Data minining and warehousing

Practical 1

Aim: To study various data mining tools and their applications. Weka, orange, natural language data mining, python.

Data Mining is a process of finding potentially useful patterns from huge data sets. It is a multi-disciplinary skill that uses machine learning, statistics, and AI to extract information to evaluate future events probability. The insights derived from Data Mining are used for marketing, fraud detection, scientific discovery, etc.

Data Mining is all about discovering hidden, unsuspected, and previously unknown yet valid relationships amongst the data. Data mining is also called Knowledge Discovery in Data (KDD), Knowledge extraction, data/pattern analysis, information harvesting, etc.

Data mining tools

1) R-Programming

R is a language for statistical computing and graphics. It also used for big data analysis. It provides a wide variety of statistical tests.

Features:

- Effective data handling and storage facility,
- It provides a suite of operators for calculations on arrays, in particular, matrices,

- It provides a coherent, integrated collection of big data tools for data analysis
- It provides graphical facilities for data analysis which display either on-screen or on hardcopy.

Download link; https://www.r-project.org/

2) Oracle BI

Oracle BI is an open source machine learning and data visualization for novice and expert. Interactive data analysis workflows with a large toolbox.

Features:

- Interactive Data Visualization.
- It Offers Interactive data exploration for rapid qualitative analysis with clean visualizations.
- Orange supports hands-on training and visual illustrations of concepts from data science.
- It offers an extensive range of add-ons to data mining from external data sources.

Download link: https://orange.biolab.si/

3) Orange

Availability: Open source

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 modeling.

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 & analyzing the data.

Click **Orange** official website.

4) Weka

Availability: Free software
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 modeling. 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.

Click **WEKA** official website.

5) KNIME

Availability: Open Source

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.

Click **KNIME** official website.

6) Oracle Data Mining

Availability: Proprietary License

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.

Click **Oracle Data Mining** official website.

Benefits of Data Mining:

- Data mining technique helps companies to get knowledge-based information.
- Data mining helps organizations to make the profitable adjustments in operation and production.
- The data mining is a cost-effective and efficient solution compared to other statistical data applications.
- Data mining helps with the decision-making process.
- Facilitates automated prediction of trends and behaviors as well as automated discovery of hidden patterns.
- It can be implemented in new systems as well as existing platforms
- It is the speedy process which makes it easy for the users to analyze huge amount of data in less time.

Disadvantages of Data Mining

- There are chances of companies may sell useful information of their customers to other companies for

money. For example, American Express has sold credit card purchases of their customers to the other companies.

- Many data mining analytics software is difficult to operate and requires advance training to work on.
- Different data mining tools work in different manners due to different algorithms employed in their design. Therefore, the selection of correct data mining tool is a very difficult task.
- The data mining techniques are not accurate, and so it can cause serious consequences in certain conditions.

Data Mining Applications

Applications	Usage
Communications	Data mining techniques are used in communication sector to predict customer behavior to offer highly targetted and relevant campaigns.
Insurance	Data mining helps insurance companies to price their products profitable and promote new offers to their nev or existing customers.
Education	Data mining benefits educators to access student data, predict achievement levels and find students or groups of students which need extra attention. For example, students who are weak in maths subject.
Manufacturing	With the help of Data Mining Manufacturers can predict wear and tear of production assets. They can

	anticipate maintenance which helps them reduce them to minimize downtime.
Banking	Data mining helps finance sector to get a view of market risks and manage regulatory compliance. It helps banks to identify probable defaulters to decide whether to issue credit cards, loans, etc.
Retail	Data Mining techniques help retail malls and grocery stores identify and arrange most sellable items in the most attentive positions. It helps store owners to comes up with the offer which encourages customers to increase their spending.
Service Providers	Service providers like mobile phone and utility industries use Data Mining to predict the reasons when a customer leaves their company. They analyze billing details, customer service interactions, complaints made to the company to assign each customer a probability score and offers incentives.
E-Commerce	E-commerce websites use Data Mining to offer cross-sells and up-sells through their websites. One of the most famous names is Amazon, who use Data mining techniques to get more customers into their eCommerc store.

Super Markets	Data Mining allows supermarket's develope rules to predict if their shoppers were likely to be expecting. By evaluating their buying pattern, they could find woman customers who are most likely pregnant. They can start targeting products like baby powder, baby shop, diaper and so on.
Crime Investigation	Data Mining helps crime investigation agencies to deploy police workforce (where is a crime most likely to happen and when?), who to search at a border crossing etc.
Bioinformatics	Data Mining helps to mine biological data from massive datasets gathered in biology and medicine.

Practical - 2

Aim: To study about various dataset from UCI repository for following:

Classification, prediction, clustering, pattern mining, time series analysis.

1. Classification:

i. Dataset: Arcene

Data Set Characteristics:	Multivariate	**Number of Instances:**	900	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	10000	**Date Donated**	2008-02-29
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	139570

- **Data Set Information:**

ARCENE was obtained by merging three mass-spectrometry datasets to obtain enough training and test data for a benchmark. The original features indicate the abundance of proteins in human sera having a given mass value. Based on those features one must separate cancer patients from healthy patients. We added a number of distractor feature called 'probes' having no predictive power. The order of the features and patterns were randomized.

ARCENE -- Positive ex. -- Negative ex. -- Total
Training set -- 44 -- 56 -- 100
Validation set -- 44 -- 56 -- 100
Test set -- 310 -- 390 -- 700
All -- 398 -- 502 -- 900

Number of variables/features/attributes:
Real: 7000
Probes: 3000
Total: 10000

This dataset is one of five datasets used in the NIPS 2003 feature selection challenge. Our website_ is still open for post-challenge submissions. Information about other related challenges are found . CLOP package includes sample code to process these data.

All details about the preparation of the data are found in our technical report: Design of experiments for the NIPS 2003 variable selection benchmark, Isabelle Guyon, July 2003, (also included in the dataset archive). Such information was made available only after the end of the challenge.

The data are split into training, validation, and test set. Target values are provided only for the 2 first sets. Test set performance results are obtained by submitting prediction results to
The data are in the following format:
dataname.param: Parameters and statistics about the data
dataname.feat: Identities of the features (withheld, to avoid biasing feature selection).
dataname_train.data: Training set (a coma delimited regular matrix, patterns in lines, features in columns).
dataname_valid.data: Validation set.
dataname_test.data: Test set.
dataname_train.labels: Labels (truth values of the classes) for training examples.
dataname_valid.labels: Validation set labels (withheld during the benchmark, but provided now).
dataname_test.labels: Test set labels (withheld, so the data can still be use as a benchmark).

- **Attribute Information:**

We do not provide attribute information to avoid biasing the feature selection process.

ii. Dataset: Crop mapping using fused optical-radar data set

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	325834	**Area:**	N/A
Attribute Characteristics:	Real	**Number of Attributes:**	175	**Date Donated**	2020-06-16
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	12167

- **Data Set Information:**
- This big data set is a fused bi-temporal optical-radar data for cropland classification. The images were collected by RapidEye satellites (optical) and the Unmanned Aerial Vehicle Synthetic Aperture Radar (UAVSAR) system (Radar) over an agricultural region near Winnipeg, Manitoba, Canada on 2012.
  There are 2 * 49 radar features and 2 * 38 optical features for two dates: 05 and 14 July 2012.
  Seven crop type classes exist for this data set as follows: 1-Corn; 2-Peas; 3- Canola; 4-Soybeans; 5- Oats; 6- Wheat; and 7-Broadleaf.
-

- **Attribute Information:**

- 175 attributes including:
  1- class;
  2- f1 to f49:Polarimetric features on 05 July 2012;
  3- f50 to f98:Polarimetric features on 14 July 2012;
  4- f99 to f136:Optical features on 05 July 2012;
  5- f137 to f174:Optical features on 14 July 2012;

  Details:
  label:crop type class
  f1:sigHH_Rad05July
  f2:sigHV_Rad05July
  f3:sigVV_Rad05July
  f4:sigRR_Rad05July
  f5:sigRL_Rad05July
  f6:sigLL_Rad05July
  f7:Rhhvv_Rad05July
  f8:Rhvhh_Rad05July
  f9:Rhvvv_Rad05July
  f10:Rrrll_Rad05July

2. Clustering:

i. A study of Asian Religious and Biblical Texts

Data Set Characteristics:	Multivariate, Text	**Number of Instances:**	590	**Area:**	Social

Attribute Characteristics:	Integer	**Number of Attributes:**	8265	**Date Donated**	2019-12-24
Associated Tasks:	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	25659

- **Data Set Information:**

Most of the sacred texts in this dataset were collected from Project Gutenberg. We herein provide the raw texts along with our pre-processed Document Term Matrices (DTM). For more details, please contact the authors

- **Attribute Information:**

The attributes are just the words from the bag of words preprocessing of the mini-corpus made up of the 8 religious books considered in this study. There are 8265 words used

ii. Bag of words:

Data Set Characteristics:	Text	**Number of Instances:**	8000000	**Area:**	N/A

Attribute Characteristics:	Integer	**Number of Attributes:**	100000	**Date Donated**	2008-03-12
Associated Tasks:	Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	322629

- **Data Set Information:**

For each text collection, D is the number of documents, W is the number of words in the vocabulary, and N is the total number of words
in the collection (below, NNZ is the number of nonzero counts in the
bag-of-words). After tokenization and removal of stopwords, the vocabulary of unique words was truncated by only keeping words that
occurred more than ten times. Individual document names (i.e. a identifier for each docID) are not provided for copyright reasons.

These data sets have no class labels, and for copyright reasons no
filenames or other document-level metadata. These data sets are ideal
for clustering and topic modeling experiments.

For each text collection we provide docword.*.txt (the bag of

words
file in sparse format) and vocab.*.txt (the vocab file).

Enron Emails:
orig source: www.cs.cmu.edu/~enron
D=39861
W=28102
N=6,400,000 (approx)

NIPS full papers:
orig source: books.nips.cc
D=1500
W=12419
N=1,900,000 (approx)

KOS blog entries:
orig source: dailykos.com
D=3430
W=6906
N=467714

NYTimes news articles:
orig source: ldc.upenn.edu
D=300000
W=102660
N=100,000,000 (approx)

PubMed abstracts:
orig source: www.pubmed.gov
D=8200000

W=141043
N=730,000,000 (approx)

- **Attribute Information:**

The format of the docword.*.txt file is 3 header lines, followed by
NNZ triples:

D
W
NNZ
docID wordID count
docID wordID count
docID wordID count
docID wordID count
...
docID wordID count
docID wordID count
docID wordID count

The format of the vocab.*.txt file is line contains wordID=n.

3. Outlier analysis:

i. Wine Quality

Data Set Characteristics:	Multivariate	**Number of Instances:**	4898	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	12	**Date Donated**	2009-10-07
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	1427531

- **Data Set Information:**

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: [Web Link] or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are

relevant. So it could be interesting to test feature selection methods.

- **Attribute Information:**

For more information, read [Cortez et al., 2009].
Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

ii. detection_ of _ IoT _ botnet _ attacks _N _ BaIoT

Data Set Characteristics:	Multivariate, Sequential	**Number of Instances:**	7062606	**Area:**	Computer

Attribute Characteristics:	Real	**Number of Attributes:**	115	**Date Donated**	2018-03-19
Associated Tasks:	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	76158

- **Data Set Information:**

(a) Attribute being predicted:
-- Originally we aimed at distinguishing between benign and Malicious traffic data by means of anomaly detection techniques.
-- However, as the malicious data can be divided into 10 attacks carried by 2 bonnets, the dataset can also be used for multi-class classification: 10 classes of attacks, plus 1 class of 'benign'.

(b) The study's results:
-- For each of the 9 IoT devices we trained and optimized a deep auto encoder on 2/3 of its benign data (i.e., the training set of each device). This was done to capture normal network traffic patterns.
-- The test data of each device comprised of the remaining 1/3 of benign data plus all the malicious data. On each test set we applied the respective trained (deep) auto encoder as an anomaly

detector. The detection of anomalies (i.e., the cyber attacks launched from each of the above IoT devices) concluded with 100% TPR.

- **Attribute Information:**

-- The following describes each of the features headers:

* Stream aggregation:
H: Stats summarizing the recent traffic from this packet's host (IP)
HH: Stats summarizing the recent traffic going from this packet's host (IP) to the packet's destination host.
HpHp: Stats summarizing the recent traffic going from this packet's host+port (IP) to the packet's destination host+port. Example 192.168.4.2:1242 -> 192.168.4.12:80
HH_jit: Stats summarizing the jitter of the traffic going from this packet's host (IP) to the packet's destination host.

* Time-frame (The decay factor Lambda used in the damped window):
How much recent history of the stream is capture in these statistics
L5, L3, L1, ...

* The statistics extracted from the packet stream:
weight: The weight of the stream (can be viewed as the number of items observed in recent history)
mean: ...

std: ...
radius: The root squared sum of the two streams' variances
magnitude: The root squared sum of the two streams' means
cov: an approximated covariance between two streams
pcc: an approximated covariance between two streams

4. Time series:

i. Bach Chorales

Data Set Characteristics:	Univariate, Time-Series	**Number of Instances:**	100	**Area:**	N/A
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	6	**Date Donated**	N/A

Associated Tasks:	N/A	**Missing Values?**	No	**Number of Web Hits:**	139071

- **Data Set Information:**

Sequential (time-series) domain. Single-line melodies of 100 Bach chorales (originally 4 voices). The melody line can be studied independently of other voices. The grand challenge is to learn a generative grammar for stylistically valid chorales (see references and discussion in "Multiple Viewpoint Systems for Music Prediction").

- **Attribute Information:**

Number of Attributes: 6 (nominal) per event

(a) start-time, measured in 16th notes from chorale beginning (time 0)
(b) pitch, MIDI number (60 = C4, 61 = C#4, 72 = C5, etc.)
(c) duration, measured in 16th notes
(d) key signature, number of sharps or flats, positive if key signature has sharps, negative if key signature has flats
(e) time signature, in 16th notes per bar
(f) fermata, true or false depending on whether event is under a fermata

Attribute domains (all integers):

(a) {0,1,2,...}
(b) {60,...,75}
(c) {1,...,16}
(d) {-4,...,+4}
(e) {12,16}
(f) {0,1}

ii. CalIt2 Building People Counts

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	10080	**Area:**	N/A
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	4	**Date Donated**	2006-12-01
Associated Tasks:	N/A	**Missing Values?**	No	**Number of**	56427

				Web Hits:	

- **Data Set Information:**

Observations come from 2 data streams (people flow in and out of the building), over 15 weeks, 48 time slices per day (half hour count aggregates).

The purpose is to predict the presence of an event such as a conference in the building that is reflected by unusually high people counts for that day/time period.

- **Attribute Information:**

1. Flow ID: 7 is out flow, 9 is in flow
2. Date: MM/DD/YY
3. Time: HH:MM:SS
4. Count: Number of counts reported for the previous half hour

Rows: Each half hour time slice is represented by 2 rows: one row for the out flow during that time period (ID=7) and one row for the in flow during that time period (ID=9)

Attributes in .events file ("ground truth")
1. Date: MM/DD/YY
2. Begin event time: HH:MM:SS (military)
3. End event time: HH:MM:SS (military)
4. Event name (anonymized)

5. Prediction:

i. Internet Firewall Data Data Set

Data Set Characteristics:	Multivariate	**Number of Instances:**	65532	**Area:**	Computer
Attribute Characteristics:	N/A	**Number of Attributes:**	12	**Date Donated**	2019-02-04
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	7080

- **Data Set Information:**

There are 12 features in total. Action feature is used as a class. There are 4 classes in total. These are allow, action, drop and reset-both classes.

- **Attribute Information:**

Source Port,Destination Port,NAT Source Port,NAT Destination Port,Action,Bytes,Bytes Sent,Bytes Received,Packets,Elapsed Time (sec),pkts_sent,pkts_received

ii. COVID-19 Surveillance Data Set

Data Set Characteristics:	Multivariate	**Number of Instances:**	14	**Area:**	Computer
Attribute Characteristics:	N/A	**Number of Attributes:**	7	**Date Donated**	2020-04-24
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	20925

- **Data Set Information:**

Guidelines for Prevention and Control of Corona virus Disease (COVID-19).

- **Attribute Information:**

Symptoms of COVID-19

6. Pattern mining:

i. Artificial Characters Data Set

Data Set Characteristics:	Multivariate	**Number of Instances:**	6000	**Area:**	Computer

Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	7	**Date Donated**	1992-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	260482

- **Data Set Information:**
- This database has been artificially generated by using a first order theory which describes the structure of ten capital letters of the English alphabet and a random choice theorem prover which accounts for etherogeneity in the instances. The capital letters represented are the following: A, C, D, E, F, G, H, L, P, R. Each instance is structured and is described by a set of segments (lines) which resemble the way an automatic program would segment an image. Each instance is stored in a separate file whose format is the following:

  CLASS OBJNUM TYPE XX1 YY1 XX2 YY2 SIZE DIAG

  where CLASS is an integer number indicating the class as described below, OBJNUM is an integer identifier of a

segment (starting from 0) in the instance and the remaining columns represent attribute values. For further details, contact the author.

-

- **Attribute Information:**
- TYPE: the first attribute describes the type of segment and is always set to the string "line". Its C language type is char.

  XX1,YY1,XX2,YY2: these attributes contain the initial and final coordinates of a segment in a Cartesian plane. Their C language type is int.

  SIZE: this is the length of a segment computed by using the geometric distance between two points A(X1,Y1) and B(X2,Y2). Its C language type is float.

  DIAG: this is the length of the diagonal of the smallest rectangle which includes the picture of the character. The value of this attribute is the same in each object. Its C language type is float.

ii. Flags Data Set

Data Set Characteristics:	Multivariate	**Number of Instances:**	194	**Area:**	N/A
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	30	**Date Donated**	1990-05-15
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	322703

- **Data Set Information:**

This data file contains details of various nations and their flags. In this file the fields are separated by spaces (not commas). With this data you can try things like predicting the religion of a country from its size and the colours in its flag.

10 attributes are numeric-valued. The remainder are either Boolean- or nominal-valued.

- **Attribute Information:**

1. name: Name of the country concerned
2. landmass: 1=N.America, 2=S.America, 3=Europe, 4=Africa, 4=Asia, 6=Oceania
3. zone: Geographic quadrant, based on Greenwich and the Equator; 1=NE, 2=SE, 3=SW, 4=NW
4. area: in thousands of square km
5. population: in round millions
6. language: 1=English, 2=Spanish, 3=French, 4=German, 5=Slavic, 6=Other Indo-European, 7=Chinese, 8=Arabic, 9=Japanese/Turkish/Finnish/Magyar, 10=Others
7. religion: 0=Catholic, 1=Other Christian, 2=Muslim, 3=Buddhist, 4=Hindu, 5=Ethnic, 6=Marxist, 7=Others
8. bars: Number of vertical bars in the flag
9. stripes: Number of horizontal stripes in the flag
10. colors: Number of different colors in the flag
11. red: 0 if red absent, 1 if red present in the flag
12. green: same for green
13. blue: same for blue
14. gold: same for gold (also yellow)
15. white: same for white
16. black: same for black
17. orange: same for orange (also brown)
18. mainhue: predominant color in the flag (tie-breaks decided by taking the topmost hue, if that fails then the most central hue, and if that fails the leftmost hue)
19. circles: Number of circles in the flag
20. crosses: Number of (upright) crosses
21. saltires: Number of diagonal crosses
22. quarters: Number of quartered sections
23. sunstars: Number of sun or star symbols
24. crescent: 1 if a crescent moon symbol present, else 0

25. triangle: 1 if any triangles present, 0 otherwise
26. icon: 1 if an inanimate image present (e.g., a boat), otherwise 0
27. animate: 1 if an animate image (e.g., an eagle, a tree, a human hand) present, 0 otherwise
28. text: 1 if any letters or writing on the flag (e.g., a motto or slogan), 0 otherwise
29. topleft: colour in the top-left corner (moving right to decide tie-breaks)
30. botright: Color in the bottom-left corner (moving left to decide tie-breaks)

Practical - 3

Aim: To study about Weka minor tool and ARFF file.

Weka minor tool

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. **Weka** contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Five features of Weka that I like to promote are:

- **Open Source**: It is released as open source software under the GNU GPL. It is dual licensed and Pentaho Corporation owns the exclusive license to use the platform for business intelligence in their own product.
- **Graphical Interface**: It has a Graphical User Interface (GUI). This allows you to complete your machine learning projects without programming.
- **Command Line Interface**: All features of the software can used from the command line. This can be very useful for scripting large jobs.
- **Java API**: It is written in Java and provides a API that is well documented and promotes integration into your own applications. Note that the GNU GPL means that in turn your software would also have to be released as GPL.
- **Documentation**: There books, manuals, wikis and MOOC courses that can train you how to use the platform effectively.

❖ Weka Knowledge Explorer

The Weka Knowledge Explorer is an easy to use graphical user interface that harnesses the power of the weka software. Each of the major weka packages Filters, Classifiers, Clusterers, Associations, and Attribute Selection is represented in the Explorer along with a Visualization tool which allows datasets and the

predictions of Classifiers and Clusterers to be visualized in two dimensions.

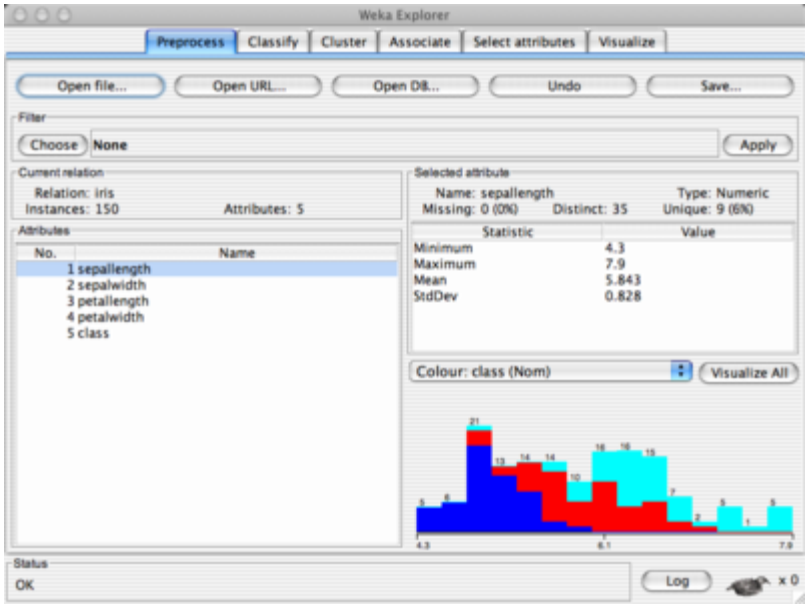
The Explorer interface is divided into 5 different tabs:

Preprocess: Load a dataset and manipulate the data into a form that you want to work with.
Classify: Select and run classification and regression algorithms to operate on your data.
Cluster: Select and run clustering algorithms on your dataset.
Associate: Run association algorithms to extract insights from your data.
Select Attributes: Run attribute selection algorithms on your data to select those attributes that are relevant to the feature you want to predict.
Visualize: Visualize the relationship between attributes.

Clicking on each of the small images below will load a full sized version.

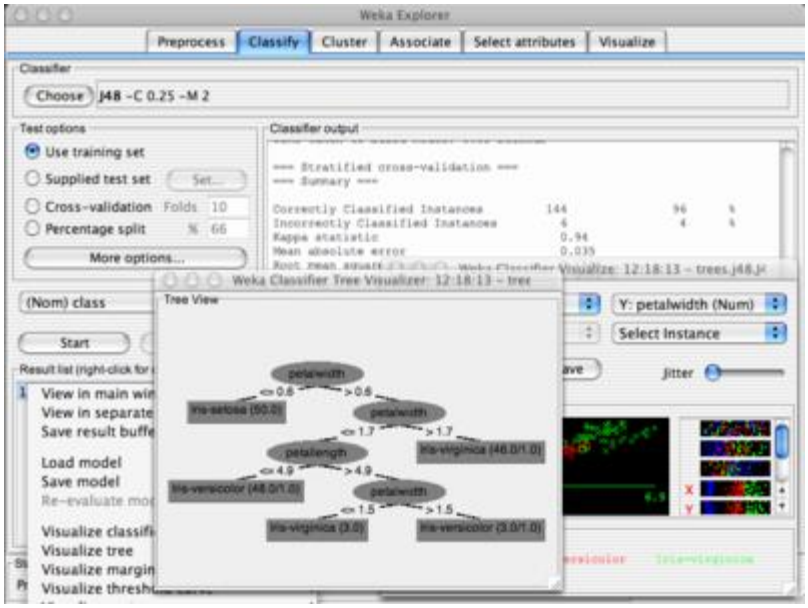
i. Preprocess Panel

The preprocess panel is the start point for knowledge exploration. From this panel you can load datasets, browse the characteristics of attributes and apply any combination of Weka's unsupervised filters. to the data.



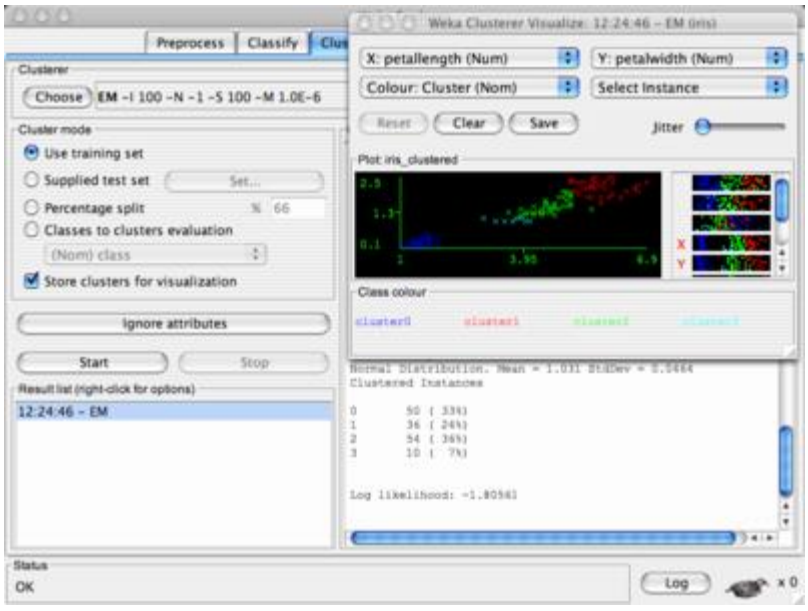
ii. Classifier Panel

The classifier panel allows you to configure and execute any of the weka classifiers on the current dataset. You can choose to perform a cross validation or test on a separate dataset. Classification errors can be visualized in a pop-up data visualization tool. If the classifier produces a decision tree it can be displayed graphically in a pop-up tree visualizer.



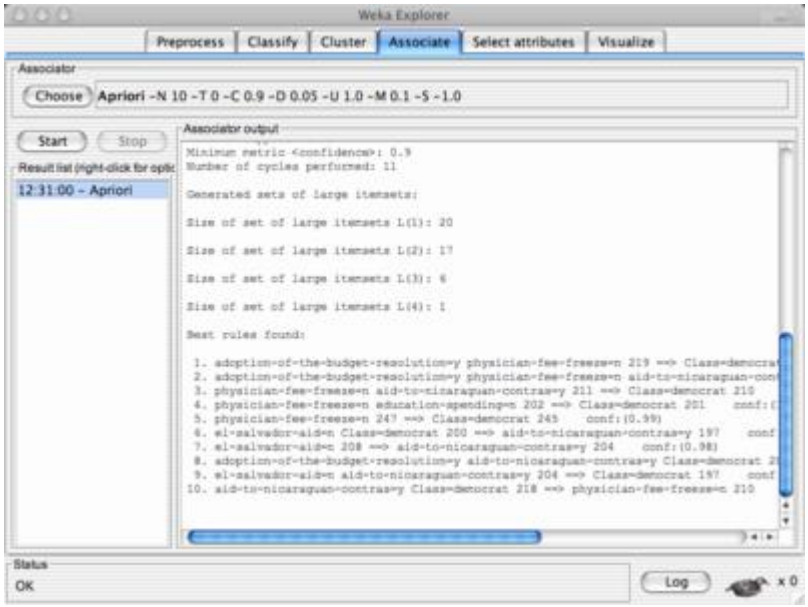
iii. Cluster Panel

From the cluster panel you can configure and execute any of the weka clusterers on the current dataset. Clusters can be visualized in a pop-up data visualization tool.



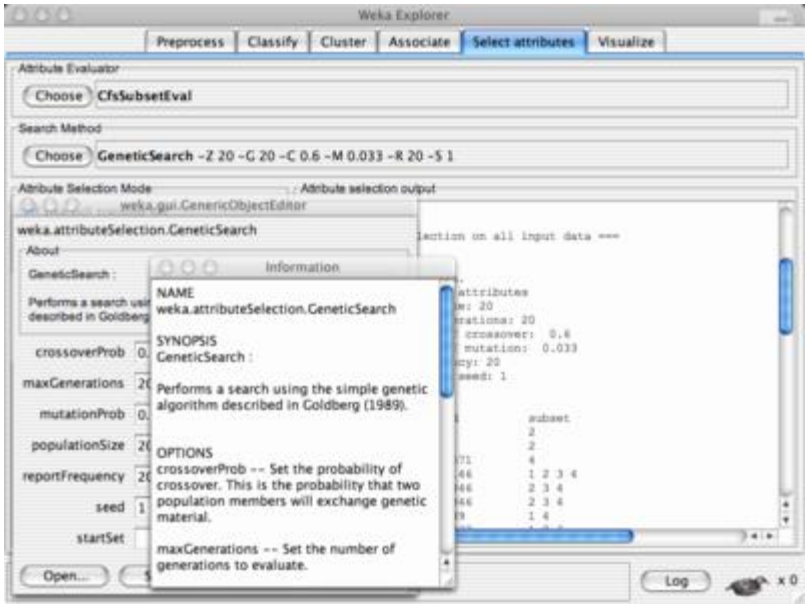
iv. Associate Panel

From the associate panel you can mine the current dataset for association rules using the weka associators.



v. Select Attributes Panel

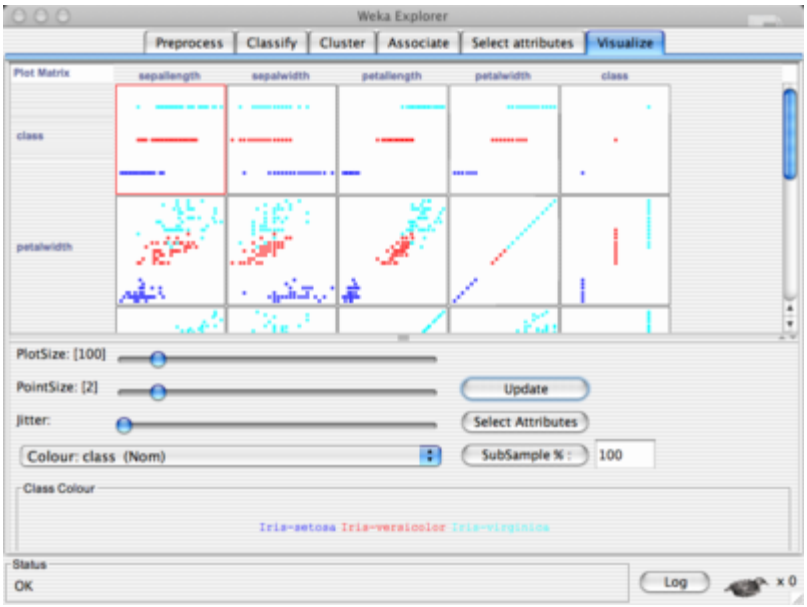
This panel allows you to configure and apply any combination of weka attribute evaluator and search method to select the most pertinent attributes in the dataset. If an attribute selection scheme transforms the data then the transformed data can be visualized in a pop-up data visualization tool.

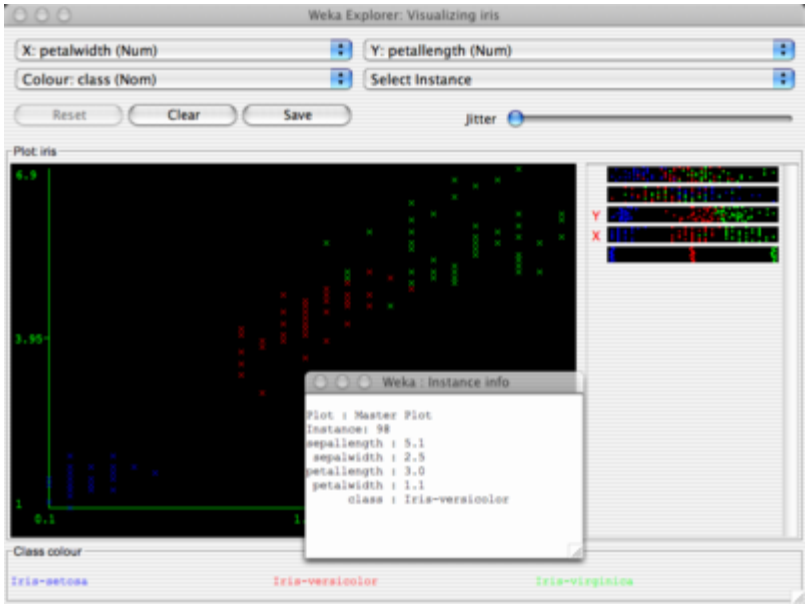


vi. Visualize Panel

This panel displays a scatter plot matrix for the current dataset. The size of the individual cells and the size of the points they display can be adjusted using the slider controls at the bottom of the panel. The number of cells in the matrix can be changed by pressing the "Select Attributes" button and then choosing those attributes to displayed. When a dataset is large, plotting performance can be improved by displaying only a subsample of the current dataset. Clicking on a cell in the matrix pops up a larger plot panel window that displays the view from that cell. This panel allows you to visualize the current dataset in one and two dimensions. When the colouring attribute is discrete, each value is displayed as a different colour; when the colouring attribute is continuous, a spectrum is used to indicate the value. Attribute "bars" (down the right hand side of the panel)

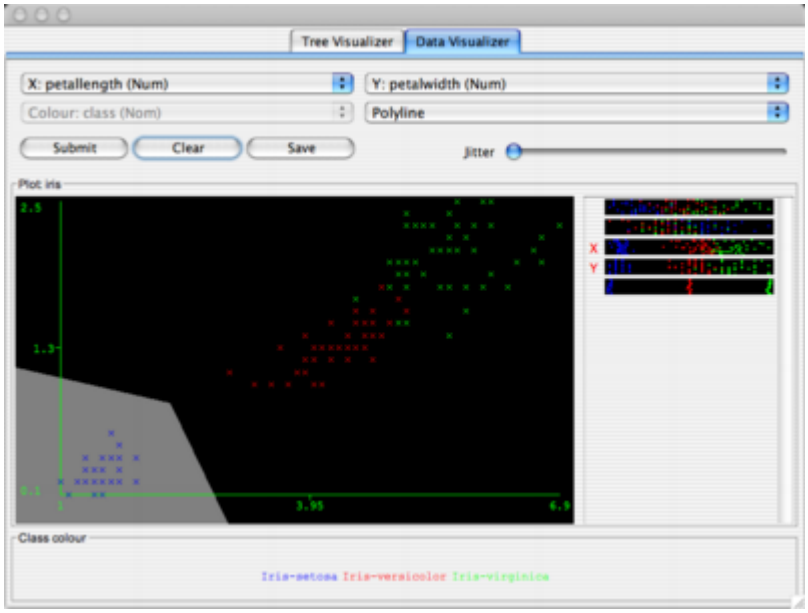
provide a convenient summary of the discriminating power of the attributes individually. This panel can also be popped up in a separate window from the classifier panel and the cluster panel to allow you to visualize predictions made by classifiers/clusterers. When the class is discrete, misclassified points are shown by a box in the colour corresponding to the class predicted by the classifier; when the class is continuous, the size of each plotted point varies in proportion to the magnitude of the error made by the classifier.

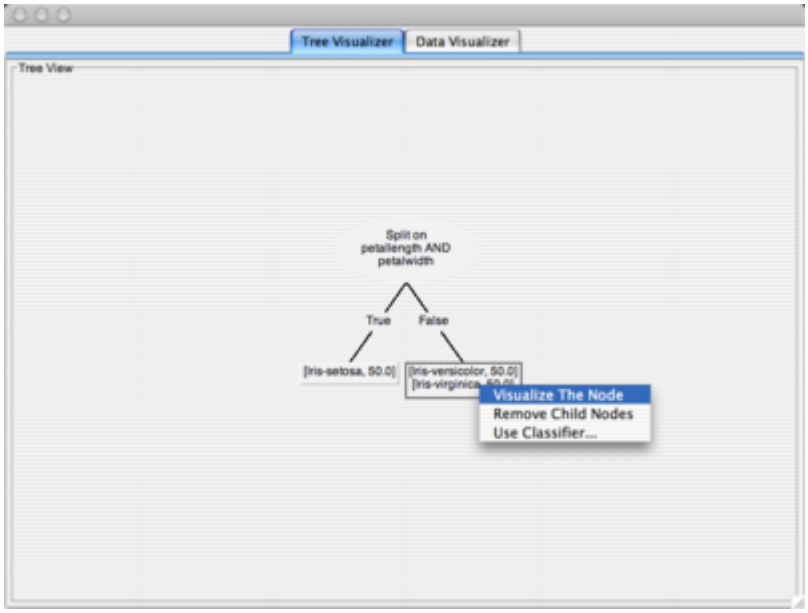




vii. Interactive decision tree construction

Weka has a novel interactive decision tree classifier (weka.classifiers.trees.UserClassifier). Through an intuitive, easy to use graphical interface, UserClassifier allows the user to manually construct a decision tree by definining bi-variate splits in the instance space. The structure of the tree can be viewed and revised at any point in the construction phase.



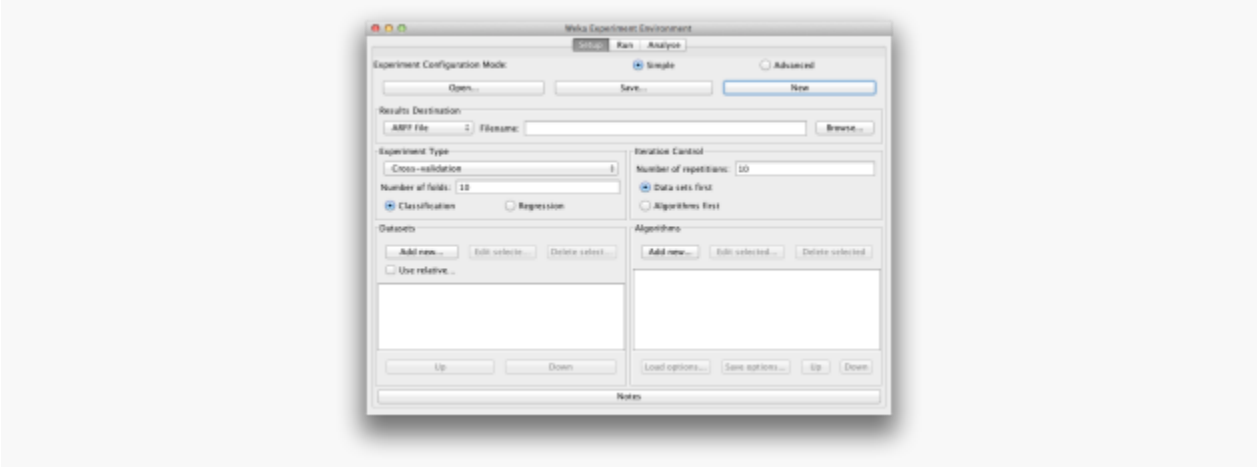


iii. Neural Network GUI

Weka also has a graphical user interface to a neural network (weka.classifiers.functions.neural.NeuralNetwork). This interface allows the user to specify the structure of a multi-layer perceptron and the parameters that control its training.

❖ Weka Experimenter

This interface is for designing experiments with your selection of algorithms and datasets, running experiments and analyzing the results.

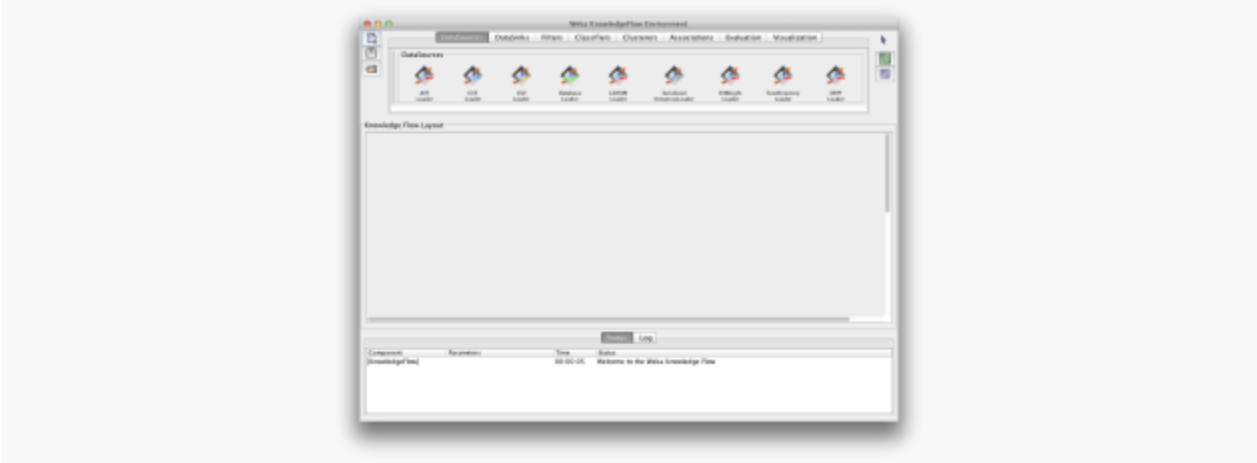


Weka Experimenter Interface

The tools for analyzing results are very powerful, allowing you to consider and compare results that are statistically significant over multiple runs.

❖ Knowledge Flow

Applied machine learning is a process and the Knowledge Flow interface allows you to graphically design that process and run the designs that you create. This includes the loading and transforming of input data, running of algorithms and the presentation of results.



Weka Knowledge Flow Interface

It's a powerful interface and metaphor for solving complex problems graphically.

ARFF FILE

❖ Synopsis

This operator is used for reading an ARFF file.

❖ Description

This operator can read ARFF (Attribute-Relation File Format) files known from the machine learning library Weka. An ARFF

file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software. Please study the attached Example Process for understanding the basics and structure of the ARFF file format. Please note that when an ARFF file is written, the roles of the attributes are not stored. Similarly when an ARFF file is read, the roles of all the attributes are set to regular.

❖ Input

- file
  An ARFF file is expected as a file object which can be created with other operators with file output ports like the Read File operator.

❖ Output

- output *(Data Table)*
  This port delivers the ARFF file in tabular form along with the meta data. This output is similar to the output of the Retrieve operator.

❖ Parameters

- **data_file** The path of the ARFF file is specified here. It can be selected using the *choose a file* button. *Range: filename*
- **encoding** This is an expert parameter. A long list of encoding is provided; users can select any of them. *Range: selection*
- **read_not_matching_values_as_missings** This is an expert parameter. If this parameter is set to true, values that do not match with the expected value type are considered as missing values and are replaced by '?'. For example if 'abc' is written in

an integer column, it will be treated as a missing value. Question mark (?) in ARFF file is also read as missing value.*Range: Boolean*

- **decimal_character** This character is used as the decimal character.*Range: char*
- **grouped_digits** This parameter decides whether grouped digits should be parsed or not. If this parameter is set to true, the *grouping character* parameter should be specified.*Range: Boolean*
- **grouping_character** This parameter is available only when the *grouped digits* parameter is set to true. This character is used as the grouping character. If it is found between numbers, the numbers are combined and this character is ignored. For example if "22-14" is present in the ARFF file and "-" is set as *grouping character*, then "2214" will be stored.*Range: char*
- **infinity_string** This parameter can be set to parse a specific infinity representation (e.g. "Infinity"). If it is not set, the local specific infinity representation will be used.*Range: string*

❖ Tutorial Processes

- The basics of the ARFF

The 'Iris' data set is loaded using the Retrieve operator. The Write ARFF operator is applied on it to write the 'Iris' data set into an ARFF file. The example set file parameter is set to 'D:\Iris'. Thus an ARFF file is created in the 'D' drive of your computer with the name 'Iris'. Open this file to see the structure of an ARFF file.

ARFF files have two distinct sections. The first section is the Header information, which is followed by the Data information. The Header of the ARFF file contains the name of the Relation

and a list of the attributes. The name of the Relation is specified after the @RELATION statement. The Relation is ignored by RapidMiner. Each attribute definition starts with the @ATTRIBUTE statement followed by the attribute name and its type. The resultant ARFF file of this Example Process starts with the Header. The name of the relation is 'RapidMinerData'. After the name of the Relation, six attributes are defined.

Attribute declarations take the form of an ordered sequence of @ATTRIBUTE statements. Each attribute in the data set has its own @ATTRIBUTE statement which uniquely defines the name of that attribute and its data type. The order of declaration of the attributes indicates the column position in the data section of the file. For example, in the resultant ARFF file of this Example Process the 'label' attribute is declared at the end of all other attribute declarations. Therefore values of the 'label' attribute are in the last column of the Data section.

The possible attribute types in ARFF are: numeric integer real {nominalValue1,nominalValue2,...} for nominal attributes string for nominal attributes without distinct nominal values (it is however recommended to use the nominal definition above as often as possible) date [date-format] (currently not supported by Rapid Miner)

You can see in the resultant ARFF file of this Example Process that the attributes 'a1', 'a2', 'a3' and 'a4' are of real type. The attributes 'id' and 'label' are of nominal type. The distinct nominal values are also specified with these nominal attributes.

The ARFF Data section of the file contains the data declaration line @DATA followed by the actual example data lines. Each example is represented on a single line, with carriage returns

denoting the end of the example. Attribute values for each example are delimited by commas. They must appear in the order that they were declared in the Header section (i.e. the data corresponding to the n-th @ATTRIBUTE declaration is always the n-th field of the example line). Missing values are represented by a single question mark (?).

A percent sign (%) introduces a comment and will be ignored during reading. Attribute names or example values containing spaces must be quoted with single quotes ('). Please note that in RapidMiner the sparse ARFF format is currently only supported for numerical attributes. Please use one of the other options for sparse data files provided by RapidMiner if you also need sparse data files for nominal attributes.

- Reading an ARFF file using the Read ARFF operator

The ARFF file that was written in the first Example Process using the Write ARFF operator is retrieved in this Example Process using the Read ARFF operator. The data file parameter is set to '%{tempdir}/Iris'. All other parameters are used with default values. Run the process. You will see that the results are very similar to the original Iris data set of RapidMiner repository. Please note that the role of all the attributes is regular in the results of the Read ARFF operator. Even the roles of 'id' and 'label' attributes are set to regular. This is so because the ARFF files do not store information about the roles of the attributes.

❖ Sample Weka Data Sets

Below are some sample WEKA data sets, in arff format.

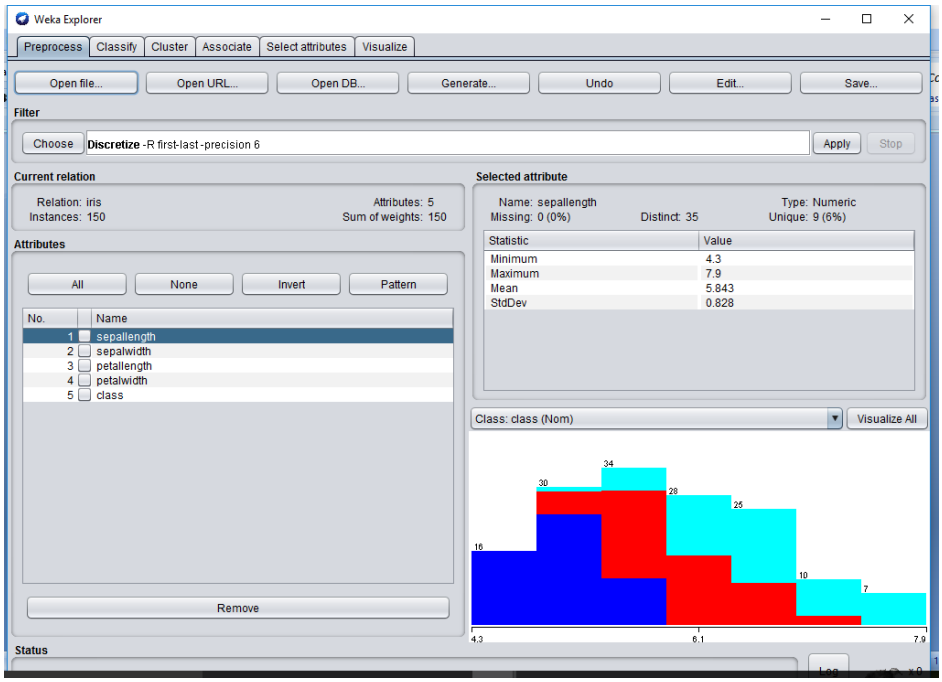
- contact-lens.arff
- cpu.arff
- <="" a="" style="font-family: Verdana, Arial, Helvetica, sans-serif; font-size: 10pt; line-height: 13pt; color: rgb(0, 68, 204); text-decoration: none;">cpu.with-vendor.arff
- diabetes.arff
- glass.arff
- ionospehre.arff
- iris.arff
- labor.arff
- ReutersCorn-train.arff
- ReutersCorn-test.arff
- ReutersGrain-train.arff
- ReutersGrain-test.a*rff*
- segment-challenge.arff
- segment-test.arff
- soybean.arff
- supermarket.arff
- vote.arff
- weather.arff
- weather.nominal.arff

Practical 4

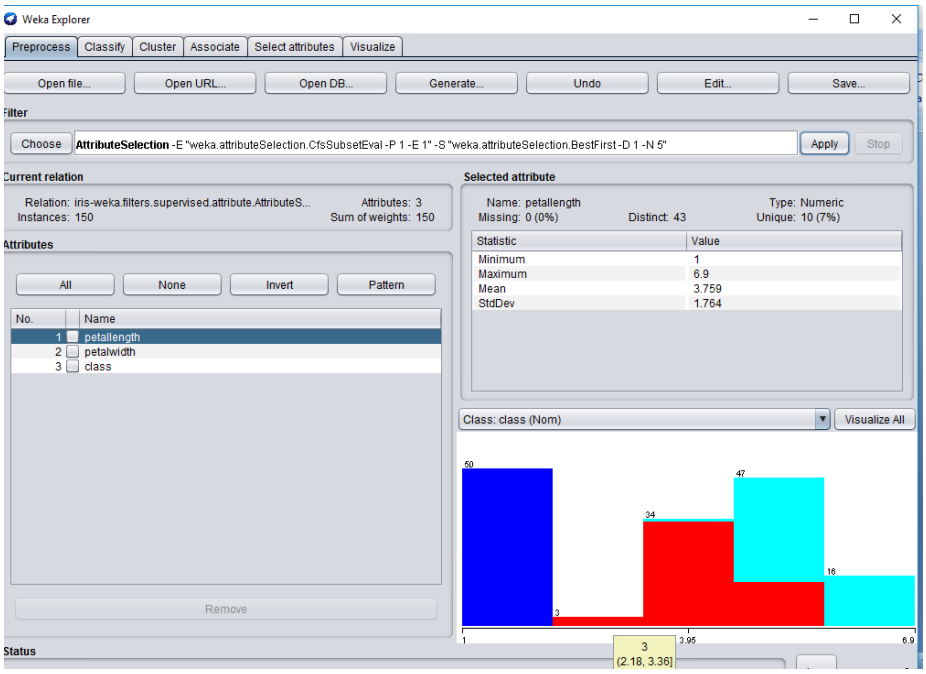
Aim: To preprocess the data using weka mining tool.

Use at least two preprocessing methods on at least three dataset.

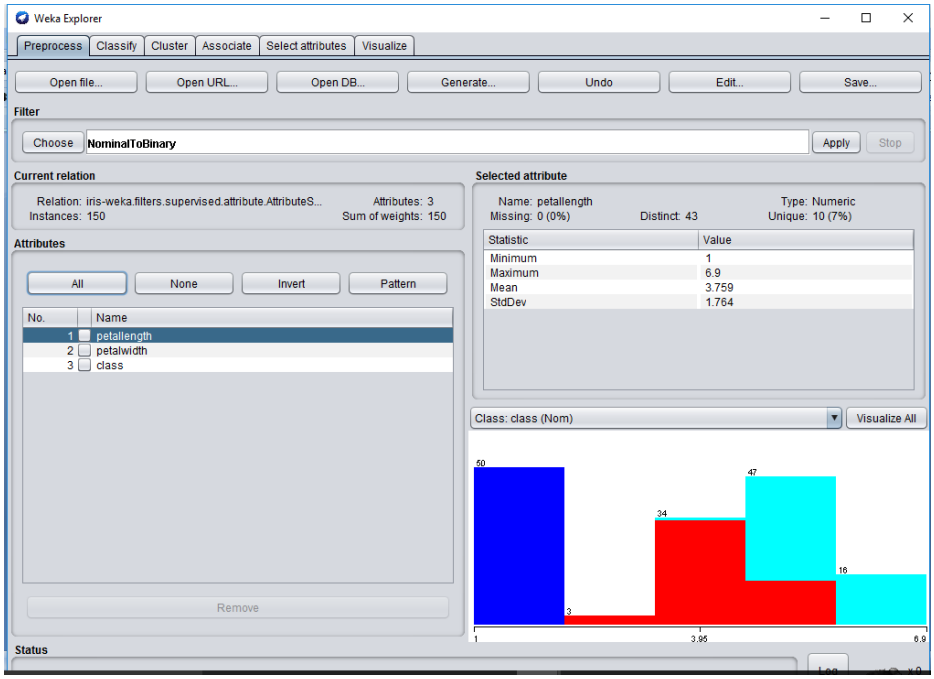
- ❖ Dataset 1: iris dataset



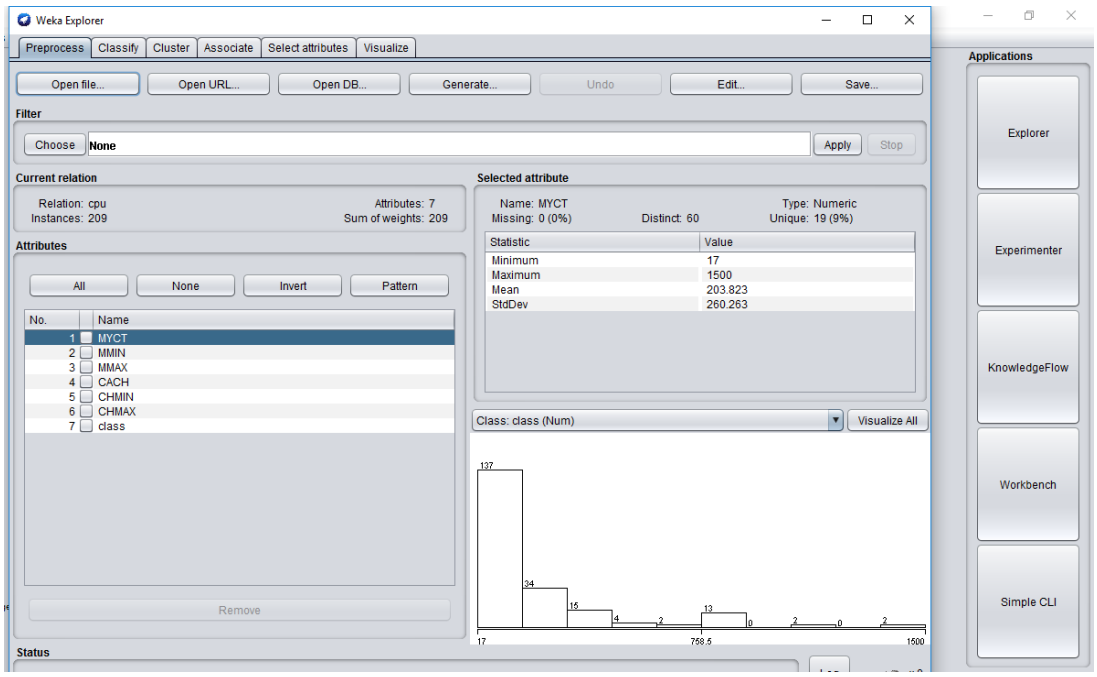
- Preprocessing Method 1: attribute selection



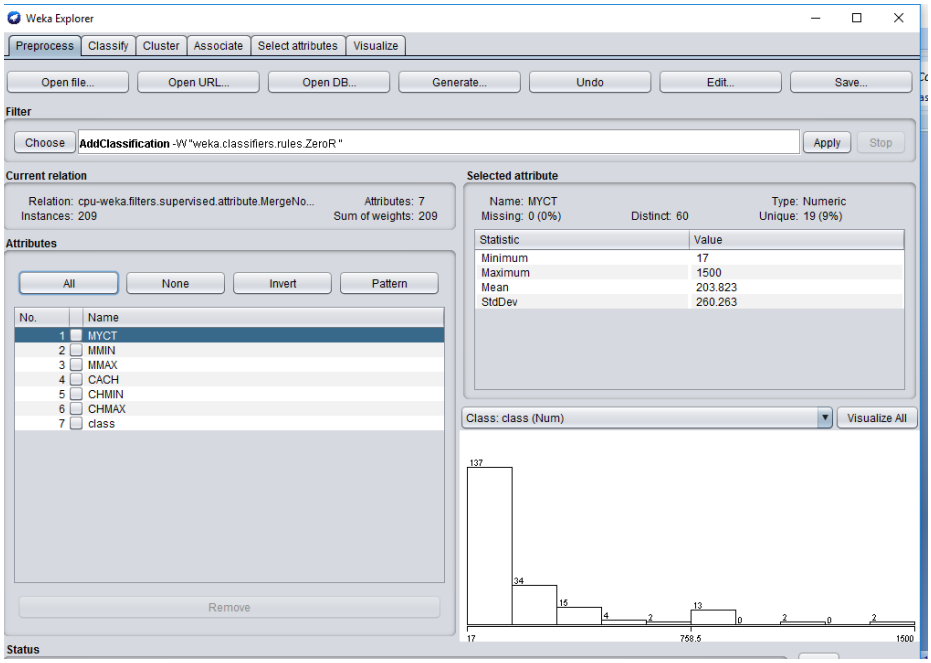
- Preprocessing Method 2: nominal to binary



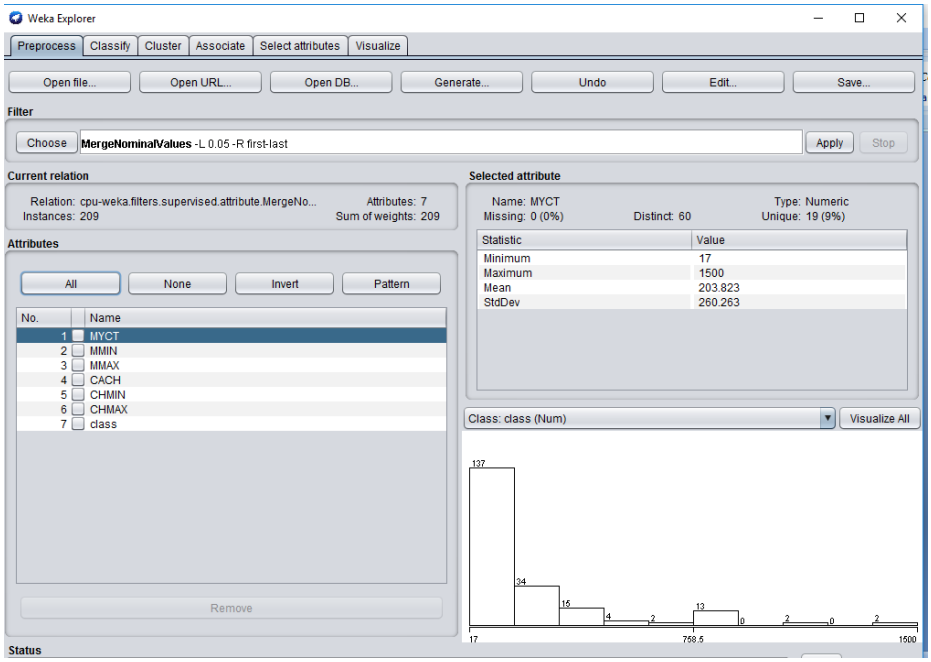
❖ Dataset 2: cpu dataset



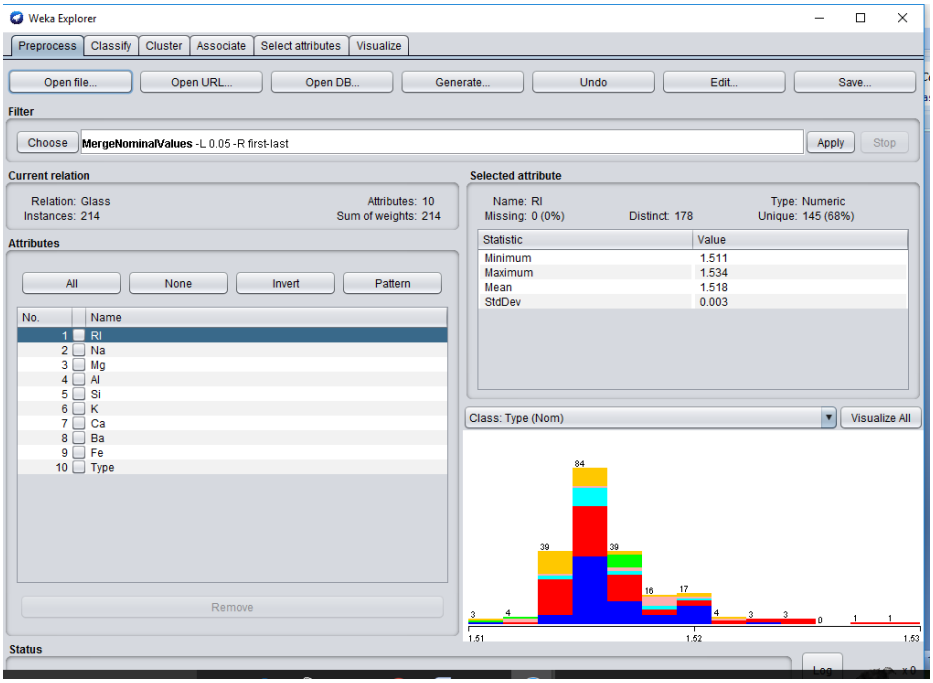
- Preprocessing Method 1: add classification



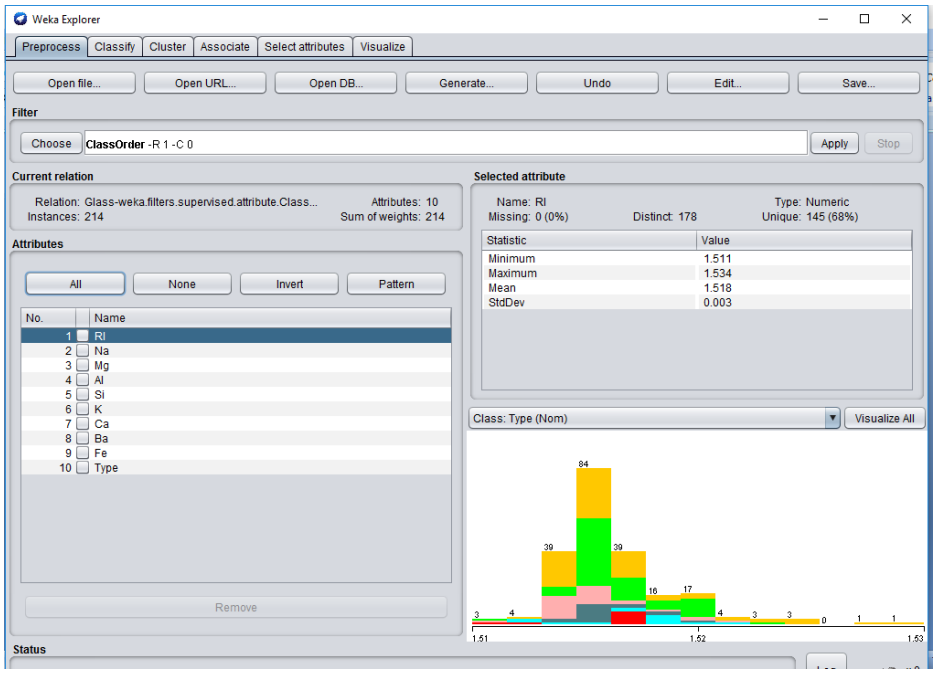
- Preprocessing Method 2: merge nominal values



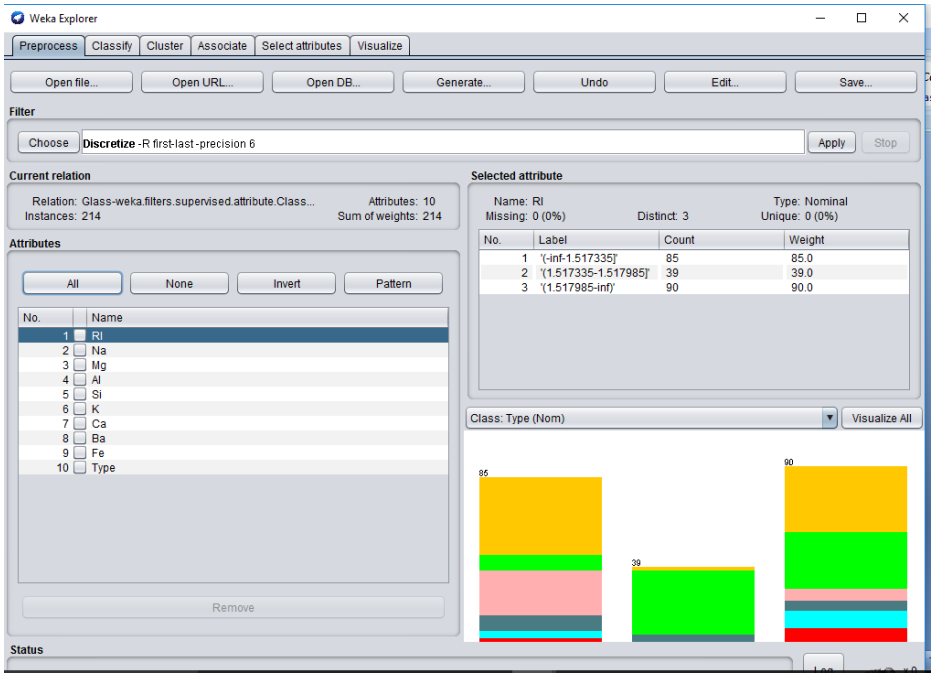
❖ Dataset 3: glass dataset



- Preprocessing Method 1: class order



- Preprocessing Method 2: discretize

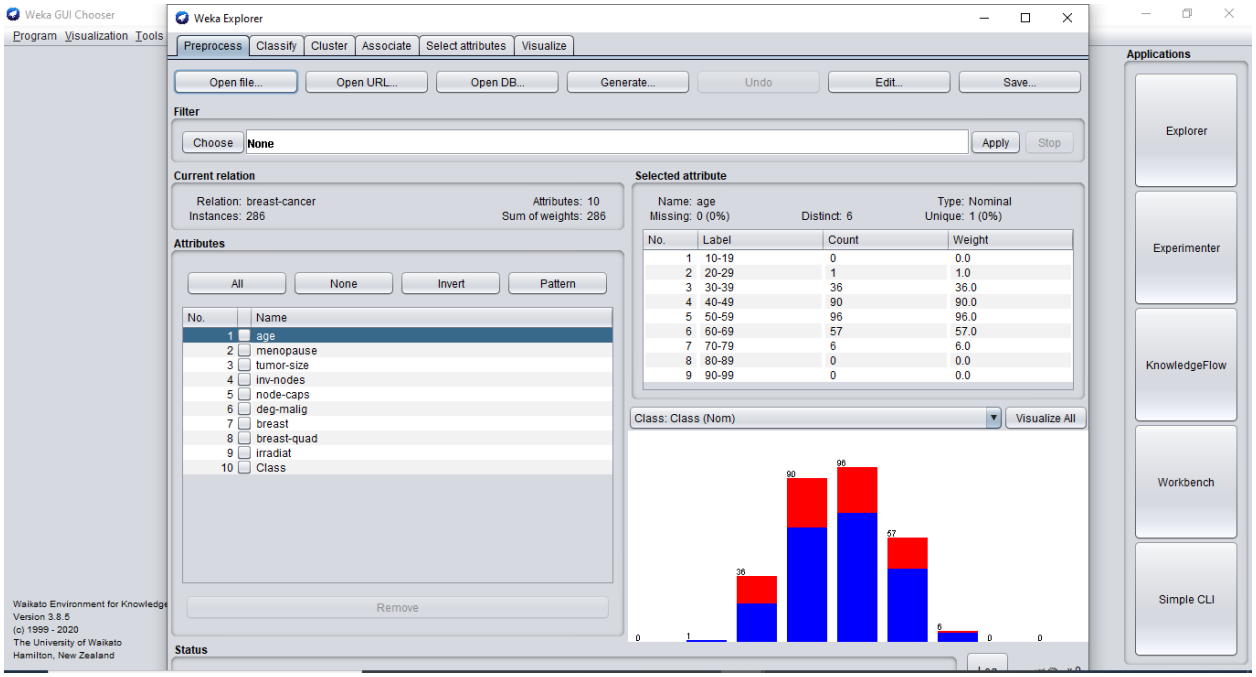


Practical 5

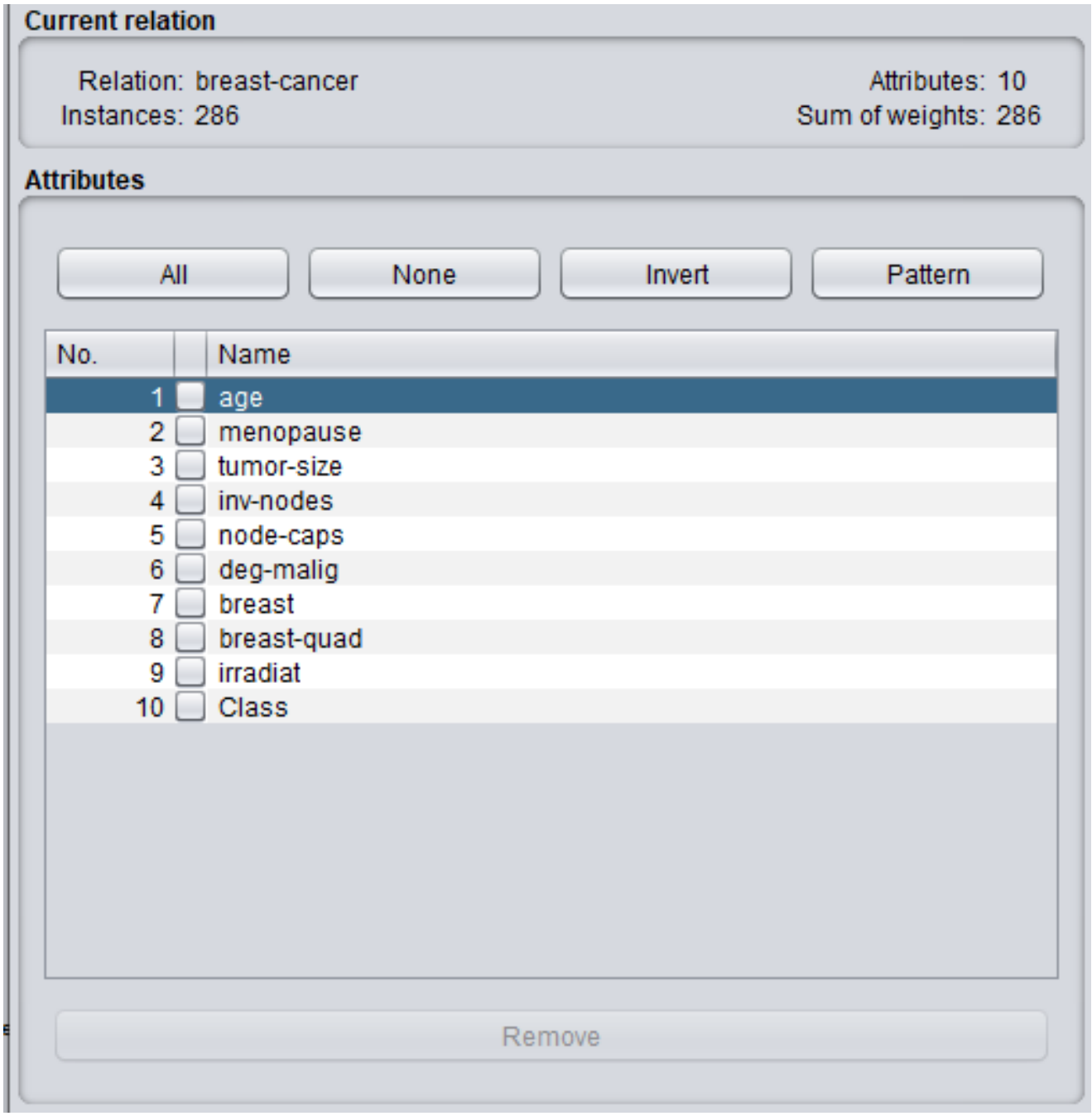
Aim: To preprocess the data using weka mining tool.

Use supervised and unsupervised preprocessing methods on at "breast cancer" dataset.

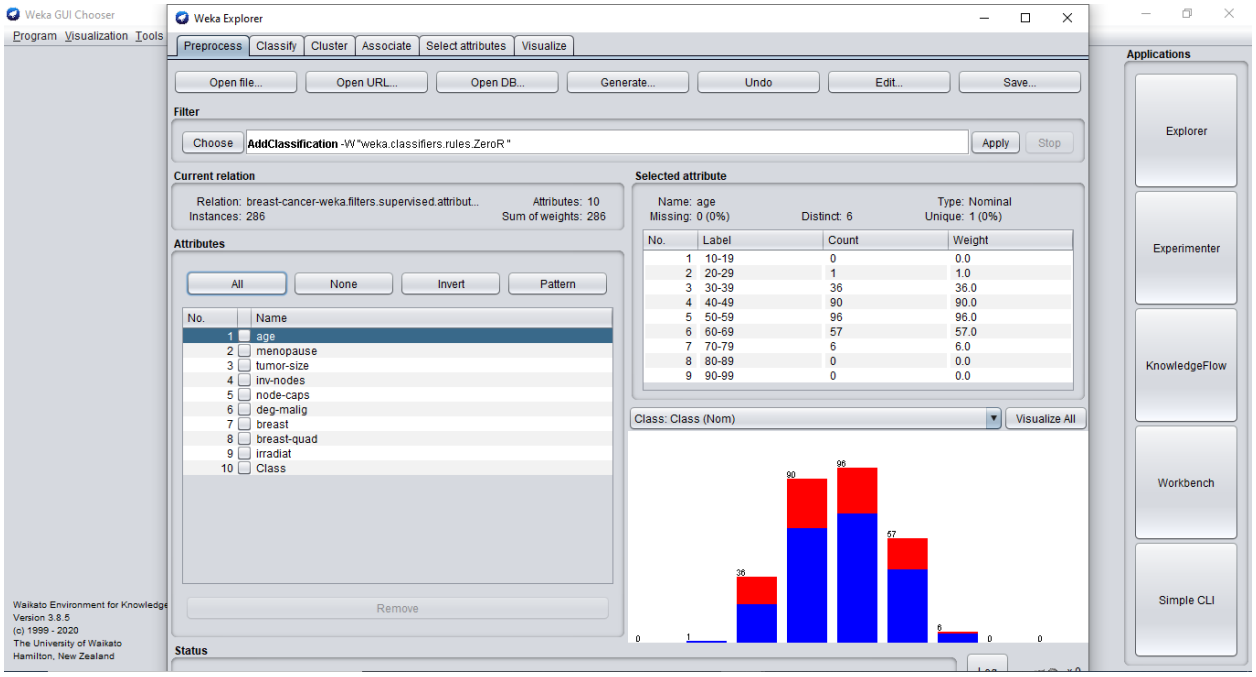
❖ Dataset: breast cancer:



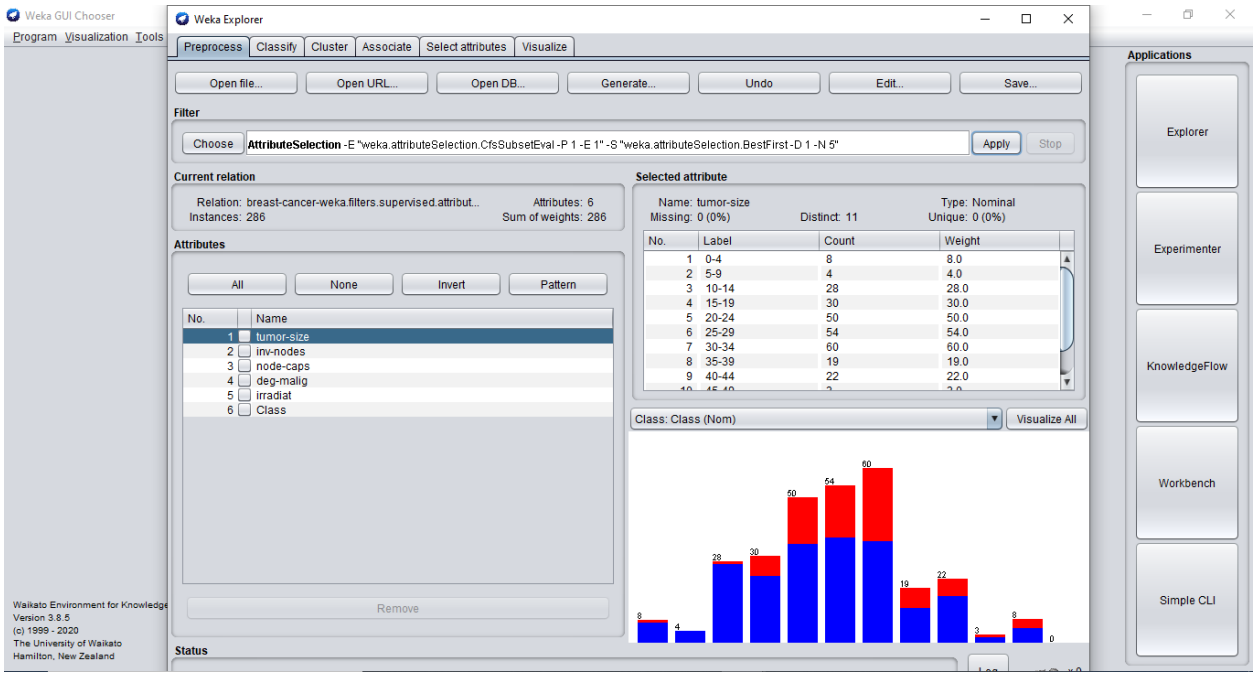
Attributes:



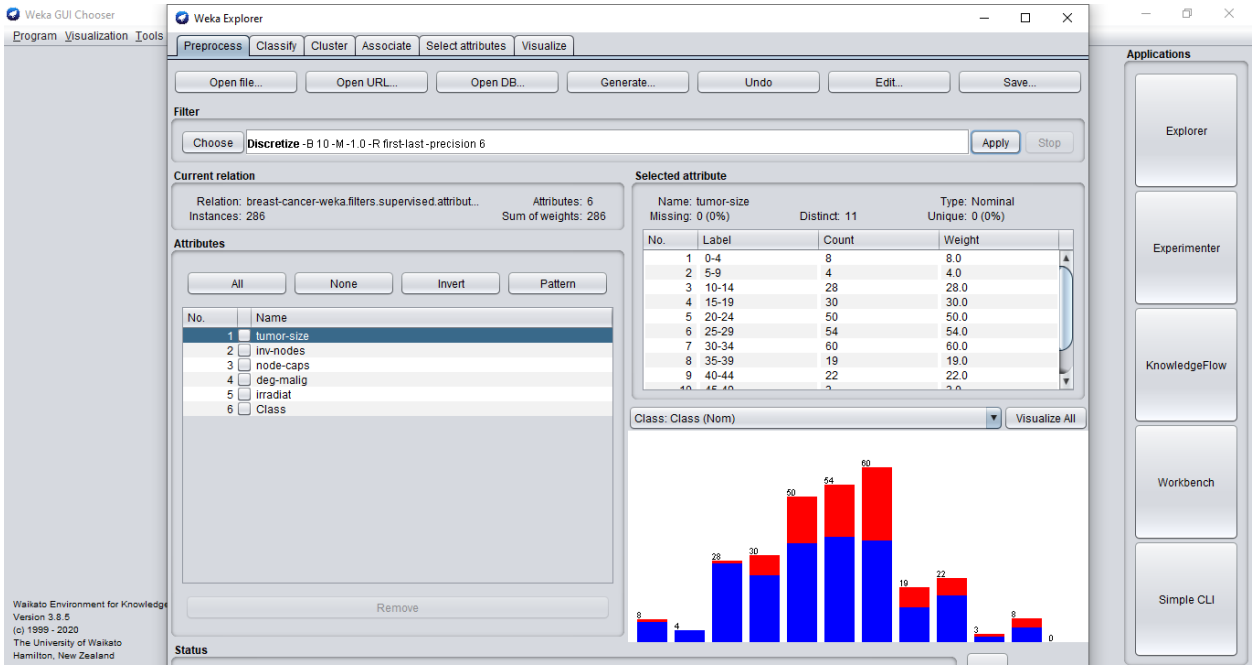
- Filter: Supervised: -> ADD classification



- Filter: Supervised: -> attribute selection



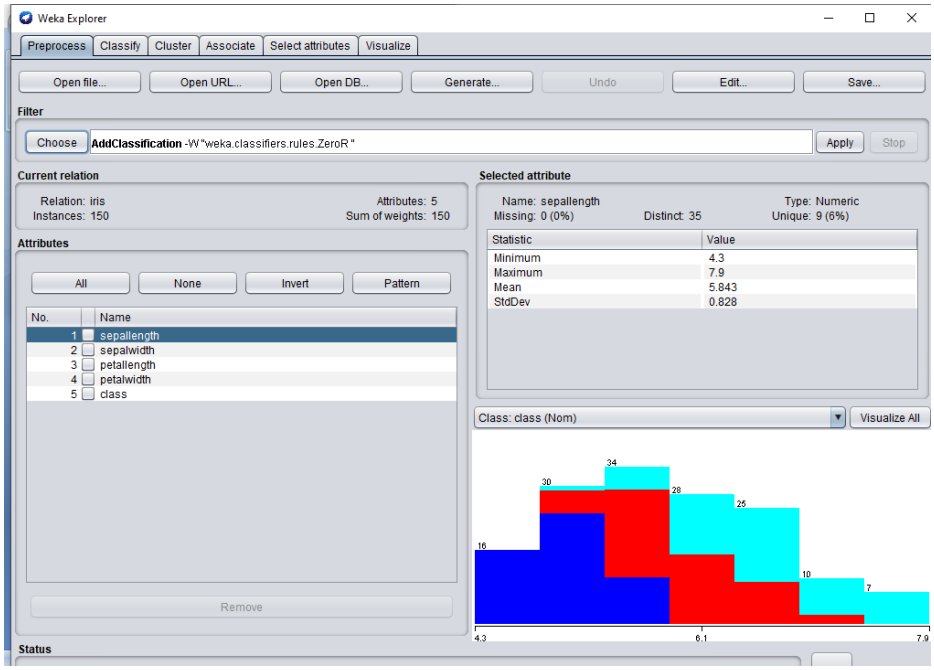
- Filter: Unsupervised: -> discretize

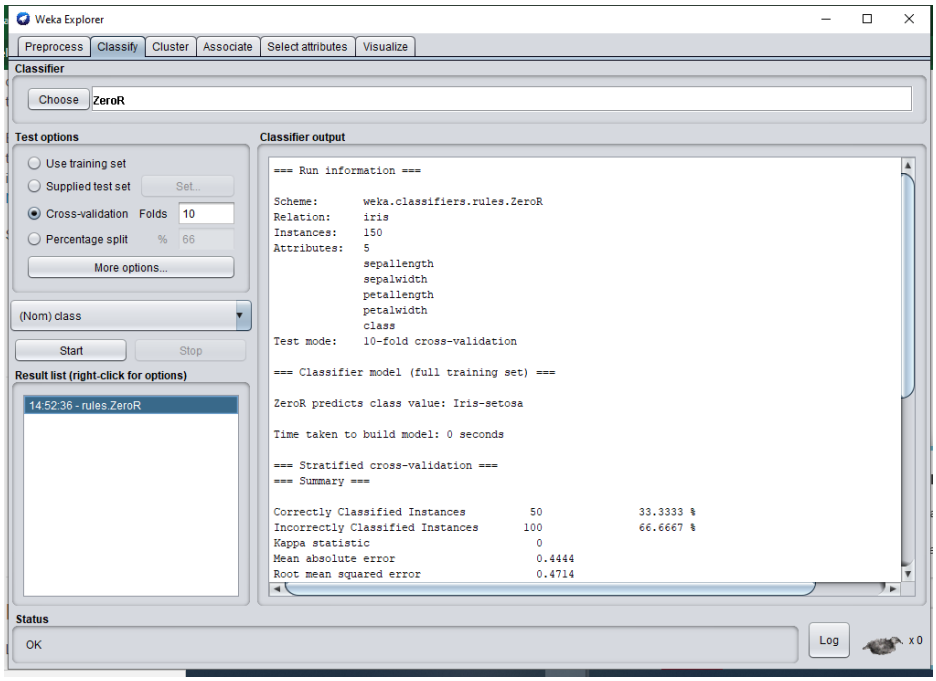


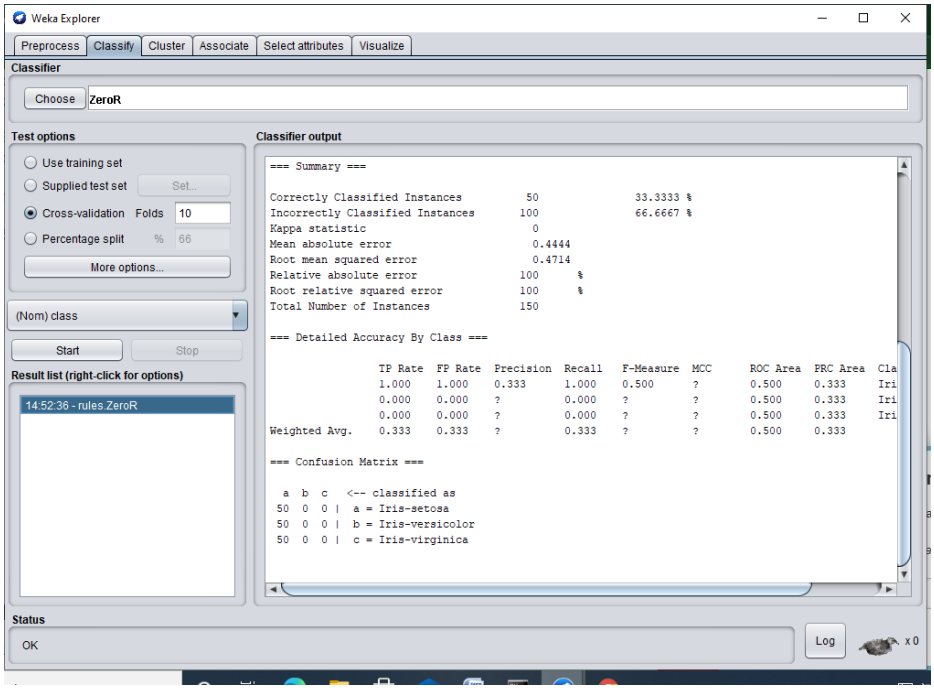
Practical 6

Aim: To perform Classification on various datasets and compare the accuracy of various algorithms.

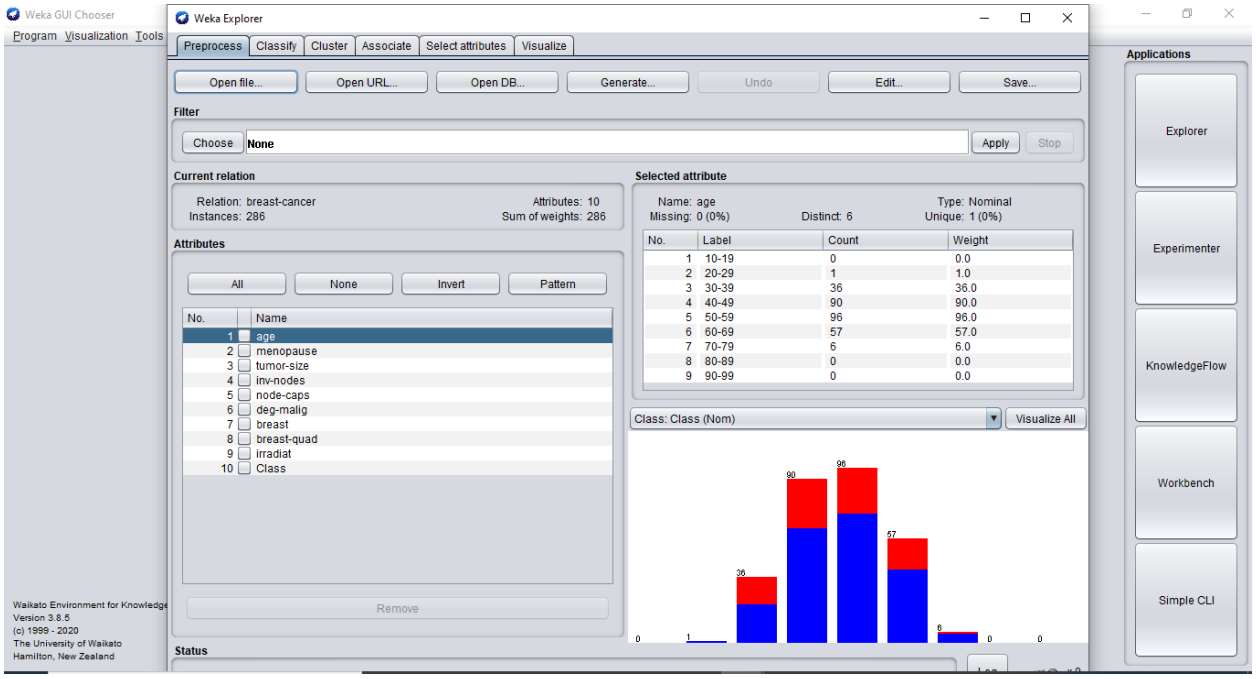
❖ Dataset 1: iris

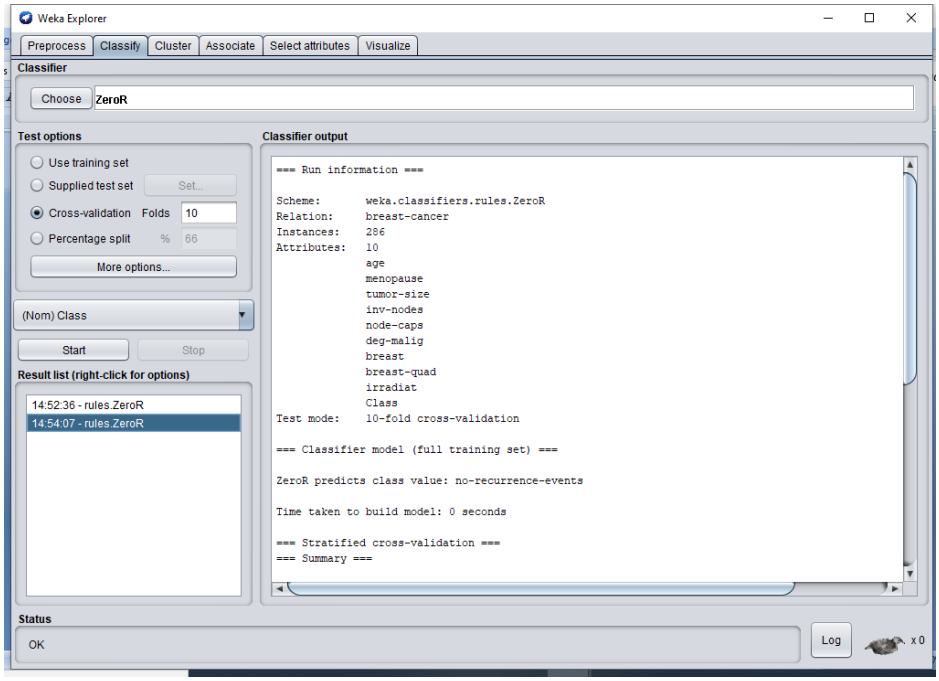


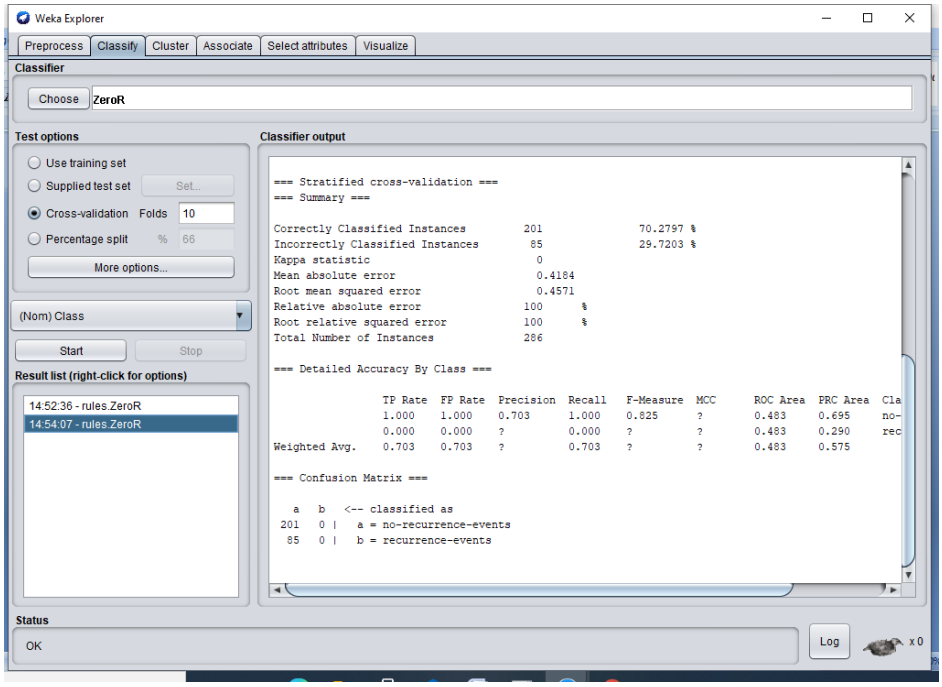




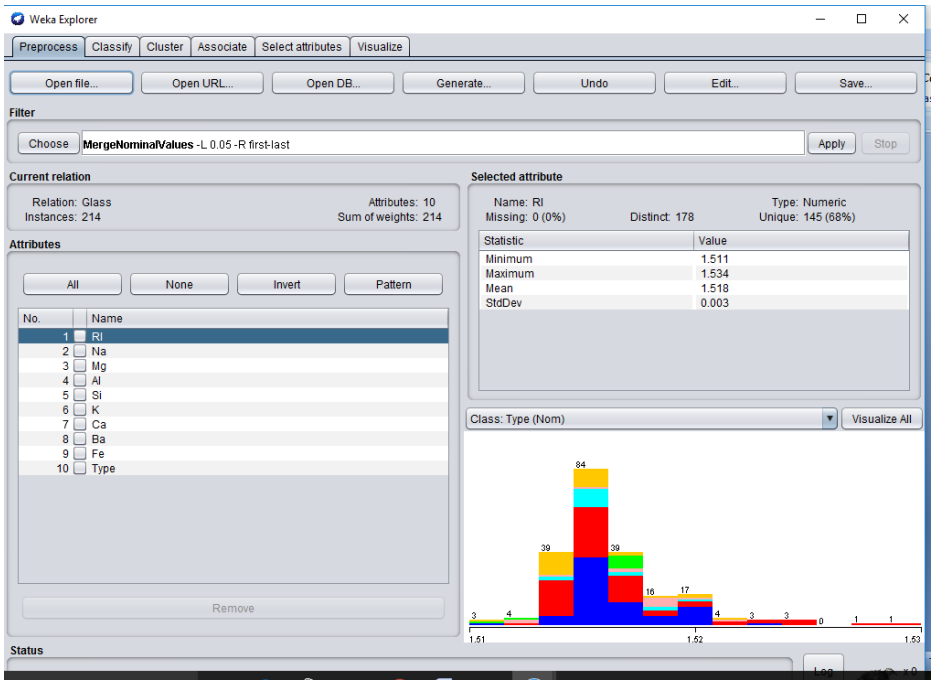
❖ Dataset 2: breast cancer:

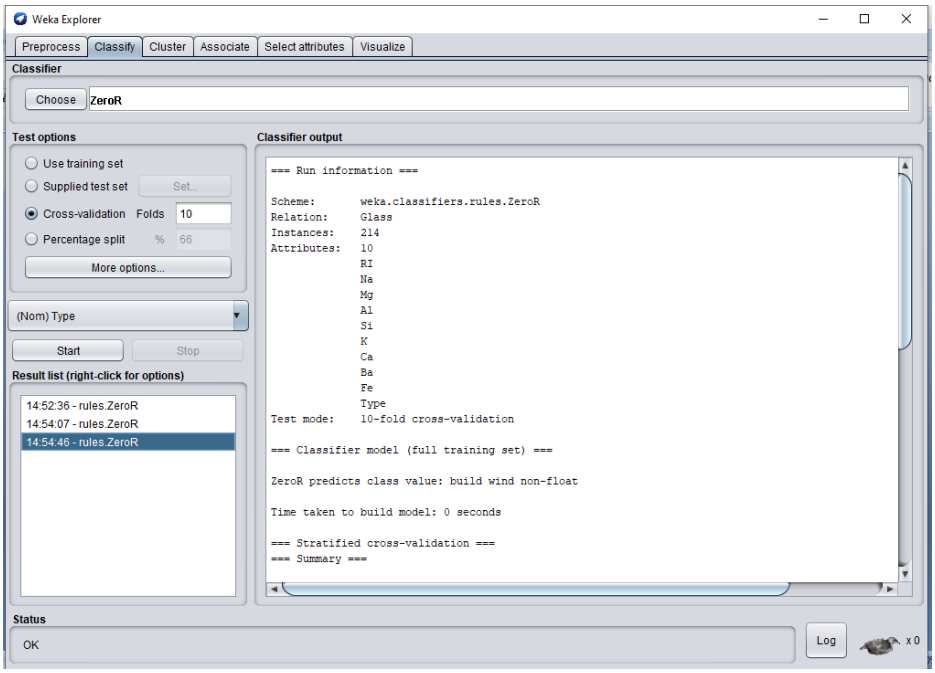

Weka GUI Chooser
Program Visualization Tools
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Filter
Choose None Apply Stop
Current relation
Relation: breast-cancer
Instances: 286
Attributes: 10
Sum of weights: 286
Selected attribute
Name: age
Missing: 0 (0%)
Distinct: 6
Type: Nominal
Unique: 1 (0%)
No. Label Count Weight
1 10-19 0 0.0
2 20-29 1 1.0
3 30-39 36 36.0
4 40-49 90 90.0
5 50-59 96 96.0
6 60-69 57 57.0
7 70-79 6 6.0
8 80-89 0 0.0
9 90-99 0 0.0
Attributes
All None Invert Pattern
No. Name
1 age
2 menopause
3 tumor-size
4 inv-nodes
5 node-caps
6 deg-malig
7 breast
8 breast-quad
9 irradiat
10 Class
Remove
Class: Class (Nom)
Visualize All
Status
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Waikato Environment for Knowledge
Version 3.8.5
(c) 1999 - 2020
The University of Waikato
Hamilton, New Zealand

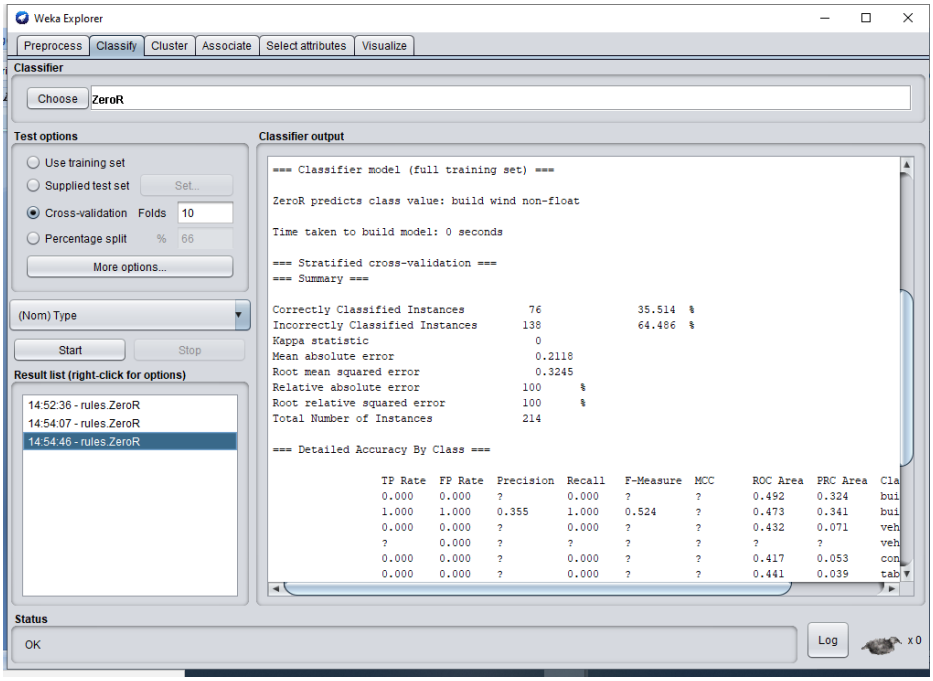

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose ZeroR
Test options
Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Class
Start Stop
Result list (right-click for options)
14:52:36 - rules.ZeroR
14:54:07 - rules.ZeroR
Classifier output
=== Run information ===
Scheme: weka.classifiers.rules.ZeroR
Relation: breast-cancer
Instances: 286
Attributes: 10
age
menopause
tumor-size
inv-nodes
node-caps
deg-malig
breast
breast-quad
irradiat
Class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
ZeroR predicts class value: no-recurrence-events
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Status
OK
Log
x 0

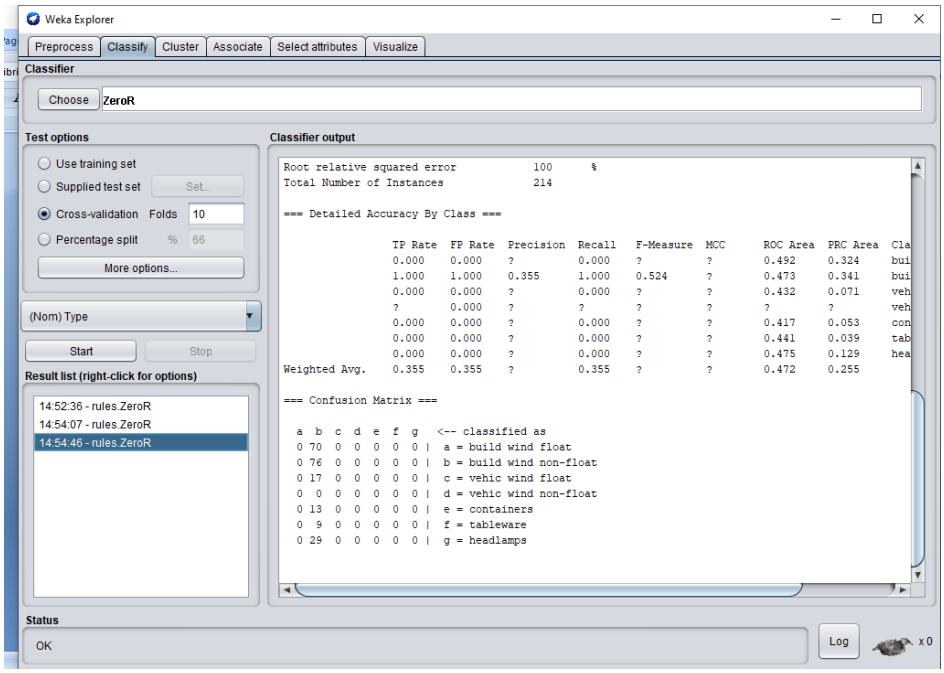


❖ Dataset 3: glass dataset




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
14:52:36 - rules.ZeroR
14:54:07 - rules.ZeroR
14:54:46 - rules.ZeroR
Classifier output
=== Run information ===
Scheme: weka.classifiers.rules.ZeroR
Relation: Glass
Instances: 214
Attributes: 10
RI
Na
Mg
Al
Si
K
Ca
Ba
Fe
Type
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
ZeroR predicts class value: build wind non-float
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
14:52:36 - rules.ZeroR
14:54:07 - rules.ZeroR
14:54:46 - rules.ZeroR
Classifier output
=== Classifier model (full training set) ===
ZeroR predicts class value: build wind non-float
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 76 35.514 %
Incorrectly Classified Instances 138 64.486 %
Kappa statistic 0
Mean absolute error 0.2118
Root mean squared error 0.3245
Relative absolute error 100 %
Root relative squared error 100 %
Total Number of Instances 214
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.000 0.000 ? 0.000 ? ? 0.492 0.324 bui
1.000 1.000 0.355 1.000 0.524 ? 0.473 0.341 bui
0.000 0.000 ? 0.000 ? ? 0.432 0.071 veh
? 0.000 ? ? ? ? ? ? veh
0.000 0.000 ? 0.000 ? ? 0.417 0.053 con
0.000 0.000 ? 0.000 ? ? 0.441 0.039 tab
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
ZeroR
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
14:52:36 - rules.ZeroR
14:54:07 - rules.ZeroR
14:54:46 - rules.ZeroR
Classifier output
Root relative squared error 100 %
Total Number of Instances 214
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.000 0.000 ? 0.000 ? ? 0.492 0.324 bui
1.000 1.000 0.355 1.000 0.524 ? 0.473 0.341 bui
0.000 0.000 ? 0.000 ? ? 0.432 0.071 veh
? 0.000 ? ? ? ? ? ? veh
0.000 0.000 ? 0.000 ? ? 0.417 0.053 con
0.000 0.000 ? 0.000 ? ? 0.441 0.039 tab
0.000 0.000 ? 0.000 ? ? 0.475 0.129 hea
Weighted Avg. 0.355 0.355 ? 0.355 ? ? 0.472 0.255
=== Confusion Matrix ===
a b c d e f g <-- classified as
0 70 0 0 0 0 0 | a = build wind float
0 76 0 0 0 0 0 | b = build wind non-float
0 17 0 0 0 0 0 | c = vehic wind float
0 0 0 0 0 0 0 | d = vehic wind non-float
0 13 0 0 0 0 0 | e = containers
0 9 0 0 0 0 0 | f = tableware
0 29 0 0 0 0 0 | g = headlamps
Status
OK
Log
x 0

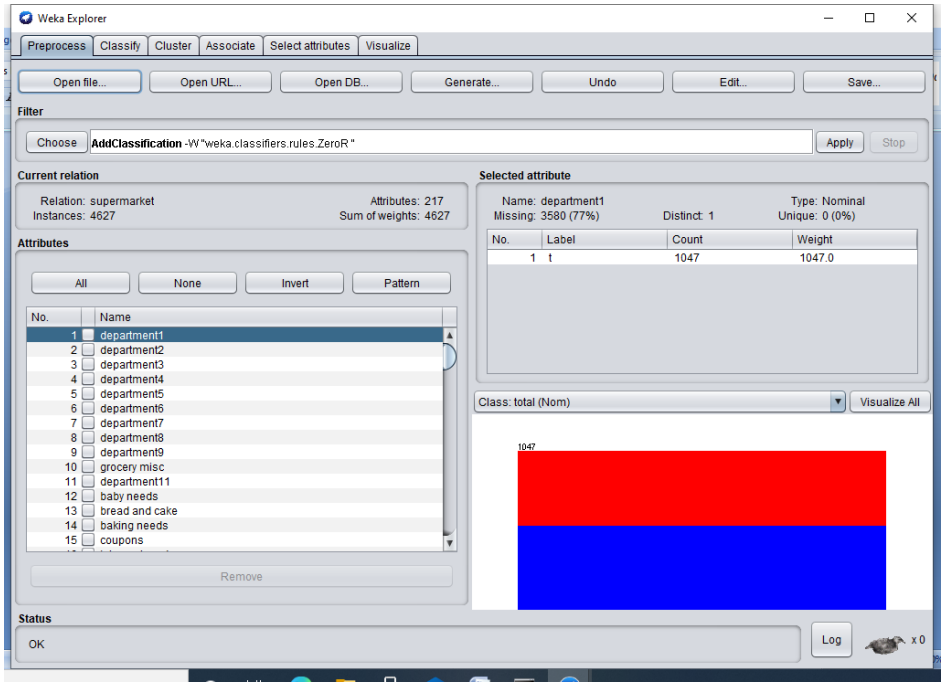
Practical 7

Aim: To perform Association rule mining on various datasets.

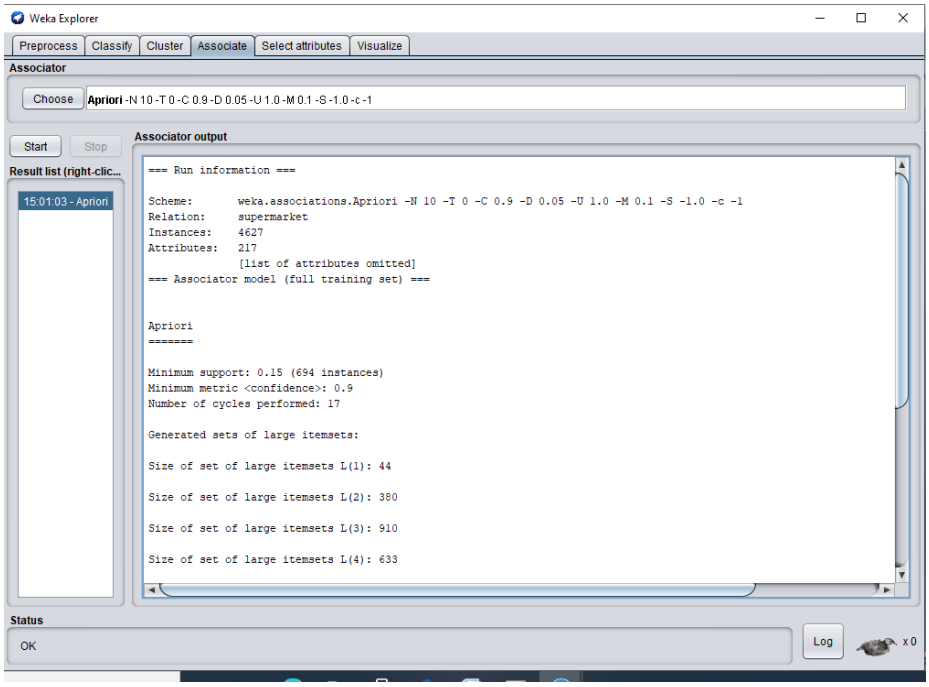
Dataset: Supermarket

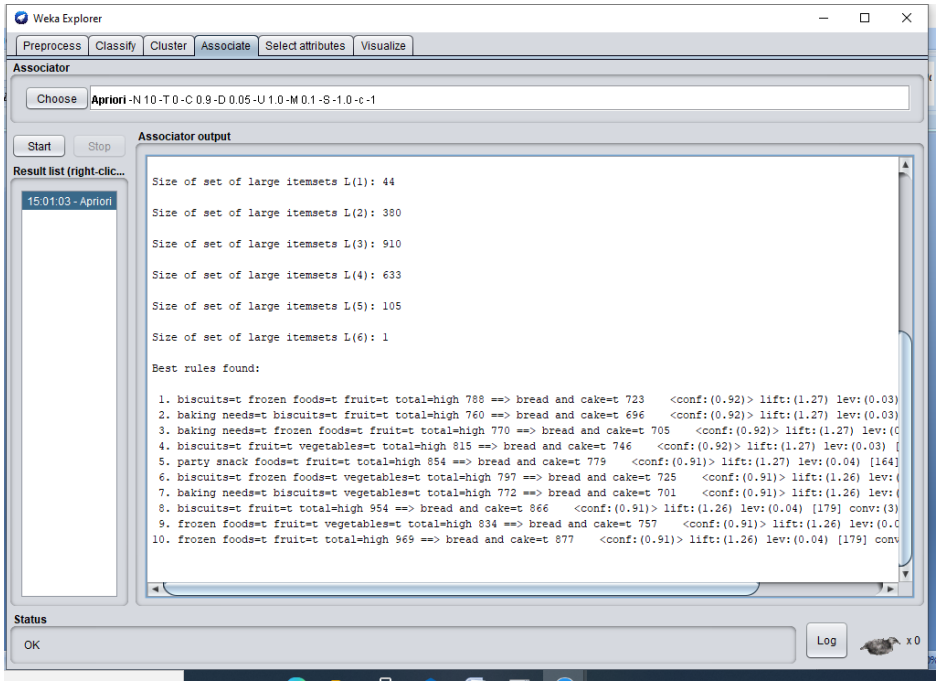
Algorithm: Apriori

- Dataset: Supermarket



- Algorithm: Apriori





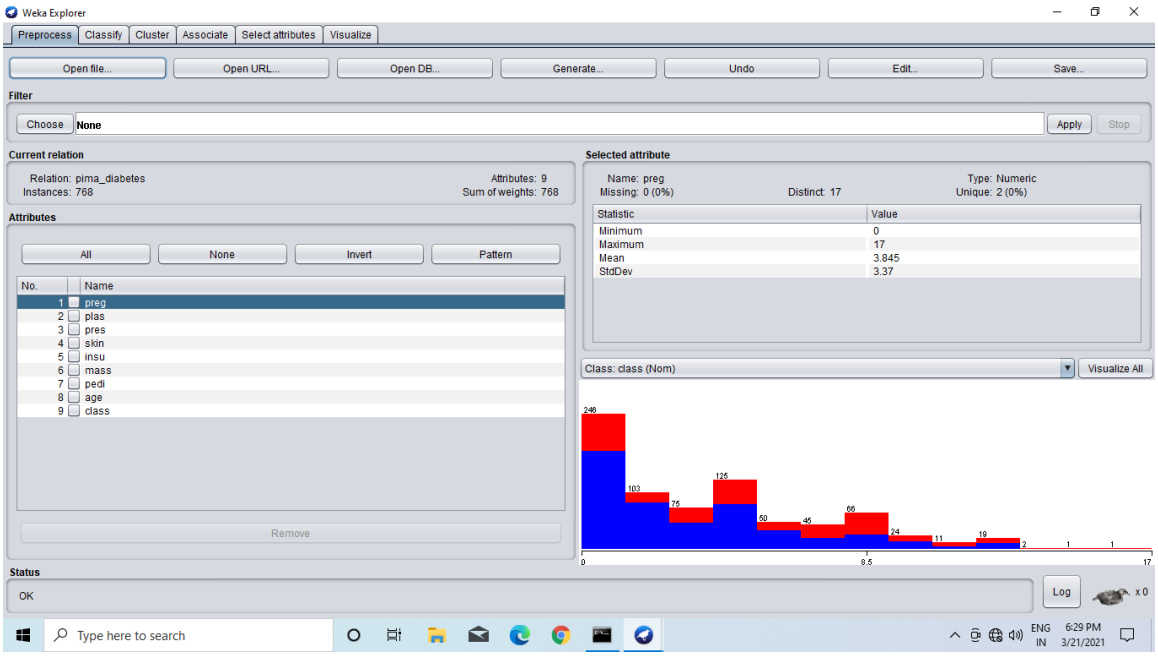
Practical 8

❖ **AIM**: to apply prediction models on various datasets and compare the accuracy of various algorithms:

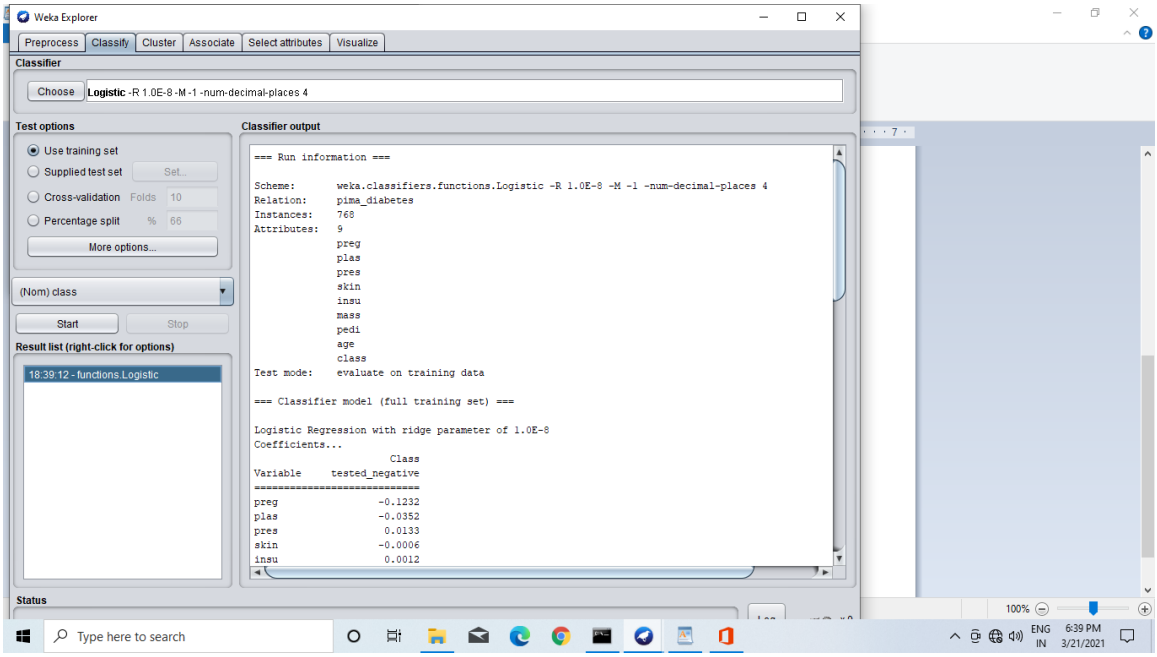
1) Dataset: Diabetes model: logistic

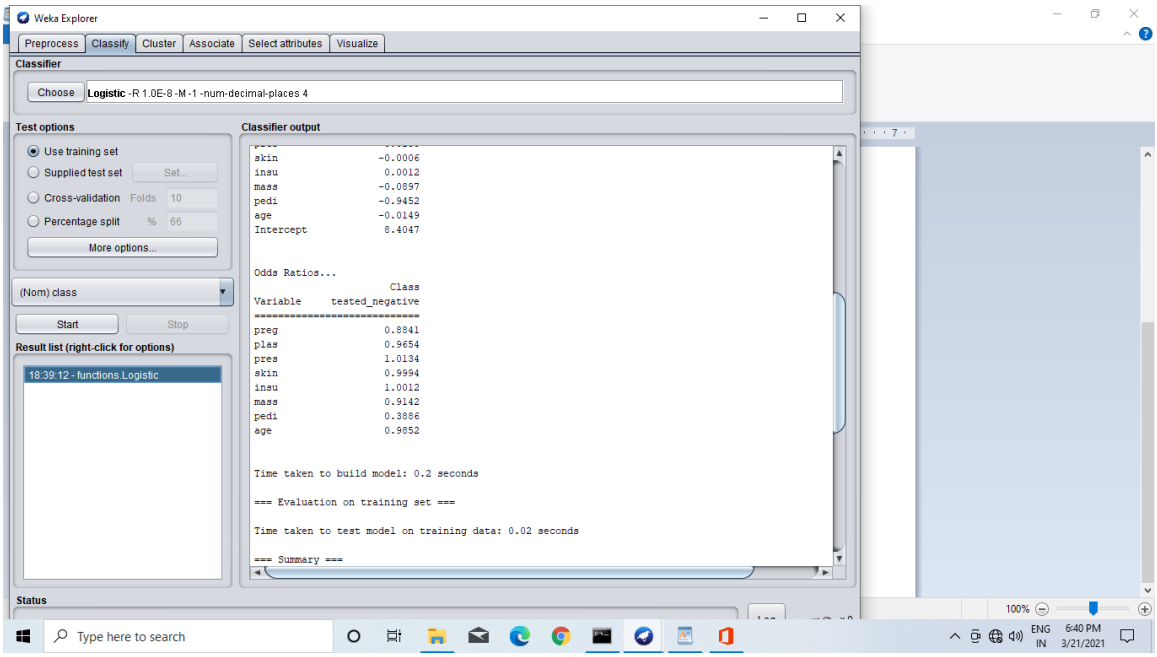
2) Dataset: Credit-g model: logistic

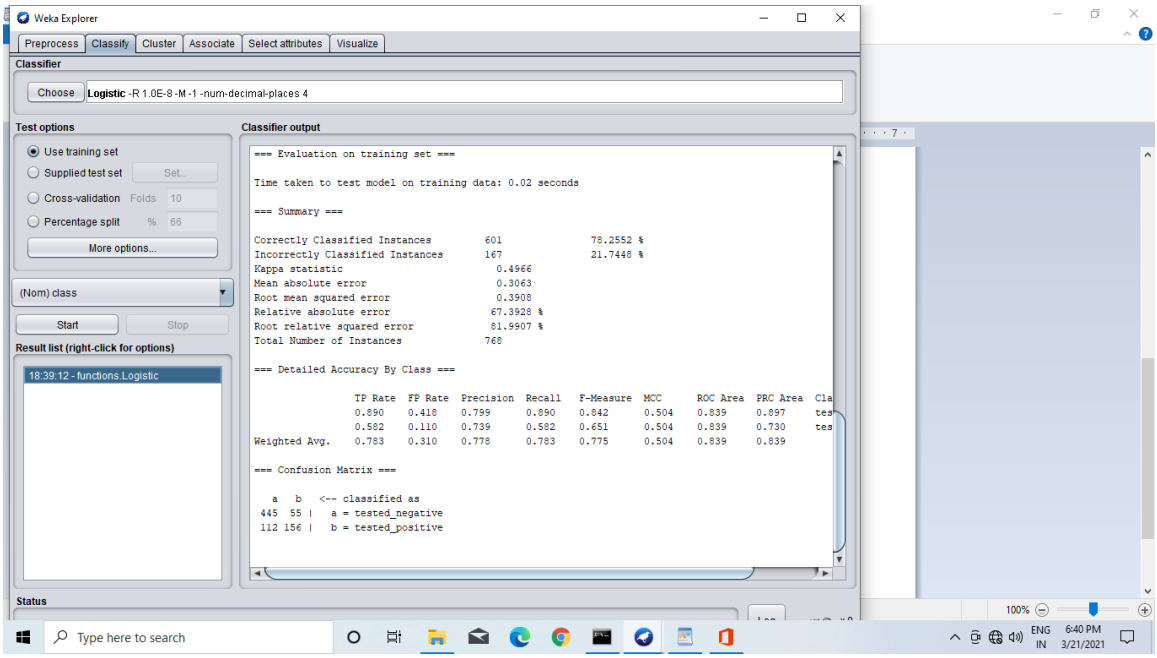
❖ Dataset: Diabetes



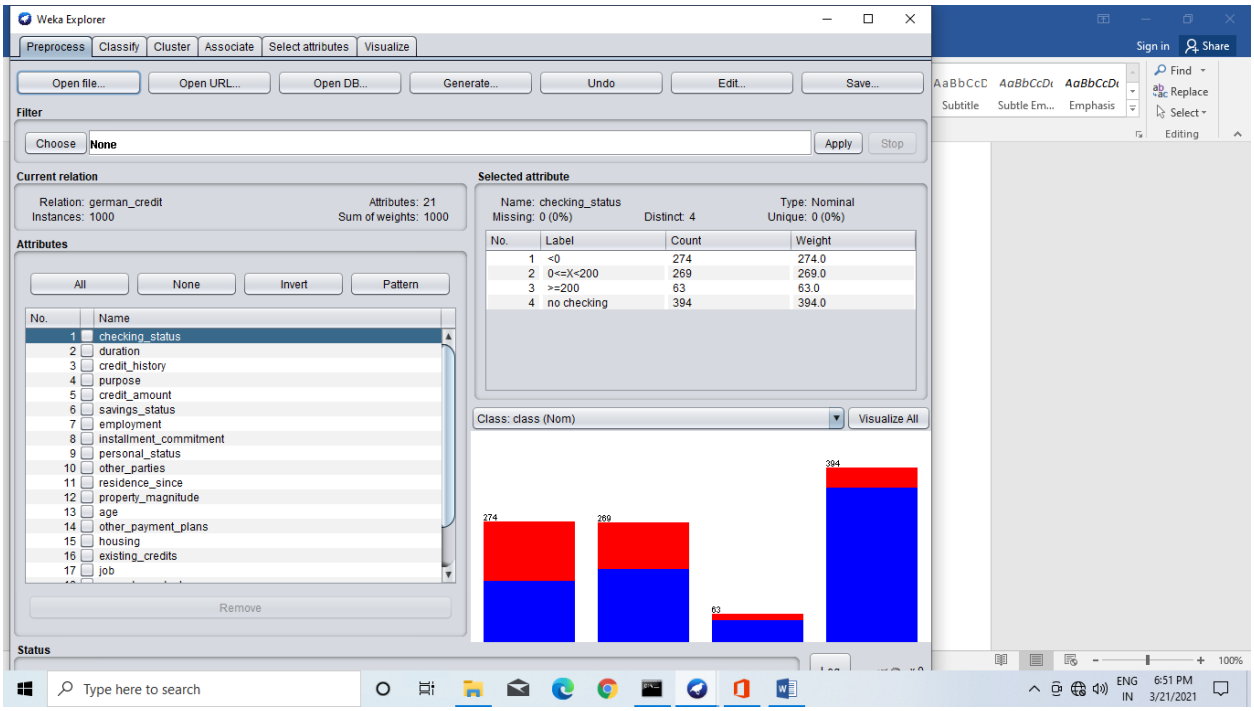
❖ Model: logistic



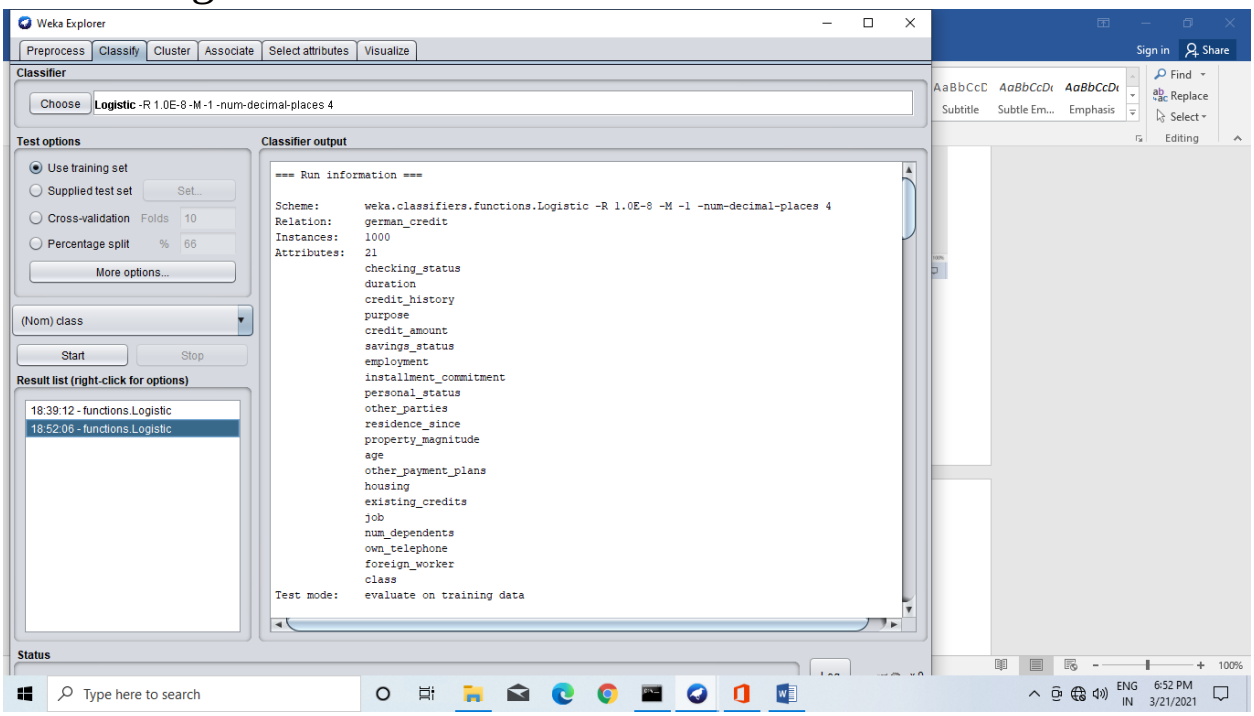


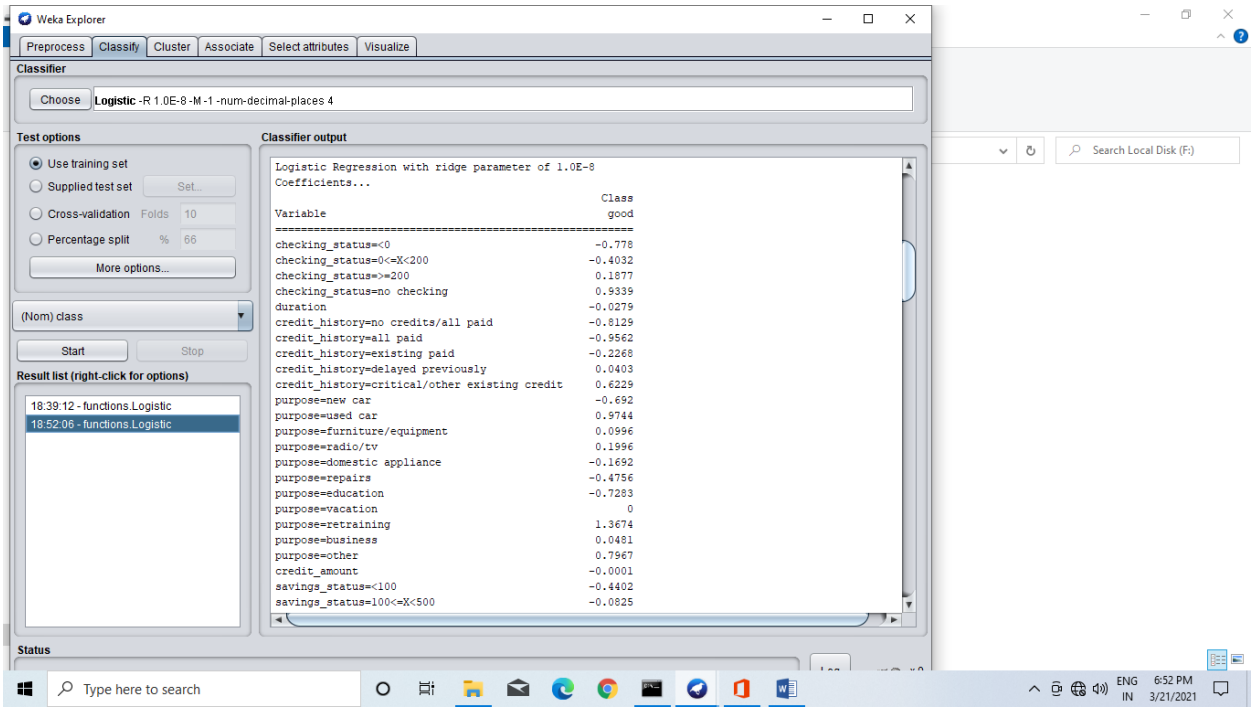


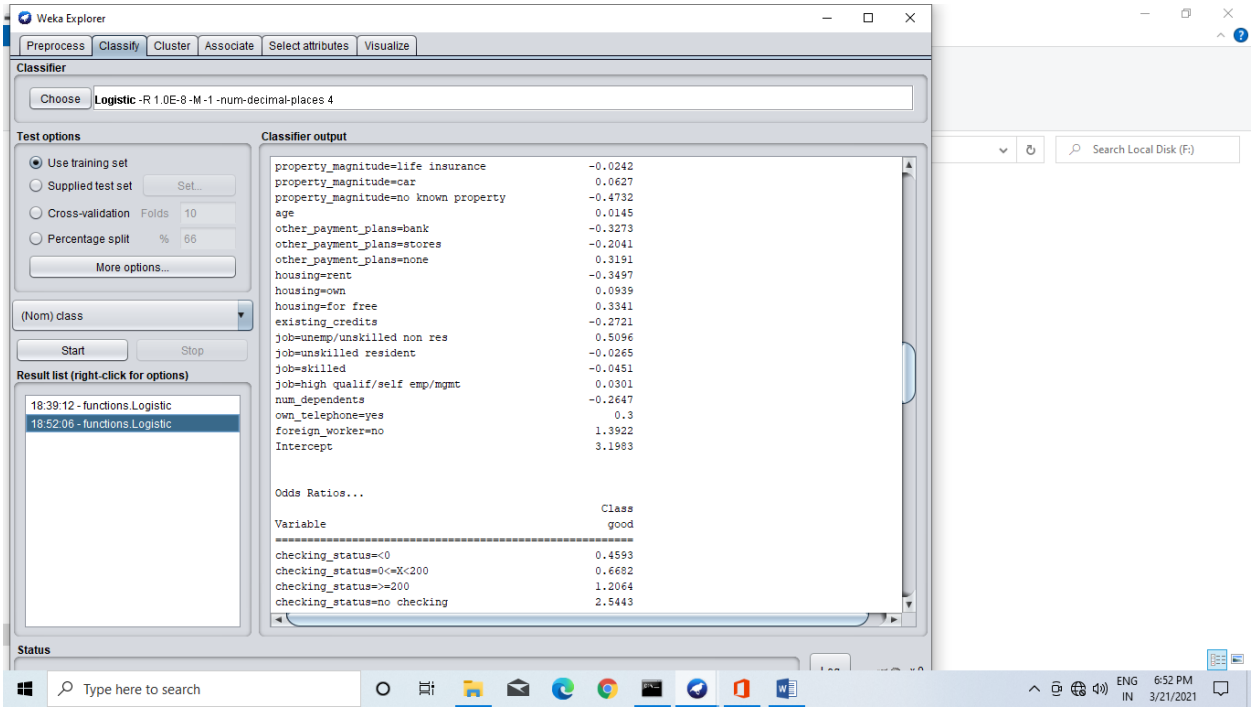
❖ Dataset: Credit-g

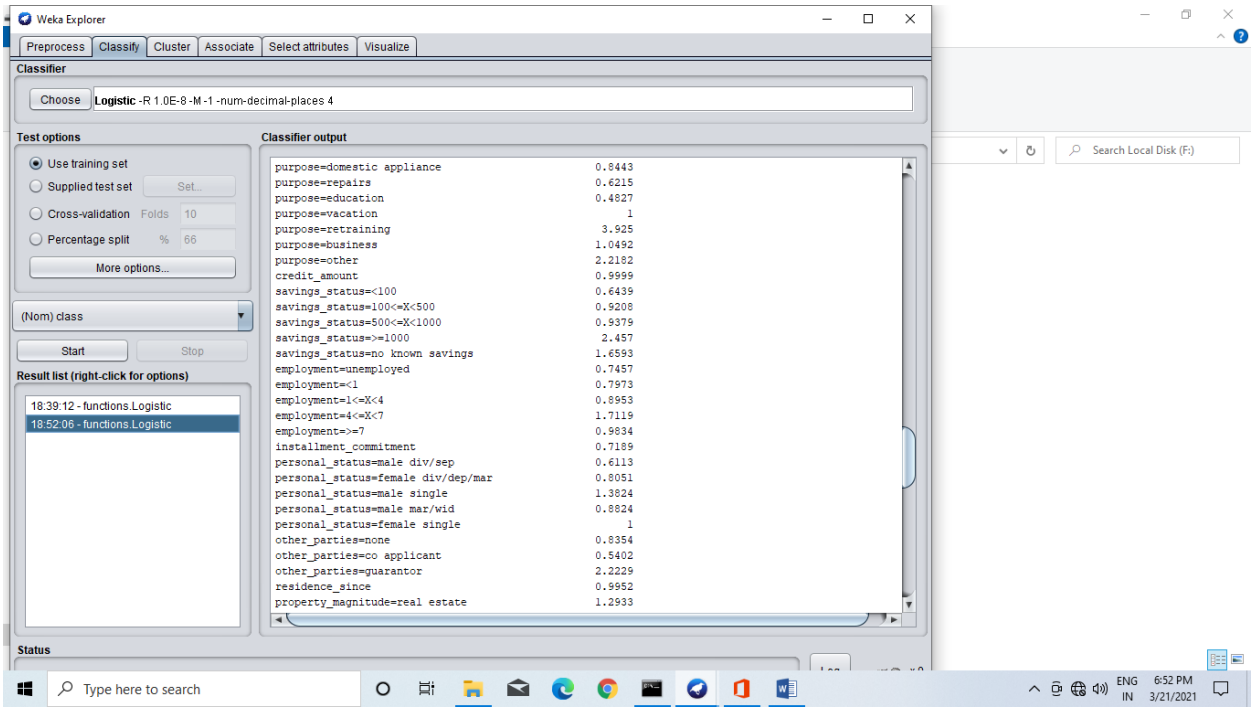


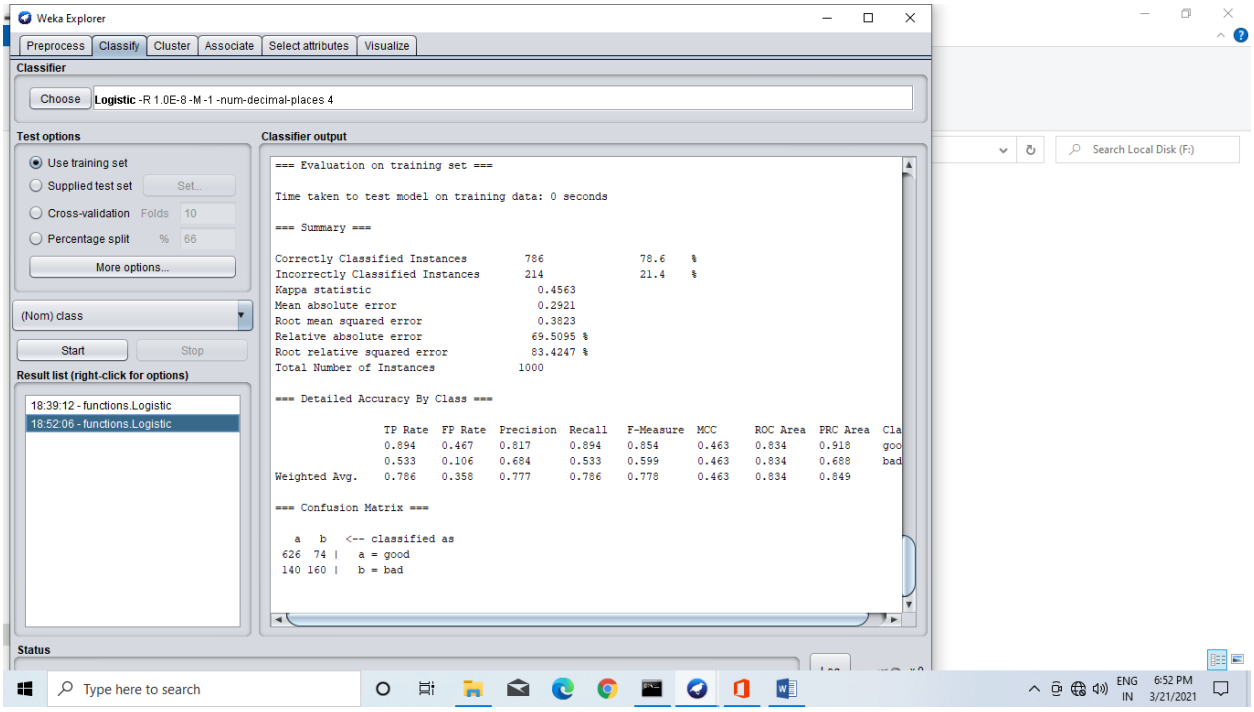
❖ Model: logistic




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
18:39:12 - functions.Logistic
18:52:06 - functions.Logistic
Classifier output
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
Class
Variable good
checking_status=<0 -0.778
checking_status=0<=X<200 -0.4032
checking_status=>=200 0.1877
checking_status=no checking 0.9339
duration -0.0279
credit_history=no credits/all paid -0.8129
credit_history=all paid -0.9562
credit_history=existing paid -0.2268
credit_history=delayed previously 0.0403
credit_history=critical/other existing credit 0.6229
purpose=new car -0.692
purpose=used car 0.9744
purpose=furniture/equipment 0.0996
purpose=radio/tv 0.1996
purpose=domestic appliance -0.1692
purpose=repairs -0.4756
purpose=education -0.7283
purpose=vacation 0
purpose=retraining 1.3674
purpose=business 0.0481
purpose=other 0.7967
credit_amount -0.0001
savings_status=<100 -0.4402
savings_status=100<=X<500 -0.0825
Status
Type here to search
ENG
IN
6:52 PM
3/21/2021


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
18:39:12 - functions.Logistic
18:52:06 - functions.Logistic
Classifier output
property_magnitude=life insurance -0.0242
property_magnitude=car 0.0627
property_magnitude=no known property -0.4732
age 0.0145
other_payment_plans=bank -0.3273
other_payment_plans=stores -0.2041
other_payment_plans=none 0.3191
housing=rent -0.3497
housing=own 0.0939
housing=for free 0.3341
existing_credits -0.2721
job=unemp/unskilled non res 0.5096
job=unskilled resident -0.0265
job=skilled -0.0451
job=high qualif/self emp/mgmt 0.0301
num_dependents -0.2647
own_telephone=yes 0.3
foreign_worker=no 1.3922
Intercept 3.1983
Odds Ratios...
Class
Variable good
checking_status=<0 0.4593
checking_status=0<=X<200 0.6682
checking_status=>=200 1.2064
checking_status=no checking 2.5443
Status
Type here to search
ENG
IN
6:52 PM
3/21/2021


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
18:39:12 - functions.Logistic
18:52:06 - functions.Logistic
Classifier output
purpose=domestic appliance 0.8443
purpose=repairs 0.6215
purpose=education 0.4827
purpose=vacation 1
purpose=retraining 3.925
purpose=business 1.0492
purpose=other 2.2182
credit_amount 0.9999
savings_status=<100 0.6439
savings_status=100<=X<500 0.9208
savings_status=500<=X<1000 0.9379
savings_status=>=1000 2.457
savings_status=no known savings 1.6593
employment=unemployed 0.7457
employment=<1 0.7973
employment=1<=X<4 0.8953
employment=4<=X<7 1.7119
employment=>=7 0.9834
installment_commitment 0.7189
personal_status=male div/sep 0.6113
personal_status=female div/dep/mar 0.8051
personal_status=male single 1.3824
personal_status=male mar/wid 0.8824
personal_status=female single 1
other_parties=none 0.8354
other_parties=co applicant 0.5402
other_parties=guarantor 2.2229
residence_since 0.9952
property_magnitude=real estate 1.2933
Status
Search Local Disk (F:)
Type here to search
ENG
IN
6:52 PM
3/21/2021


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
18:39:12 - functions.Logistic
18:52:06 - functions.Logistic
Classifier output
=== Evaluation on training set ===
Time taken to test model on training data: 0 seconds
=== Summary ===
Correctly Classified Instances 786 78.6 %
Incorrectly Classified Instances 214 21.4 %
Kappa statistic 0.4563
Mean absolute error 0.2921
Root mean squared error 0.3823
Relative absolute error 69.5095 %
Root relative squared error 83.4247 %
Total Number of Instances 1000
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.894 0.467 0.817 0.894 0.854 0.463 0.834 0.918 goo
0.533 0.106 0.684 0.533 0.599 0.463 0.834 0.688 bad
Weighted Avg. 0.786 0.358 0.777 0.786 0.778 0.463 0.834 0.849
=== Confusion Matrix ===
a b <-- classified as
626 74 | a = good
140 160 | b = bad
Status
Search Local Disk (F:)
Type here to search
ENG
IN
6:52 PM
3/21/2021

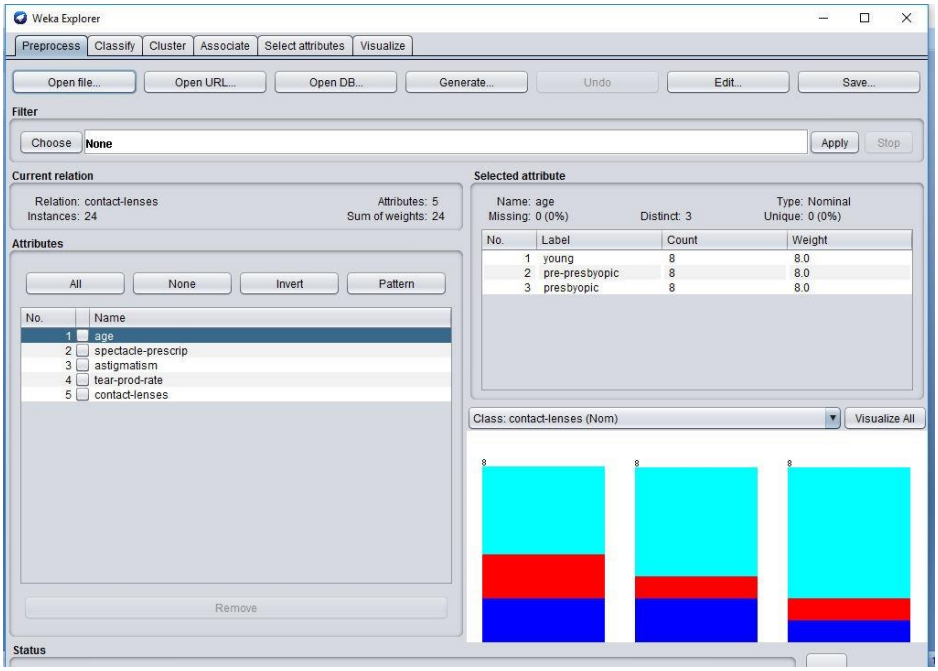
Practical 9(1)

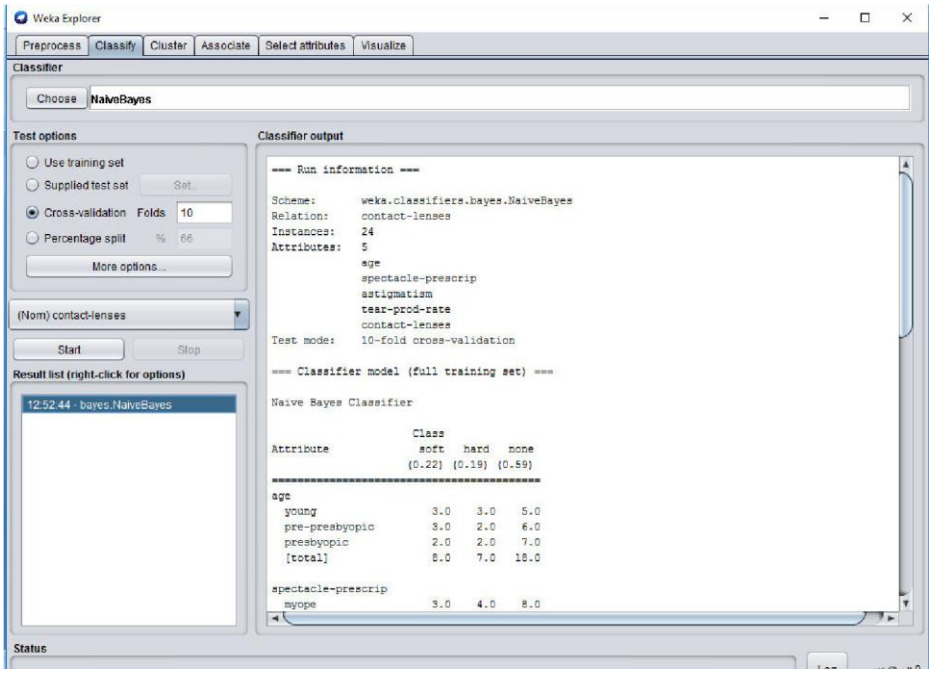
Aim: To perform classification on various datasets and the naive base algorithm.

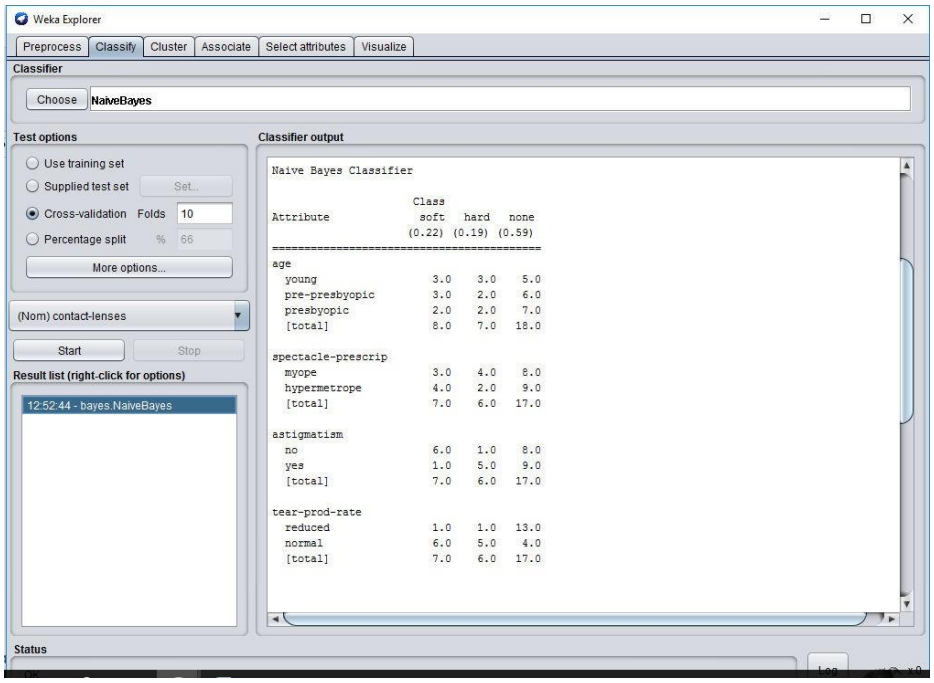
9-1-Naive Bayes :-

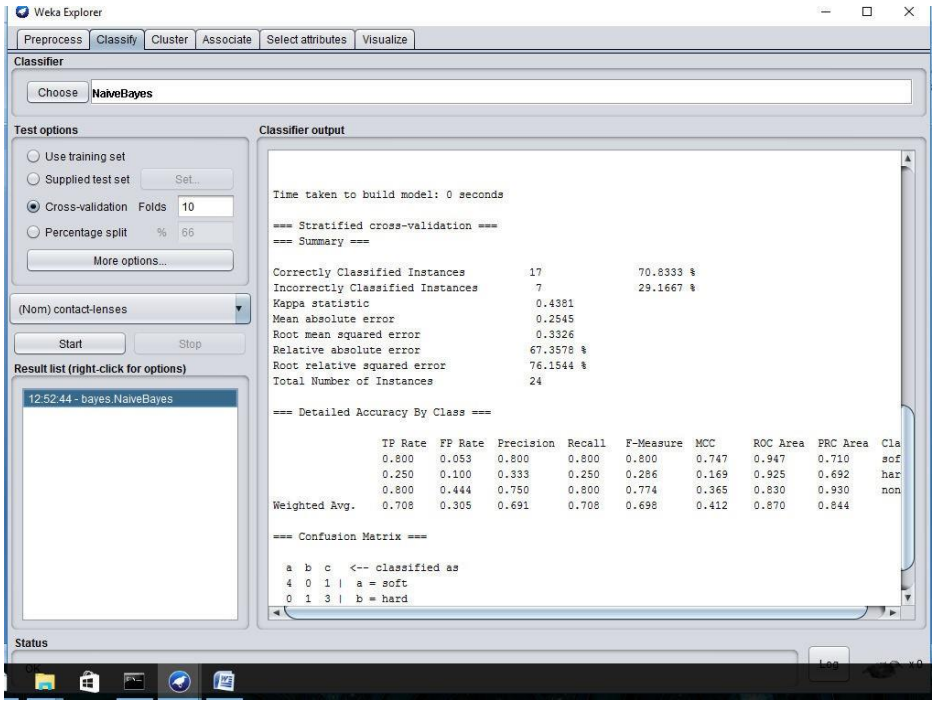
1. DataSet:-contact-lenses
2. DataSet:-soybean
3. DataSet:-Glass
4. DataSet:-segment-challenge
5. Dataset:-weather Nominal

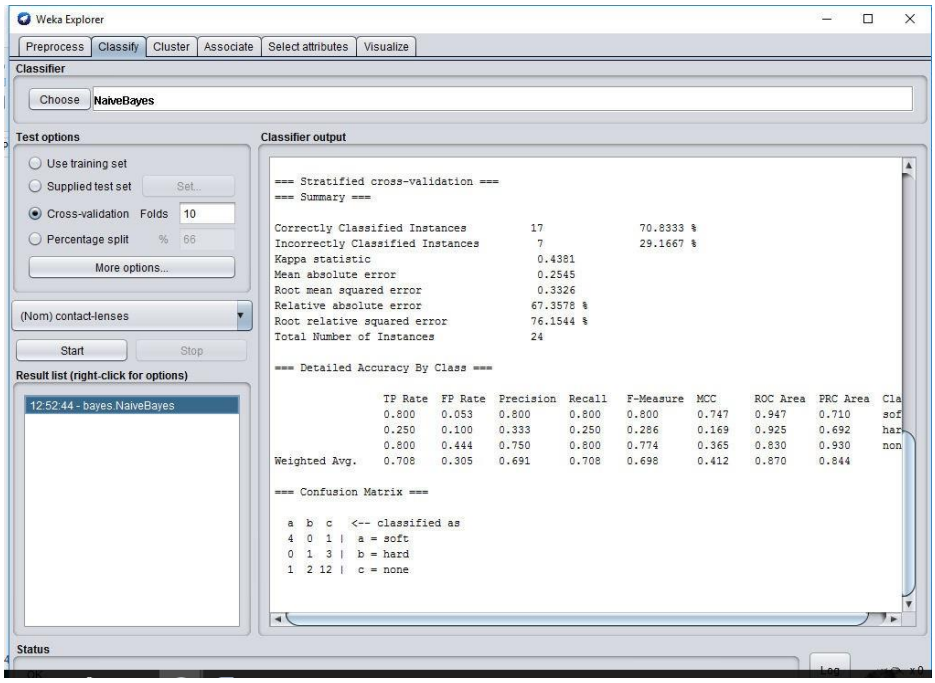
❖ Dataset 1: contact-lenses


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: contact-lenses
Instances: 24
Attributes: 5
Sum of weights: 24
Selected attribute
Name: age
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)
No. Label Count Weight
1 young 8 8.0
2 pre-presbyopic 8 8.0
3 presbyopic 8 8.0
Attributes
All
None
Invert
Pattern
No. Name
1 age
2 spectacle-prescrip
3 astigmatism
4 tear-prod-rate
5 contact-lenses
Remove
Class: contact-lenses (Nom)
Visualize All
Status

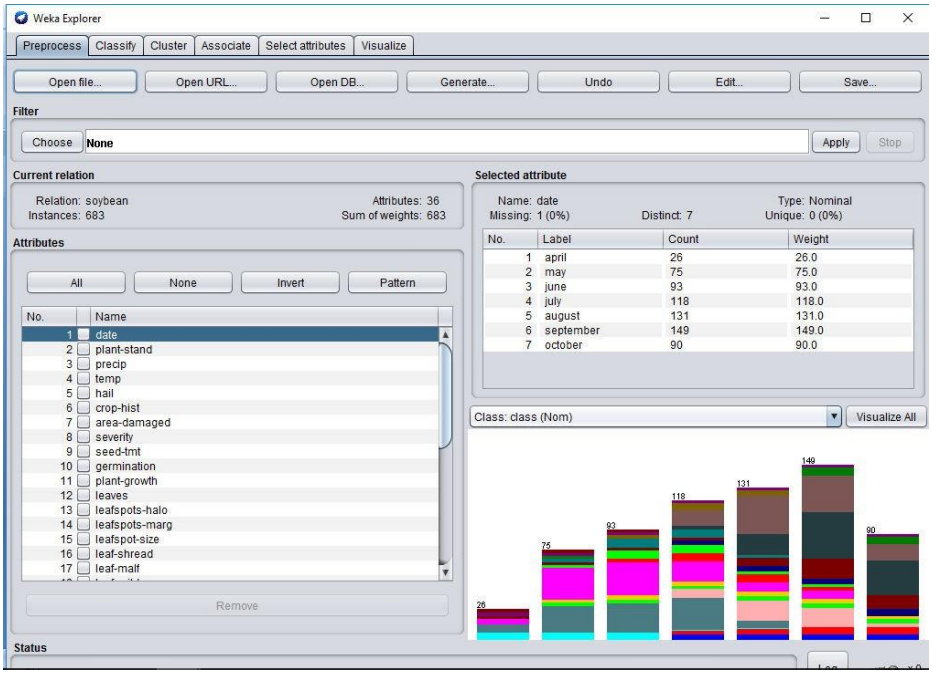

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) contact-lenses
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
Classifier output
=== Run information ===
Scheme: weka.classifiers.bayes.NaiveBayes
Relation: contact-lenses
Instances: 24
Attributes: 5
age
spectacle-prescrip
astigmatism
tear-prod-rate
contact-lenses
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class
Attribute soft hard none
(0.22) (0.19) (0.59)
age
young 3.0 3.0 5.0
pre-presbyopic 3.0 2.0 6.0
presbyopic 2.0 2.0 7.0
[total] 8.0 7.0 18.0
spectacle-prescrip
myope 3.0 4.0 8.0
Status

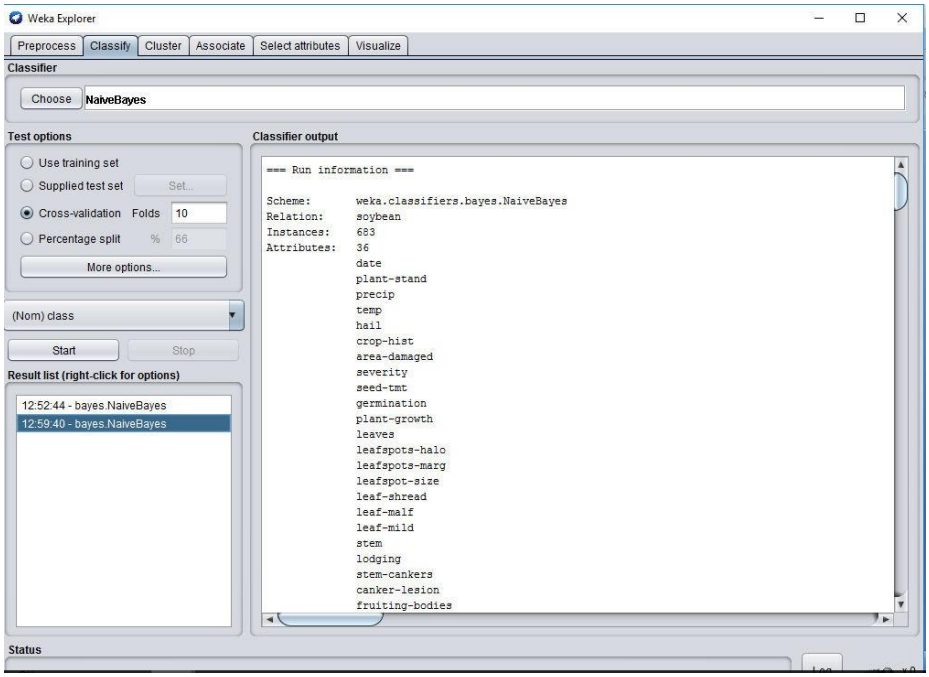

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) contact-lenses
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
Classifier output
Naive Bayes Classifier
Class
Attribute soft hard none
(0.22) (0.19) (0.59)
age
young 3.0 3.0 5.0
pre-presbyopic 3.0 2.0 6.0
presbyopic 2.0 2.0 7.0
[total] 8.0 7.0 18.0
spectacle-prescrip
myope 3.0 4.0 8.0
hypermetrope 4.0 2.0 9.0
[total] 7.0 6.0 17.0
astigmatism
no 6.0 1.0 8.0
yes 1.0 5.0 9.0
[total] 7.0 6.0 17.0
tear-prod-rate
reduced 1.0 1.0 13.0
normal 6.0 5.0 4.0
[total] 7.0 6.0 17.0
Status

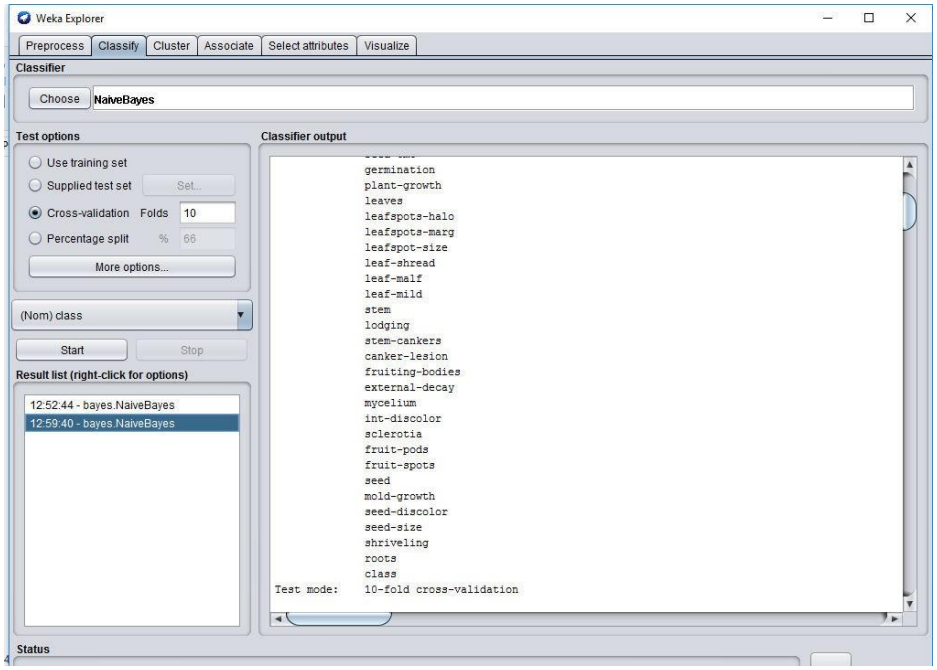

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) contact-lenses
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
Classifier output
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 17 70.8333 %
Incorrectly Classified Instances 7 29.1667 %
Kappa statistic 0.4381
Mean absolute error 0.2545
Root mean squared error 0.3326
Relative absolute error 67.3578 %
Root relative squared error 76.1544 %
Total Number of Instances 24
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.800 0.053 0.800 0.800 0.800 0.747 0.947 0.710 sof
0.250 0.100 0.333 0.250 0.286 0.169 0.925 0.692 har
0.800 0.444 0.750 0.800 0.774 0.365 0.830 0.930 non
Weighted Avg. 0.708 0.305 0.691 0.708 0.698 0.412 0.870 0.844
=== Confusion Matrix ===
a b c <-- classified as
4 0 1 | a = soft
0 1 3 | b = hard
Status

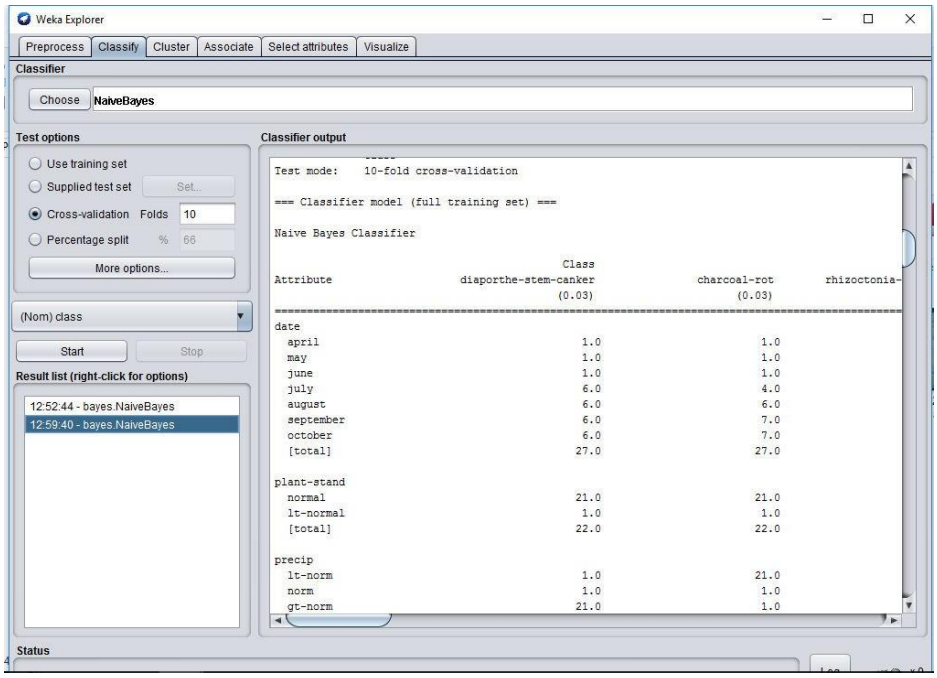


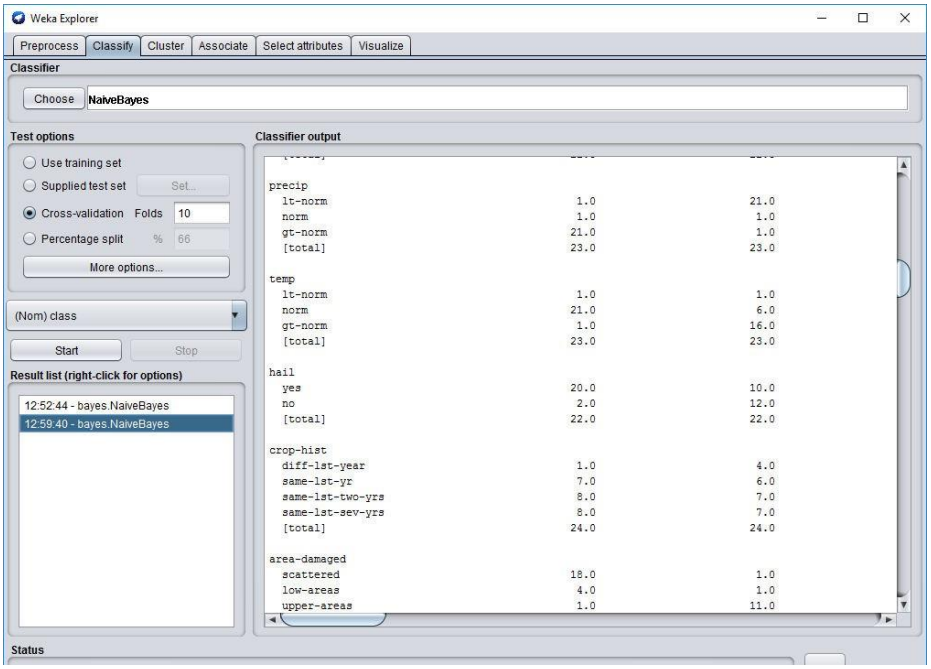
❖ Dataset 2 : soybean

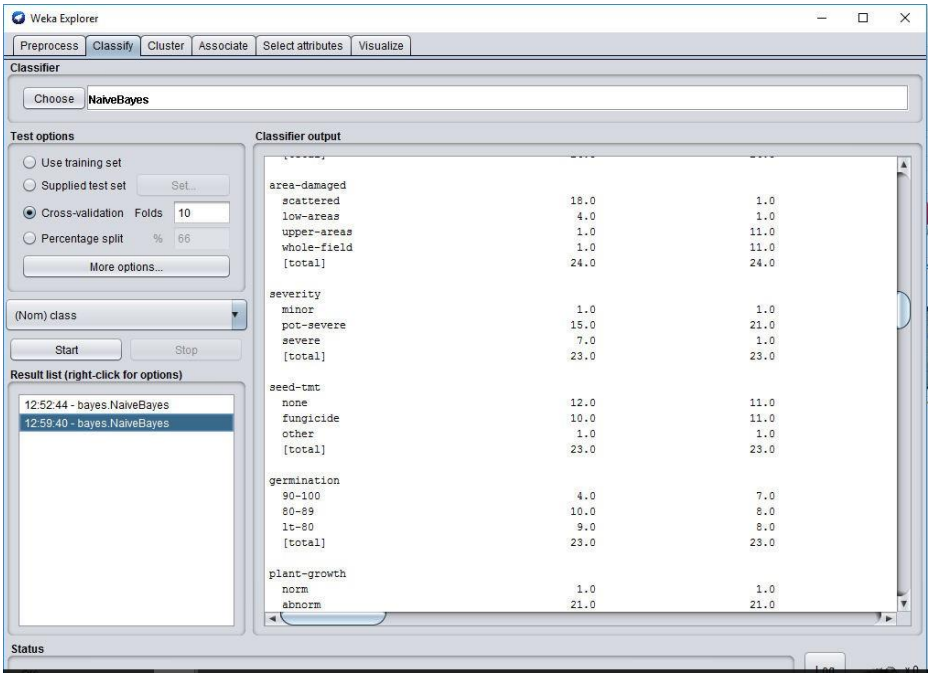


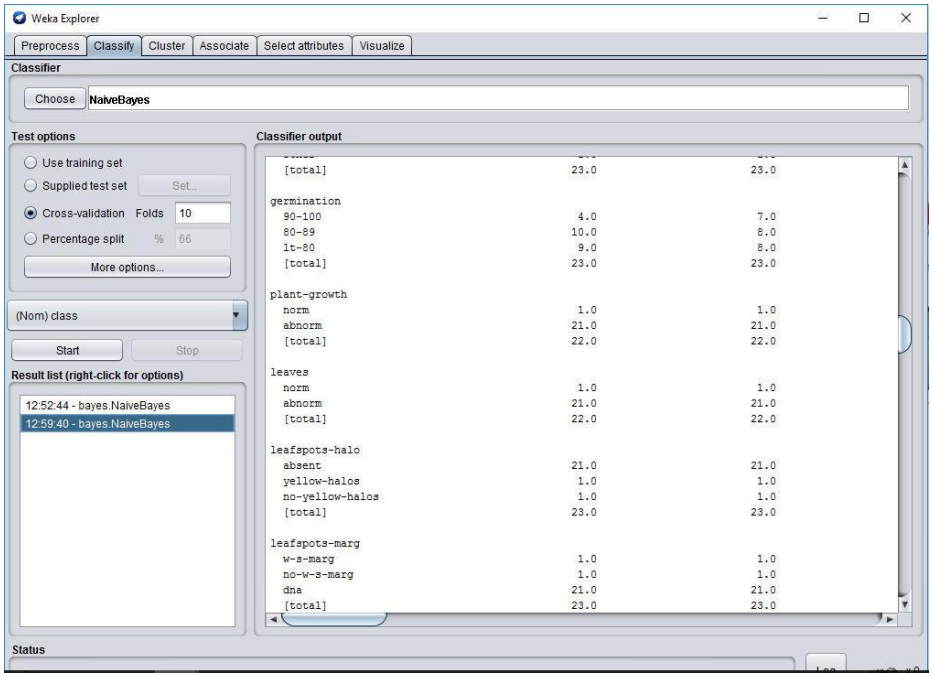

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
Classifier output
=== Run information ===
Scheme: weka.classifiers.bayes.NaiveBayes
Relation: soybean
Instances: 683
Attributes: 36
date
plant-stand
precip
temp
hail
crop-hist
area-damaged
severity
seed-tmt
germination
plant-growth
leaves
leafspots-halo
leafspots-marg
leafspot-size
leaf-shread
leaf-malf
leaf-mild
stem
lodging
stem-cankers
canker-lesion
fruiting-bodies
Status

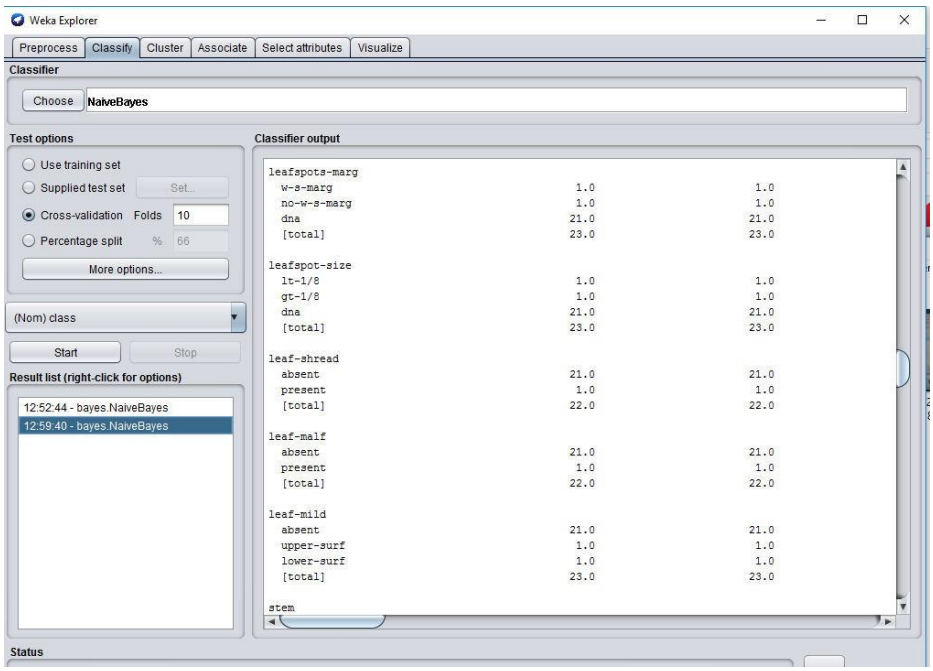

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
Classifier output
germination
plant-growth
leaves
leafspots-halo
leafspots-marg
leafspot-size
leaf-shread
leaf-malf
leaf-mild
stem
lodging
stem-cankers
canker-lesion
fruiting-bodies
external-decay
mycelium
int-discolor
sclerotia
fruit-pods
fruit-spots
seed
mold-growth
seed-discolor
seed-size
shriveling
roots
class
Test mode: 10-fold cross-validation
Status

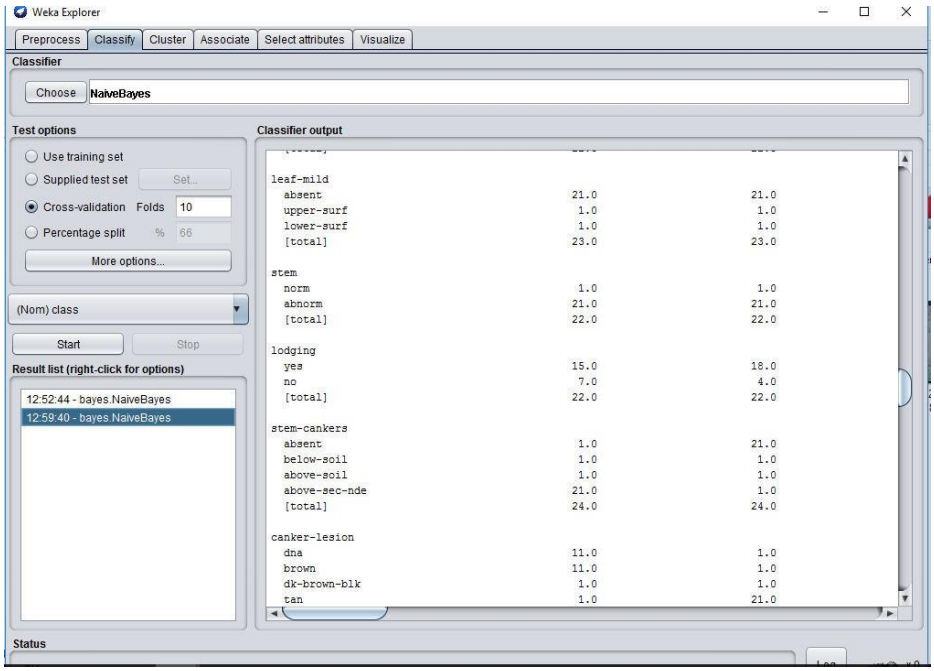

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
Classifier output
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class
Attribute diaporthe-stem-canker charcoal-rot rhizoctonia-
(0.03) (0.03)
date
april 1.0 1.0
may 1.0 1.0
june 1.0 1.0
july 6.0 4.0
august 6.0 6.0
september 6.0 7.0
october 6.0 7.0
[total] 27.0 27.0
plant-stand
normal 21.0 21.0
lt-normal 1.0 1.0
[total] 22.0 22.0
precip
lt-norm 1.0 21.0
norm 1.0 1.0
gt-norm 21.0 1.0
Status

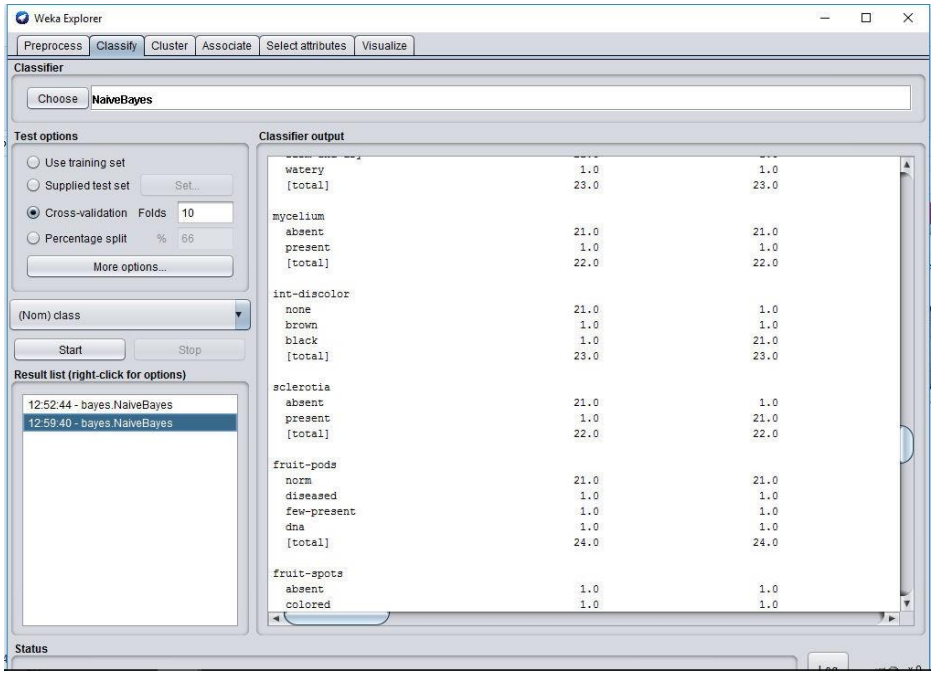

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
Classifier output
precip
lt-norm 1.0 21.0
norm 1.0 1.0
gt-norm 21.0 1.0
[total] 23.0 23.0
temp
lt-norm 1.0 1.0
norm 21.0 6.0
gt-norm 1.0 16.0
[total] 23.0 23.0
hail
yes 20.0 10.0
no 2.0 12.0
[total] 22.0 22.0
crop-hist
diff-lst-year 1.0 4.0
same-lst-yr 7.0 6.0
same-lst-two-yrs 8.0 7.0
same-lst-sev-yrs 8.0 7.0
[total] 24.0 24.0
area-damaged
scattered 18.0 1.0
low-areas 4.0 1.0
upper-areas 1.0 11.0
Status

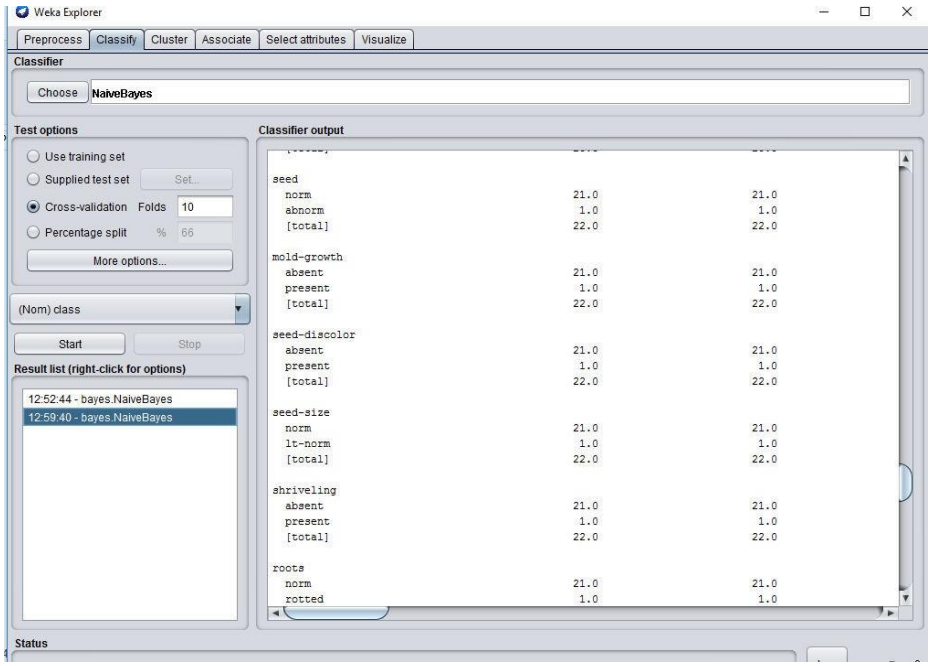

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
Classifier output
area-damaged
scattered 18.0 1.0
low-areas 4.0 1.0
upper-areas 1.0 11.0
whole-field 1.0 11.0
[total] 24.0 24.0
severity
minor 1.0 1.0
pot-severe 15.0 21.0
severe 7.0 1.0
[total] 23.0 23.0
seed-tmt
none 12.0 11.0
fungicide 10.0 11.0
other 1.0 1.0
[total] 23.0 23.0
germination
90-100 4.0 7.0
80-89 10.0 8.0
lt-80 9.0 8.0
[total] 23.0 23.0
plant-growth
norm 1.0 1.0
abnorm 21.0 21.0
Status

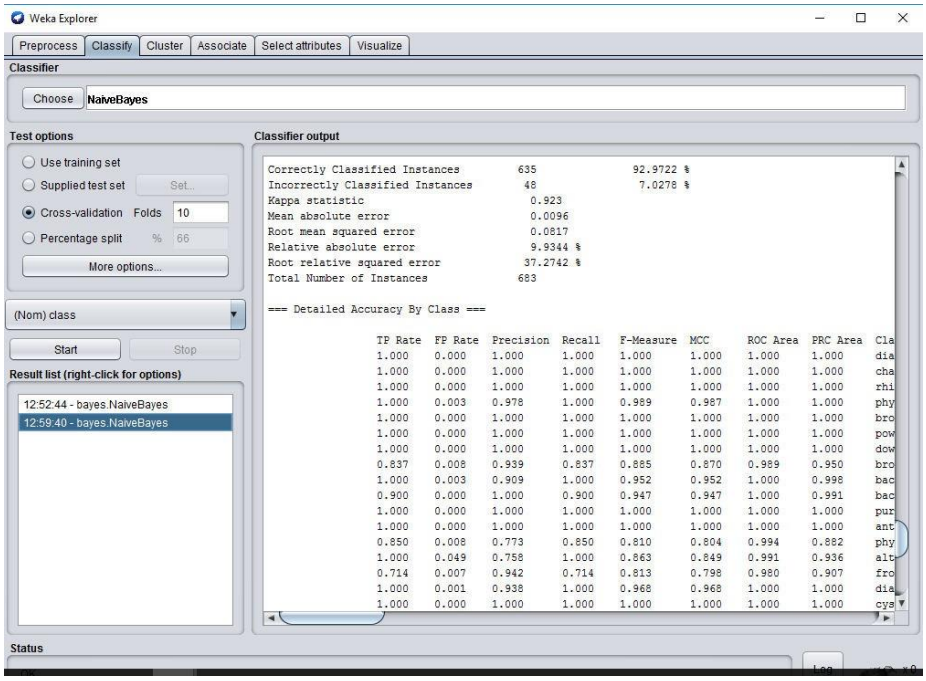

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
Classifier output
[total] 23.0 23.0
germination
90-100 4.0 7.0
80-89 10.0 8.0
lt-80 9.0 8.0
[total] 23.0 23.0
plant-growth
norm 1.0 1.0
abnorm 21.0 21.0
[total] 22.0 22.0
leaves
norm 1.0 1.0
abnorm 21.0 21.0
[total] 22.0 22.0
leafspots-halo
absent 21.0 21.0
yellow-halos 1.0 1.0
no-yellow-halos 1.0 1.0
[total] 23.0 23.0
leafspots-marg
w-s-marg 1.0 1.0
no-w-s-marg 1.0 1.0
dna 21.0 21.0
[total] 23.0 23.0
Status

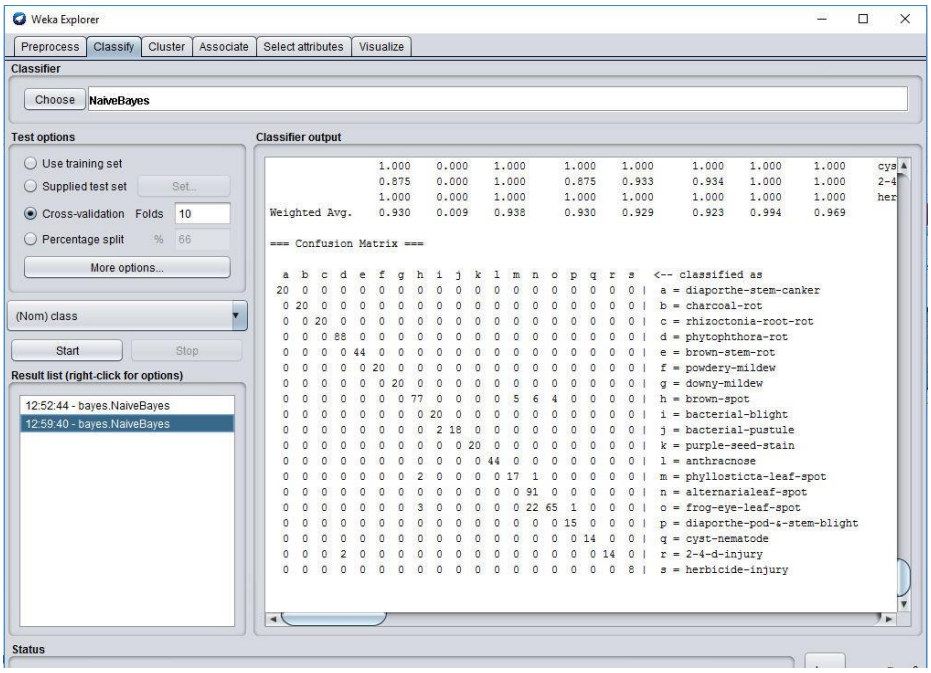

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
Classifier output
leafspots-marg
w-s-marg 1.0 1.0
no-w-s-marg 1.0 1.0
dna 21.0 21.0
[total] 23.0 23.0
leafspot-size
lt-1/8 1.0 1.0
gt-1/8 1.0 1.0
dna 21.0 21.0
[total] 23.0 23.0
leaf-shread
absent 21.0 21.0
present 1.0 1.0
[total] 22.0 22.0
leaf-malf
absent 21.0 21.0
present 1.0 1.0
[total] 22.0 22.0
leaf-mild
absent 21.0 21.0
upper-surf 1.0 1.0
lower-surf 1.0 1.0
[total] 23.0 23.0
stem
Status


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
Classifier output
leaf-mild
absent 21.0 21.0
upper-surf 1.0 1.0
lower-surf 1.0 1.0
[total] 23.0 23.0
stem
norm 1.0 1.0
abnorm 21.0 21.0
[total] 22.0 22.0
lodging
yes 15.0 18.0
no 7.0 4.0
[total] 22.0 22.0
stem-cankers
absent 1.0 21.0
below-soil 1.0 1.0
above-soil 1.0 1.0
above-sec-nde 21.0 1.0
[total] 24.0 24.0
canker-lesion
dna 11.0 1.0
brown 11.0 1.0
dk-brown-blk 1.0 1.0
tan 1.0 21.0
Status

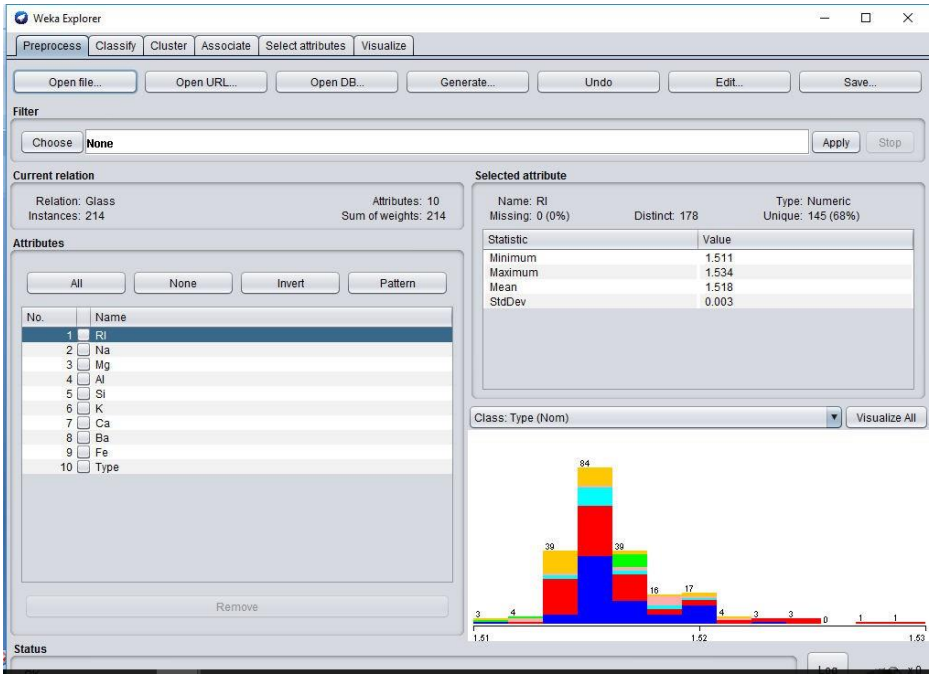

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
Classifier output
watery 1.0 1.0
[total] 23.0 23.0
mycelium
absent 21.0 21.0
present 1.0 1.0
[total] 22.0 22.0
int-discolor
none 21.0 1.0
brown 1.0 1.0
black 1.0 21.0
[total] 23.0 23.0
sclerotia
absent 21.0 1.0
present 1.0 21.0
[total] 22.0 22.0
fruit-pods
norm 21.0 21.0
diseased 1.0 1.0
few-present 1.0 1.0
dna 1.0 1.0
[total] 24.0 24.0
fruit-spots
absent 1.0 1.0
colored 1.0 1.0
Status

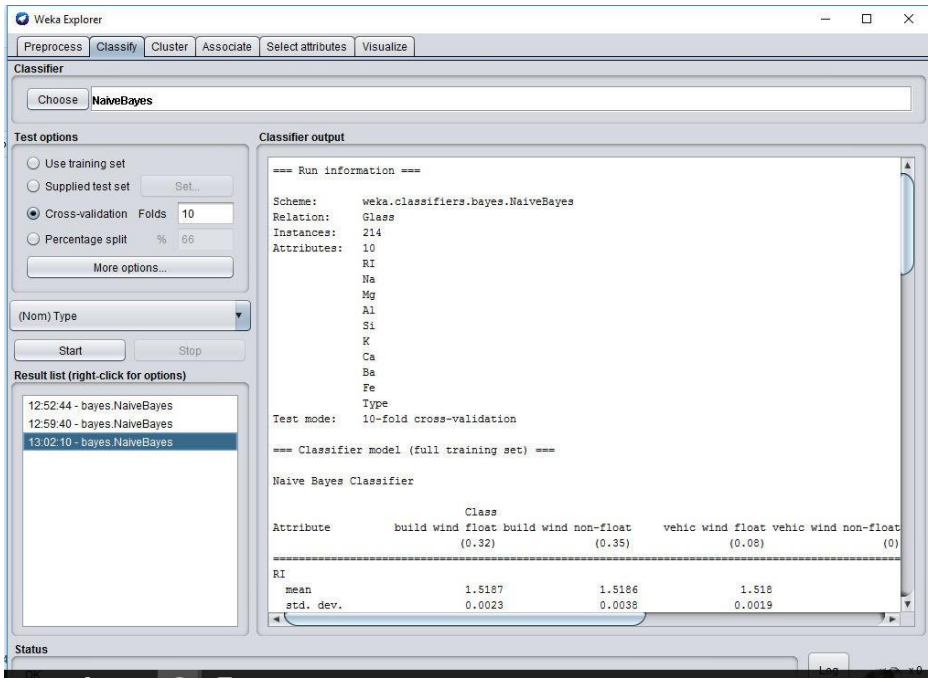

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
Classifier output
seed
norm 21.0 21.0
abnorm 1.0 1.0
[total] 22.0 22.0
mold-growth
absent 21.0 21.0
present 1.0 1.0
[total] 22.0 22.0
seed-discolor
absent 21.0 21.0
present 1.0 1.0
[total] 22.0 22.0
seed-size
norm 21.0 21.0
lt-norm 1.0 1.0
[total] 22.0 22.0
shriveling
absent 21.0 21.0
present 1.0 1.0
[total] 22.0 22.0
roots
norm 21.0 21.0
rotted 1.0 1.0
Status

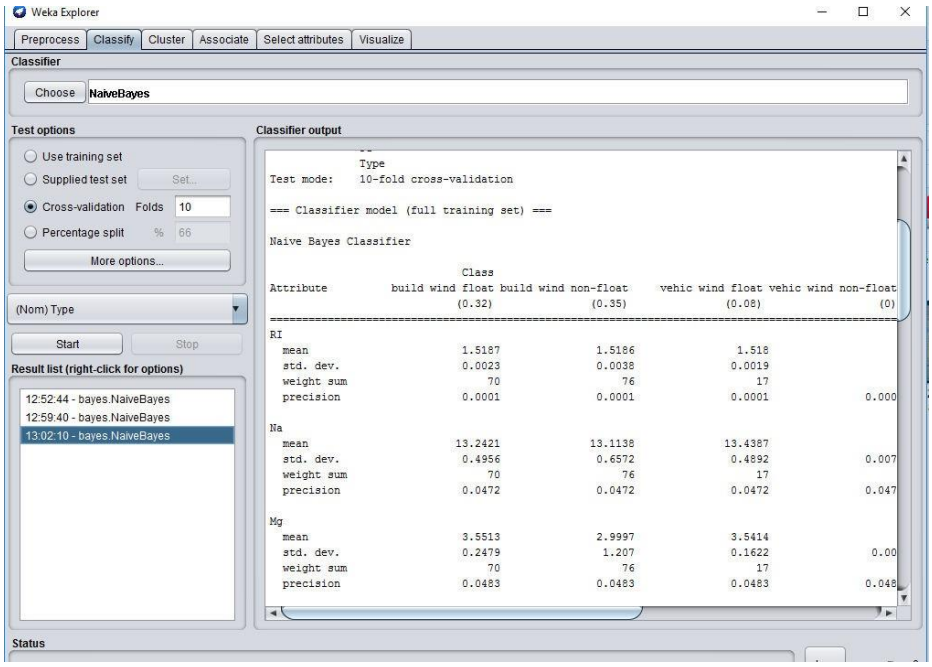

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
Classifier output
Correctly Classified Instances 635 92.9722 %
Incorrectly Classified Instances 48 7.0278 %
Kappa statistic 0.923
Mean absolute error 0.0096
Root mean squared error 0.0817
Relative absolute error 9.9344 %
Root relative squared error 37.2742 %
Total Number of Instances 683
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 dia
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 cha
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 rhi
1.000 0.003 0.978 1.000 0.989 0.987 1.000 1.000 phy
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 bro
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 pow
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 dow
0.837 0.008 0.939 0.837 0.885 0.870 0.989 0.950 bro
1.000 0.003 0.909 1.000 0.952 0.952 1.000 0.998 bac
0.900 0.000 1.000 0.900 0.947 0.947 1.000 0.991 bac
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 pur
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 ant
0.850 0.008 0.773 0.850 0.810 0.804 0.994 0.882 phy
1.000 0.049 0.758 1.000 0.863 0.849 0.991 0.936 alt
0.714 0.007 0.942 0.714 0.813 0.798 0.980 0.907 fro
1.000 0.001 0.938 1.000 0.968 0.968 1.000 1.000 dia
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 cys
Status

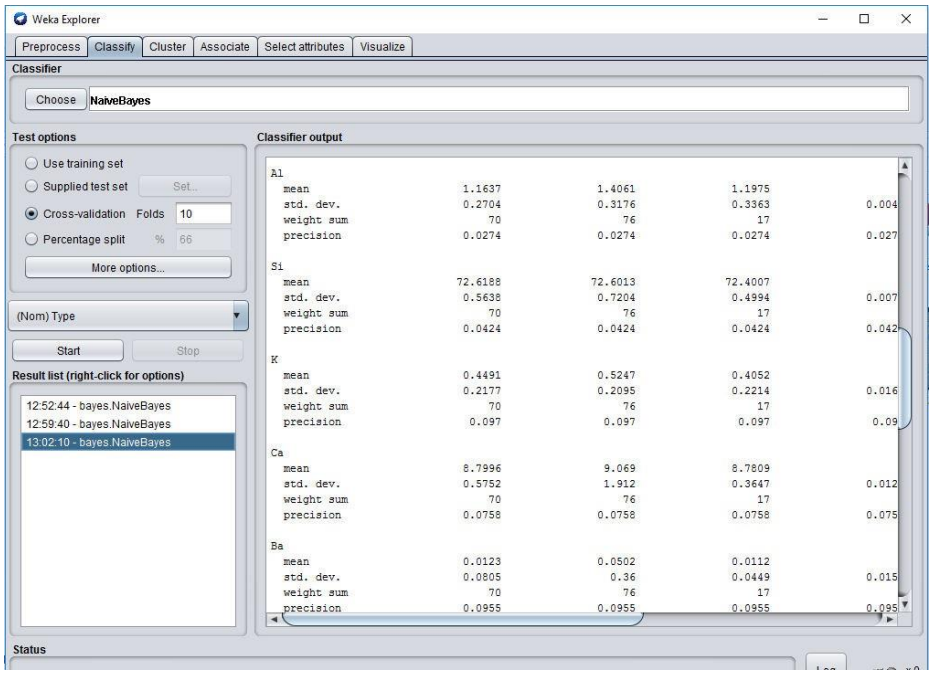


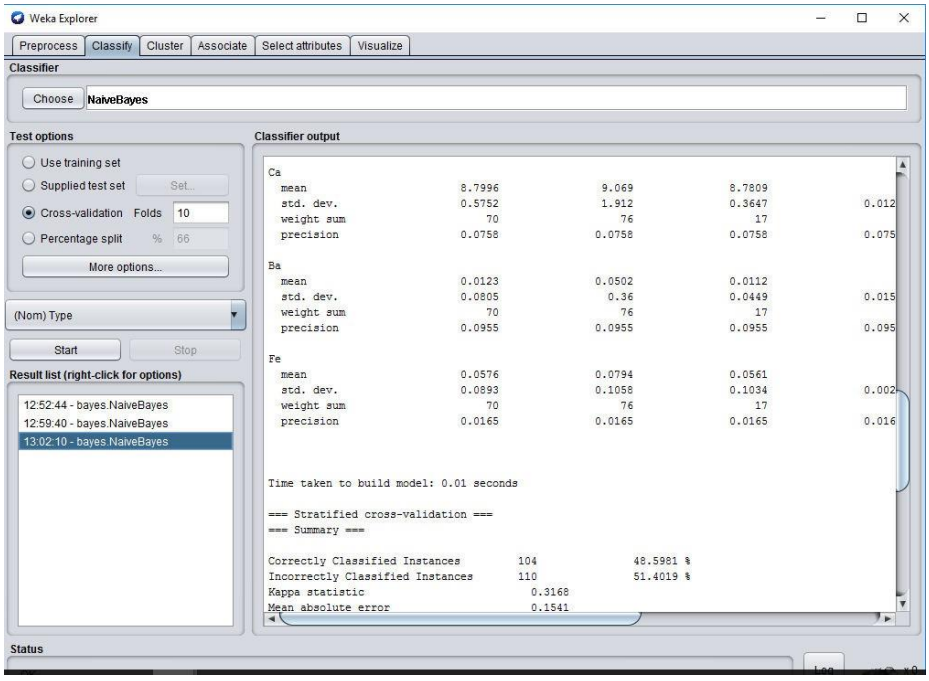
❖ Dataset 3: glass

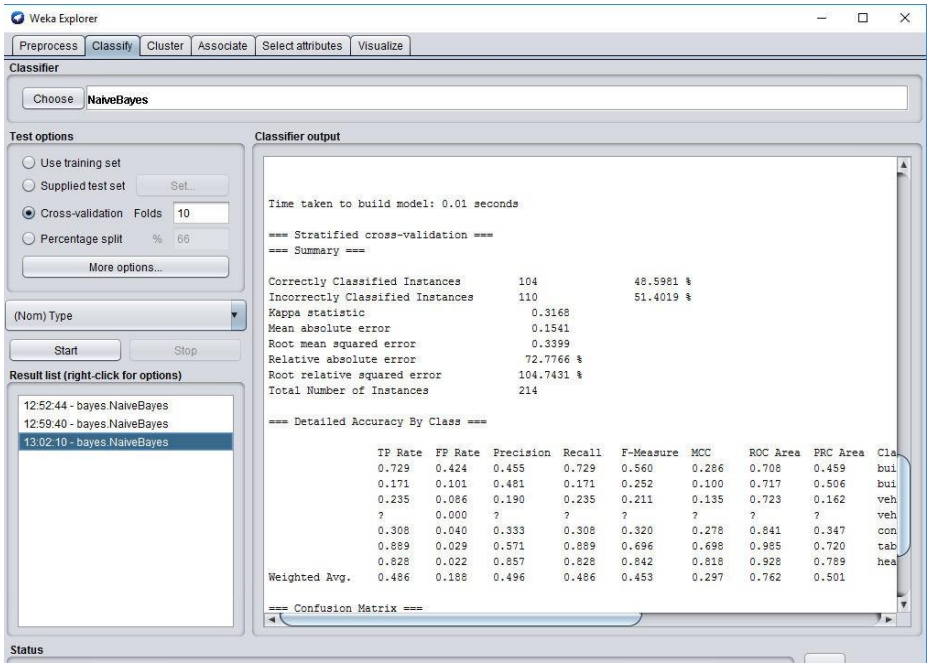


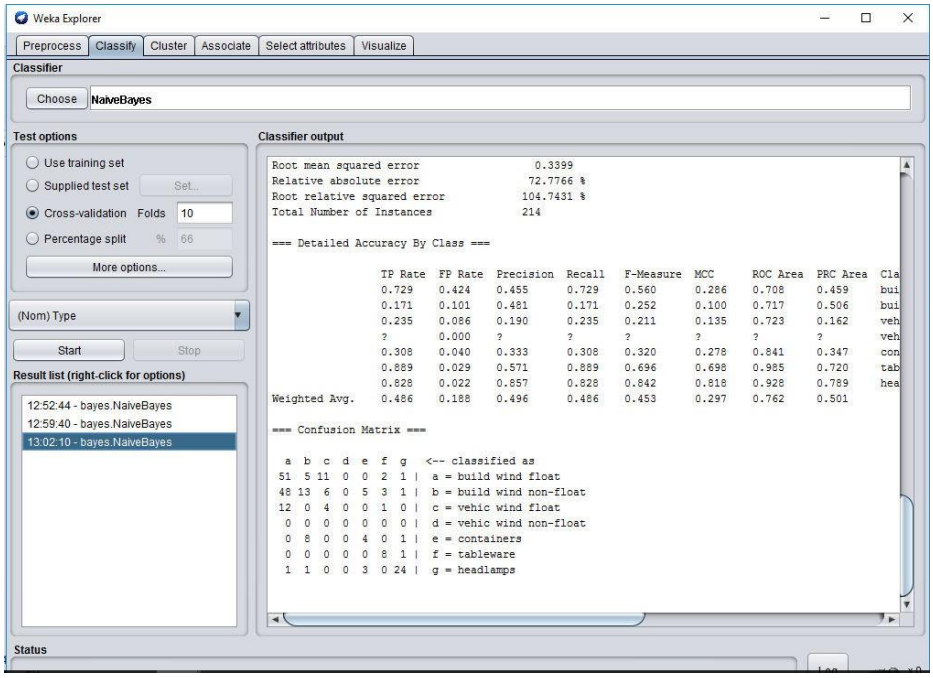

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
Classifier output
=== Run information ===
Scheme: weka.classifiers.bayes.NaiveBayes
Relation: Glass
Instances: 214
Attributes: 10
RI
Na
Mg
Al
Si
K
Ca
Ba
Fe
Type
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class
Attribute build wind float build wind non-float vehic wind float vehic wind non-float
(0.32) (0.35) (0.08) (0)
RI
mean 1.5187 1.5186 1.518
std. dev. 0.0023 0.0038 0.0019
Status


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
Classifier output
Type
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class
Attribute build wind float build wind non-float vehic wind float vehic wind non-float
(0.32) (0.35) (0.08) (0)
RI
mean 1.5187 1.5186 1.518
std. dev. 0.0023 0.0038 0.0019
weight sum 70 76 17
precision 0.0001 0.0001 0.0001 0.000
Na
mean 13.2421 13.1138 13.4387
std. dev. 0.4956 0.6572 0.4892 0.007
weight sum 70 76 17
precision 0.0472 0.0472 0.0472 0.047
Mg
mean 3.5513 2.9997 3.5414
std. dev. 0.2479 1.207 0.1622 0.00
weight sum 70 76 17
precision 0.0483 0.0483 0.0483 0.048
Status

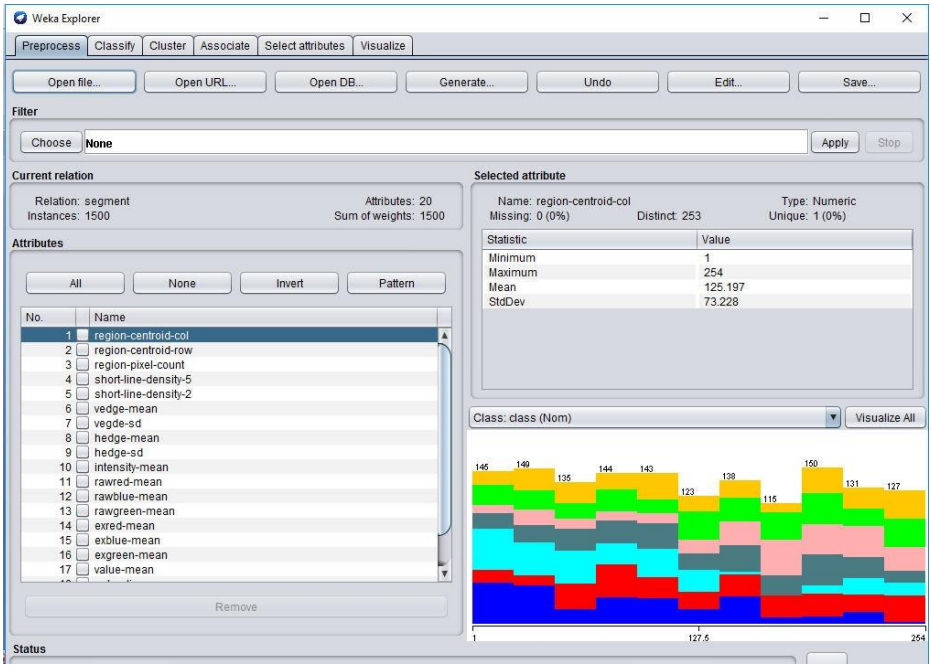

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
Classifier output
Al
mean 1.1637 1.4061 1.1975
std. dev. 0.2704 0.3176 0.3363 0.004
weight sum 70 76 17
precision 0.0274 0.0274 0.0274 0.027
Si
mean 72.6188 72.6013 72.4007
std. dev. 0.5638 0.7204 0.4994 0.007
weight sum 70 76 17
precision 0.0424 0.0424 0.0424 0.042
K
mean 0.4491 0.5247 0.4052
std. dev. 0.2177 0.2095 0.2214 0.016
weight sum 70 76 17
precision 0.097 0.097 0.097 0.09
Ca
mean 8.7996 9.069 8.7809
std. dev. 0.5752 1.912 0.3647 0.012
weight sum 70 76 17
precision 0.0758 0.0758 0.0758 0.075
Ba
mean 0.0123 0.0502 0.0112
std. dev. 0.0805 0.36 0.0449 0.015
weight sum 70 76 17
precision 0.0955 0.0955 0.0955 0.095
Status

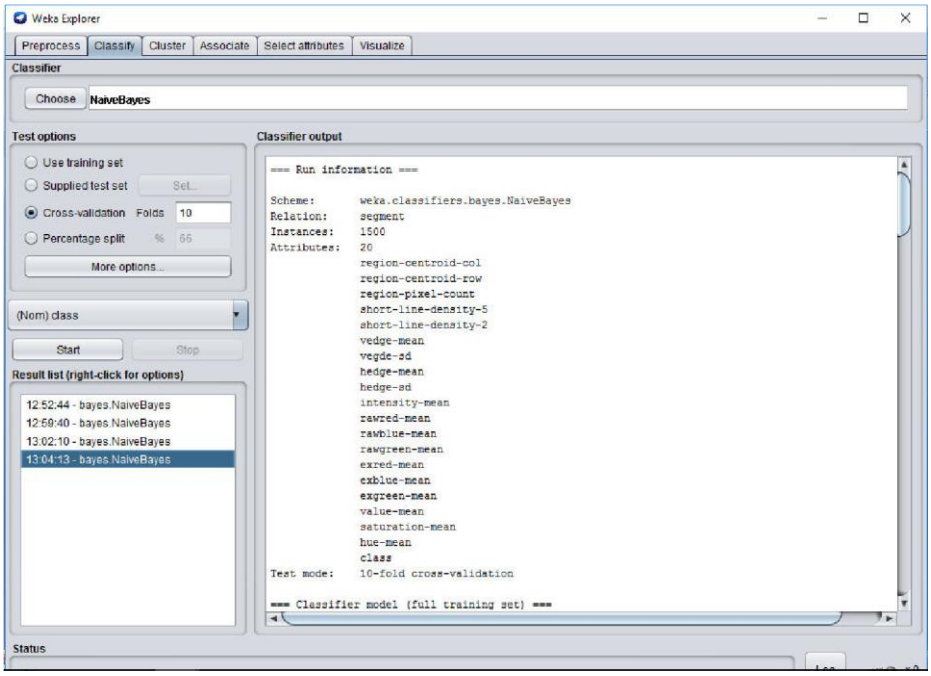

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
Classifier output
Ca
mean 8.7996 9.069 8.7809
std. dev. 0.5752 1.912 0.3647 0.012
weight sum 70 76 17
precision 0.0758 0.0758 0.0758 0.075
Ba
mean 0.0123 0.0502 0.0112
std. dev. 0.0805 0.36 0.0449 0.015
weight sum 70 76 17
precision 0.0955 0.0955 0.0955 0.095
Fe
mean 0.0576 0.0794 0.0561
std. dev. 0.0893 0.1058 0.1034 0.002
weight sum 70 76 17
precision 0.0165 0.0165 0.0165 0.016
Time taken to build model: 0.01 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 104 48.5981 %
Incorrectly Classified Instances 110 51.4019 %
Kappa statistic 0.3168
Mean absolute error 0.1541
Status

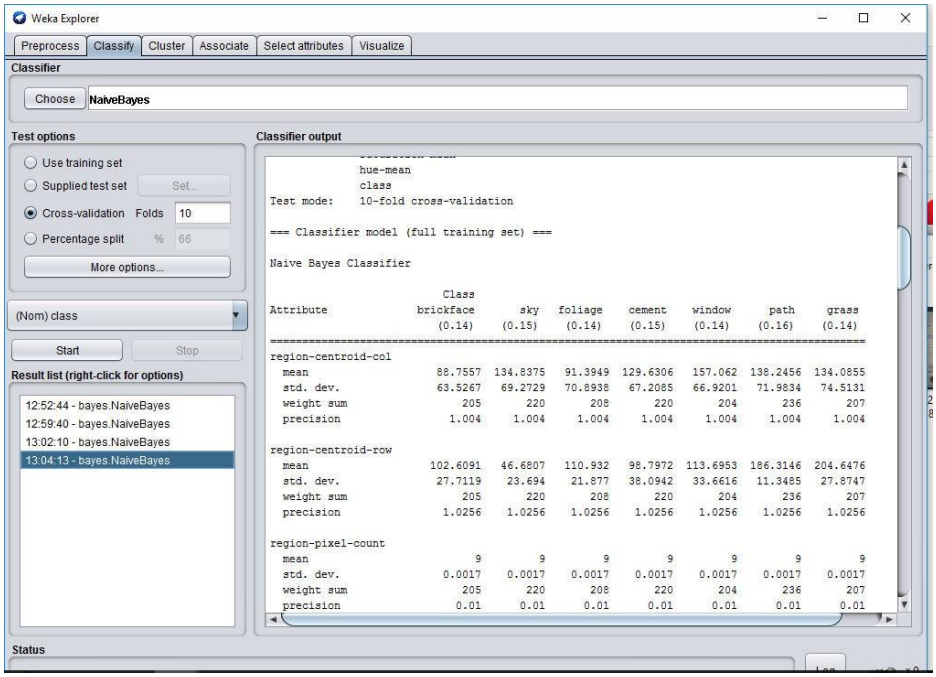

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
Classifier output
Time taken to build model: 0.01 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 104 48.5981 %
Incorrectly Classified Instances 110 51.4019 %
Kappa statistic 0.3168
Mean absolute error 0.1541
Root mean squared error 0.3399
Relative absolute error 72.7766 %
Root relative squared error 104.7431 %
Total Number of Instances 214
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.729 0.424 0.455 0.729 0.560 0.286 0.708 0.459 bui
0.171 0.101 0.481 0.171 0.252 0.100 0.717 0.506 bui
0.235 0.086 0.190 0.235 0.211 0.135 0.723 0.162 veh
? 0.000 ? ? ? ? ? ? veh
0.308 0.040 0.333 0.308 0.320 0.278 0.841 0.347 con
0.889 0.029 0.571 0.889 0.696 0.698 0.985 0.720 tab
0.828 0.022 0.857 0.828 0.842 0.818 0.928 0.789 hea
Weighted Avg. 0.486 0.188 0.496 0.486 0.453 0.297 0.762 0.501
=== Confusion Matrix ===
Status

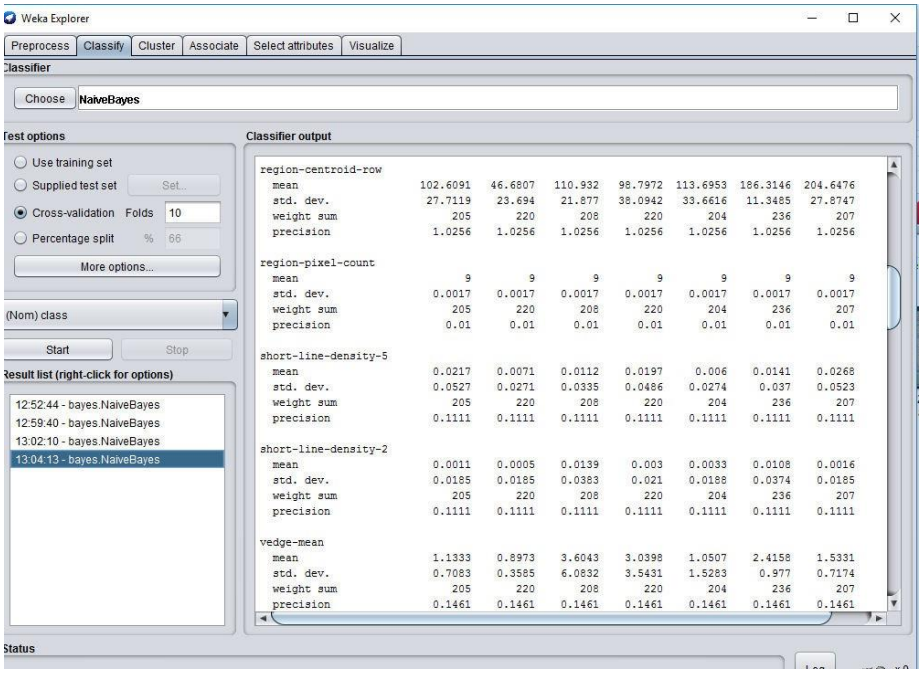

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
Classifier output
Root mean squared error 0.3399
Relative absolute error 72.7766 %
Root relative squared error 104.7431 %
Total Number of Instances 214
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.729 0.424 0.455 0.729 0.560 0.286 0.708 0.459 bui
0.171 0.101 0.481 0.171 0.252 0.100 0.717 0.506 bui
0.235 0.086 0.190 0.235 0.211 0.135 0.723 0.162 veh
? 0.000 ? ? ? ? ? ? veh
0.308 0.040 0.333 0.308 0.320 0.278 0.841 0.347 con
0.889 0.029 0.571 0.889 0.696 0.698 0.985 0.720 tab
0.828 0.022 0.857 0.828 0.842 0.818 0.928 0.789 hea
Weighted Avg. 0.486 0.188 0.496 0.486 0.453 0.297 0.762 0.501
=== Confusion Matrix ===
a b c d e f g <-- classified as
51 5 11 0 0 2 1 | a = build wind float
48 13 6 0 5 3 1 | b = build wind non-float
12 0 4 0 0 1 0 | c = vehic wind float
0 0 0 0 0 0 0 | d = vehic wind non-float
0 8 0 0 4 0 1 | e = containers
0 0 0 0 0 8 1 | f = tableware
1 1 0 0 3 0 24 | g = headlamps
Status

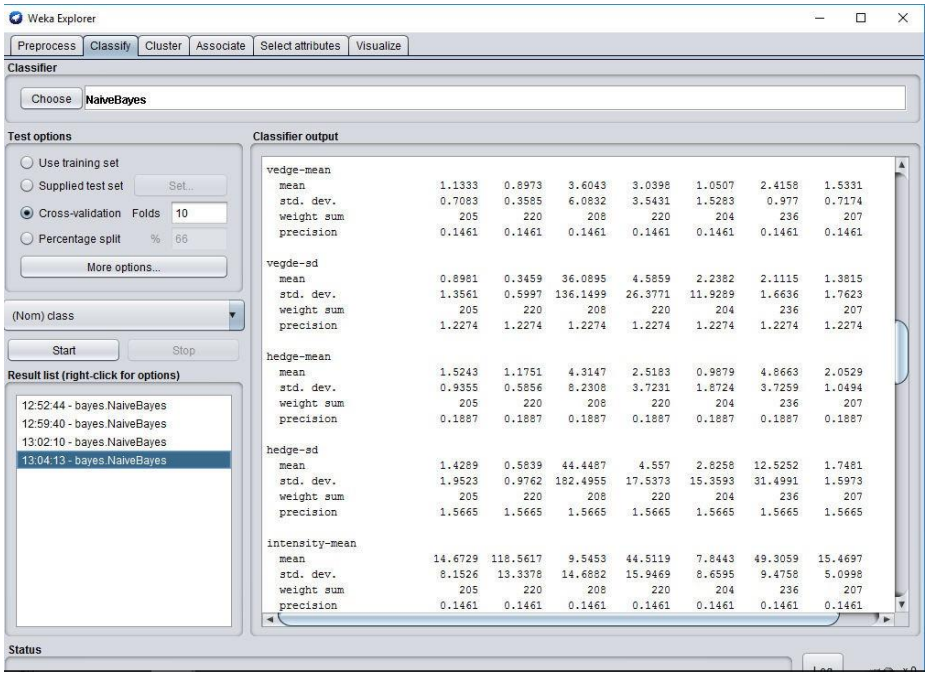
❖ Dataset 4: segment-challenge

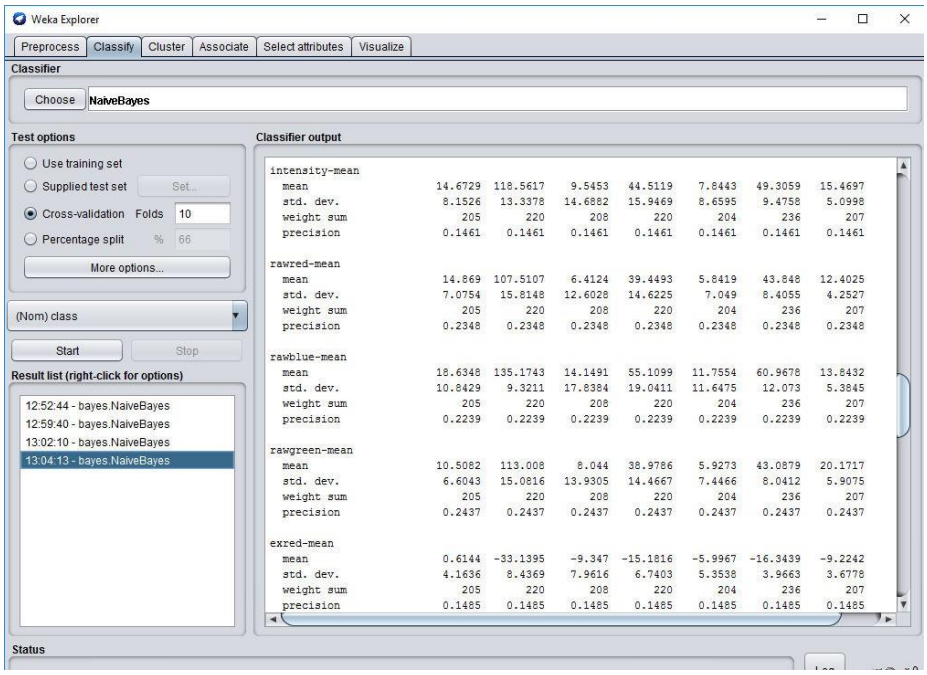


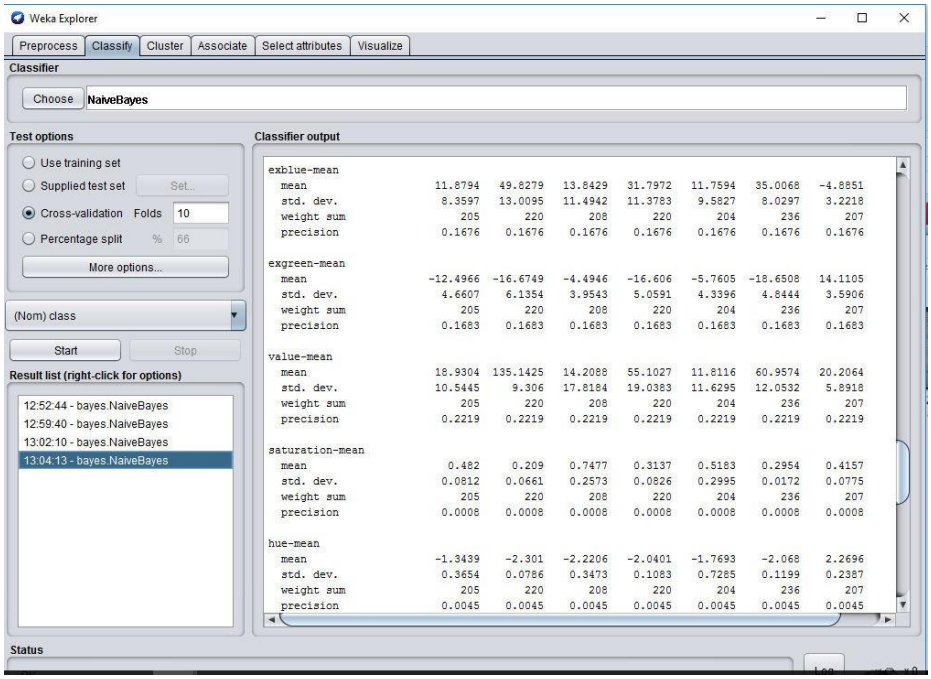


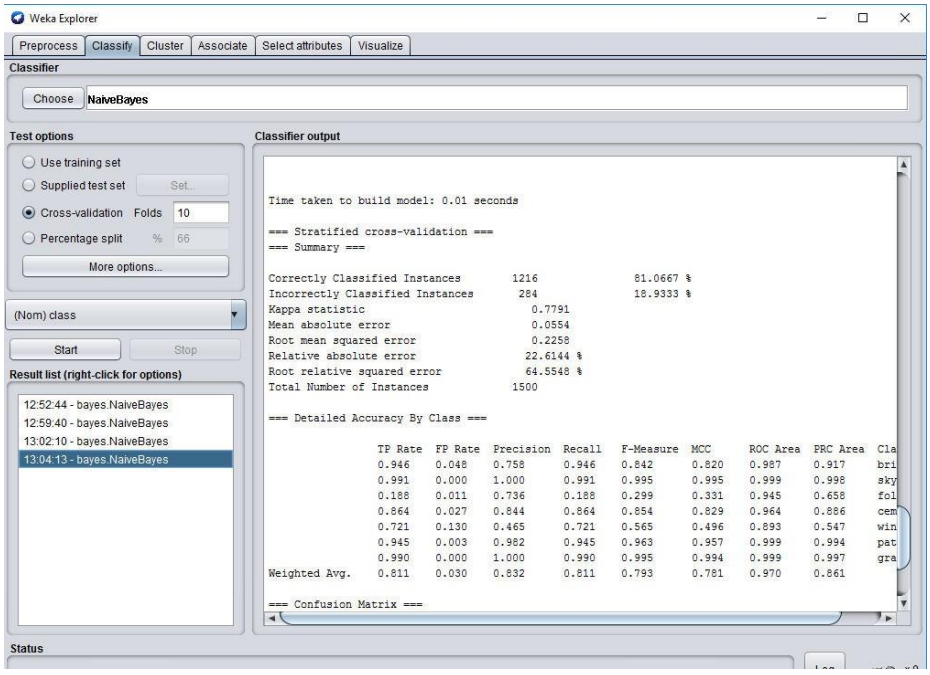

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
hue-mean
class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class
Attribute brickface sky foliage cement window path grass
(0.14) (0.15) (0.14) (0.15) (0.14) (0.16) (0.14)
region-centroid-col
mean 88.7557 134.8375 91.3949 129.6306 157.062 138.2456 134.0855
std. dev. 63.5267 69.2729 70.8938 67.2085 66.9201 71.9834 74.5131
weight sum 205 220 208 220 204 236 207
precision 1.004 1.004 1.004 1.004 1.004 1.004 1.004
region-centroid-row
mean 102.6091 46.6807 110.932 98.7972 113.6953 186.3146 204.6476
std. dev. 27.7119 23.694 21.877 38.0942 33.6616 11.3485 27.8747
weight sum 205 220 208 220 204 236 207
precision 1.0256 1.0256 1.0256 1.0256 1.0256 1.0256 1.0256
region-pixel-count
mean 9 9 9 9 9 9 9
std. dev. 0.0017 0.0017 0.0017 0.0017 0.0017 0.0017 0.0017
weight sum 205 220 208 220 204 236 207
precision 0.01 0.01 0.01 0.01 0.01 0.01 0.01
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
13:04:13 - bayes.NaiveBayes
Status

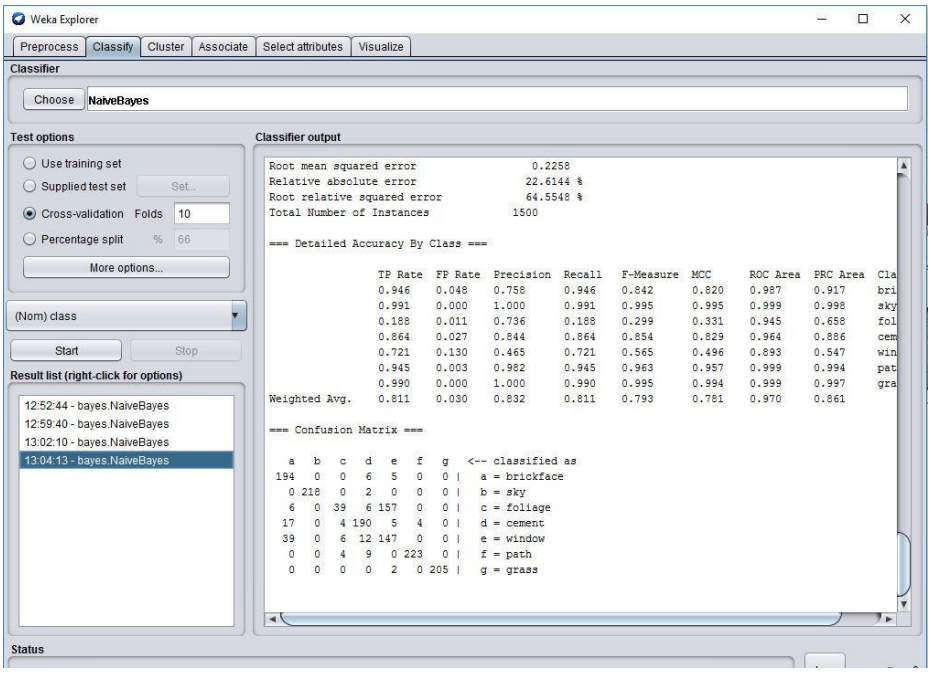

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
region-centroid-row
mean 102.6091 46.6807 110.932 98.7972 113.6953 186.3146 204.6476
std. dev. 27.7119 23.694 21.877 38.0942 33.6616 11.3485 27.8747
weight sum 205 220 208 220 204 236 207
precision 1.0256 1.0256 1.0256 1.0256 1.0256 1.0256 1.0256
region-pixel-count
mean 9 9 9 9 9 9 9
std. dev. 0.0017 0.0017 0.0017 0.0017 0.0017 0.0017 0.0017
weight sum 205 220 208 220 204 236 207
precision 0.01 0.01 0.01 0.01 0.01 0.01 0.01
short-line-density-5
mean 0.0217 0.0071 0.0112 0.0197 0.006 0.0141 0.0268
std. dev. 0.0527 0.0271 0.0335 0.0486 0.0274 0.037 0.0523
weight sum 205 220 208 220 204 236 207
precision 0.1111 0.1111 0.1111 0.1111 0.1111 0.1111 0.1111
short-line-density-2
mean 0.0011 0.0005 0.0139 0.003 0.0033 0.0108 0.0016
std. dev. 0.0185 0.0185 0.0383 0.021 0.0188 0.0374 0.0185
weight sum 205 220 208 220 204 236 207
precision 0.1111 0.1111 0.1111 0.1111 0.1111 0.1111 0.1111
vedge-mean
mean 1.1333 0.8973 3.6043 3.0398 1.0507 2.4158 1.5331
std. dev. 0.7083 0.3585 6.0832 3.5431 1.5283 0.977 0.7174
weight sum 205 220 208 220 204 236 207
precision 0.1461 0.1461 0.1461 0.1461 0.1461 0.1461 0.1461
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
13:04:13 - bayes.NaiveBayes
Status


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
13:04:13 - bayes.NaiveBayes
Classifier output
vedge-mean
mean 1.1333 0.8973 3.6043 3.0398 1.0507 2.4158 1.5331
std. dev. 0.7083 0.3585 6.0832 3.5431 1.5283 0.977 0.7174
weight sum 205 220 208 220 204 236 207
precision 0.1461 0.1461 0.1461 0.1461 0.1461 0.1461 0.1461
vegde-sd
mean 0.8981 0.3459 36.0895 4.5859 2.2382 2.1115 1.3815
std. dev. 1.3561 0.5997 136.1499 26.3771 11.9289 1.6636 1.7623
weight sum 205 220 208 220 204 236 207
precision 1.2274 1.2274 1.2274 1.2274 1.2274 1.2274 1.2274
hedge-mean
mean 1.5243 1.1751 4.3147 2.5183 0.9879 4.8663 2.0529
std. dev. 0.9355 0.5856 8.2308 3.7231 1.8724 3.7259 1.0494
weight sum 205 220 208 220 204 236 207
precision 0.1887 0.1887 0.1887 0.1887 0.1887 0.1887 0.1887
hedge-sd
mean 1.4289 0.5839 44.4487 4.557 2.8258 12.5252 1.7481
std. dev. 1.9523 0.9762 182.4955 17.5373 15.3593 31.4991 1.5973
weight sum 205 220 208 220 204 236 207
precision 1.5665 1.5665 1.5665 1.5665 1.5665 1.5665 1.5665
intensity-mean
mean 14.6729 118.5617 9.5453 44.5119 7.8443 49.3059 15.4697
std. dev. 8.1526 13.3378 14.6882 15.9469 8.6595 9.4758 5.0998
weight sum 205 220 208 220 204 236 207
precision 0.1461 0.1461 0.1461 0.1461 0.1461 0.1461 0.1461
Status

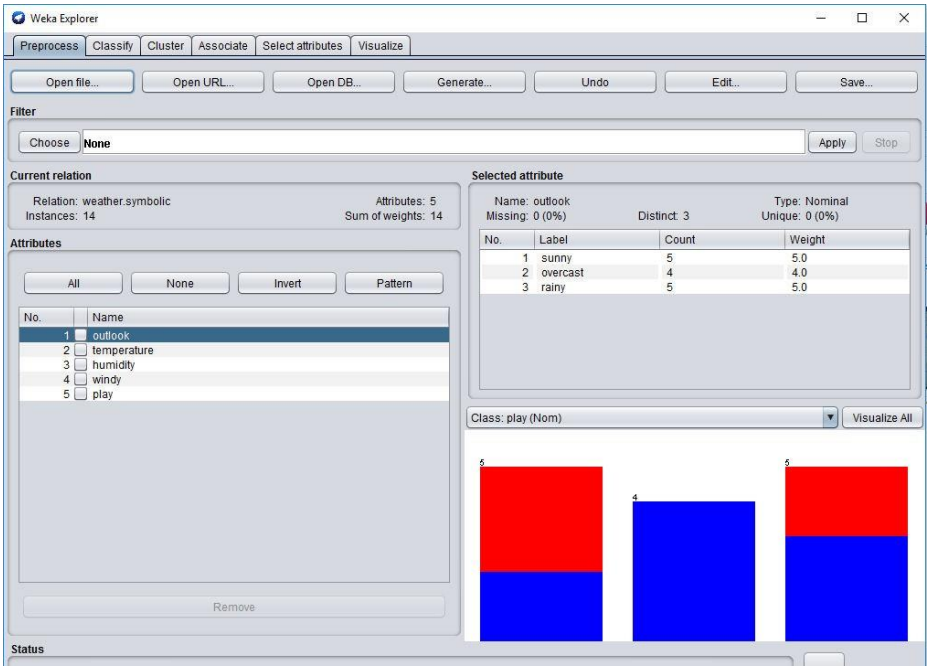

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
13:04:13 - bayes.NaiveBayes
Classifier output
intensity-mean
mean 14.6729 118.5617 9.5453 44.5119 7.8443 49.3059 15.4697
std. dev. 8.1526 13.3378 14.6882 15.9469 8.6595 9.4758 5.0998
weight sum 205 220 208 220 204 236 207
precision 0.1461 0.1461 0.1461 0.1461 0.1461 0.1461 0.1461
rawred-mean
mean 14.869 107.5107 6.4124 39.4493 5.8419 43.848 12.4025
std. dev. 7.0754 15.8148 12.6028 14.6225 7.049 8.4055 4.2527
weight sum 205 220 208 220 204 236 207
precision 0.2348 0.2348 0.2348 0.2348 0.2348 0.2348 0.2348
rawblue-mean
mean 18.6348 135.1743 14.1491 55.1099 11.7554 60.9678 13.8432
std. dev. 10.8429 9.3211 17.8384 19.0411 11.6475 12.073 5.3845
weight sum 205 220 208 220 204 236 207
precision 0.2239 0.2239 0.2239 0.2239 0.2239 0.2239 0.2239
rawgreen-mean
mean 10.5082 113.008 8.044 38.9786 5.9273 43.0879 20.1717
std. dev. 6.6043 15.0816 13.9305 14.4667 7.4466 8.0412 5.9075
weight sum 205 220 208 220 204 236 207
precision 0.2437 0.2437 0.2437 0.2437 0.2437 0.2437 0.2437
exred-mean
mean 0.6144 -33.1395 -9.347 -15.1816 -5.9967 -16.3439 -9.2242
std. dev. 4.1636 8.4369 7.9616 6.7403 5.3538 3.9663 3.6778
weight sum 205 220 208 220 204 236 207
precision 0.1485 0.1485 0.1485 0.1485 0.1485 0.1485 0.1485
Status

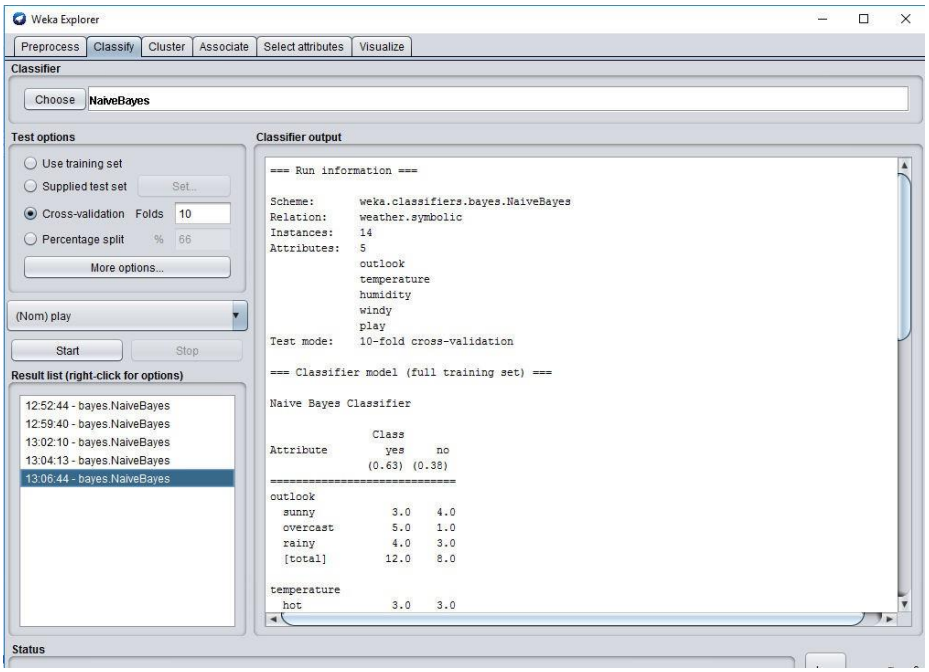

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
13:04:13 - bayes.NaiveBayes
Classifier output
exblue-mean
mean 11.8794 49.8279 13.8429 31.7972 11.7594 35.0068 -4.8851
std. dev. 8.3597 13.0095 11.4942 11.3783 9.5827 8.0297 3.2218
weight sum 205 220 208 220 204 236 207
precision 0.1676 0.1676 0.1676 0.1676 0.1676 0.1676 0.1676
exgreen-mean
mean -12.4966 -16.6749 -4.4946 -16.606 -5.7605 -18.6508 14.1105
std. dev. 4.6607 6.1354 3.9543 5.0591 4.3396 4.8444 3.5906
weight sum 205 220 208 220 204 236 207
precision 0.1683 0.1683 0.1683 0.1683 0.1683 0.1683 0.1683
value-mean
mean 18.9304 135.1425 14.2088 55.1027 11.8116 60.9574 20.2064
std. dev. 10.5445 9.306 17.8184 19.0383 11.6295 12.0532 5.8918
weight sum 205 220 208 220 204 236 207
precision 0.2219 0.2219 0.2219 0.2219 0.2219 0.2219 0.2219
saturation-mean
mean 0.482 0.209 0.7477 0.3137 0.5183 0.2954 0.4157
std. dev. 0.0812 0.0661 0.2573 0.0826 0.2995 0.0172 0.0775
weight sum 205 220 208 220 204 236 207
precision 0.0008 0.0008 0.0008 0.0008 0.0008 0.0008 0.0008
hue-mean
mean -1.3439 -2.301 -2.2206 -2.0401 -1.7693 -2.068 2.2696
std. dev. 0.3654 0.0786 0.3473 0.1083 0.7285 0.1199 0.2387
weight sum 205 220 208 220 204 236 207
precision 0.0045 0.0045 0.0045 0.0045 0.0045 0.0045 0.0045
Status

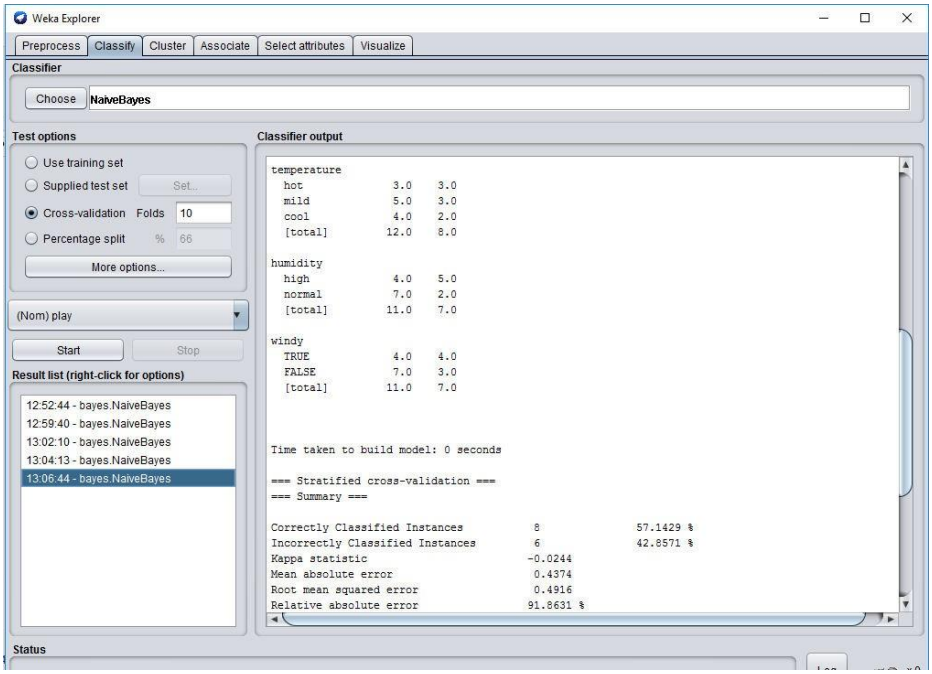

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
13:04:13 - bayes.NaiveBayes
Classifier output
Time taken to build model: 0.01 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 1216 81.0667 %
Incorrectly Classified Instances 284 18.9333 %
Kappa statistic 0.7791
Mean absolute error 0.0554
Root mean squared error 0.2258
Relative absolute error 22.6144 %
Root relative squared error 64.5548 %
Total Number of Instances 1500
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.946 0.048 0.758 0.946 0.842 0.820 0.987 0.917 bri
0.991 0.000 1.000 0.991 0.995 0.995 0.999 0.998 sky
0.188 0.011 0.736 0.188 0.299 0.331 0.945 0.658 fol
0.864 0.027 0.844 0.864 0.854 0.829 0.964 0.886 cem
0.721 0.130 0.465 0.721 0.565 0.496 0.893 0.547 win
0.945 0.003 0.982 0.945 0.963 0.957 0.999 0.994 pat
0.990 0.000 1.000 0.990 0.995 0.994 0.999 0.997 gra
Weighted Avg. 0.811 0.030 0.832 0.811 0.793 0.781 0.970 0.861
=== Confusion Matrix ===
Status

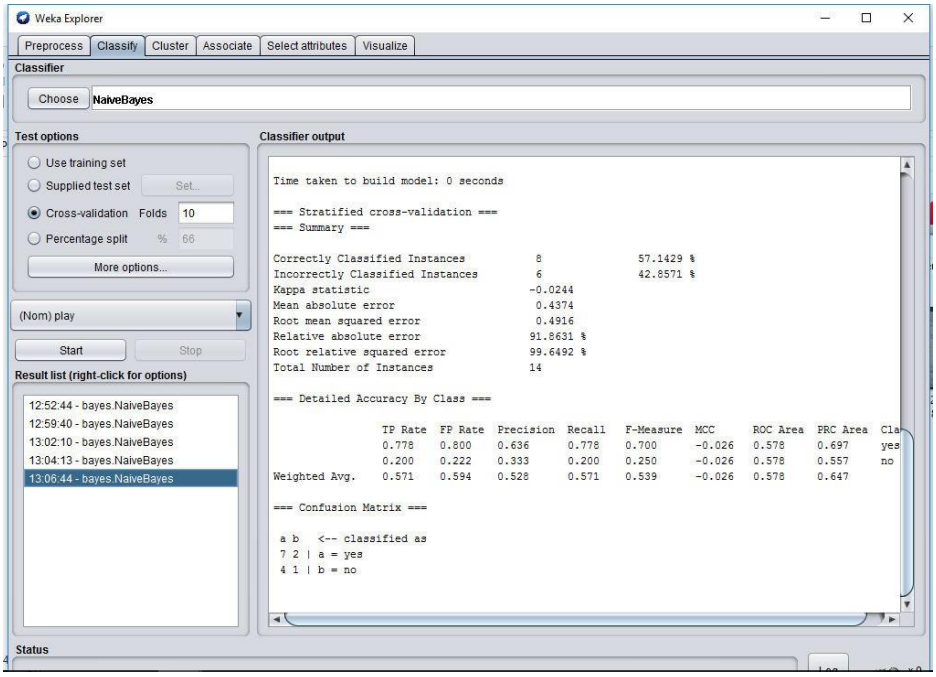


❖ Dataset 5: weather Nominal




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
13:04:13 - bayes.NaiveBayes
13:06:44 - bayes.NaiveBayes
Classifier output
=== Run information ===
Scheme: weka.classifiers.bayes.NaiveBayes
Relation: weather.symbolic
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
Class
Attribute yes no
(0.63) (0.38)
outlook
sunny 3.0 4.0
overcast 5.0 1.0
rainy 4.0 3.0
[total] 12.0 8.0
temperature
hot 3.0 3.0
Status


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
13:04:13 - bayes.NaiveBayes
13:06:44 - bayes.NaiveBayes
Classifier output
temperature
hot 3.0 3.0
mild 5.0 3.0
cool 4.0 2.0
[total] 12.0 8.0
humidity
high 4.0 5.0
normal 7.0 2.0
[total] 11.0 7.0
windy
TRUE 4.0 4.0
FALSE 7.0 3.0
[total] 11.0 7.0
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 8 57.1429 %
Incorrectly Classified Instances 6 42.8571 %
Kappa statistic -0.0244
Mean absolute error 0.4374
Root mean squared error 0.4916
Relative absolute error 91.8631 %
Status


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
12:52:44 - bayes.NaiveBayes
12:59:40 - bayes.NaiveBayes
13:02:10 - bayes.NaiveBayes
13:04:13 - bayes.NaiveBayes
13:06:44 - bayes.NaiveBayes
Classifier output
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 8 57.1429 %
Incorrectly Classified Instances 6 42.8571 %
Kappa statistic -0.0244
Mean absolute error 0.4374
Root mean squared error 0.4916
Relative absolute error 91.8631 %
Root relative squared error 99.6492 %
Total Number of Instances 14
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.778 0.800 0.636 0.778 0.700 -0.026 0.578 0.697 yes
0.200 0.222 0.333 0.200 0.250 -0.026 0.578 0.557 no
Weighted Avg. 0.571 0.594 0.528 0.571 0.539 -0.026 0.578 0.647
=== Confusion Matrix ===
a b <-- classified as
7 2 | a = yes
4 1 | b = no
Status

Practical 9(2)

Aim: To perform classification on various datasets and the logistic algorithm.

9-2-2.Logistic:-

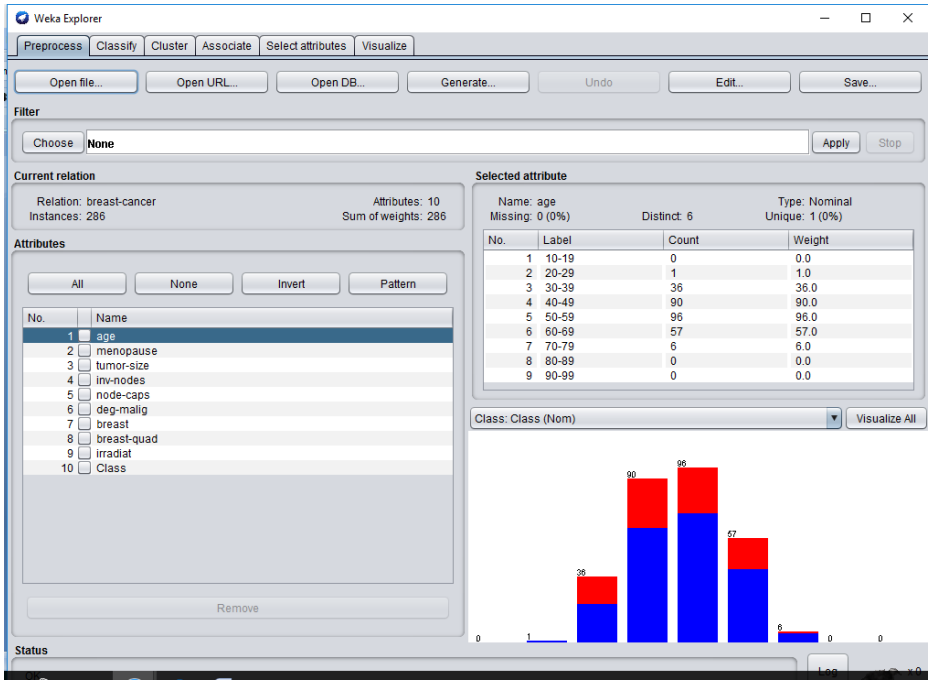
1 DataSet:- breast-cancer

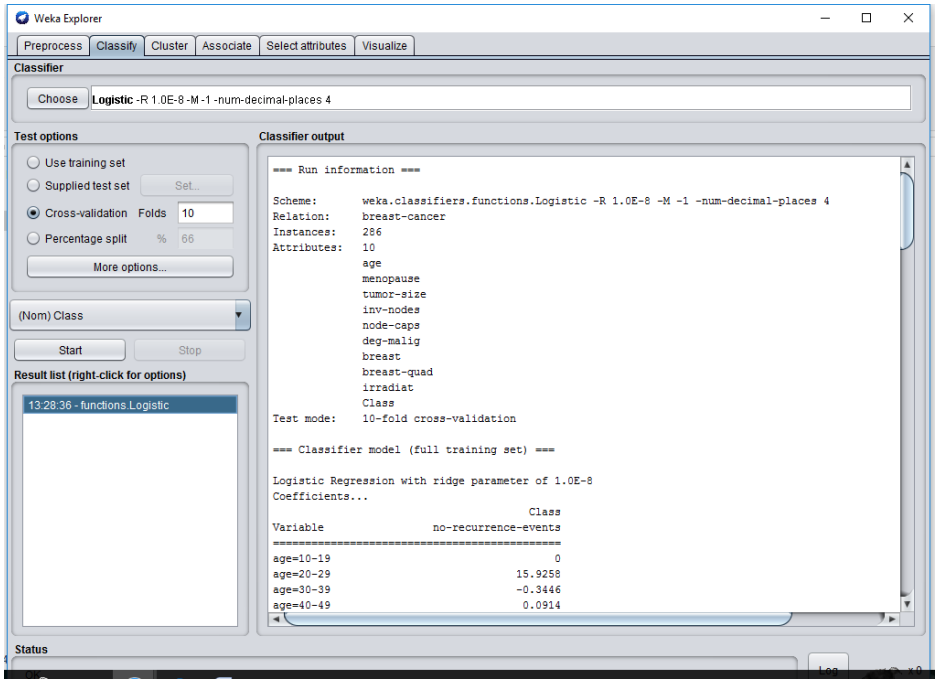
2 DataSet:-Glass

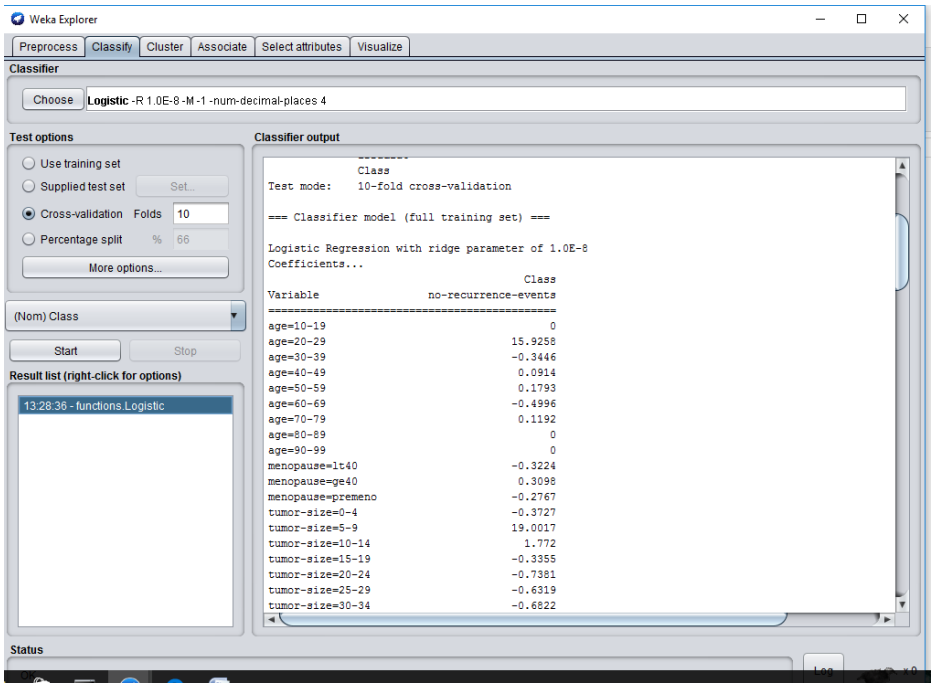
3 DataSet:- segment-challenge

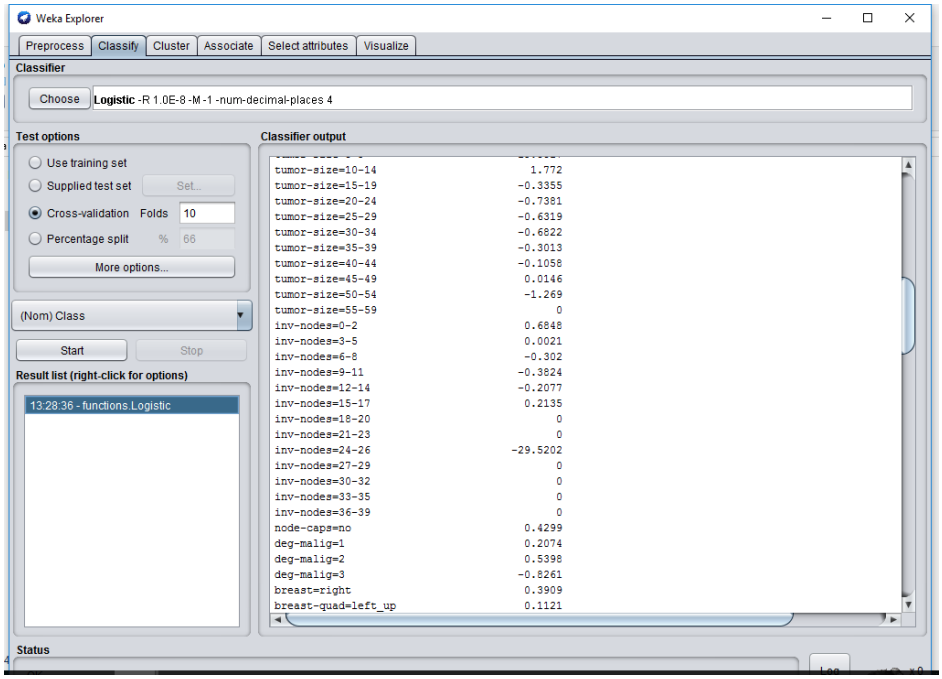
4 DataSet:- weather. Nominal

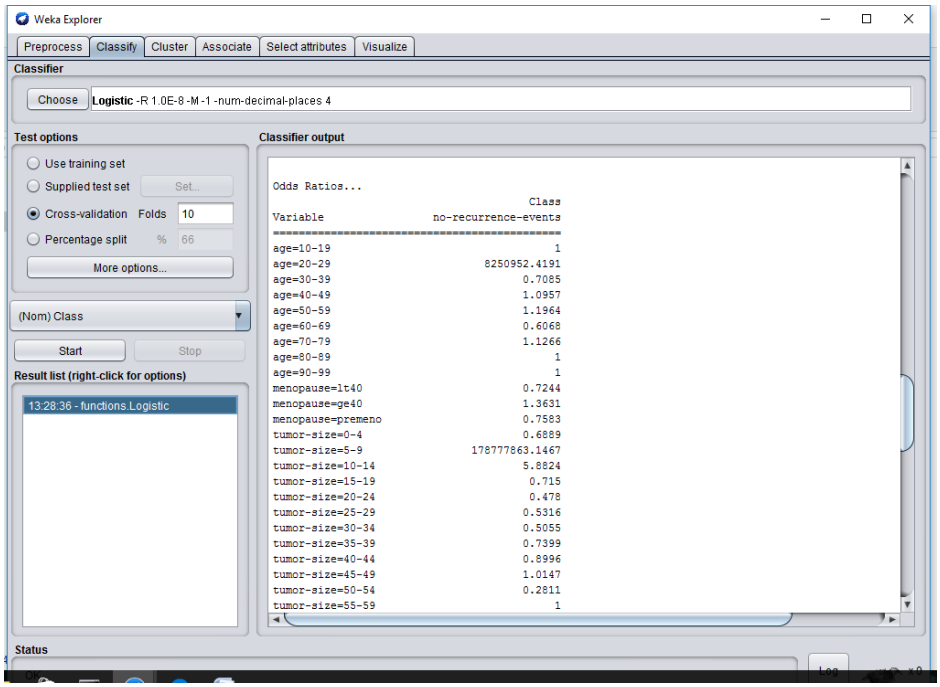
- Dataset 1: breast-cancer

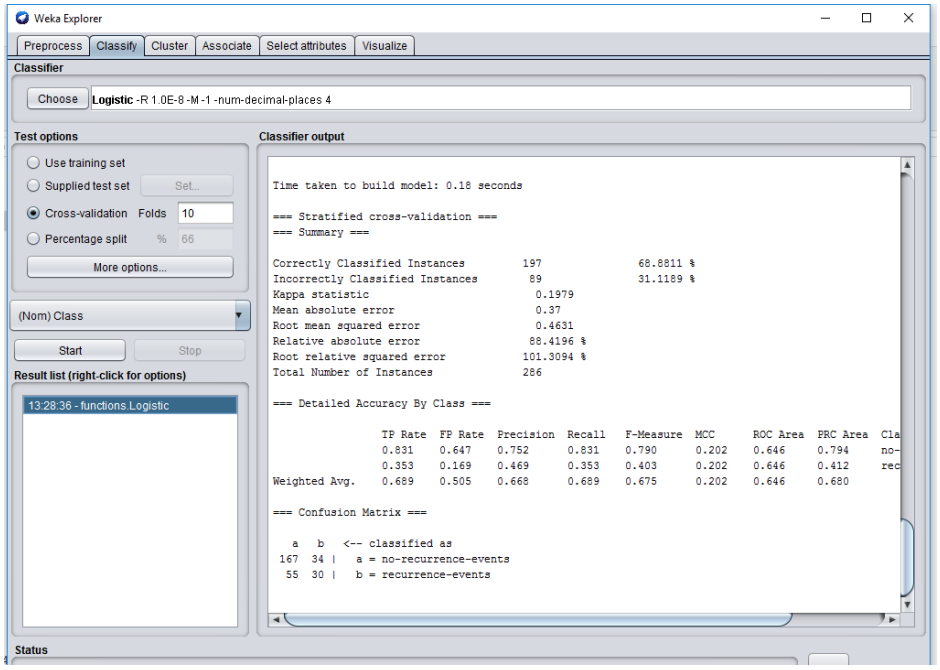

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: breast-cancer
Instances: 286
Attributes: 10
Sum of weights: 286
Selected attribute
Name: age
Missing: 0 (0%)
Distinct: 6
Type: Nominal
Unique: 1 (0%)
Attributes
All
None
Invert
Pattern
No. Name
1 age
2 menopause
3 tumor-size
4 inv-nodes
5 node-caps
6 deg-malig
7 breast
8 breast-quad
9 irradiat
10 Class
Remove
No. Label Count Weight
1 10-19 0 0.0
2 20-29 1 1.0
3 30-39 36 36.0
4 40-49 90 90.0
5 50-59 96 96.0
6 60-69 57 57.0
7 70-79 6 6.0
8 80-89 0 0.0
9 90-99 0 0.0
Class: Class (Nom)
Visualize All
Status


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
13:28:36 - functions.Logistic
Classifier output
=== Run information ===
Scheme: weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Relation: breast-cancer
Instances: 286
Attributes: 10
age
menopause
tumor-size
inv-nodes
node-caps
deg-malig
breast
breast-quad
irradiat
Class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
Class
Variable no-recurrence-events
age=10-19 0
age=20-29 15.9258
age=30-39 -0.3446
age=40-49 0.0914
Status

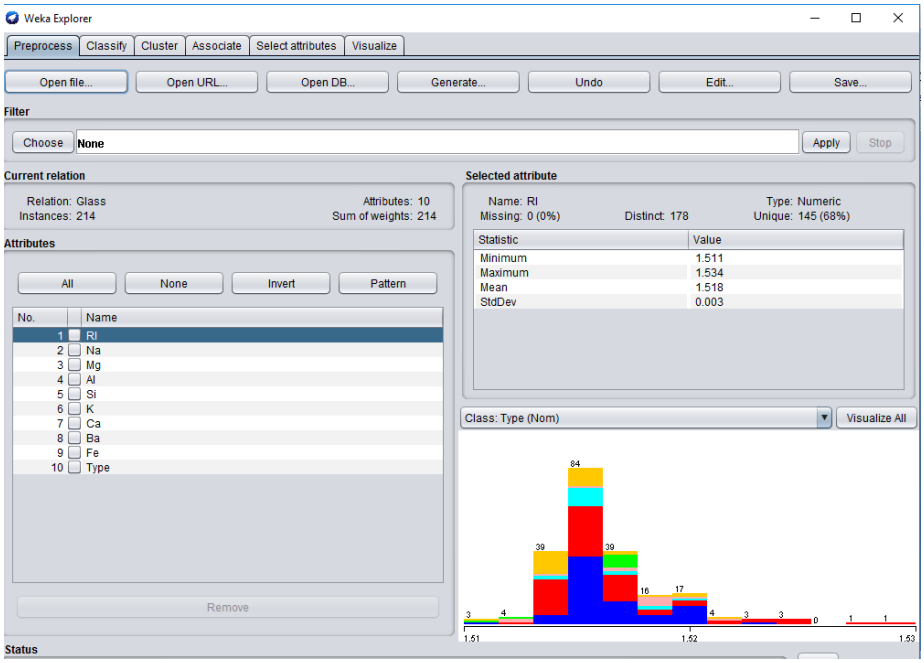

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
13:28:36 - functions.Logistic
Classifier output
Class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
Class
Variable no-recurrence-events
age=10-19 0
age=20-29 15.9258
age=30-39 -0.3446
age=40-49 0.0914
age=50-59 0.1793
age=60-69 -0.4996
age=70-79 0.1192
age=80-89 0
age=90-99 0
menopause=lt40 -0.3224
menopause=ge40 0.3098
menopause=premeno -0.2767
tumor-size=0-4 -0.3727
tumor-size=5-9 19.0017
tumor-size=10-14 1.772
tumor-size=15-19 -0.3355
tumor-size=20-24 -0.7381
tumor-size=25-29 -0.6319
tumor-size=30-34 -0.6822
Status

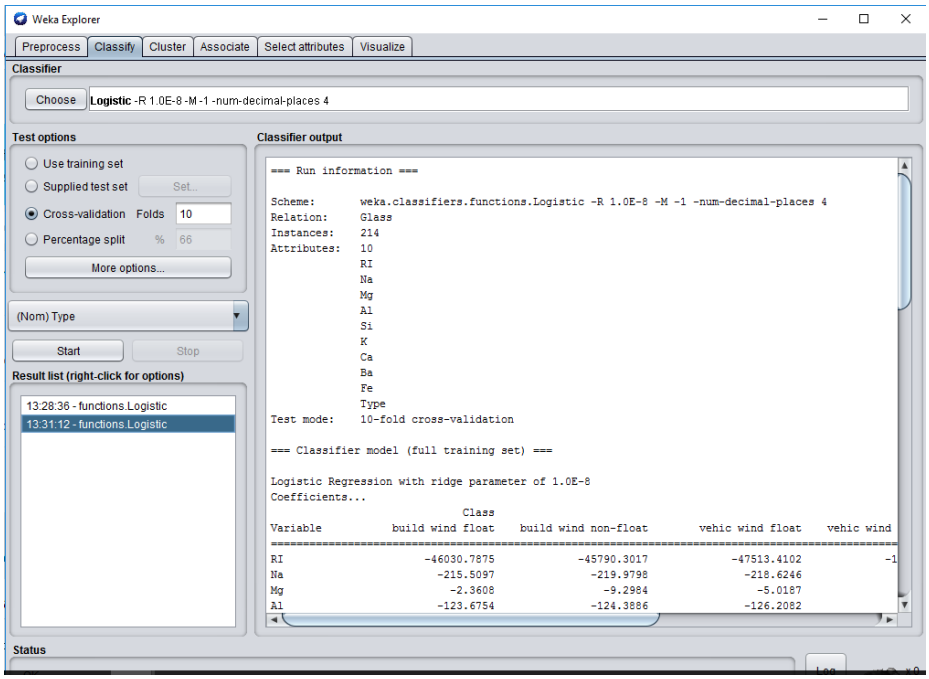

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
13:28:36 - functions.Logistic
Classifier output
tumor-size=10-14 1.772
tumor-size=15-19 -0.3355
tumor-size=20-24 -0.7381
tumor-size=25-29 -0.6319
tumor-size=30-34 -0.6822
tumor-size=35-39 -0.3013
tumor-size=40-44 -0.1058
tumor-size=45-49 0.0146
tumor-size=50-54 -1.269
tumor-size=55-59 0
inv-nodes=0-2 0.6848
inv-nodes=3-5 0.0021
inv-nodes=6-8 -0.302
inv-nodes=9-11 -0.3824
inv-nodes=12-14 -0.2077
inv-nodes=15-17 0.2135
inv-nodes=18-20 0
inv-nodes=21-23 0
inv-nodes=24-26 -29.5202
inv-nodes=27-29 0
inv-nodes=30-32 0
inv-nodes=33-35 0
inv-nodes=36-39 0
node-caps=no 0.4299
deg-malig=1 0.2074
deg-malig=2 0.5398
deg-malig=3 -0.8261
breast=right 0.3909
breast-quad=left_up 0.1121
Status

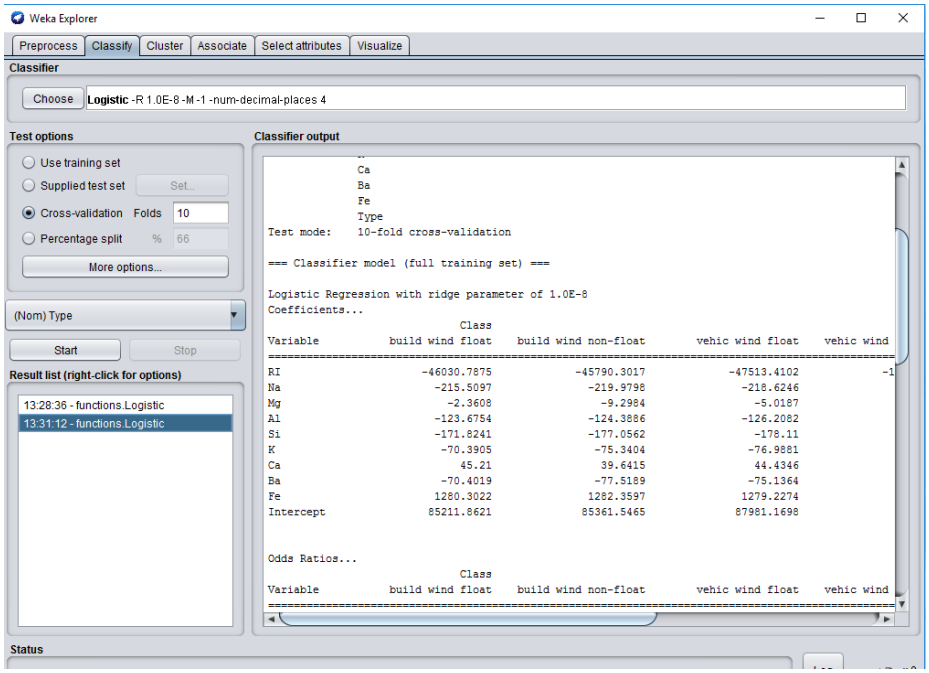

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
13:28:36 - functions.Logistic
Classifier output
Odds Ratios...
Class
Variable no-recurrence-events
age=10-19 1
age=20-29 8250952.4191
age=30-39 0.7085
age=40-49 1.0957
age=50-59 1.1964
age=60-69 0.6068
age=70-79 1.1266
age=80-89 1
age=90-99 1
menopause=lt40 0.7244
menopause=ge40 1.3631
menopause=premeno 0.7583
tumor-size=0-4 0.6889
tumor-size=5-9 178777863.1467
tumor-size=10-14 5.8824
tumor-size=15-19 0.715
tumor-size=20-24 0.478
tumor-size=25-29 0.5316
tumor-size=30-34 0.5055
tumor-size=35-39 0.7399
tumor-size=40-44 0.8996
tumor-size=45-49 1.0147
tumor-size=50-54 0.2811
tumor-size=55-59 1
Status

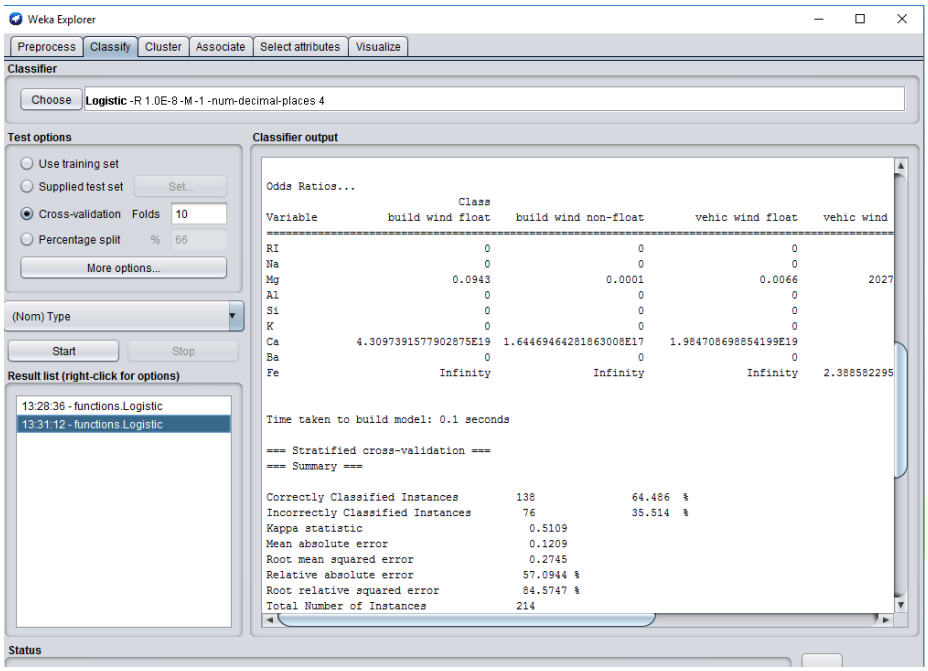

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Class
Start
Stop
Result list (right-click for options)
13:28:36 - functions.Logistic
Classifier output
Time taken to build model: 0.18 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 197 68.8811 %
Incorrectly Classified Instances 89 31.1189 %
Kappa statistic 0.1979
Mean absolute error 0.37
Root mean squared error 0.4631
Relative absolute error 88.4196 %
Root relative squared error 101.3094 %
Total Number of Instances 286
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.831 0.647 0.752 0.831 0.790 0.202 0.646 0.794 no-
0.353 0.169 0.469 0.353 0.403 0.202 0.646 0.412 rec
Weighted Avg. 0.689 0.505 0.668 0.689 0.675 0.202 0.646 0.680
=== Confusion Matrix ===
a b <-- classified as
167 34 | a = no-recurrence-events
55 30 | b = recurrence-events
Status

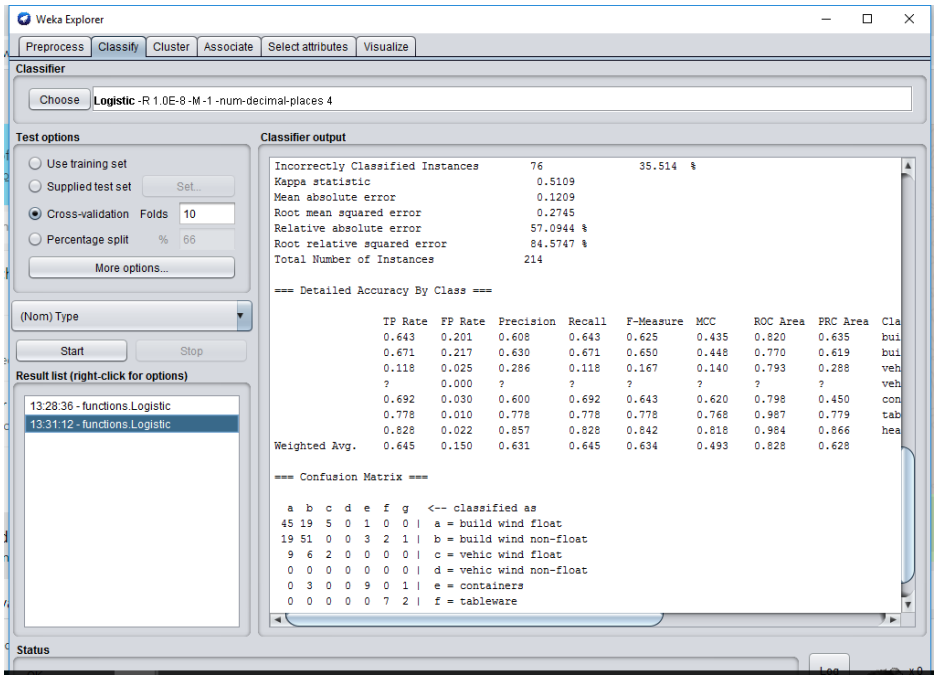
❖ Dataset 2: glass



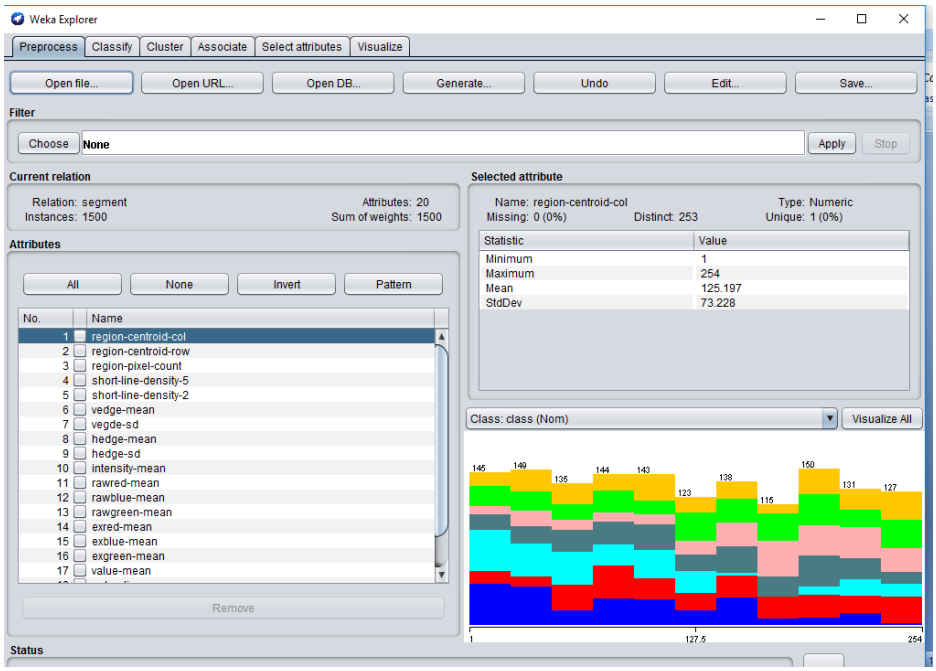

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
13:28:36 - functions.Logistic
13:31:12 - functions.Logistic
Classifier output
=== Run information ===
Scheme: weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Relation: Glass
Instances: 214
Attributes: 10
RI
Na
Mg
Al
Si
K
Ca
Ba
Fe
Type
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
Class
Variable build wind float build wind non-float vehic wind float vehic wind
RI -46030.7875 -45790.3017 -47513.4102
Na -215.5097 -219.9798 -218.6246
Mg -2.3608 -9.2984 -5.0187
Al -123.6754 -124.3886 -126.2082
Status

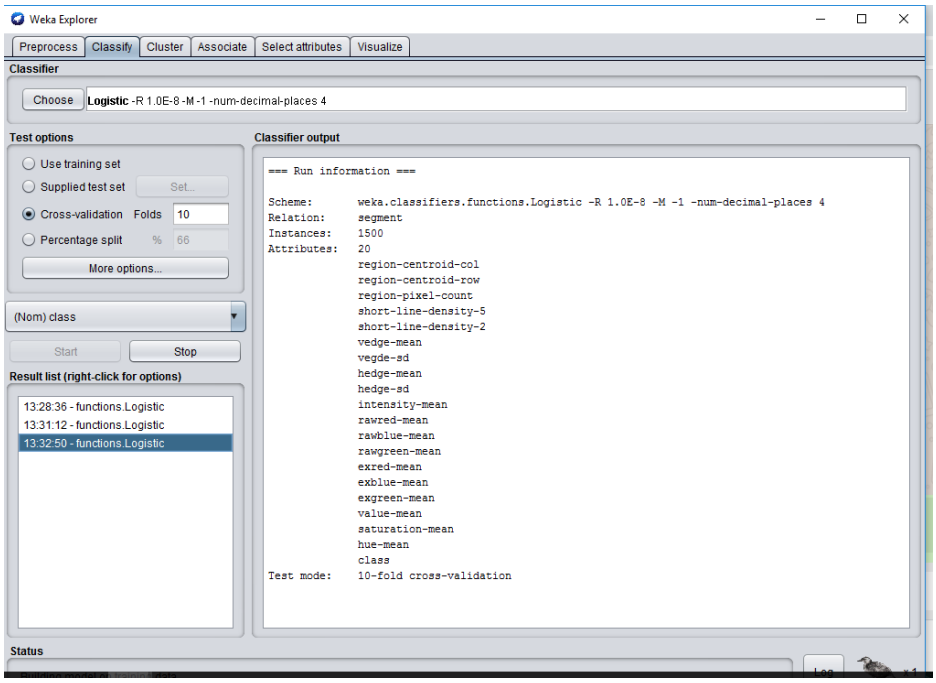

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) Type
Start
Stop
Result list (right-click for options)
13:28:36 - functions.Logistic
13:31:12 - functions.Logistic
Classifier output
Ca
Ba
Fe
Type
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
Class
Variable build wind float build wind non-float vehic wind float vehic wind
RI -46030.7875 -45790.3017 -47513.4102
Na -215.5097 -219.9798 -218.6246
Mg -2.3608 -9.2984 -5.0187
Al -123.6754 -124.3886 -126.2082
Si -171.8241 -177.0562 -178.11
K -70.3905 -75.3404 -76.9881
Ca 45.21 39.6415 44.4346
Ba -70.4019 -77.5189 -75.1364
Fe 1280.3022 1282.3597 1279.2274
Intercept 85211.8621 85361.5465 87981.1698
Odds Ratios...
Class
Variable build wind float build wind non-float vehic wind float vehic wind
Status



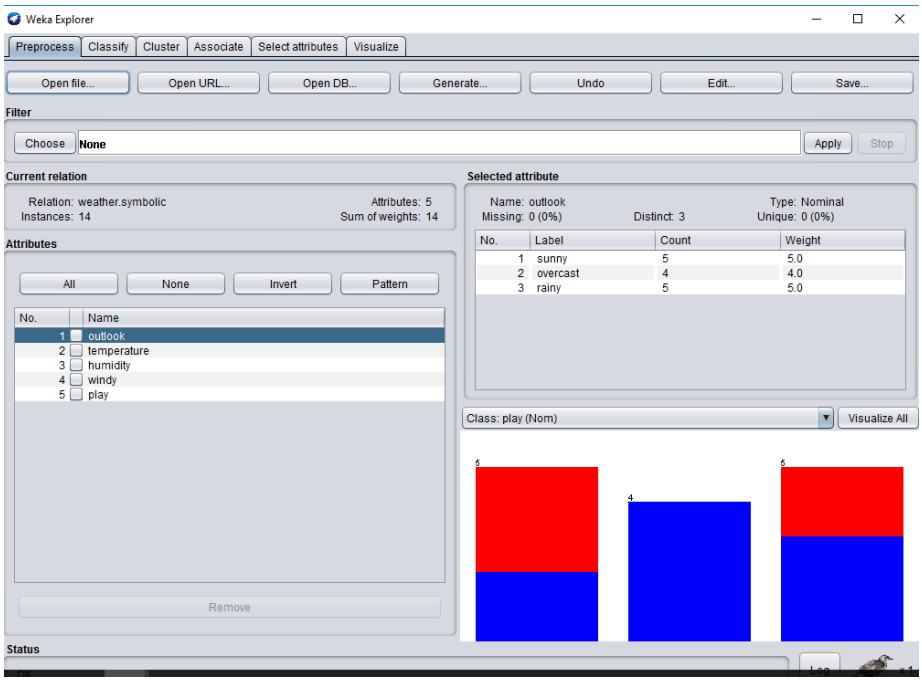


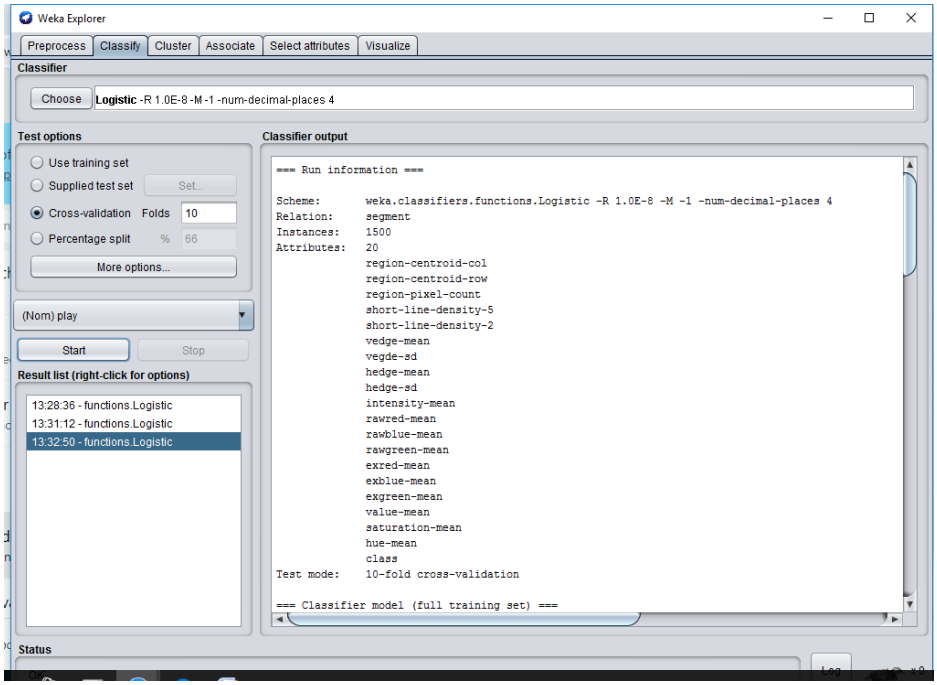
❖ Dataset 3: segment-challenge

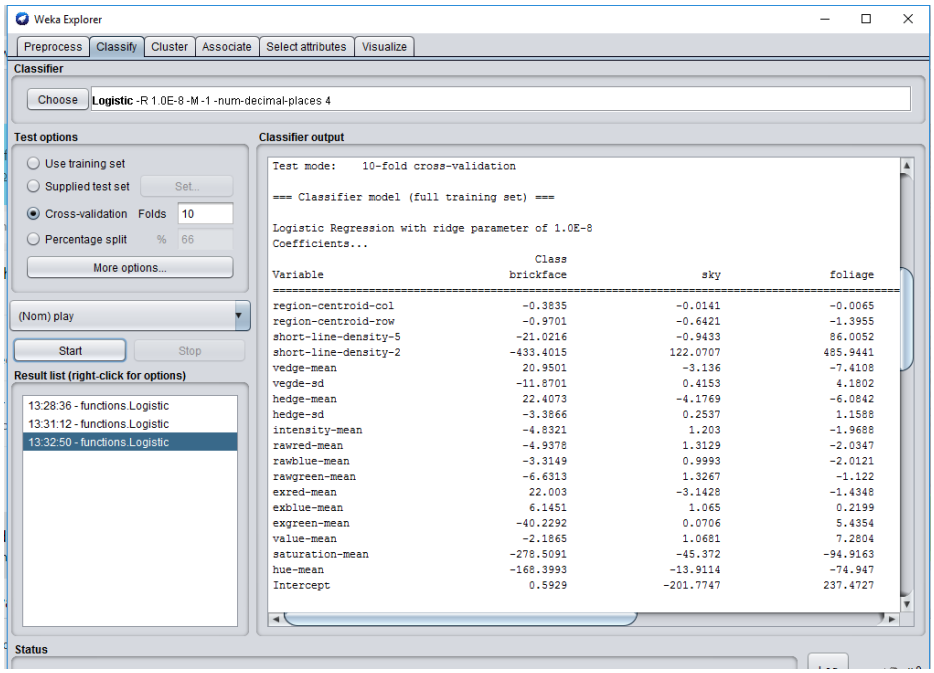


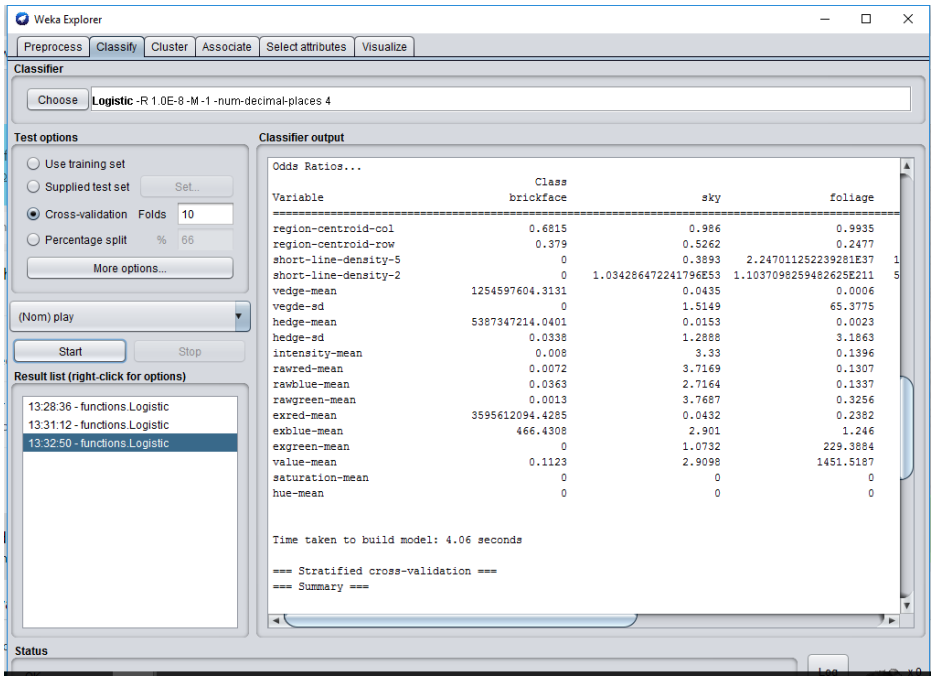


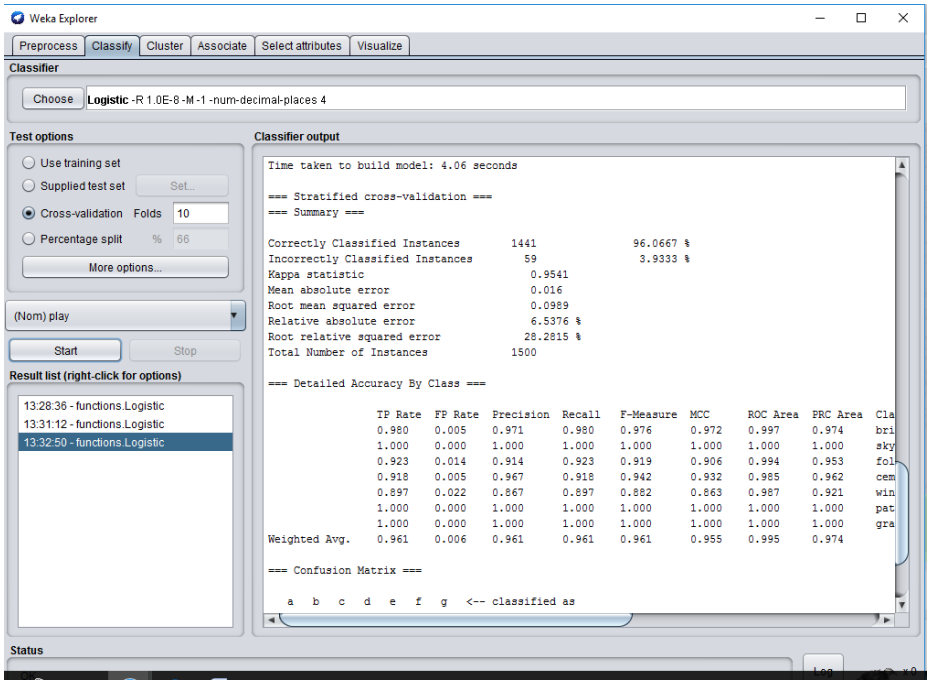
❖ Dataset 3: weather. Nominal

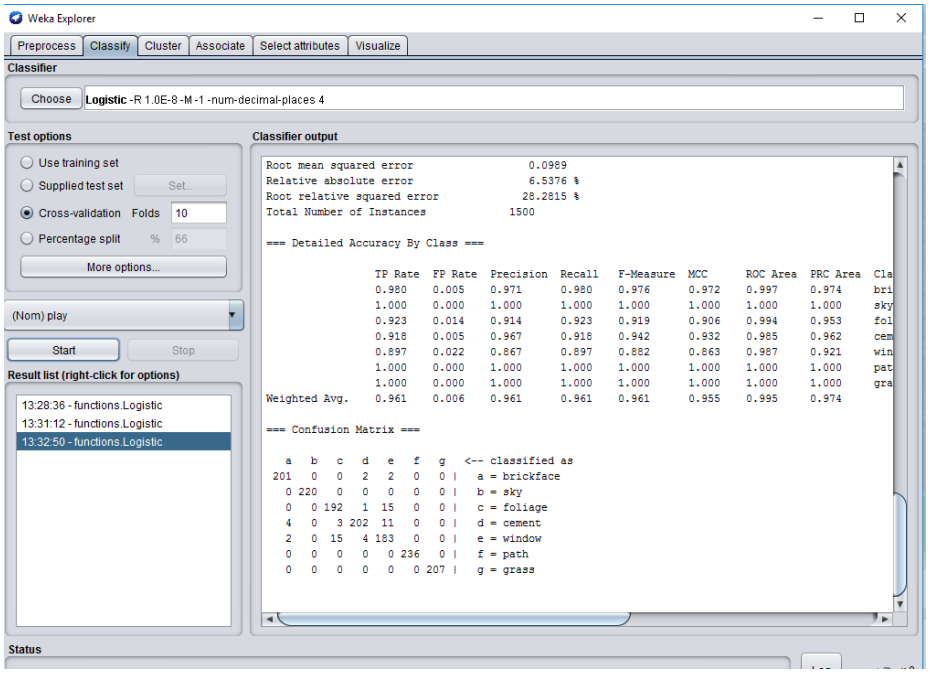

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: weather.symbolic
Instances: 14
Attributes: 5
Sum of weights: 14
Selected attribute
Name: outlook
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)
No. Label Count Weight
1 sunny 5 5.0
2 overcast 4 4.0
3 rainy 5 5.0
Attributes
All
None
Invert
Pattern
No. Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Class: play (Nom)
Visualize All
Status


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
13:28:36 - functions.Logistic
13:31:12 - functions.Logistic
13:32:50 - functions.Logistic
Classifier output
=== Run information ===
Scheme: weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Relation: segment
Instances: 1500
Attributes: 20
region-centroid-col
region-centroid-row
region-pixel-count
short-line-density-5
short-line-density-2
vedge-mean
vegde-sd
hedge-mean
hedge-sd
intensity-mean
rawred-mean
rawblue-mean
rawgreen-mean
exred-mean
exblue-mean
exgreen-mean
value-mean
saturation-mean
hue-mean
class
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Status


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
13:28:36 - functions.Logistic
13:31:12 - functions.Logistic
13:32:50 - functions.Logistic
Classifier output
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
Class
Variable brickface sky foliage
region-centroid-col -0.3835 -0.0141 -0.0065
region-centroid-row -0.9701 -0.6421 -1.3955
short-line-density-5 -21.0216 -0.9433 86.0052
short-line-density-2 -433.4015 122.0707 485.9441
vedge-mean 20.9501 -3.136 -7.4108
vegde-sd -11.8701 0.4153 4.1802
hedge-mean 22.4073 -4.1769 -6.0842
hedge-sd -3.3866 0.2537 1.1588
intensity-mean -4.8321 1.203 -1.9688
rawred-mean -4.9378 1.3129 -2.0347
rawblue-mean -3.3149 0.9993 -2.0121
rawgreen-mean -6.6313 1.3267 -1.122
exred-mean 22.003 -3.1428 -1.4348
exblue-mean 6.1451 1.065 0.2199
exgreen-mean -40.2292 0.0706 5.4354
value-mean -2.1865 1.0681 7.2804
saturation-mean -278.5091 -45.372 -94.9163
hue-mean -168.3993 -13.9114 -74.947
Intercept 0.5929 -201.7747 237.4727
Status


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
13:28:36 - functions.Logistic
13:31:12 - functions.Logistic
13:32:50 - functions.Logistic
Classifier output
Odds Ratios...
Class
Variable brickface sky foliage
region-centroid-col 0.6815 0.986 0.9935
region-centroid-row 0.379 0.5262 0.2477
short-line-density-5 0 0.3893 2.2470112522392 81E37
short-line-density-2 0 1.034286472241796E53 1.1037098259482625E211
vedge-mean 1254597604.3131 0.0435 0.0006
vegde-sd 0 1.5149 65.3775
hedge-mean 5387347214.0401 0.0153 0.0023
hedge-sd 0.0338 1.2888 3.1863
intensity-mean 0.008 3.33 0.1396
rawred-mean 0.0072 3.7169 0.1307
rawblue-mean 0.0363 2.7164 0.1337
rawgreen-mean 0.0013 3.7687 0.3256
exred-mean 3595612094.4285 0.0432 0.2382
exblue-mean 466.4308 2.901 1.246
exgreen-mean 0 1.0732 229.3884
value-mean 0.1123 2.9098 1451.5187
saturation-mean 0 0 0
hue-mean 0 0 0
Time taken to build model: 4.06 seconds
=== Stratified cross-validation ===
=== Summary ===
Status


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
13:28:36 - functions.Logistic
13:31:12 - functions.Logistic
13:32:50 - functions.Logistic
Classifier output
Time taken to build model: 4.06 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 1441 96.0667 %
Incorrectly Classified Instances 59 3.9333 %
Kappa statistic 0.9541
Mean absolute error 0.016
Root mean squared error 0.0989
Relative absolute error 6.5376 %
Root relative squared error 28.2815 %
Total Number of Instances 1500
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.980 0.005 0.971 0.980 0.976 0.972 0.997 0.974 bri
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 sky
0.923 0.014 0.914 0.923 0.919 0.906 0.994 0.953 fol
0.918 0.005 0.967 0.918 0.942 0.932 0.985 0.962 cem
0.897 0.022 0.867 0.897 0.882 0.863 0.987 0.921 win
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 pat
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 gra
Weighted Avg. 0.961 0.006 0.961 0.961 0.961 0.955 0.995 0.974
=== Confusion Matrix ===
a b c d e f g <-- classified as
Status


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Test options
Use training set
Supplied test set
Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
13:28:36 - functions.Logistic
13:31:12 - functions.Logistic
13:32:50 - functions.Logistic
Classifier output
Root mean squared error 0.0989
Relative absolute error 6.5376 %
Root relative squared error 28.2815 %
Total Number of Instances 1500
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.980 0.005 0.971 0.980 0.976 0.972 0.997 0.974 bri
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 sky
0.923 0.014 0.914 0.923 0.919 0.906 0.994 0.953 fol
0.918 0.005 0.967 0.918 0.942 0.932 0.985 0.962 cem
0.897 0.022 0.867 0.897 0.882 0.863 0.987 0.921 win
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 pat
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 gra
Weighted Avg. 0.961 0.006 0.961 0.961 0.961 0.955 0.995 0.974
=== Confusion Matrix ===
a b c d e f g <-- classified as
201 0 0 2 2 0 0 | a = brickface
0 220 0 0 0 0 0 | b = sky
0 0 192 1 15 0 0 | c = foliage
4 0 3 202 11 0 0 | d = cement
2 0 15 4 183 0 0 | e = window
0 0 0 0 0 236 0 | f = path
0 0 0 0 0 0 207 | g = grass
Status

Practical 10(1)

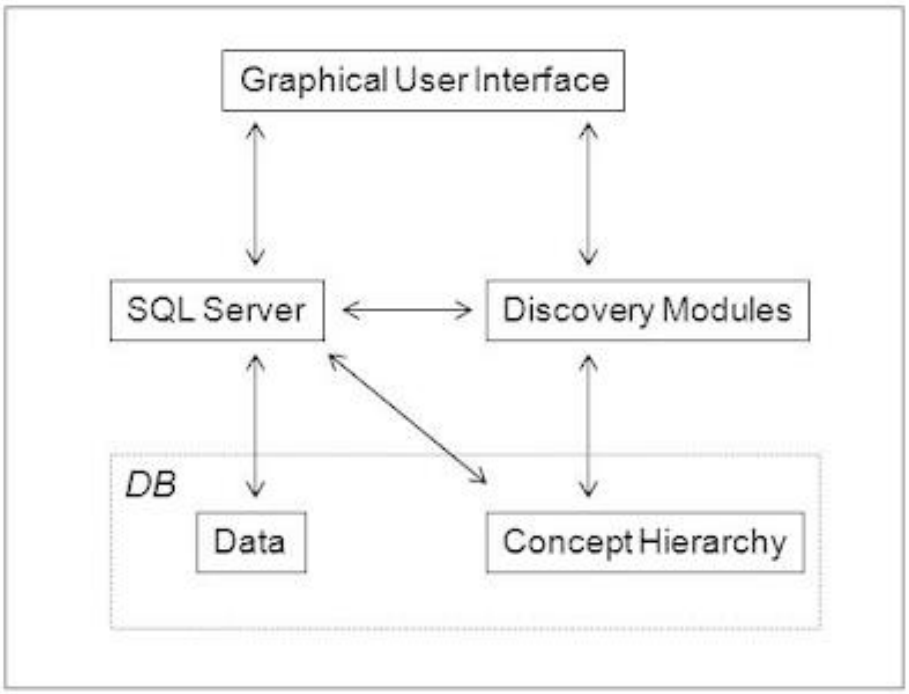
Aim: introduction to DB miner tool.

Introduction

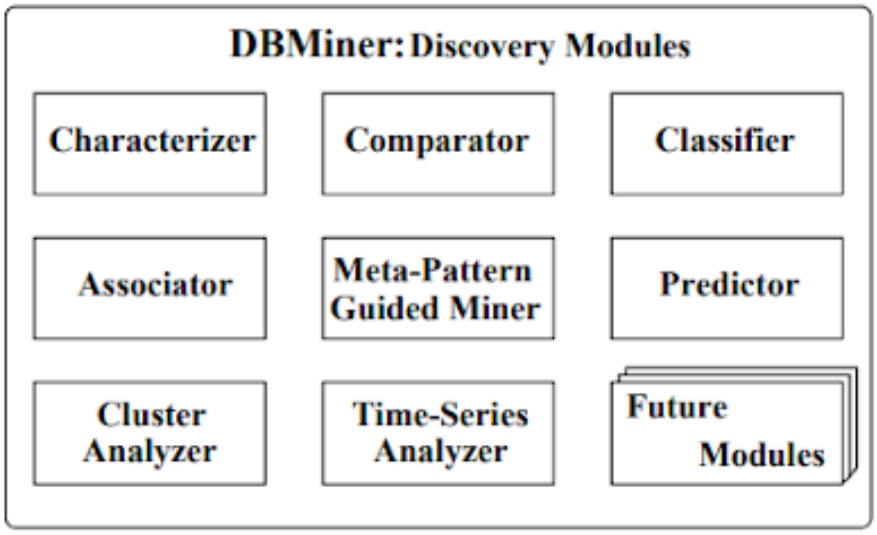
DB Miner, a data mining system for interactive mining of multiple-level knowledge in large relational databases, has been developed based on our years-of-research. The system implements a wide spectrum of data mining functions, including generalization, characterization, discrimination, association, classification, and prediction. By incorporation of several interesting data mining techniques, including attribute-oriented induction, progressive deepening for mining multiple-level rules, and meta-rule guided knowledge mining, the system provides a user-friendly, interactive data mining environment with good performance.

A data mining system, DB Miner, has been developed for interactive mining of multiple-level knowledge in large relational databases. It is based on studies of data mining techniques and experience in the development of an early system prototype, DBLearn. The system implements a wide spectrum of data mining functions, including generalization, characterization, association, classification, and prediction. By incorporation of several interesting data mining techniques, including attribute-oriented induction, statistical analysis, progressive deepening for mining multiple level knowledge, and meta-rule guided mining, the system provides a user-friendly, interactive data mining environment with good performance.

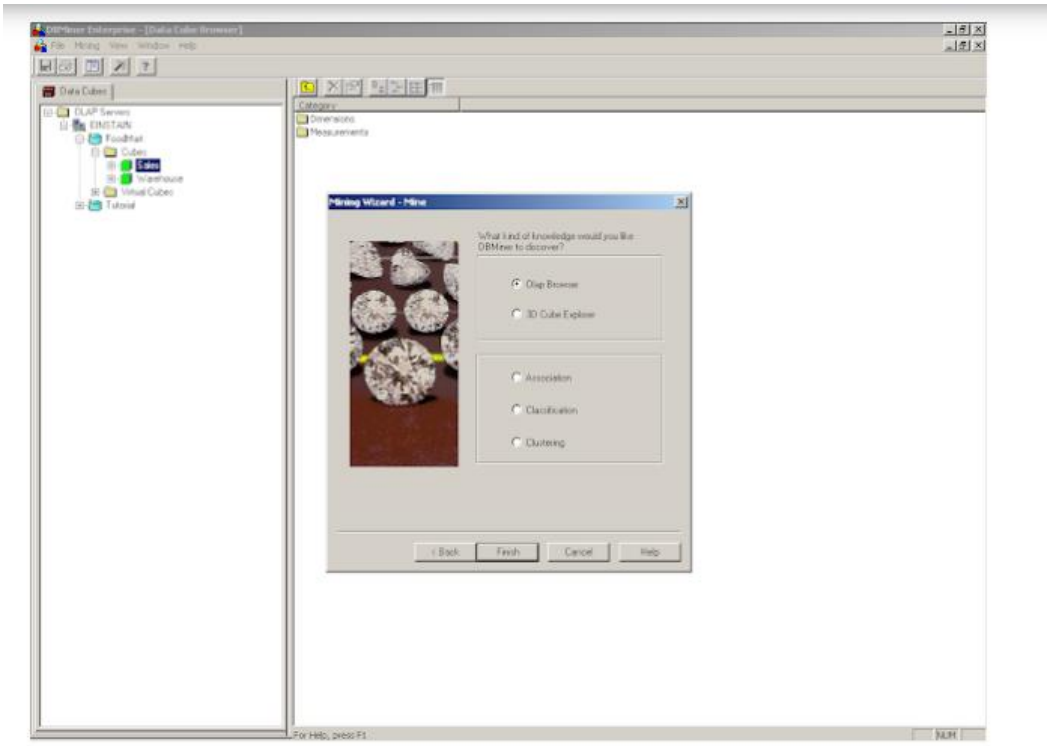
❖ General architecture of DB Miner



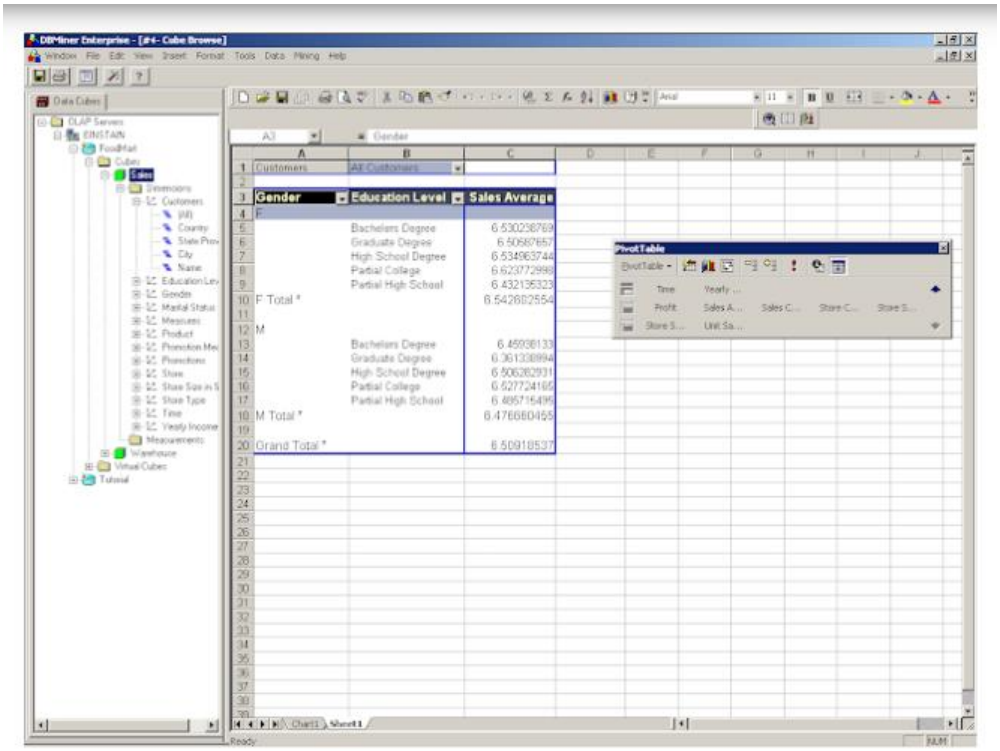
❖ Discovery Modules

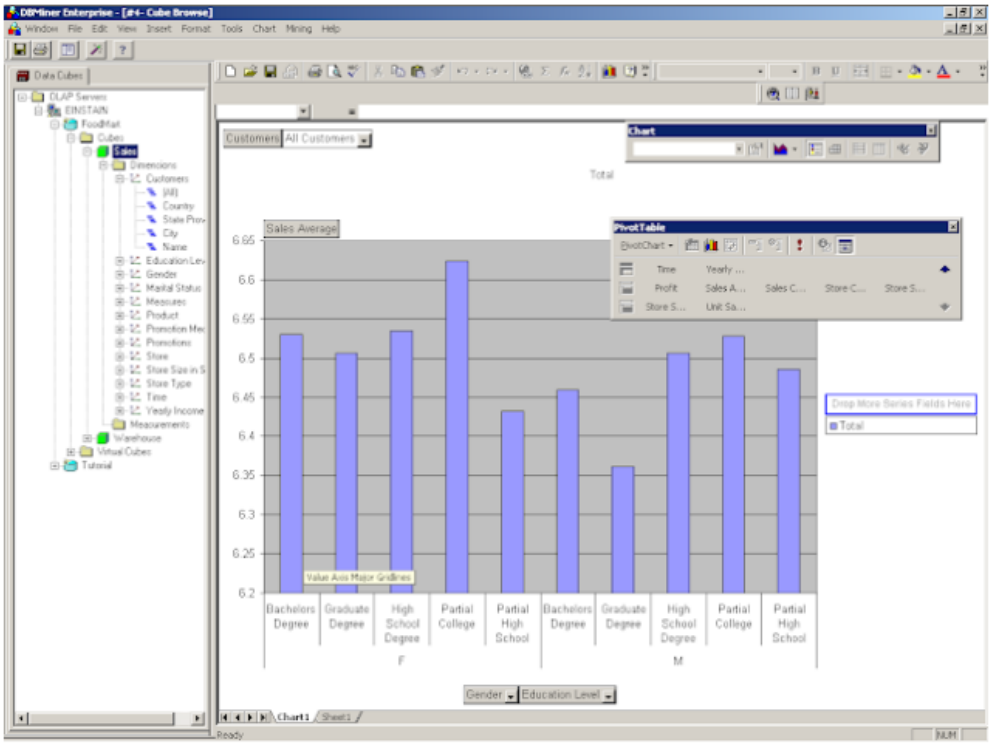


- **DB Miner user interfaces**
  - UNIX-based
  - Windows/NT-based
  - WWW/Netscape-based
- **DB Miner Wizard**

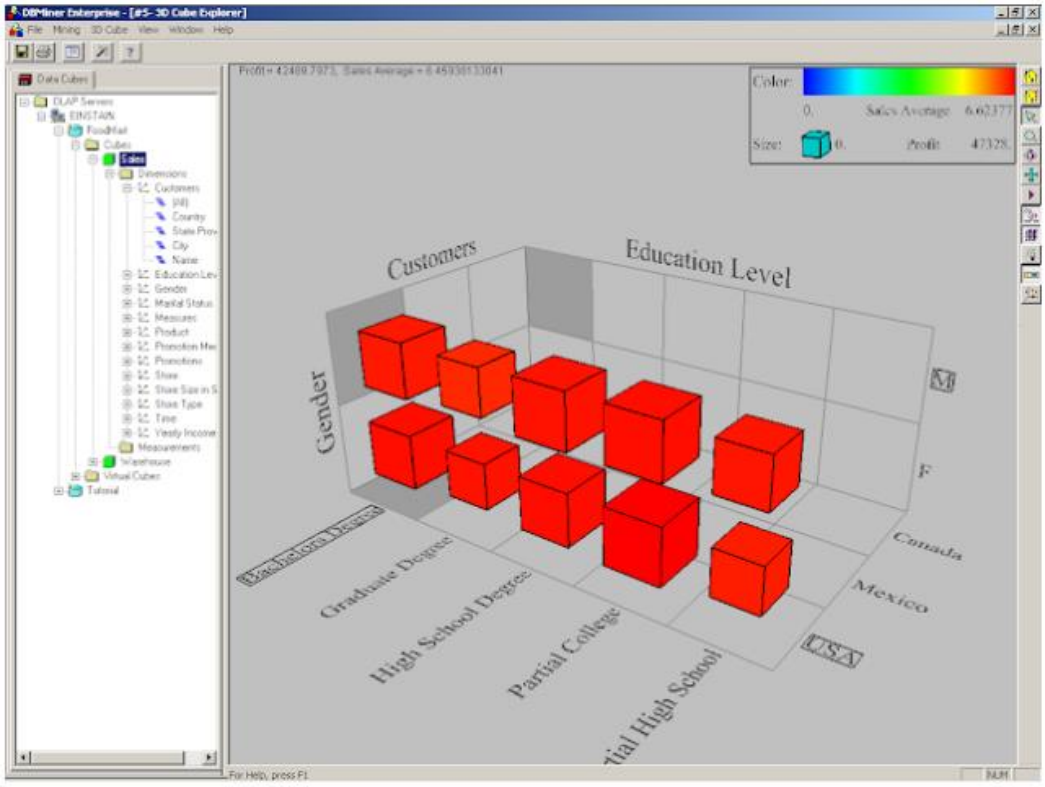


❖ OLAP Browser

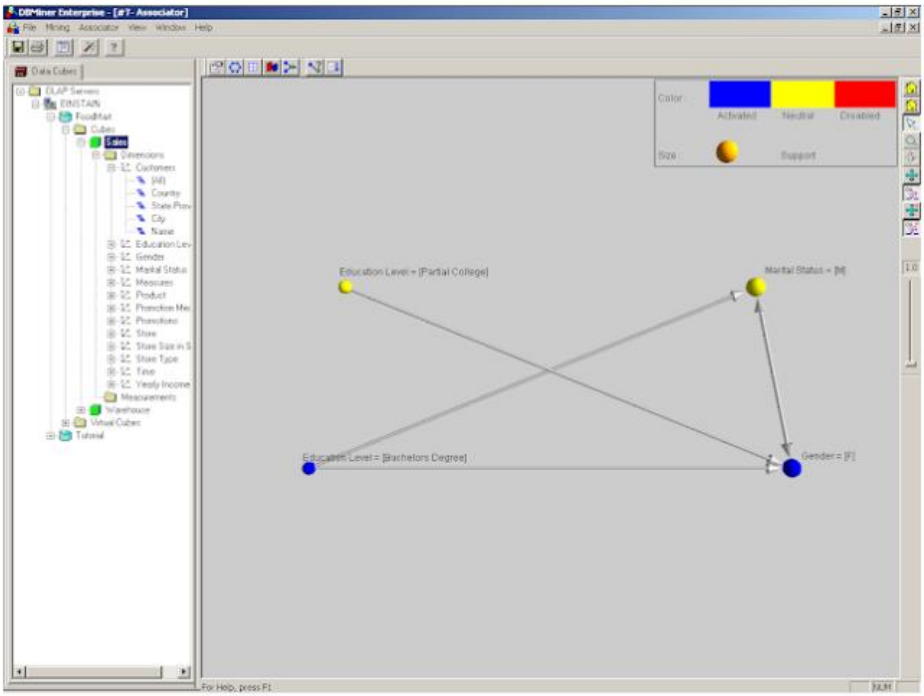




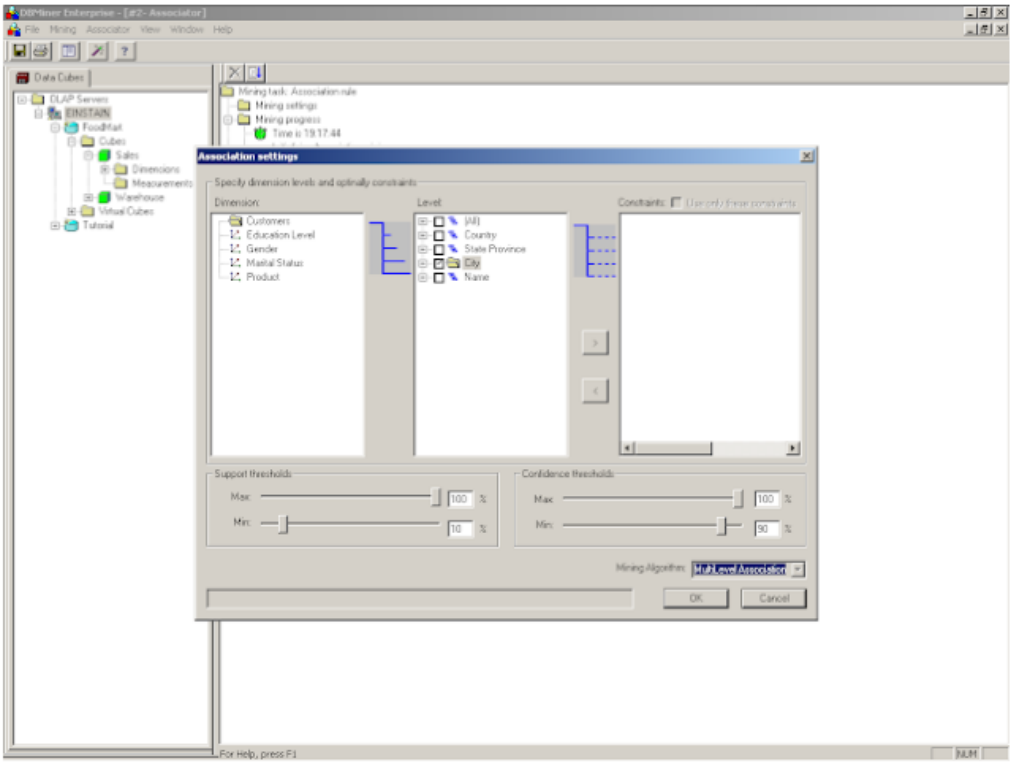
❖ 3-D Cube



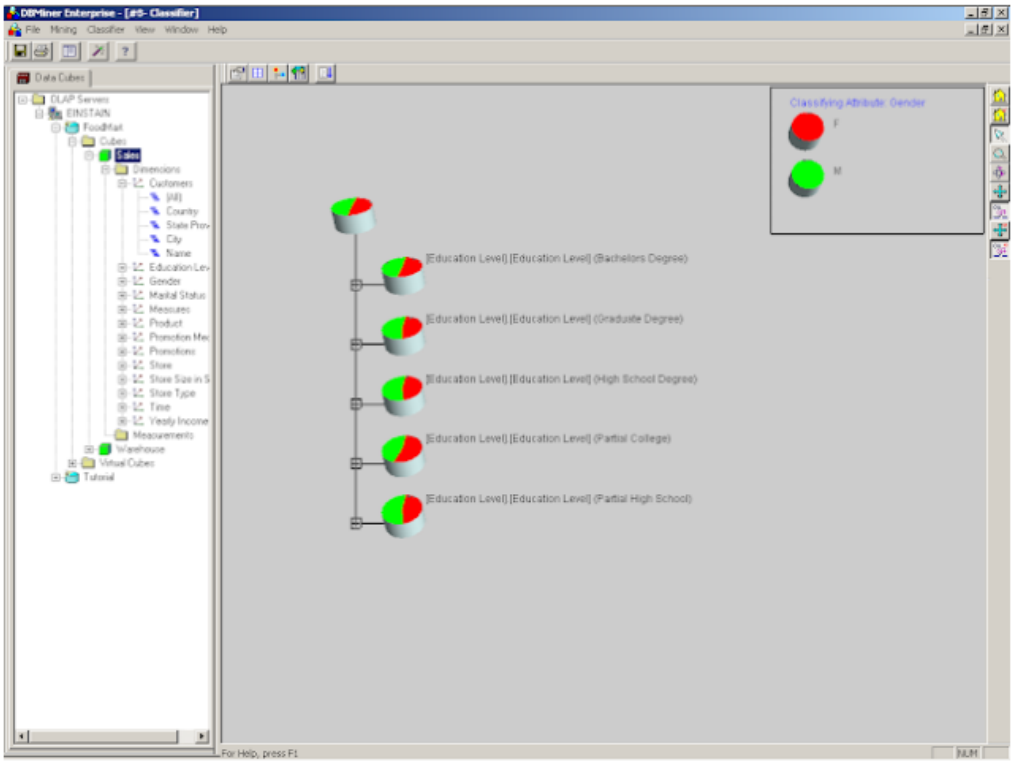
❖ Association



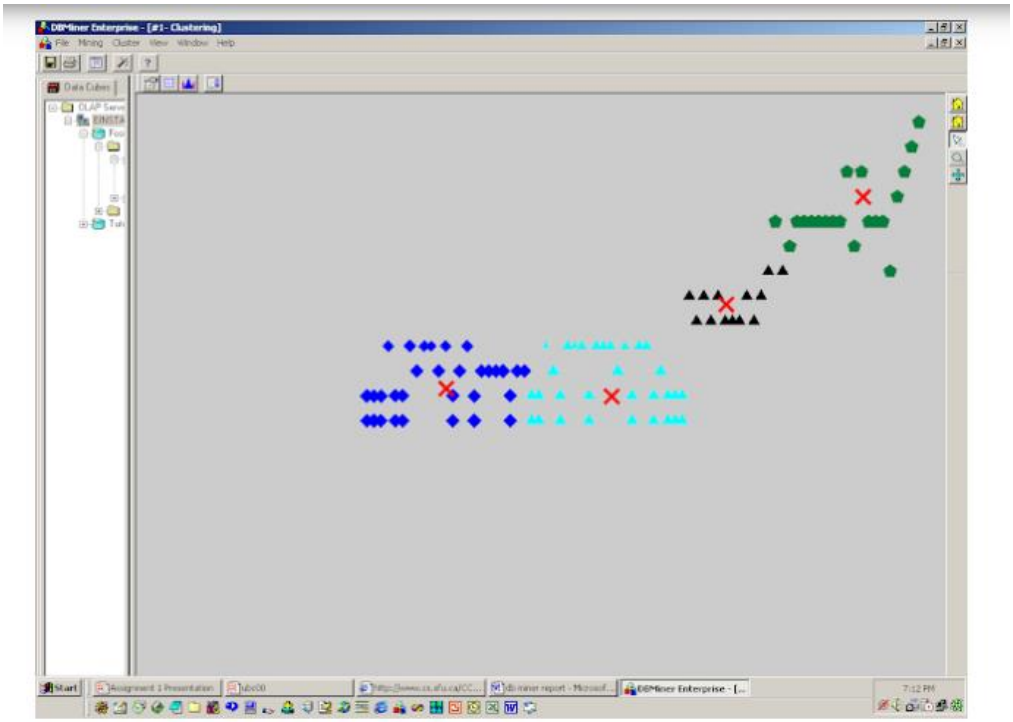
❖ Association Settings



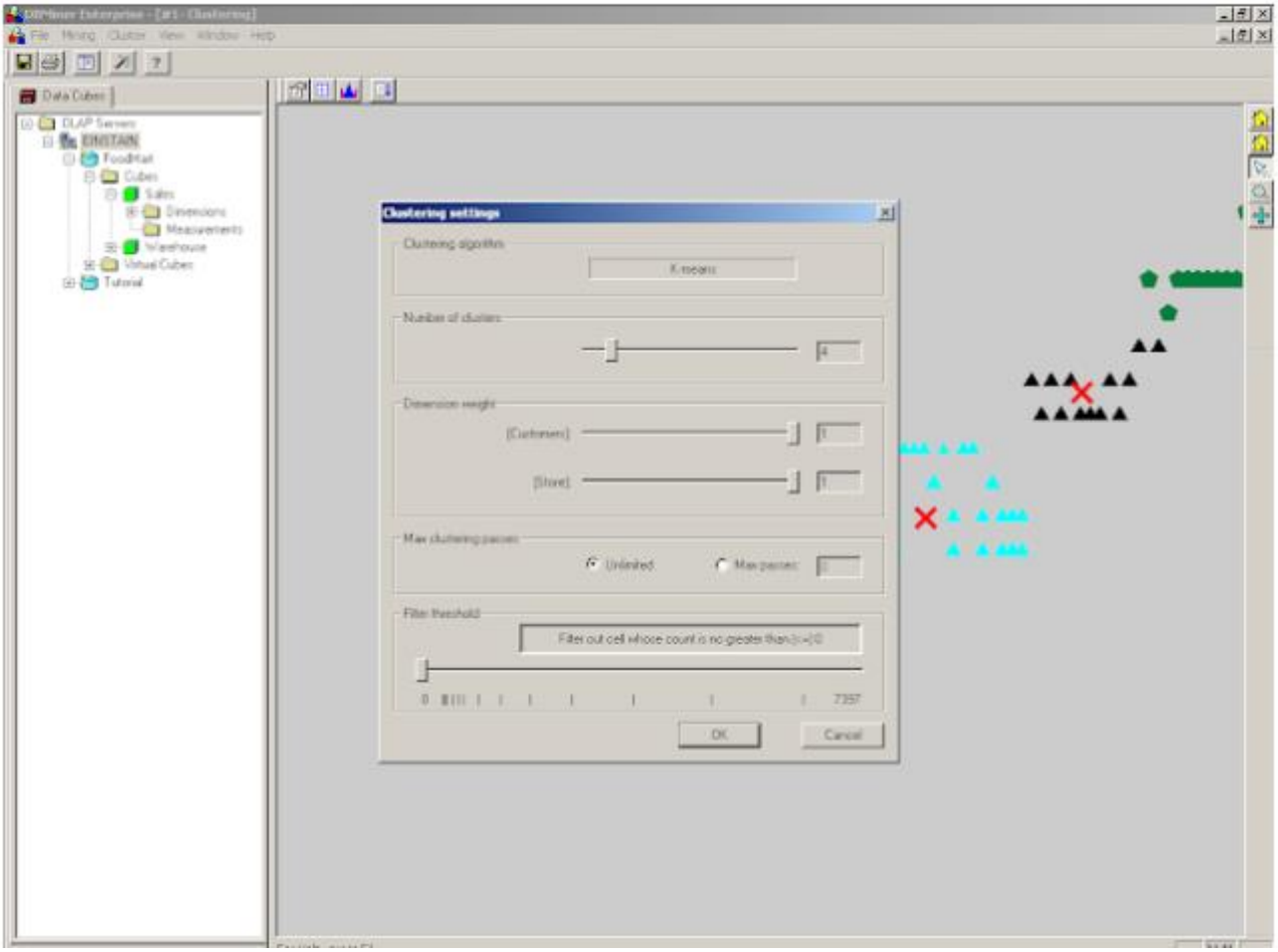
❖ Classification



❖ Classification Settings



❖ Clustering Settings


Clustering settings
Clustering algorithm
K-means
Number of clusters
Dimension weight
[Customers]
[Store]
Max clustering passes
Unlimited
Max passes:
Filter threshold
Filter out cell whose count is no greater than
OK
Cancel