**VIETNAM NATIONAL UNIVERSITY**

**UNIVERSITY OF SCIENCE**

**DEPARTMENT OF COMPUTER SCIENCE**

**FACULTY OF INFORMATION TECHNOLOGY**

\*\*\*



**LAB 01: DATA PREPROCESSING AND DATA EXPLORATION**

|  |  |
| --- | --- |
| Lecturer: | Lê Hoài Bắc  Nguyễn Bảo Long  Nguyễn Ngọc Đức |
|  |  |
| Class: | 21CLC9 |
|  |  |
| Student: | 21127007 – Nguyễn Quốc Anh  21127418 – Phạm Phi Sơn |

*Ho Chi Minh City, November 2023*

**Table of Contents**

[0. Collaborators 3](#_Toc149601925)

[1. The contribution rate of each member 3](#_Toc149601926)

[1.1.1. Task assignment 3](#_Toc149601927)

[1.1.2. Contribution rate 3](#_Toc149601928)

[2. Assess the level of project completion 3](#_Toc149601929)

[3. The answers to each question and requirement 4](#_Toc149601930)

[3.1. Install WEKA 4](#_Toc149601931)

[3.1.1. Requirement 1 4](#_Toc149601932)

[3.1.2. Requirement 2 4](#_Toc149601933)

[3.2. Getting Acquainted With WEKA 4](#_Toc149601934)

[3.2.1. Exploring Breast Cancer data set 4](#_Toc149601935)

[3.2.2. Exploring Weather data set 9](#_Toc149601936)

[3.2.3. Exploring Credit in Germany data set 11](#_Toc149601937)

[3.3. Preprocessing Data in Python 17](#_Toc149601938)

[3.3.1. Extract columns with missing values 17](#_Toc149601939)

[3.3.2. Count the number of lines with missing data. 17](#_Toc149601940)

[3.3.3. Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute). 18](#_Toc149601941)

[3.3.4. Deleting rows containing more than a particular number of missing values (Example: delete rows with the number of missing values is more than 50% of the number of attributes). 18](#_Toc149601942)

[3.3.5. Deleting columns containing more than a particular number of missing values (Example: delete columns with the number of missing values is more than 50% of the number of samples). 18](#_Toc149601943)

[3.3.6. Delete duplicate samples. 18](#_Toc149601944)

[3.3.7. Normalize a numeric attribute using min-max and Z-score methods. 18](#_Toc149601945)

[3.3.8. Performing addition, subtraction, multiplication, and division between two numerical attributes 18](#_Toc149601946)

[4. References 18](#_Toc149601947)

# Collaborators

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **ID** | **Full name** | **Email address** |
| 1 | 21127007 | Nguyễn Quốc Anh | nqanh212@clc.fitus.edu.vn |
| 2 | 21127418 | Phạm Phi Sơn | ppson21@clc.fitus.edu.vn |

# The contribution rate of each member

### Task assignment

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | | **Task description** | **Member in charge** |
| 1 | | Install WEKA & answer questions | 21127007 |
| 2 | 2.1 | Exploring Breast Cancer data set | 21127007 |
| 2.2 | Exploring Weather data set | 21127418 |
| 2.3 | Exploring Credit in Germany data set | 21127418 |
| 3 | | Implement requirements 1, 2, 3, 5, 7 | 21127007 |
| Implement requirements 4, 6, 8 | 21127418 |

### Contribution rate

Nguyễn Quốc Anh: 50%

Phạm Phi Sơn: 50%

# Assess the level of project completion

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | | **Task description** | **Complete percentage** |
| 1 | | Install WEKA & answer questions | 100% |
| 2 | 2.1 | Exploring Breast Cancer data set | 100% |
| 2.2 | Exploring Weather data set | 100% |
| 2.3 | Exploring Credit in Germany data set | 100% |
| 3 | | Preprocessing data in Python | 100% |

**Total percentage of requirements completed: 100%**

# The answers to each question and requirement

## Install WEKA

### Requirement 1

A screenshot of a computer

Description automatically generated

### Requirement 2

* ***Current relation***: Shows the name of the database that is currently loaded and number of instances, attributes, sum of weight.
  + Instances: Number of rows in the database.
  + Sum of weight: The same as Instances.
  + Attributes: Number of fields in the database.
* ***Attributes***: Displays the various fields in the database and the “Remove” button at the bottom to remove an attribute, select and click the button.
* ***Selected Attributes***: You can observe the following information:
* The name and the type of the attribute are displayed.
* The type for the attribute.
* The number of Missing values.
* Distinct values mean that there is no unique value.
* The table underneath shows information about the values.
* It also shows the count and weight in terms of a percentage for each nominal value.

## Getting Acquainted With WEKA

### Exploring Breast Cancer data set

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

A white rectangular frame with a black border

Description automatically generated

Instances: 286

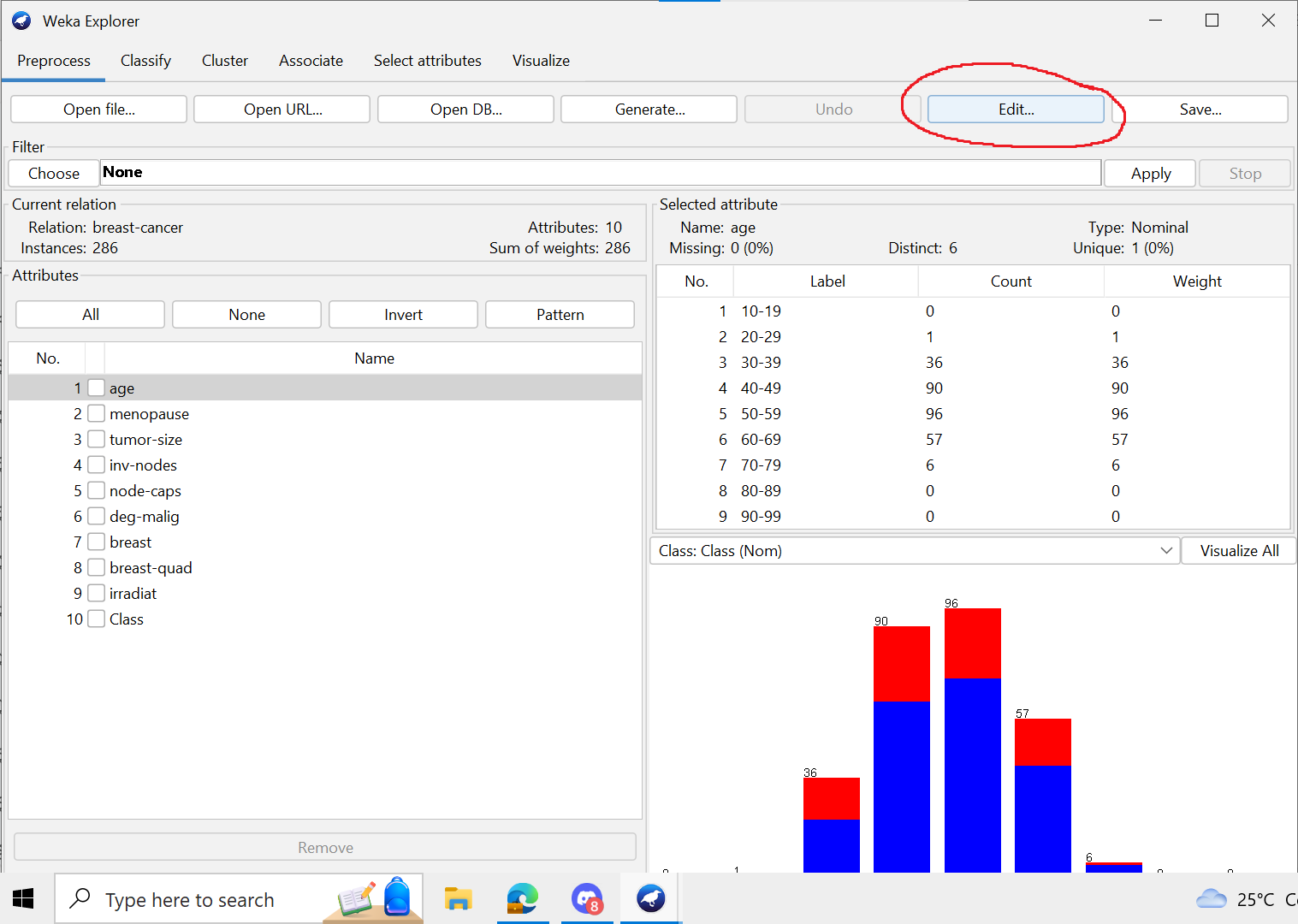
#### How many attributes does this data set have?

A white rectangular frame with a black border

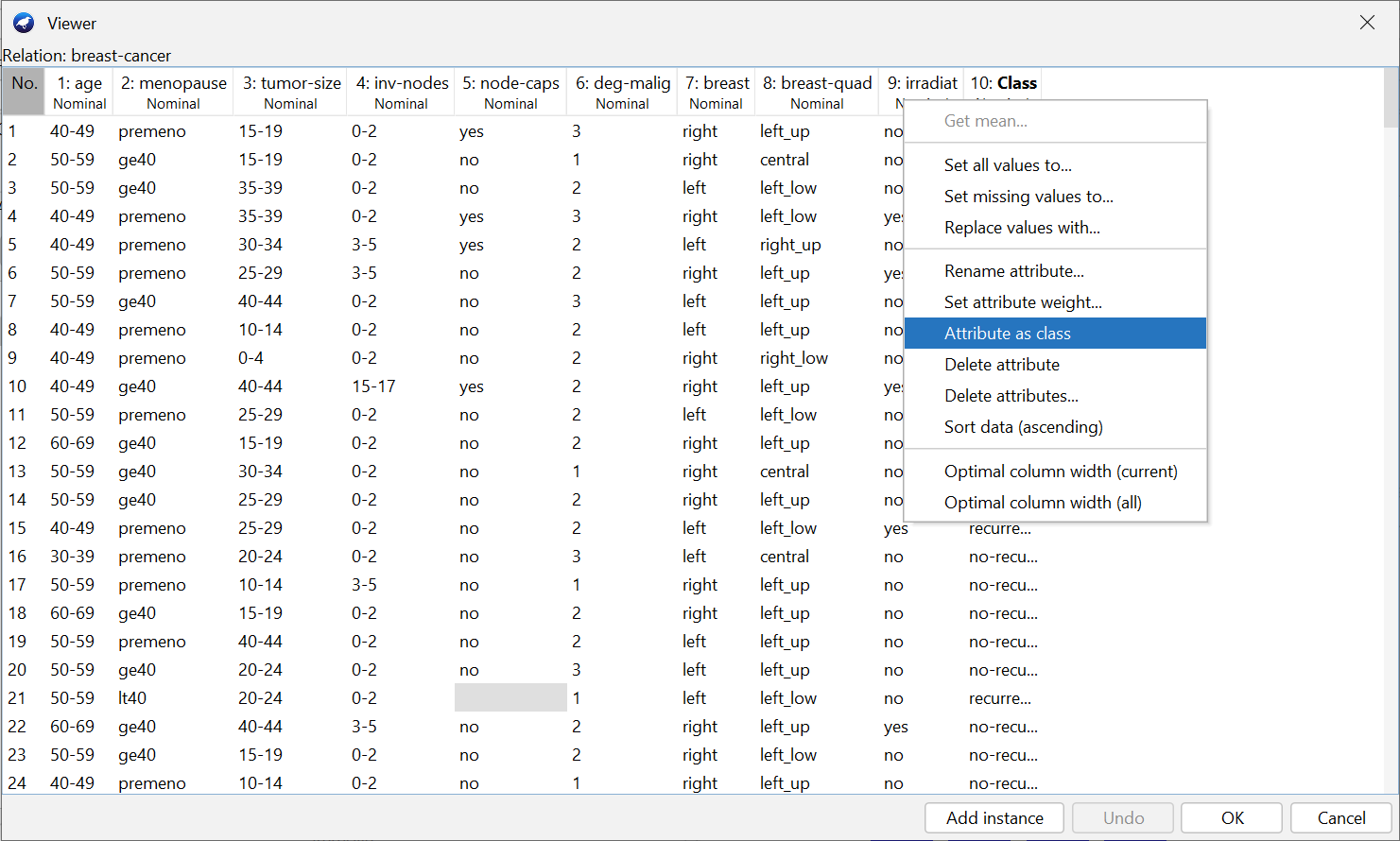
Description automatically generated

Attributes: 10

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



In Weka, the attribute used for the label is typically the last attribute in the dataset.



2 ways to change the attribute used for the label, you can follow these steps:

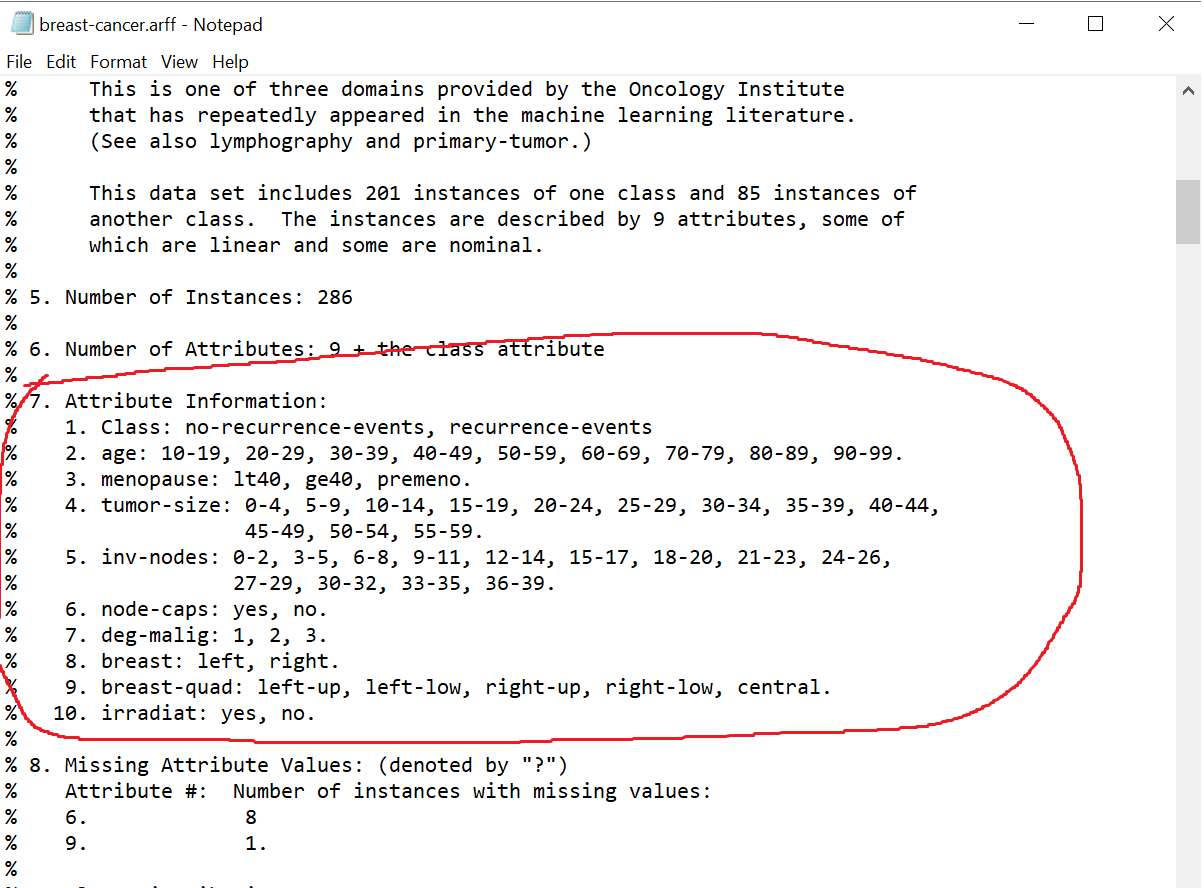
* S1:
* Open the dataset in explorer.
* Perform any required filtering (if necessary).
* Click the “Edit” button on the top panel.
* Right-click on the respective attributes you wish to assign as a label.
* Once you’ve implemented your changes, click “OK”.
* S2:
* Change it directly in the drop-down list below the “Selected attributes” region.

A screenshot of a computer

Description automatically generated

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

You can see the meaning by open the file with notepad



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

A screenshot of a computer

Description automatically generated

Investigate the missing value status in each attribute:

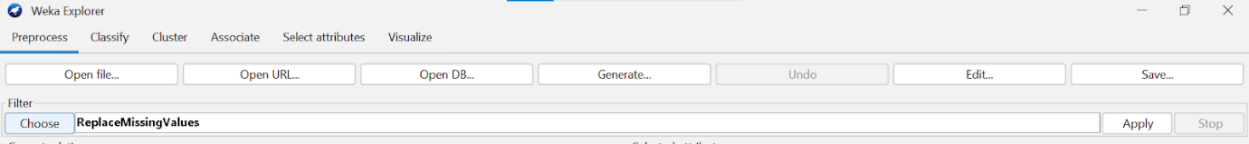
* In WEKA Explorer, you can check the status of missing values in each attribute by looking at the “***Preprocess***” tab. The “***Attributes***” section provides a list of all attributes in your dataset, and for each attribute, it shows the number of missing values.

Solve the problem of missing values:

* Ignore the records with missing values.
* Replace them with a global constant.
* Fill in missing values manually based on your domain ***knowledgeReplace*** them with the variable mean (if numerical) or the most frequent value (if categorical)
* Use a filter like ***NumericalCleaner*** or ***RemoveWithValues***.

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

* For the node-caps attribute we can use the “***ReplaceMissingValues***” filter. This filter can mark or remove missing values based on certain criteria.



A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

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

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

A screenshot of a computer

Description automatically generated

* The chart in the WEKA Explorer is a scatter plot that shows the relationship between two attributes of the data set. The points on the plot represent the instances of the data set, and they are colored according to their class values. You can see the legend of the colors at the bottom of the plot.
* The scatter plot can help you visualize the distribution, correlation, and outliers of the data. You can also zoom in and out by dragging a rectangle on the plot or select a subset of instances by clicking on them.

### Exploring Weather data set

#### How many attributes does this data set have? How many samples? Which attributes have data type that is categorical? Which attributes have a data type that is numerical? Which attribute is used for the label?

A screenshot of a computer

Description automatically generated

There are 5 attributes. Number of instances representing number of samples and is 14. Attribute ***outlook***, ***windy*** and ***play*** are categorical. Attribute ***temperature*** and ***humidity*** are numerical. In Weka, the attribute used for the label is typically the last attribute in the dataset, but we can easily change the label by choosing the attribute in the class box above the chart.

#### Let’s list a five-number summary of two attributes temperature and humidity. Does WEKA provide these values?

Five-number summary of two attributes consists of: ***Maximum*** value, ***Minimum*** value, first quartile (***Q1***), Median (***Q2***), third quartile (***Q3***). However, Weka does not provide all these 5 values since ***Q1*** and ***Q3*** are not shown in the selected attribute window.

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

A screenshot of a graph

Description automatically generated

Each chart represents the percentage of the yes-no values to each value of an attribute. The chart for numerical attributes is continuous while it is discrete for categorical attributes.

#### Let’s move to the Visualize tag. What’s the name of this chart? Do you think there are any pairs of different attributes that have correlated?

A screenshot of a computer screen

Description automatically generated

The name of the chart is data visualization scatter plot. There are two correlated pairs in the dataset:

1. ***Temperature*** - ***humidity***: The higher the temperature, the lower the humidity.
2. ***Temperature*** - ***outlook***: the temperature depends on the value of the attribute outlook.

### 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?

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

When opening that file with any text editor, we can easily specify the structure with the detail presented within. The content of the comments section includes 7 parts and 2 short explaining paragraphs by the followed order:

* Title of the dataset.
* Source information of the dataset.
* Number of instances.
* The first paragraph provides more information about the dataset. The next paragraph provides more information about attributes.
* Number of attributes for the main dataset and the sub dataset which is used for the numerical required dataset algorithms.
* Attribute description for the dataset
* Cost matrix
* Relabeled values and samples of the dataset.

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

A white rectangular object with a black border

Description automatically generated

The dataset “Credit in Germany data set” has 1000 samples with 21 attributes.

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

1. “***checking\_status***”: discrete attribute, account status checking on number of Deutsche Mark units per salary assignments for at least 1 year. Including labels: “<0”, “0<=X<200”, “>=200”, “no checking”.

A screenshot of a computer

Description automatically generated

1. “***duration***”: continuous attribute, duration until expiration in month. Minimum: 4, maximum: 72, mean: 20.903, standard deviation: 12.059.

A screenshot of a computer

Description automatically generated

1. “***credit\_history***”: discrete attribute, describe the current credit status of the customer. Including:  “no credits/all paid”, “all paid”, “existing paid”, “delayed previously”, “critical/other existing credit”.

A screenshot of a graph

Description automatically generated

1. “***purpose***”: discrete attribute, purposes of using credit card, including many values.

A screenshot of a computer

Description automatically generated

1. “**credit\_amount**”: continuous attribute, the amount of the credit. Minimum: 250, maximum: 18424, mean: 3271.258, standard deviation: 2822.737.

A screenshot of a computer

Description automatically generated

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

The attribute is used for the label is the name of class.

A screenshot of a computer

Description automatically generated

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

“***duration***”: left skewed.

A graph with numbers and squares

Description automatically generated

“***credit\_amount***”: left skewed.

A graph with numbers and a line

Description automatically generated

“***installment\_commitment***”: right skewed.

A graph with numbers and a line

Description automatically generated

“***resident\_since***”: right skewed.

A graph with numbers and a bar

Description automatically generated

“***age***”: left skewed.

A graph with numbers and a bar

Description automatically generated

“***existing\_credits***”: left skewed.

A graph with numbers and a square

Description automatically generated

“***num\_dependents***”: left skewed.

A graph with numbers and a line

Description automatically generated with medium confidence

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

Each chart represents the percentage of the yes-no values to each value of an attribute. The chart for numerical attributes is continuous while it is discrete for categorical attributes.

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

“***CfsSubsetEval***”: Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.

“***ClassifierAttributeEval***”: Evaluates the worth of an attribute by using a user-specified classifier.

“***ClassifierSubsetEval***”: Evaluates attribute subsets on training data or a separate hold out testing set.

“***CorrelationAttributeEval***”: Evaluates the worth of an attribute by measuring the correlation (Person’s) between it and the class.

“***GainRatioAttributeEval***”: Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.

“***InfoGainAttributeEval***”: Evaluates the worth of an attribute by measuring the information gained with respect to the class.

“***OneRAttributeEval***”: Evaluates the worth of an attribute by using the OneR classifier.

“***PrincipalComponents***”: Performs a principal components analysis and transformation of data.

“***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 class 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)

Step 1: Select the “***Select attribute***” tab.

A screenshot of a computer

Description automatically generated

Step 2: Click choose under “***Attribute Evaluator***”, select “***CorrelationAttributeEval***”

A screenshot of a computer

Description automatically generated

Step 3: Click on the text box with the word “***Ranker***”. Change to **5** in “***numToSelect***” field to select the 5 attributes that have the highest correlation with the class.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

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

A screen shot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generated

## Preprocessing Data in Python

### Extract columns with missing values

* Command line: py extract\_missing.py --input=<input> --output=<output>
* Example:

py extract\_missing.py --input=data.csv --output=output.csv

### Count the number of lines with missing data.

* Command line: py count\_missing\_line.py --input=<input>
* Example:

py count\_missing\_line.py --input=data.csv

### Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute).

* Command line: py Fill\_missing.py --input=<input> --columns= <\*> --method=<\*\*> --output=<output>

(\*): columns separated by ‘,’ character.

(\*\*): including mean, median, and mode.

* Example:

py Fill\_missing.py --input=data.csv --columns= Id,MSSubClass --method=mean --output=output.csv

### Deleting rows containing more than a particular number of missing values (Example: delete rows with the number of missing values is more than 50% of the number of attributes).

* Command line:
* Example:

### Deleting columns containing more than a particular number of missing values (Example: delete columns with the number of missing values is more than 50% of the number of samples).

* Command line: py delete\_columns.py --input=<input> --output=<output>
* Example:

py delete\_columns.py --input=data.csv --output=output.csv

### Delete duplicate samples.

* Command line:
* Example:

### Normalize a numeric attribute using min-max and Z-score methods.

* Command line: py normalize.py --input=<input> --columns= <\*> --method=<\*\*> --output=<output>

(\*): multiple columns separated by ‘,’ character.

(\*\*): including min-max and z-score.

* Example:

py normalize.py --input=data.csv --columns= Id,MSSubClass --method=min-max --output=output.csv

### Performing addition, subtraction, multiplication, and division between two numerical attributes

* Command line:
* Example:

# References

Lecture slides provided via <https://courses.fit.hcmus.edu.vn>

[Working with csv files in Python - GeeksforGeeks](https://www.geeksforgeeks.org/working-csv-files-python/)

Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.