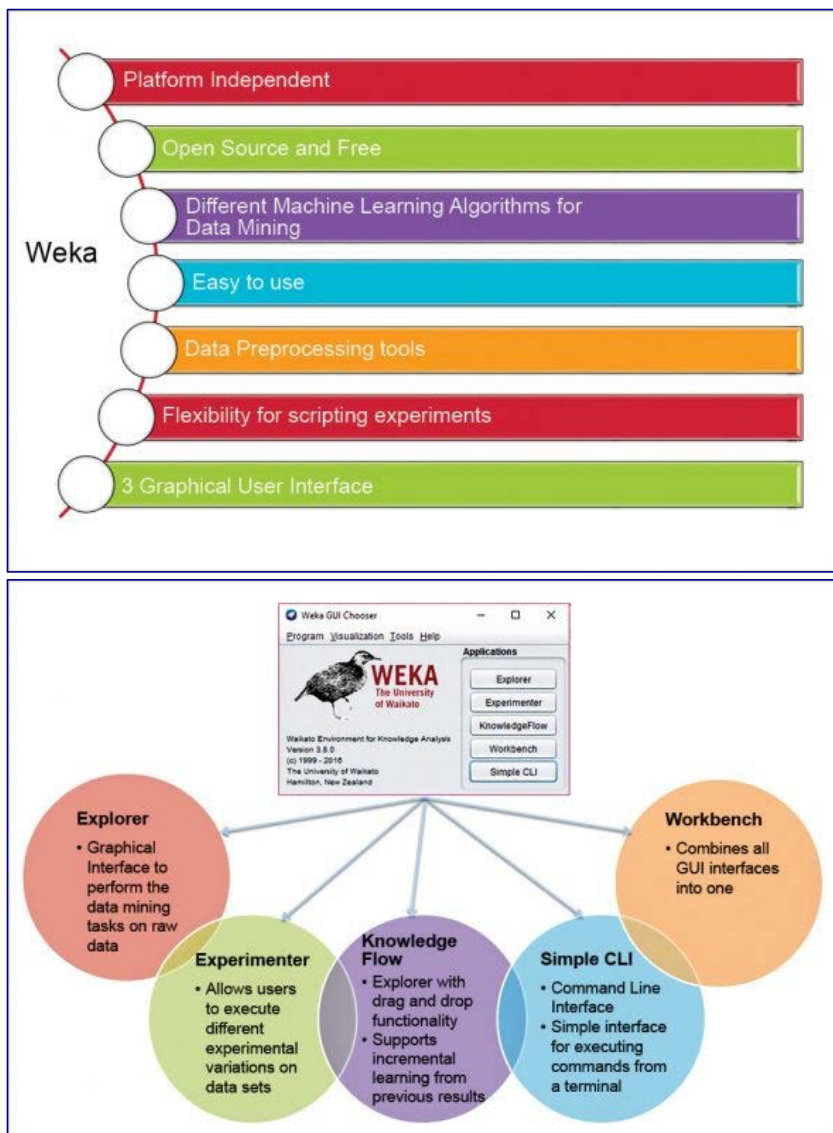
4. Brief Review of WEKA TOOL

Weka is data mining software that uses a collection of machine learning algorithms. These algorithms can be applied directly to the data or called from the Java code.
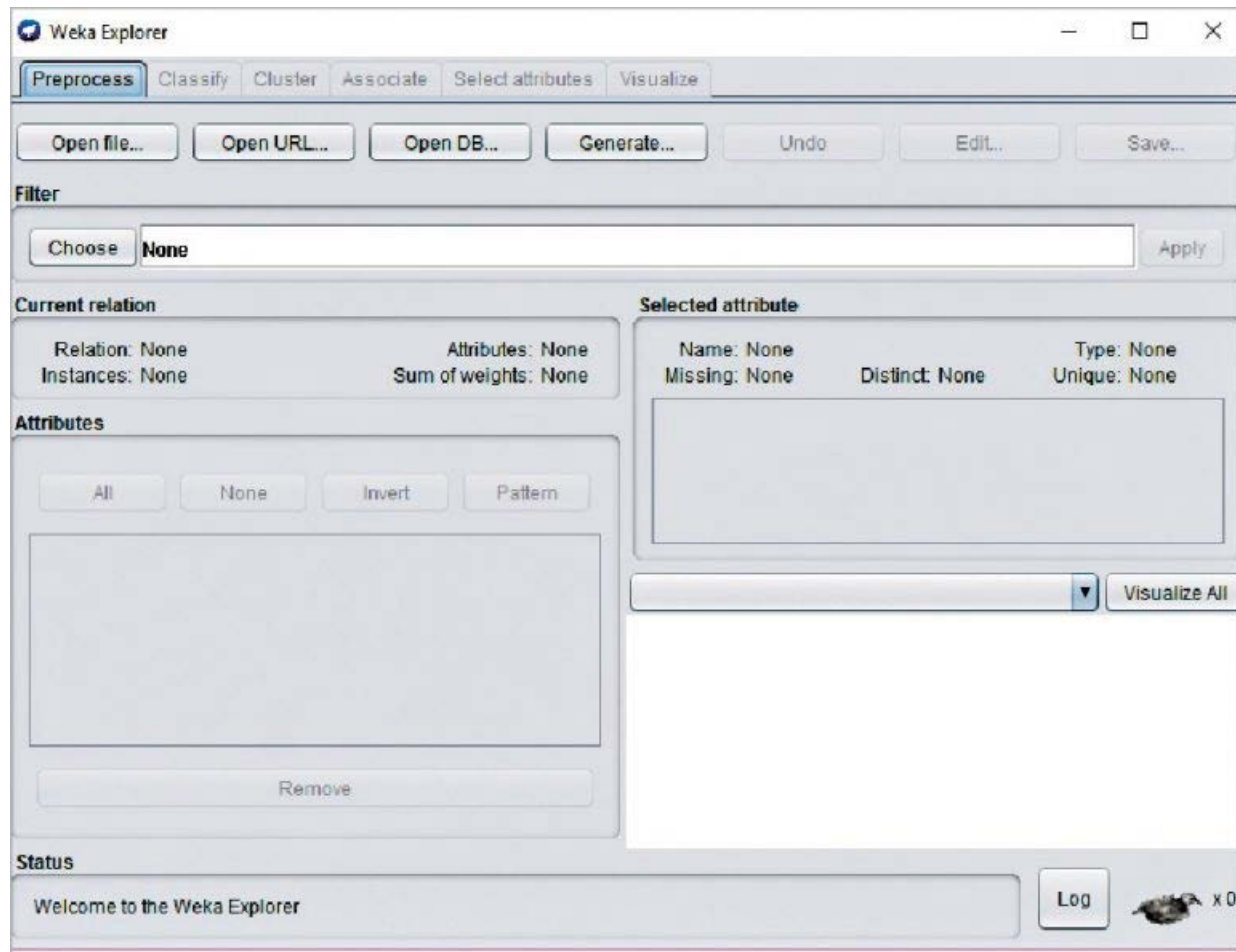
Weka is a collection of tools for:

- Regression
- Clustering
- Association
- Data pre-processing
- Classification
- Visualisation

The features of Weka are shown in Figure:

Weka Explorer

The Weka Explorer is illustrated in Figure and contains a total of six tabs. The tabs are as follows.

1) *Preprocess:* This allows us to choose the data file.

2) *Classify:* This allows us to apply and experiment with different algorithms on preprocessed data files.

3) *Cluster:* This allows us to apply different clustering tools, which identify clusters within the data file.

4) *Association:* This allows us to apply association rules, which identify the association within the data.

5) *Select attributes:* These allow us to see the changes on the inclusion and exclusion of attributes from the experiment.

6) *Visualize:* This allows us to see the possible visualisation produced on the data set in a 2D format, in scatter plot and bar graph output.

The user cannot move between the different tabs until the initial preprocessing of the data set has been completed.

Preprocessing

Data preprocessing is a must. There are three ways to inject the data for preprocessing:

- Open File – enables the user to select the file from the local machine
- Open URL – enables the user to select the data file from different locations
- Open Database – enables users to retrieve a data file from a database source

A screen for selecting a file from the local machine to be preprocessed is shown in Figure 5.

After loading the data in Explorer, we can refine the data by selecting different options. We can also select or remove the attributes as per our need and even apply filters on data to refine the result.
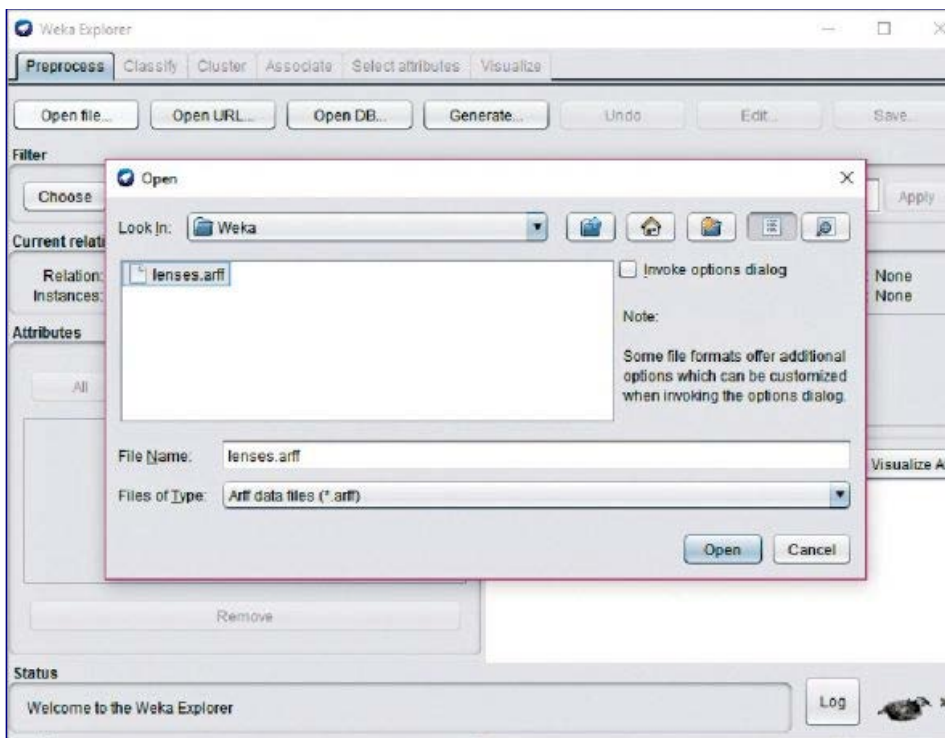
Classification

To predict nominal or numeric quantities, we have classifiers in Weka. Available learning schemes are decision-trees and lists, support vector machines, instance-based classifiers, logistic regression and Bayes' nets. Once the data has been loaded, all the tabs are enabled. Based on the requirements and by trial and error, we can find out the most suitable algorithm to produce an easily understandable representation of data.

Before running any classification algorithm, we need to set test options. Available test options are listed below.

Use training set: Evaluation is based on how well it can predict the class of the instances it was trained on.

Supplied training set: Evaluation is based on how well it can predict the class of a set of instances loaded from a file.

Cross-validation: Evaluation is based on cross-validation by using the number of folds entered in the 'Folds' text field.

Split percentage: Evaluation is based on how well it can predict a certain percentage of the data, held out for testing by using the values entered in the '%' field.

To classify the data set based on the characteristics of attributes, Weka uses classifiers.

Clustering

The cluster tab enables the user to identify similarities or groups of occurrences within the data set. Clustering can provide data for the user to analyse. The training set, percentage split, supplied test set and classes are used for clustering, for which the user can ignore some attributes from the data set, based on the requirements. Available clustering schemes in Weka are k-Means, EM, Cobweb, X-means and FarthestFirst.

Association

The only available scheme for association in Weka is the Apriori algorithm. It identifies statistical dependencies between clusters of attributes, and only works with discrete data. The Apriori algorithm computes all the rules having minimum support and exceeding a given confidence level.

Attribute selection

Attribute selection crawls through all possible combinations of attributes in the data to decide which of these will best fit the desired calculation—which subset of attributes works best for prediction. The attribute selection method contains two parts.

- *Search method:* Best-first, forward selection, random, exhaustive, genetic algorithm, ranking algorithm
- *Evaluation method:* Correlation-based, wrapper, information gain, chi-squared

All the available attributes are used in the evaluation of the data set by default. But it enables users to exclude some of them if they want to.

Visualisation

The user can see the final piece of the puzzle, derived throughout the process. It allows users to visualise a 2D representation of data, and is used to determine the difficulty of the learning problem. We can visualise single attributes (1D) and pairs of attributes (2D), and rotate 3D visualisations in Weka. It has the Jitter option to deal with nominal attributes and to detect 'hidden' data points.

5. Practical of Classification

Steps: Open Weka > Workbench > Open File > Classify > Choose Classifier > Start

Classifier Type: NaiveBayes

6. Practical of Regression Analysis

Steps: Open Weka > Workbench > Open File > Classify > Choose Classifier > Start

Classifier: Linear Regression
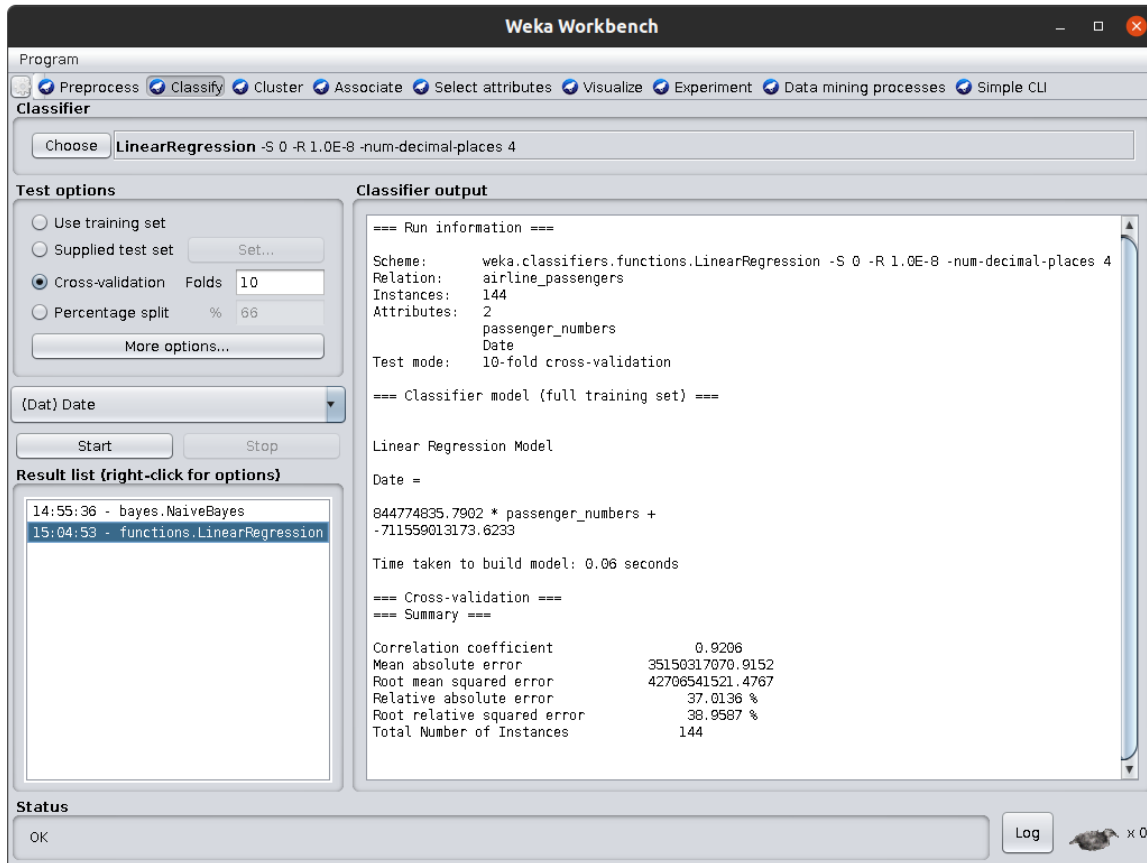
7. Practical of Clustering

Steps: Open Weka > Workbench > Open File > Cluster > Choose Clusterer > Start

Clusterer : Hierarchical Clusterer
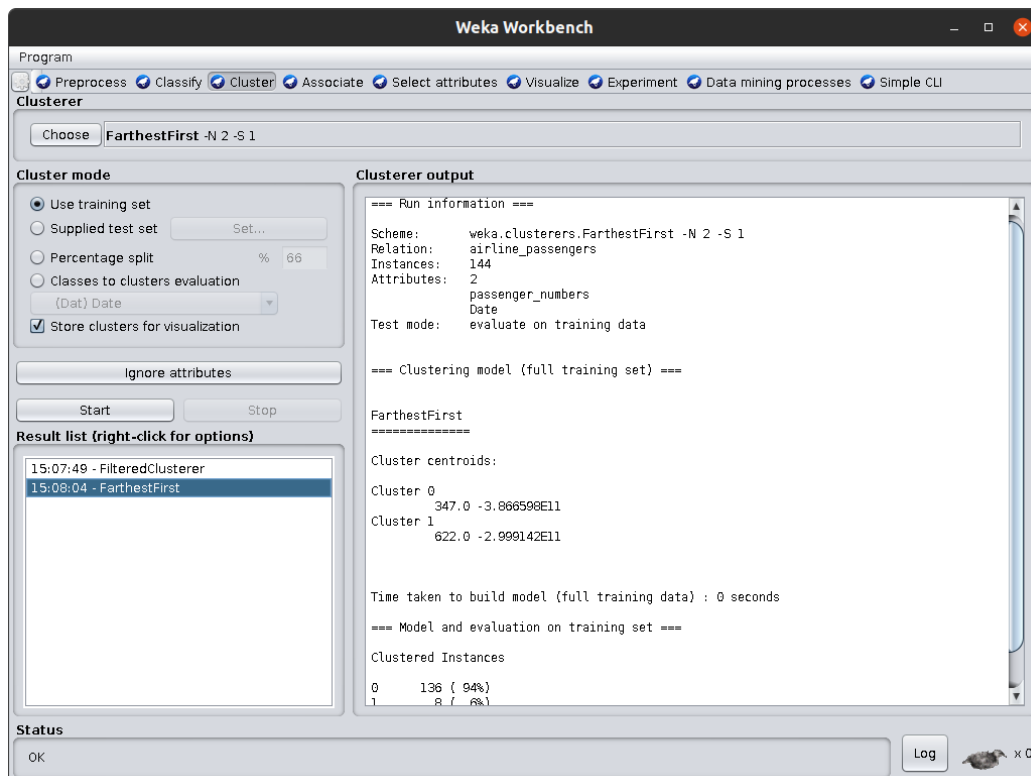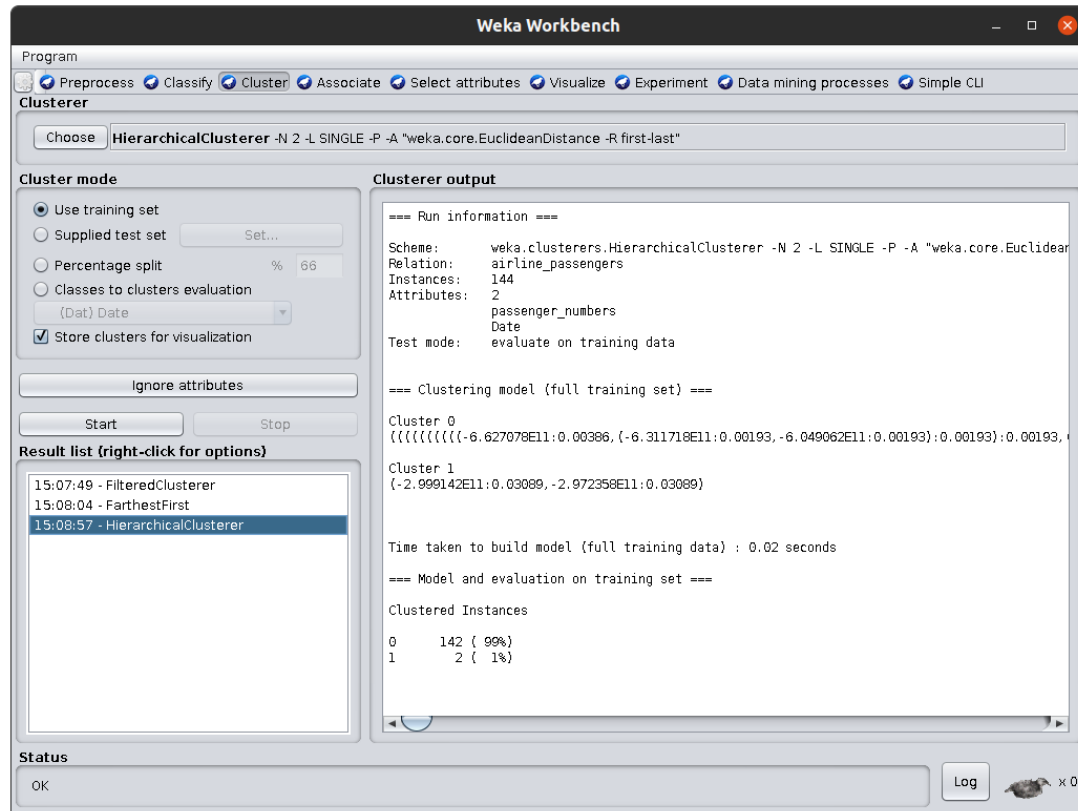
8. Practical of Prediction

Steps: Open Weka > Workbench > Open File > Classify > Choose Classifier > Start

Classifier: NaiveBayes
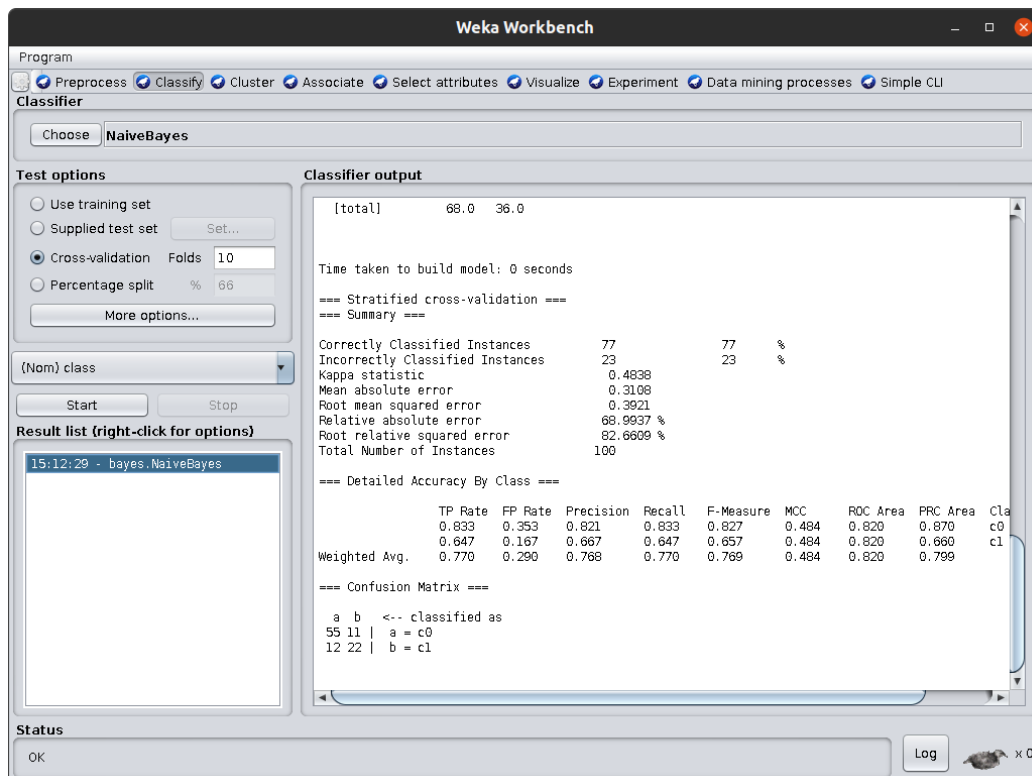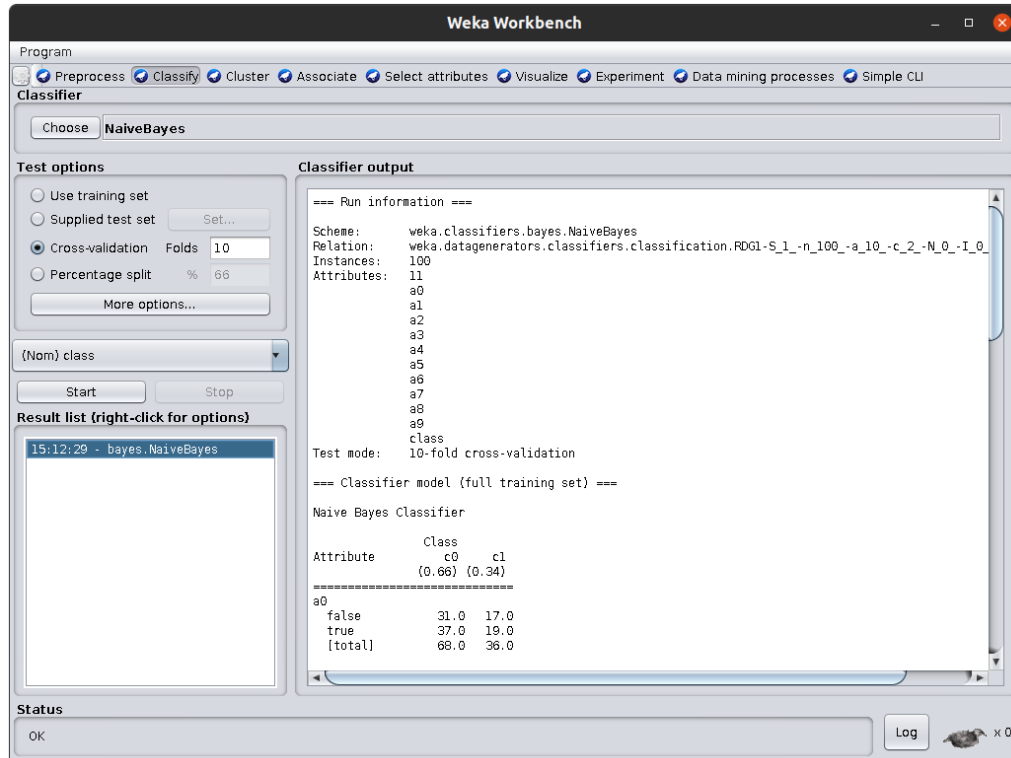
9. Practical of Association Rules

Steps: Open Weka > Workbench > Open File > Associate > Choose Associator > Start

Associator: Filtered Associator
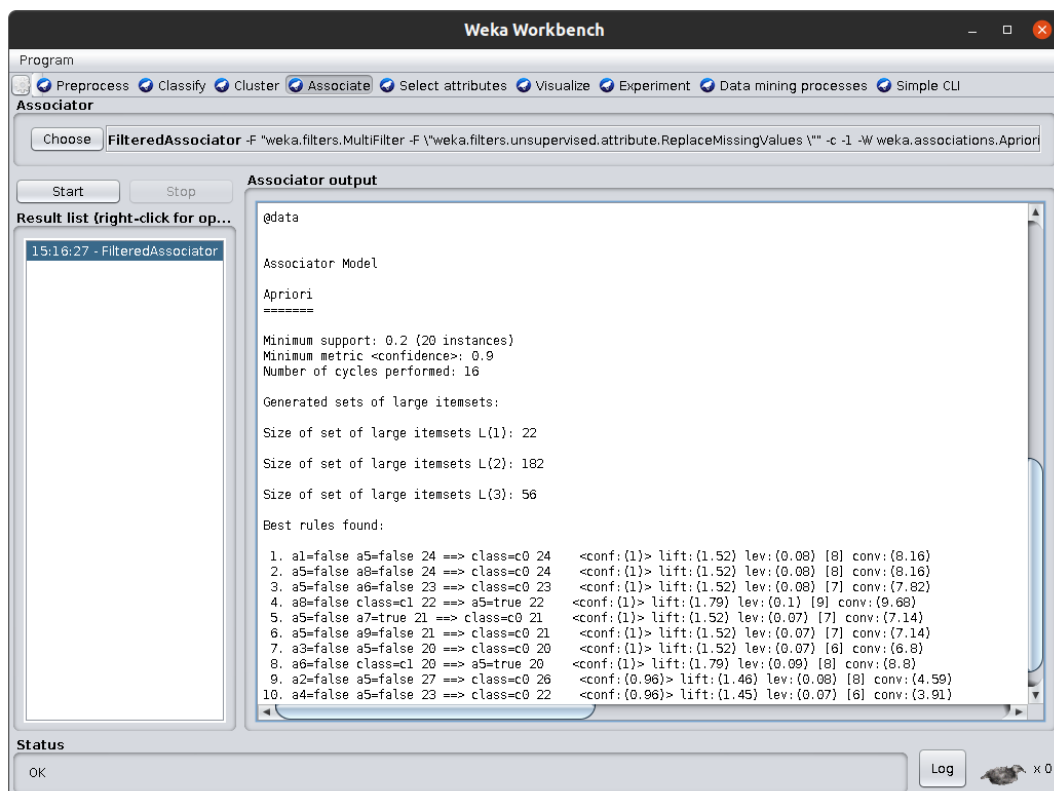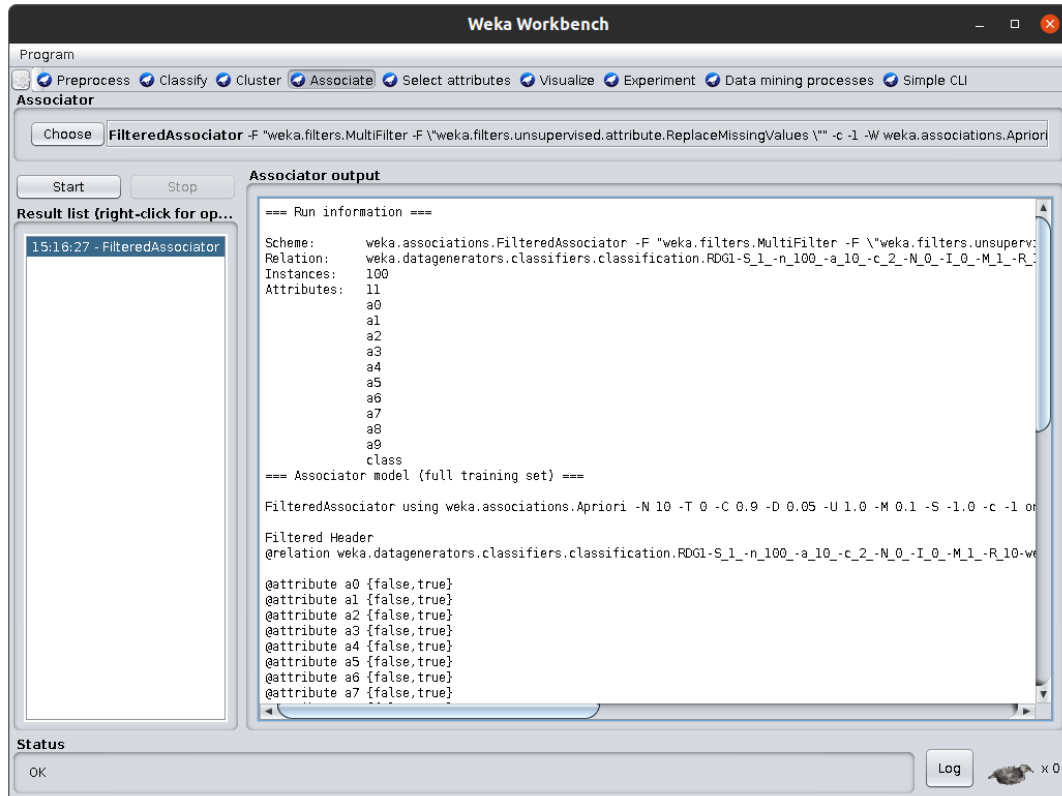
10. Practical of Outer detection

Steps: Open Weka > Workbench > Open File > Classify > Choose Classifier > Start

Classifier: J48
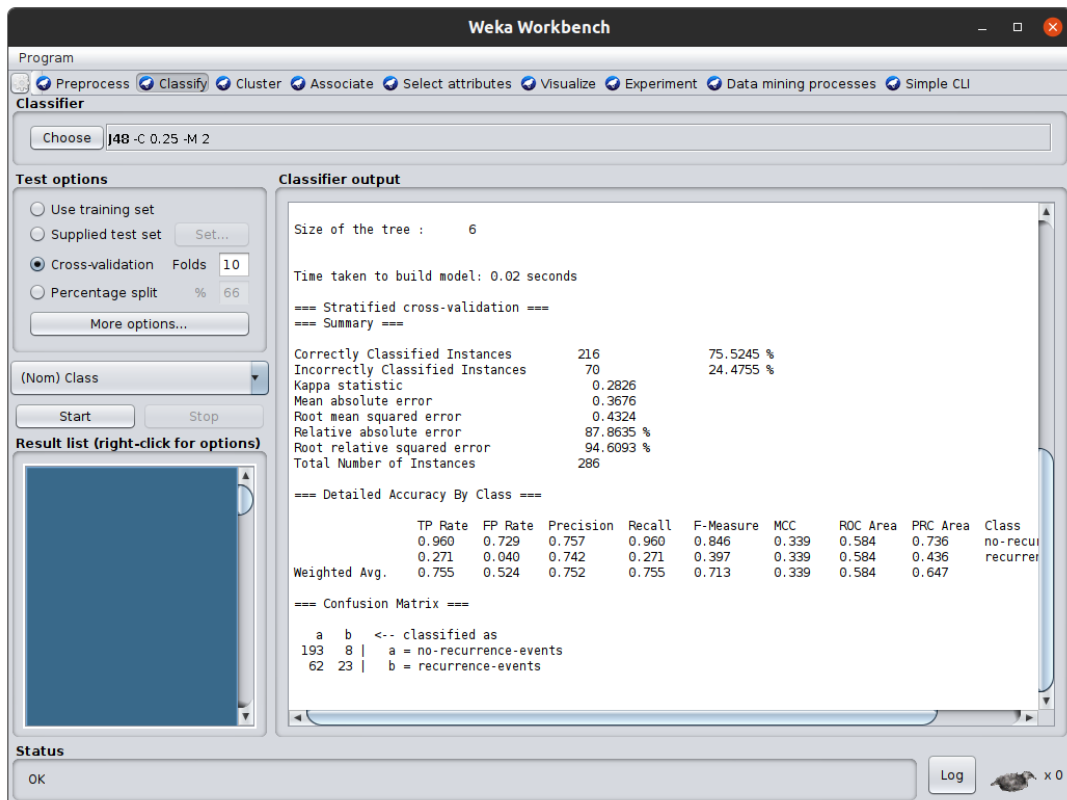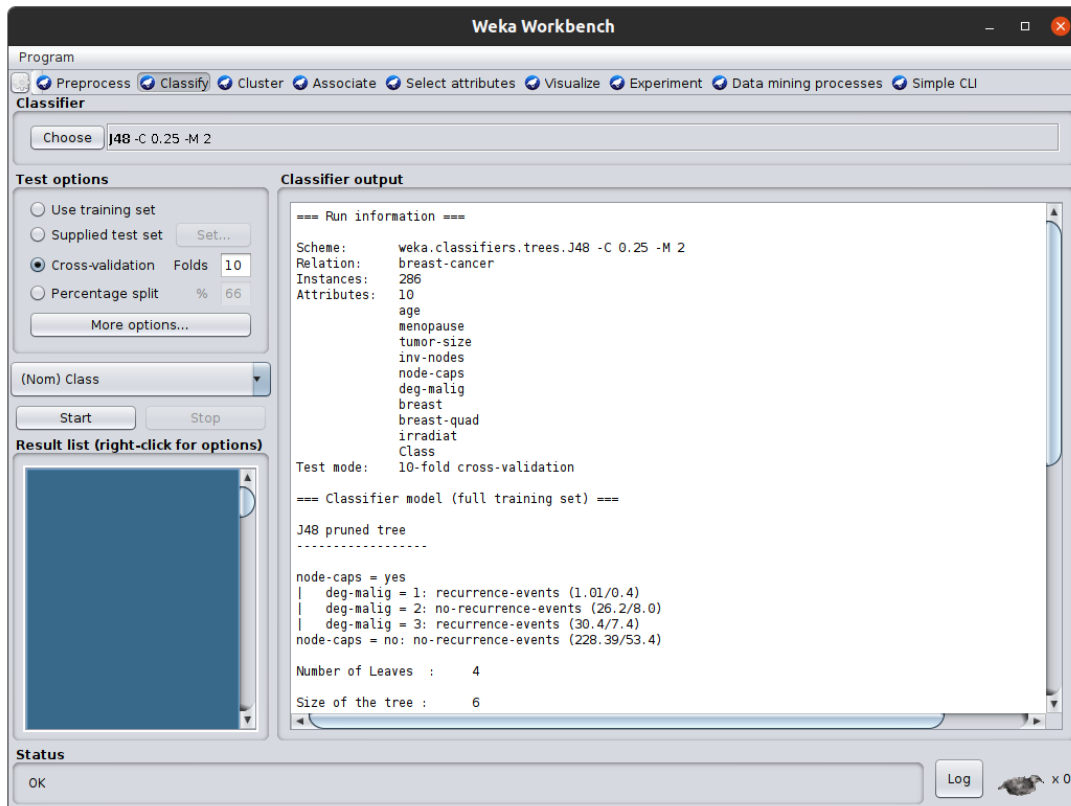
11. Practical of Sequential Patterns

Steps: Open Weka > Workbench > Open File > Associate > Choose Associator > Start

Associator: Apriori
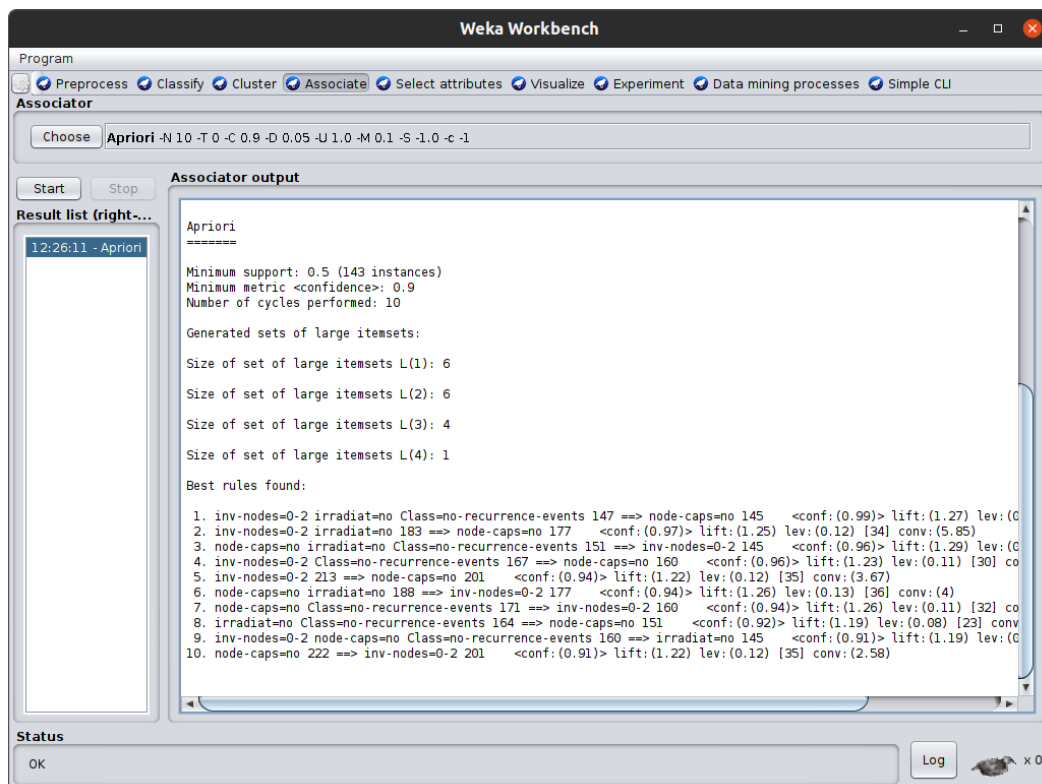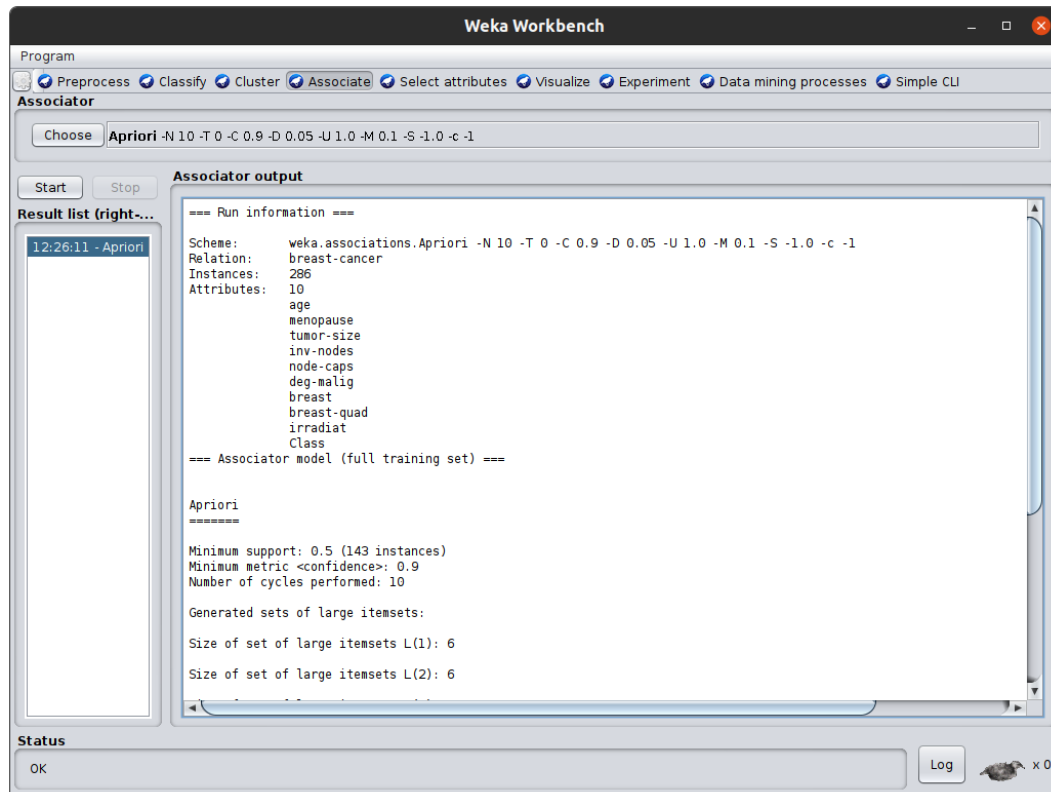
12. Data Visualization

Steps: Open Weka > Workbench > Open File > Visualize all
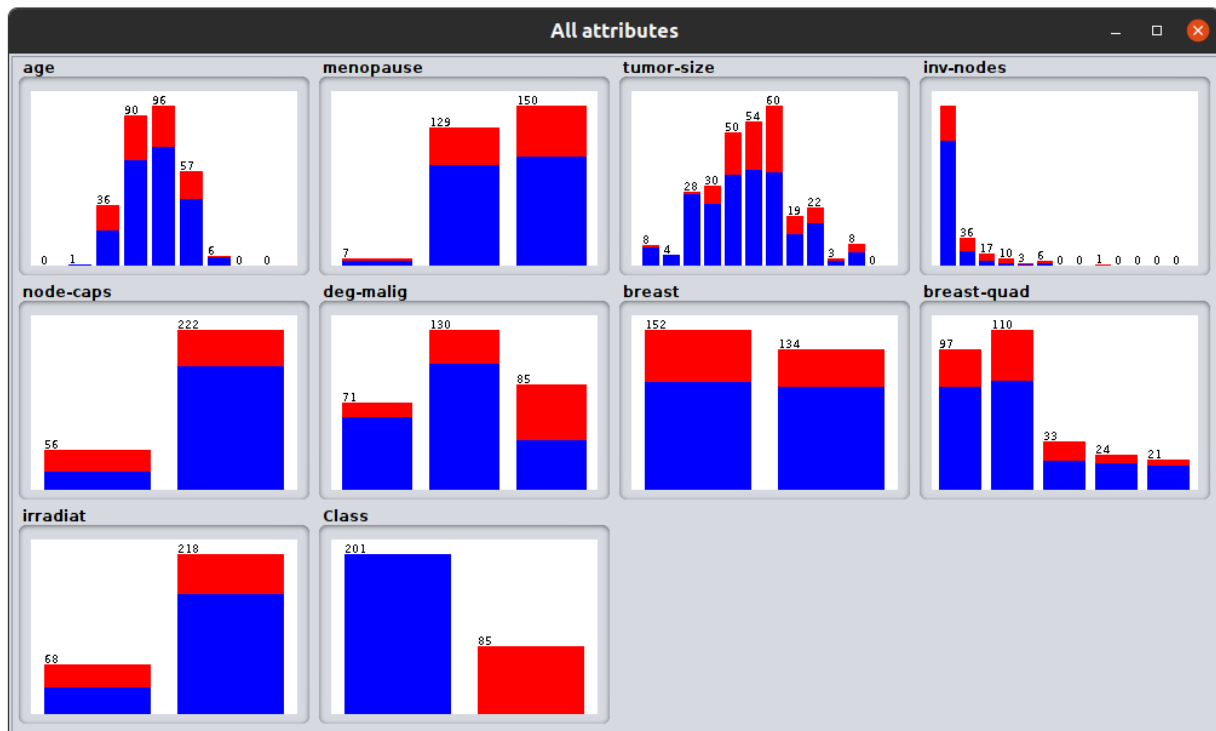
13. Enlist all the available tools for Data Mining

1. Xplenty

Xplenty provides a platform that has functionalities to integrate, process, and prepare data for analytics. Businesses will be able to make most of the opportunities offered by big data with the help of Xplenty and that too without investing in related personnel, hardware, and software. It is a complete toolkit for building data pipelines.

You will be able to implement complex data preparation functions through rich expression language. It has an intuitive interface to implement ETL, ELT, or a replication solution. You will be able to orchestrate and schedule pipelines through a workflow engine.

- Xplenty is the data integration platform for all. It offers the no-code and low-code options.
- An API component will provide advanced customization and flexibility.
- It has functionalities to transfer and transform data between databases and data warehouses.
- It provides support through email, chat, phone, and an online meeting.

Availability: Licensed tools.

2. Rapid Miner

Rapid Miner is one of the best predictive analysis system developed by the company with the same name as the Rapid Miner. It is written in JAVA programming language. It provides an integrated environment for deep learning, text mining, machine learning & predictive analysis.

The tool can be used for over a vast range of applications including for business applications, commercial applications, training, education, research, application development, machine learning.

Rapid Miner offers the server as both on premise & in public/private cloud infrastructures. It has a client/server model as its base. Rapid Miner comes with template-based frameworks that enable speedy delivery with reduced number of errors (which are quite commonly expected in manual code writing process).

Availability: Open source

Rapid Miner constitutes of three modules, namely

1. Rapid Miner Studio: This module is for workflow design, prototyping, validation etc.
2. Rapid Miner Server: To operate predictive data models created in studio
3. Rapid Miner Radoop: Executes processes directly in the Hadoop cluster to simplify predictive analysis.

3.Orange

Orange is a perfect software suite for machine learning & data mining. It best aids the data visualization and is a component-based software. It has been written in Python computing language.

As it is a component-based software, the components of orange are called 'widgets. These widgets range from data visualization & pre-processing to an evaluation of algorithms and predictive modelling.

Widgets offer major functionalities like

* Showing data table and allowing to select features
* Reading the data
* Training predictors and to compare learning algorithms
* Visualizing data elements etc.

Additionally, Orange brings a more interactive and fun vibe to the dull analytic tools. It is quite interesting to operate.

Data coming to Orange gets quickly formatted to the desired pattern and it can be easily moved where needed by simply moving/flipping the widgets. Users are quite fascinated by Orange. Orange allows users to make smarter decisions in short time by quickly comparing & analysing the data.

Availability: Open source

4. Weka

Also known as Waikato Environment is a machine learning software developed at the **University of Waikato** in New Zealand. It is best suited for data analysis and predictive modelling. It contains algorithms and visualization tools that support machine learning.

Weka has a GUI that facilitates easy access to all its features. It is written in JAVA programming language.

Weka supports major data mining tasks including data mining, processing, visualization, regression etc. It works on the assumption that data is available in the form of a flat file.

Weka can provide access to SQL Databases through database connectivity and can further process the data/results returned by the query.

Availability: Free software

5. KNIME

KNIME is the best integration platform for data analytics and reporting developed by KNIME.com AG. It operates on the concept of the modular data pipeline. KNIME constitutes of various machine learning and data mining components embedded together.

KNIME has been used widely for pharmaceutical research. In addition, it performs excellently for customer data analysis, financial data analysis, and business intelligence.

KNIME has some brilliant features like quick deployment and scaling efficiency. Users get familiar with KNIME in quite lesser time and it has made predictive analysis accessible to even naive users. KNIME utilizes the assembly of nodes to pre-process the data for analytics and visualization.

Availability: Open Source

6. Sisense

Sisense is extremely useful and best suited BI software when it comes to reporting purposes within the organization. It is developed by the company of same name 'Sisense'. It has a brilliant capability to handle and process data for the small scale/large scale organizations.

It allows combining data from various sources to build a common repository and further, refines data to generate rich reports that get shared across departments for reporting.

Sisense got awarded as best BI software is 2016 and still, holds a good position.
Sisense generates reports which are highly visual. It is specially designed for users that are non-technical. It allows drag & drop facility as well as widgets.

Different widgets can be selected to generate the reports in form of pie charts, line charts, bar graphs etc. based on the purpose of an organization. Reports can be further drilled down by simply clicking to check details and comprehensive data.

Availability: Licensed

7. SSDT (SQL Server Data Tools)

SSDT is a universal, declarative model that expands all the phases of database development in the Visual Studio IDE. BIDS was the former environment developed by Microsoft to do data analysis and provide business intelligence solutions. Developers use SSDT transact- a design capability of SQL, to build, maintain, debug and refactor databases.

A user can work directly with a database or can work directly with a connected database, thus, providing on or off-premise facility.

Users can use visual studio tools for development of databases like IntelliSense, code navigation tools, and programming support via C#, visual basic etc. SSDT provides **Table Designer** to create new tables as well as edit tables in direct databases as well as connected databases.
Deriving its base from BIDS, which was not compatible with Visual Studio2010, the SSDT BI came into existence and it replaced BIDS.

Availability: Licensed

8. Apache Mahout

Apache Mahout is a project developed by Apache Foundation that serves the primary purpose of creating machine learning algorithms. It focuses mainly on data clustering, classification, and collaborative filtering.
Mahout is written in JAVA and includes JAVA libraries to perform mathematical operations like linear algebra and statistics. Mahout is growing continuously as the algorithms implemented inside Apache Mahout are continuously growing. The algorithms of Mahout have implemented a level above Hadoop through mapping/reducing templates.

To key up, Mahout has following major features
- Extensible programming environment
- Pre-made algorithms
- Math experimentation environment
- GPU computes for performance improvement.

Availability: Open source

9. Oracle Data Mining

A component of Oracle Advance Analytics, Oracle data mining software provides excellent data mining algorithms for data classification, prediction, regression and specialized analytics that enables analysts to analyze insights, make better predictions, target best customers, identify cross-selling opportunities & detect fraud.

The algorithms designed inside ODM leverage the potential strengths of Oracle database. The data mining feature of SQL can dig data out of database tables, views, and schemas.

The GUI of Oracle data miner is an extended version of Oracle SQL Developer. It provides a facility of direct 'drag & drop' of data inside the database to users thus giving better insight.

Availability: Proprietary License

10.   Rattle

Rattle is GUI based data mining tool that uses R stats programming language. Rattle exposes the statistical power of R by providing considerable data mining functionality. Although Rattle has an extensive and well-developed UI, it has an inbuilt log code tab that generates duplicate code for any activity happening at GUI.

The data set generated by Rattle can be viewed as well as edited. Rattle gives the additional facility to review the code, use it for numerous purposes and extend the code without restriction.

Availability: Open source

14. Case Study on Web Data Mining

The case study is a performance study of two classes that are using online course over the web and the web generating a mining algorithm using a variety of parameters. A number of experiments were conducted, and their results are presented in this part. The experiments were based on different parameter of interest. The parameters varied form the number of input attributes, the sample size of the class, and so on. The results of these experiments prove that the data mining algorithm are very efficient and scalable.

The Clustering Algorithm used

The Clustering algorithm is based on the Expectation and Maximization algorithm (Microsoft, 2003). This algorithm iterates between two steps. In the first step, called the "expectation" step, the cluster membership of each case is calculated. In the second step, called the "Maximization" step, the parameters of the models are re-estimated using these cluster memberships, which has the following major steps:

1. Assign initial means

2. Assign cases to each mean using some distance measure

3. Compute new means based on members of each cluster

4. Cycle until convergence.

 A case is assigned to each cluster with a certain probability and the means of each cluster is shifted based on that iteration. The following table shows a set of data that could be used to predict best achievement. In this study, information was generated on users that included the following list of measurements:

1. Most requested pages

2. Least requested pages

3. Top exit pages

4. Most accessed directories

5. Most downloaded files

6. New versus returning visitors

7. Summary of activity for exam period

8. Summary of activity by time increment

9. Number of views per each page

10. Page not found

RESULTS

This brief case study gives a look at what statistics are commonly measured on web sites. The results of these statistics can be used to alter the web site, thereby altering the next user's experience. Table 1 displays some basic statistics that relate to frequency, length and kind of visitor. In Table 1 additional insights are gained with a breakdown of visitors by each day. This behaviour might reflect variation in activity related to exams time or other issues. Monitoring and understanding visitor behaviour are the first step in evaluating and improving the web site. Another relevant measurement is how many pages are viewed. This can reflect content as well as antigravity. If a majority of visitors viewed only one page, it may imply that they did not find it easy to determine how to take the next step.

CONCLUSION

The web offers prospecting and user relationship management opportunities that are limited only by the imagination. Data mining is a tool that can extract predictive information from large quantities of data, and is data driven. It uses mathematical and statistical calculations to uncover trends and correlations among the large quantities of data stored in a database.

15. Major or Mini project Ideas on Data mining

Some major or mini project ideas on Data Mining is following:

Electronic Mail Server

The 'electronic mail server' has been developed to simplify and enhance inter personnel communication in an intra-organizational setup. This follows a client-server approach where the website works on the server and the user form the client.

The client requests the server for services and the server responds by transferring required information to the user. The site is going to be operated by the Internet users and administered by the organizational administrator who will be server based.

The administrator will have the privilege to limit the number of users that can register as members. He can limit the number of mail a particular user can store for himself. The administrator can also restore vital information in case of a serious breakdown.

The user will operate from the client side by accessing the provisions provided by the website. This can be done irrespective of the geographical location.

However, an internet connection with its entire infrastructure is a must. The provisions include, getting registered as a member in the official website, composing and sending e-mail to his counterparts thus sharing official information and official status on a particular aspect.

Also is included the option of deleting unwanted messages from ones mailbox and the facility to modify/change personal information provided in the website like ones password, telephone number etc.

Low-Cost Wireless Internet

This paper describes an innovative concept using tethered Aerostats as a platform for raising wireless communication payload, which overcomes the two main limitations of high towers listed above. Tethered aerostats are an outcome of Lighter-Than-Air Technology, where static lift production mechanism is based on the Archimedes Principle [1]. An aerostat does not require any additional energy to reach to a certain height.

For a given volume of envelope that contains the lighter than air gas, displaced weight of air creates a vertically upward buoyant force that leads to

the lift. One or more Ballonets are provided inside the envelope to adjust the buoyancy.

The envelope volume is large enough to ensure that the displaced air should be able to produce sufficient lift, under the entire range of operating conditions, to balance all the weight groups of the aerostat system, viz., envelope, fin, nose battens, ballonets, pivot mechanism, payload, tether, recovery system, gas filling ports, and safety valves.

Aerostats are used all over the globe as a platform to house high-resolution sensors for applications such as aerial surveillance, regional atmospheric data collection and balloon barrage system. Depending on the payload, range of surveillance, and operational time, these aerostats can be launched to any desired altitude from a few meters above ground level to as high as 5000 m above ground level. Of course, the payload carrying capacity of an aerostat is reduced as the operational height is increased.

Aerostats can easily be deployed at high altitudes, ensuring disturbance free LOS for the communications payload. Once they are deployed, there is very little recurring additional expenditure to keep them afloat, except in the form of small amounts of lighter-than-air gas, just to top-up for the leakages through the fabric over a period of time.

Due to its aerodynamic shape as well as provision of fins, an aerostat can remain fairly steady even in strong winds and hence can provide stable line of sight connectivity. An omni-directional antenna mounted below the aerostat, leads to a relaxation in the antenna direction alignment requirement.

802.11b [3] uses the ISM (Industrial Scientific Medical) band from 2.400 to 2.495GHz. Due to the ubiquity of equipment and unlicensed nature of the 2.4 GHz ISM band, our work is focused on building a network using 802.11b. It makes use of Direct Sequence Spread Spectrum (DSSS) modulation and has a maximum rate of 11 Mbps, with actual usable data speeds up to about 5 Mbps.

Hospital Management system

The project Hospital Management system includes registration of patients, storing their details into the system, and also computerized billing in the pharmacy, and labs. The software has the facility to give a unique id for every patient and stores the details of every patient and the staff automatically. It includes a search facility to know the current status of each room. User can search availability of a doctor and the details of a patient using the id.

The Hospital Management System can be entered using a username and password. It is accessible either by an administrator or receptionist. Only they can add data into the database. The data can be retrieved easily. The interface is very user-friendly. The data are well protected for personal use and makes the data processing very fast.

The purpose of the project entitled as "HOSPITAL MANAGEMENT SYSTEM" is to computerize the Front Office Management of Hospital to develop software which is user friendly, simple, fast, and cost – effective. It deals with the collection of patient's information, diagnosis details, etc.

Traditionally, it was done manually. The main function of the system is to register and store patient details and doctor details and retrieve these details as and when required, and also to manipulate these details meaningfully System input contains patient details, diagnosis details; while system output is to get these details on to the CRT screen.

A form is a major part of Visual Basic application, which allows the user to enter the data as well as view the result. A control is an object that we draw on a form to enable or enhance user interaction with an application.

Hence a Visual Basic application is a combination of object like forms and controls, procedures that can respond to Events and other general-purpose procedure. Event procedures are where we do the actual computer programming and are saved with the form in the file with the form extension. These procedures are where we write BASIC language statements.

16. Website analysis using Alexa

- **SEO keyword opportunities**, including keywords that are driving traffic to competing sites, but not yours. Find your next content idea and incorporate it into your strategy.

- **Competitive benchmarking**, including how your site's traffic from search, keywords, and backlinks compare to that of your biggest online competitors.

- **Website traffic statistics**, including Alexa Rank, engagement metrics, referral sources, and more for any site in our database.

- **Audience insights**, including sites that share an audience with your site, topics your audience cares most about, and keywords they search for.