# Collaborators

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **ID** | **Full name** | **Email address** |
| 1 | 20120238 | Nguyễn Ngọc Khánh Vy | [20120238@student.hcmus.edu.vn](mailto:20120238@student.hcmus.edu.vn) |
| 2 | 20120246 | Nguyễn Hoàng Anh | [20120246@student.hcmus.edu.vn](mailto:20120246@student.hcmus.edu.vn) |

# The contribution rate of each member

### Task assignment

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | | **Task description** | **Member in charge** |
| 1 | | Install WEKA & answer questions | 20120246 |
| 2 | 2.1 | Exploring Breast Cancer data set | 20120238 |
| 2.2 | Exploring Weather data set | 20120246 |
| 2.3 | Exploring Credit in Germany data set | 20120238 |
| 3 | | Implement requirements marked even numbers | 20120246 |
| Implement requirements marked odd numbers | 20120238 |

### Contribution rate

# Assess the level of project completion

# The answers to each question and requirement

## Install WEKA

### Requirement 1

* Download WEKA from their website provided.
* Verification: Screenshot that contains the ”Explorer” function in the desktop background.

Graphical user interface, application

Description automatically generated

### Requirement 2

* The selected data set is opened from the file “airline.arff”.

A screenshot of a computer

Description automatically generated

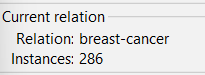
* The meaning of boxes in Preprocess tag is described as follows:
  + **Current Relation**: It gives information about the currently loaded data. Specifically, it shows the name of the relation (or table) as given in the file it was loaded from, how many instances (or rows) there are in the data set, the number of attributes and the total sum of weights across all instances. In this case, the current relation is airline\_passengers with 144 instances, 2 attributes and the sum of weights is 144.
  + **Attributes**: This is the area displaying all attributes of the selected data set, which users can view or modify them by using buttons such as Remove, All, None, Invert, and Pattern. As shown above, the airline data set has 2 attributes which are passenger\_numbers and Date.
  + **Selected Attribute**: When an attribute showed in attributes window is selected, this window will display the information of that attribute, including name, type, missing, distinct, and unique. In addition, other statistical values such as min, max, mean and sd are also shown below. In the bottom-right corner, there is a histogram as a visualization for this attribute.
* The meaning of other tags in WEKA Explorer:
  + **Classify:** This is used for building and evaluating classification models. We can choose different algorithms, parameters and evaluation methods to apply to a data set.
  + **Cluster:**
  + **Associate:**
  + **Select attributes:**
  + **Visualize:**

## Getting Acquainted With WEKA

### Exploring Breast Cancer data set

#### How many instances does this data set have?

* 286 instances



#### How many attributes does this data set have?

* 10 attributes

A picture containing graphical user interface

Description automatically generated

#### Which attribute is used for the label? Can it be changed? How?

* Class attribute is used for the label.
* It can be changed by click at Class in drop-down list under Selected attributes and choose with attributes you want to used for label.

Graphical user interface, text, application

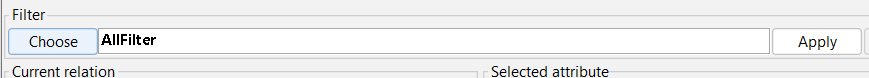
Description automatically generated

#### What is the meaning of each attribute?

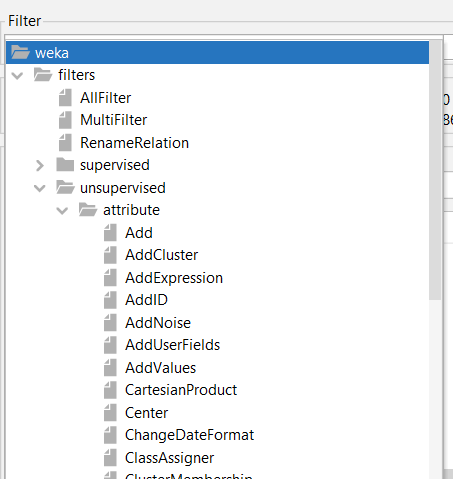
* Class: no-recurrence-events, recurrence-events
* age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
* menopause: lt40, ge40, premeno.
* tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
* inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
* node-caps: yes, no.
* deg-malig: 1, 2, 3.
* breast: left, right.
* breast-quad: left-up, left-low, right-up, right-low, central.
* irradiat: yes, no.

#### Let’s investigate the missing value status in each attribute and describe in general ways to solve the problem of missing values.

* Dataset have 2 attributes have missing value: node-caps with 8 missing, breast-quad with only one missing.
* In Weka, to fill missing value automatically, you can choose ‘Choose’ button in Filter area.



* Expand folder to attribute to find ReplaceMissingValues/ ReplaceMissingWithUserConstant.



Graphical user interface, application, table, Word

Description automatically generated

* Click apply to fill missing value.

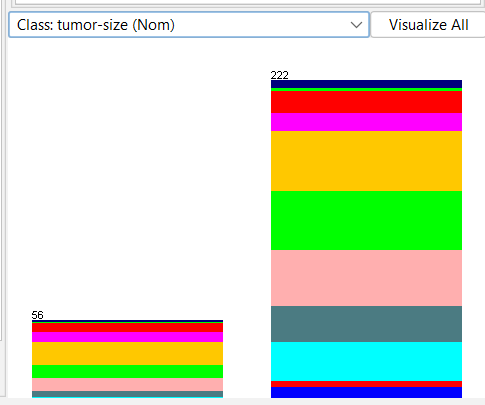
#### Let’s propose solutions to the problem of missing values in the specific attribute.

#### Let’s explain the meaning of the chart in the WEKA Explorer. Setting the title for it and describing its legend.

* The graph here is a histogram - showing the frequency of occurrence of samples (instances) with corresponding classes (Class) for continuous attributes, and a bar plot with categorical properties (similar functionality).
* If considering a single column, the blue and red colors represent the proportion of instances with the label value of that column that have the corresponding yes, no classes.

Chart, bar chart

Description automatically generated



### Exploring Weather data set

### Exploring Credit in Germany data set

#### What is the content of the comments section in credit-g.arff (when opened with any text editor) about?

* Content of comments section have 278 lines, about description of German credit dataset.
* Content include:
  + Title: German Credit data
  + Source Information: information of author
  + Number of instances, number of attributes in 2 dataset (german and german.numer)
  + Description for each attributes in german dataset
  + Cost matrix
  + Relabeled law for 14 attributes: checking\_status, credit\_history, ...

#### How many samples does the data set have?

* 1000 samples

Graphical user interface, text

Description automatically generated

#### How many attributes?

* 21 attributes

A picture containing table

Description automatically generated

#### Describe any five attributes (must have both discrete and continuous attributes).

* Age: continuous attribute, age in year, maximum is 75, minimum is 19, with mean is 35.546 and standard ... is 11.375.
* Property\_magnitude: discrete attribute, property, sample data: ‘real estate’, ‘car’, ‘life insurance’, ‘no known property’
* Residence\_since: continuous attribute, present residence since, max is 4, min is 1, with mean is 2.845, standard ... is 1.104.
* Credit\_history: discrete attribute, credit history, sample data: ‘ no credits/all paid’, ‘all paid’, ‘existing paid’, ‘delayed previously’, ‘critical/other existing credit’.
* Employment: discrete attribute, present employment since, sample data: ‘unemployment’, ‘<1’, ‘1<=X<4’, ‘4<=X<7’, ‘>=7’.

#### Which attribute is used for the label?

* The name of the class attribute is class.
* Evaluation of the distribution: skewed to one class (good). The good class has twice as many instances as the bad class.

#### Let’s describe the distribution of continuous attributes. (Left skewed or right skewed ?)

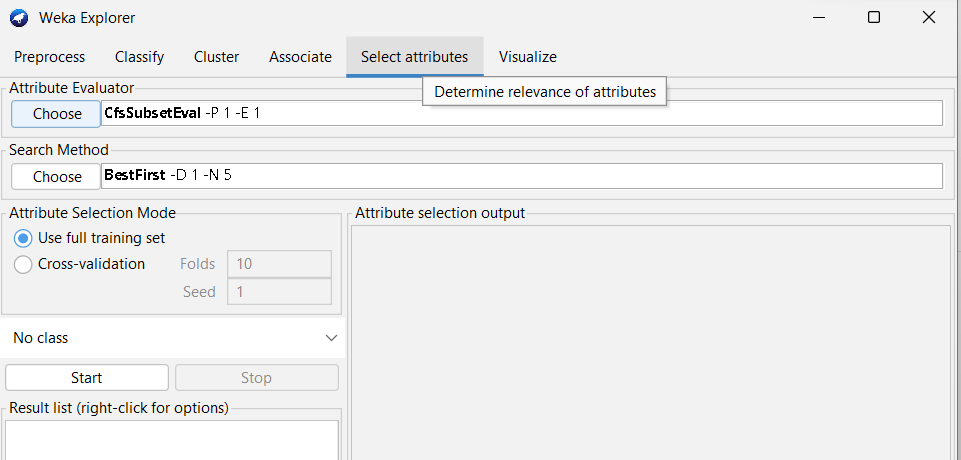
#### Let’s explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.

#### Let’s move to the Select attributes tag. Describe all of the options for attribute selection.

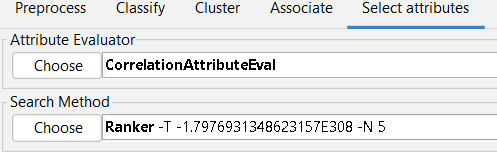
* CfsSubsetEval: Evaluates the importance of a subset of attributes by looking at the individual predictability of each feature along with the degree of redundancy between them (the set of attributes is highly correlated with the class attribute and intercorrelate is preferred).
* ClassifierAttributeEval: evaluates the importance of an attribute using a user-specified classifier.
* CorrelationAttributeEval: evaluates the importance of an attribute by measuring the correlation between it and the class using the Pearson correlation coefficient.
* ClassifierSubsetEvaluator: evaluates attribute values on training data or a separate set of cross validation (hold-out).
* PrincipleComponents: use PCA (principal components analysis) algorithm to reduce the number of dimensions (columns) of the data.
* GainRatioAttributeEval: Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.
* IfnoGainAttributeEval: Evaluates the worth of an attribute by measuring the information gain with respect to the class.
* OneRAttributeEval: Evaluates the worth of an attribute by using the OneR classifier.
* ReliefFAttributeEval: Evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class.
* SymmetricalUncertAttributeEval: Evaluates the worth of an attribute by measuring the symmetrical uncertainty with respect to the class.
* WrapperSubsetEval: Evaluates attribute sets by using a learning scheme.

#### Which options should be used to select the 5 attributes with the highest correlation?(Step-by-step description, with step-by-step photos and final results)

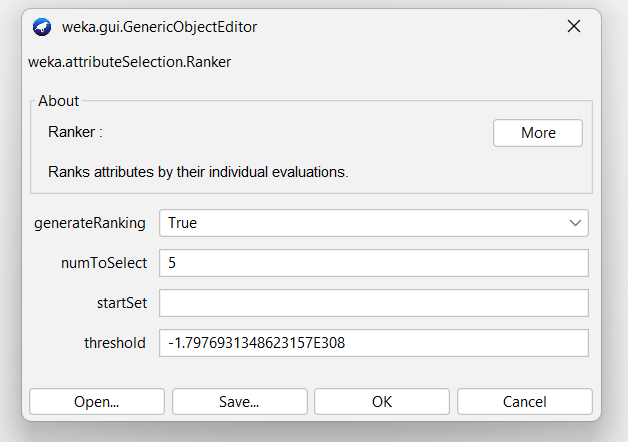
* To select the 5 attributes with the highest correlation with the class attribute, use the CorrelationAttributeEval filter.
  + Step 1: Select the Select attribute tab.



* + Step 2: Click choose in Attribute Evaluator, select CorrelationAttributeEval.



* + Step 3: Click on the text box with the word Ranker to refine the attribute search algorithm. Enter 5 in the numToSelect field to select the 5 attributes that have the highest correlation with the class attribute. Click OK.



* + Step 4: Select the class attribute in the drop-down list below the Search Method field.

Graphical user interface, text, application

Description automatically generated

* + Step 5: Click Start and wait for the algorithm to finish running. Observe the results in the text box on the right.

Graphical user interface, text, application

Description automatically generated

* + Thus the attributes selected by this algorithm are:
    - checking\_status: 0.233
    - duration: 0.215
    - credit\_amount: 0.155
    - savings\_status: 0.132
    - housing: 0.121

## Preprocessing Data in Python

# References