**VIETNAM NATIONAL UNIVERSITY**

**UNIVERSITY OF SCIENCE**

**DEPARTMENT OF COMPUTER SCIENCE**

**FACULTY OF INFORMATION TECHNOLOGY**

\*\*\*



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

Instructor: Nguyễn Thị Thu Hằng ([ntthuhang0131@gmail.com](mailto:ntthuhang0131@gmail.com))

Class: 20\_21

Academic year: 2022-2023

Group of students:

1. Nguyễn Ngọc Khánh Vy ID: 20120238
2. Nguyễn Hoàng Anh ID: 20120246

*Ho Chi Minh City, March 2023*

**Table of Contents**

[0. Collaborators 3](#_Toc130733130)

[1. The contribution rate of each member 3](#_Toc130733131)

[1.1.1. Task assignment 3](#_Toc130733132)

[1.1.2. Contribution rate 3](#_Toc130733133)

[2. Assess the level of project completion 3](#_Toc130733134)

[3. The answers to each question and requirement 4](#_Toc130733135)

[3.1. Install WEKA 4](#_Toc130733136)

[3.1.1. Requirement 1 4](#_Toc130733137)

[3.1.2. Requirement 2 4](#_Toc130733138)

[3.2. Getting Acquainted With WEKA 5](#_Toc130733139)

[3.2.1. Exploring Breast Cancer data set 5](#_Toc130733140)

[3.2.2. Exploring Weather data set 9](#_Toc130733141)

[3.2.3. Exploring Credit in Germany data set 11](#_Toc130733142)

[3.3. Preprocessing Data in Python 18](#_Toc130733143)

[3.3.1. Extract columns with missing values 18](#_Toc130733144)

[3.3.2. Count the number of lines with missing data. 19](#_Toc130733145)

[3.3.3. Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute). 19](#_Toc130733146)

[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). 19](#_Toc130733147)

[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). 19](#_Toc130733148)

[3.3.6. Delete duplicate samples. 19](#_Toc130733149)

[3.3.7. Normalize a numeric attribute using min-max and Z-score methods. 20](#_Toc130733150)

[3.3.8. Performing addition, subtraction, multiplication, and division between two numerical attributes 20](#_Toc130733151)

[4. References 21](#_Toc130733152)

# 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

Nguyễn Hoàng Anh: 50%

Nguyễn Ngọc Khánh Vy: 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

* Download WEKA from the 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

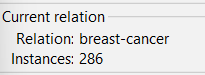
* The meaning of boxes in Preprocess tag is described as follows:
  + **Current Relation**: It gives information about the currently loaded data. Specifically, it shows the name of the relation (or table) as given in the file it was loaded from, how many instances (or rows) there are in the data set, 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, and the number of missing, distinct, and unique values. In addition, other statistical values depending on the data type of that attribute are also shown below (min, max, mean, sd for numeric data and frequency of distinct values for categorical data). 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:** This is similar to the Classify tab but applied to clustering models.
  + **Associate:** This is used for automatically finding associations in a dataset.
  + **Select attributes:** This is used for selecting the most relevant attributes in the data.
  + **Visualize:** This is used for reviewing pairwise scatterplot matrix of each attribute plotted against every other attribute in the loaded dataset. It is useful to get an idea of the shape and relationship of attributes that may aid in data filtering.

## 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 which attributes we want to use for label.

Graphical user interface, text, application

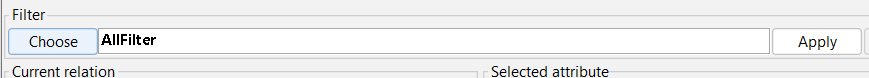
Description automatically generated

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

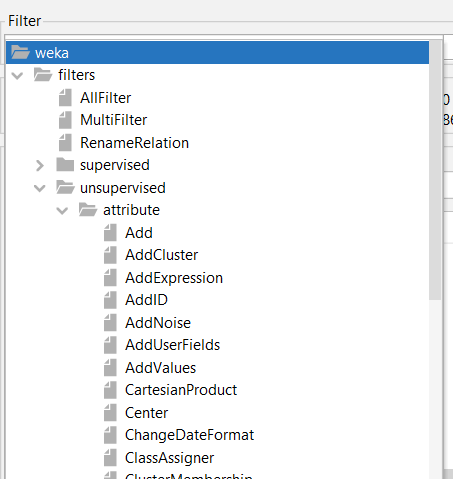
* Class: no-recurrence-events, recurrence-events
* age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
* menopause: lt40, ge40, premeno.
* tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
* inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 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.

* The dataset has 2 attributes having missing values: node-caps with 8 missing, breast-quad with only one missing.
* In Weka, to fill missing values automatically, we can choose the ‘Choose’ button in the 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.

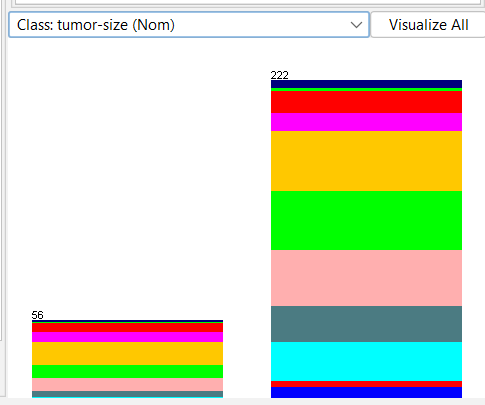
* With node\_caps attribute, it’s discrete attribute, so use mode method to fill missing value in this attribute.

#### 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

#### 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?

* The number of attributes: 5
* The number of samples: 14
* Attributes that are categorical: outlook, windy, and play.
* Attributes that are numerical: temperature and humidity.
* Label: play.

Graphical user interface, application, PowerPoint

Description automatically generated

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

A five-number summary consists of five values: the minimum value, the first quartile (Q1), the median (Q2), the third quartile (Q3), and the maximum value. However, WEKA does not provide all these 5 values as Q1, Q2, and Q3 are not shown in the selected attribute window.

Other available values:

* Temperature attribute: the minimum is 64 and maximum is 85.
* Humidity attribute: the minimum is 65 and maximum is 96.

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

Chart, bar chart

Description automatically generated

Each chart represents the percentage of yes and no values (green for yes and red for no) with respect to each value in the range of an attribute. For numerical attribute, the chart is continuous while it is discrete in categorical attribute charts.

#### 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 picture containing calendar

Description automatically generated

* The chart used to visualize this data set is the scatter plot.
* Overall, there are two correlated pairs:
  + Temperature and humidity: inversely proportional (the higher the temperature, the lower the humidity).
  + Temperature and outlook: when it’s sunny, the temperature increases and vice versa.

### 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 includes:
  + 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 attribute 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.

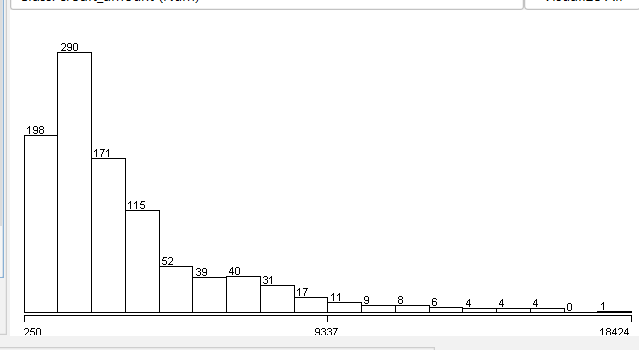
#### Let’s describe the distribution of continuous attributes. (Left skewed or right skewed?)

* Duration: left skewed

Chart, histogram

Description automatically generated

* Credit\_amount: left skewed



* Installment\_commitment: right skewed

Chart, histogram, box and whisker chart

Description automatically generated

* Resident\_since: right skewed

Chart, histogram

Description automatically generated

* Age: left skewed

Chart, histogram

Description automatically generated

* Existing\_credits: left skewed

Chart, histogram

Description automatically generated

* Num\_dependents: left skewed

Chart, histogram

Description automatically generated

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

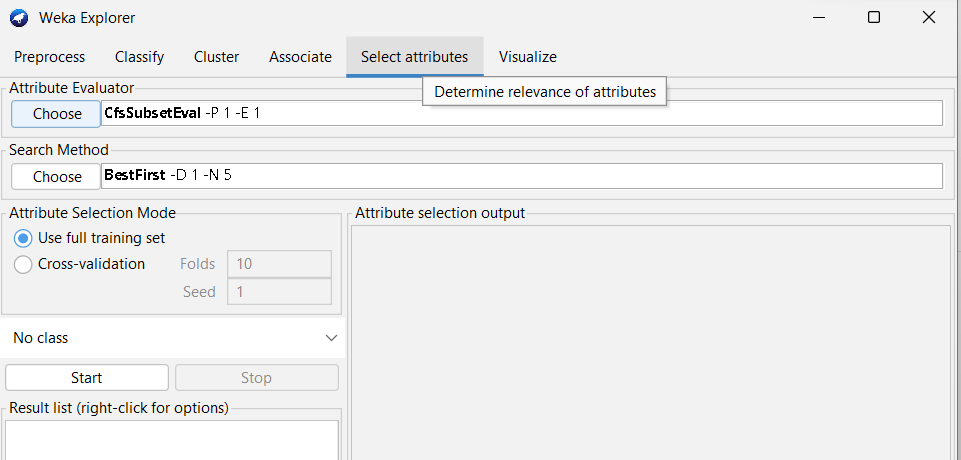
* Each chart represents the percentage of yes and no values (green for yes and red for no) with respect to each value in the range of an attribute. For numerical attribute, the chart is continuous while it is discrete in categorical attribute charts.

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

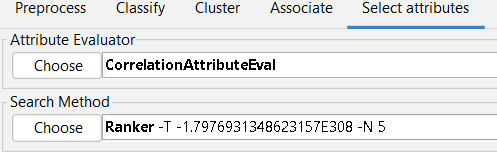
* CfsSubsetEval: Evaluates the importance of a subset of attributes by looking at the individual predictability of each feature along with the degree of redundancy between them (the set of attributes is highly correlated with the class attribute and intercorrelate is preferred).
* ClassifierAttributeEval: 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)

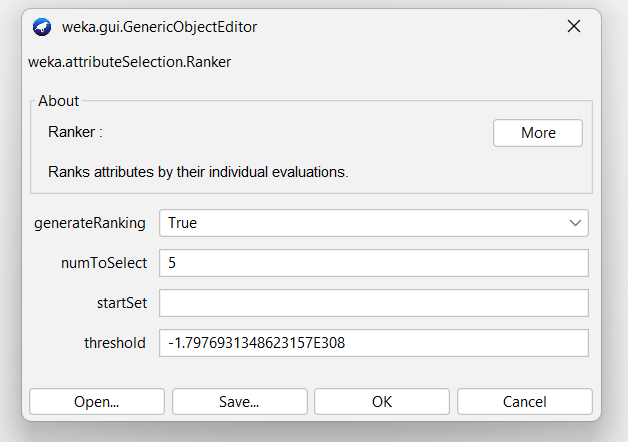
* To select the 5 attributes with the highest correlation with the class attribute, use the CorrelationAttributeEval filter.
  + Step 1: Select the Select attribute tab.



* + 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

### Extract columns with missing values

* Command line: python list\_missing.py <intput> <--print\_data> <--out>
* Example:

$ python list\_missing.py '..\house-prices.csv'

$ python list\_missing.py '..\house-prices.csv' --print\_data --out 'list\_missing\_print\_data.csv'

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

* Command line: **python count\_row\_nan.py <input>**
* Example:

$ python count\_row\_nan.py '..\house-prices.csv'

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

* Command line: **python fill\_nan.py <input> <output> <columns (\*)> <method (\*\*)>**

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

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

* Example:

$ python fill\_nan.py '..\house-prices.csv' 'fill\_na\_single\_column\_mean.csv' LotArea mean

$ python fill\_nan.py '..\house-prices.csv' 'fill\_na\_multi\_column\_median.csv' LotArea,MSSubClass median

$ python fill\_nan.py '..\house-prices.csv' 'fill\_na\_single\_column\_mode.csv' MSZoning mode

### 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: **python delete\_nan\_row.py <input> <output> <threshold (\*)>**

(\*): threshold = 0.5 by default.

* Example:

$ python delete\_nan\_row.py '..\house-prices.csv' 'delete\_nan\_row.csv' 0.5

### 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).

* Command line: **python delete\_nan\_row.py <input> <output> <threshold (\*)>**

(\*): threshold = 0.5 by default.

* Example:

$ python delete\_nan\_column.py '..\house-prices.csv' 'delete\_nan\_column.csv' 0.5

### Delete duplicate samples.

* Command line: **python drop\_dup.py <input> --out <output>**
* Example:

$ python drop\_dup.py '..\house-prices.csv' -–out 'house-prices-drop-dup.csv'

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

* Command line: **python norm.py <input> <output> <columns (\*)> <method (\*\*)>**

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

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

* Example:

$ python norm.py '..\house-prices.csv' 'norm\_single\_column\_min\_max.csv' LotArea min-max

$ python norm.py '..\house-prices.csv' 'norm\_multi\_column\_min\_max.csv' LotArea,MSSubClass min-max

$ python norm.py '..\house-prices.csv' 'norm\_single\_column\_min\_max.csv' LotArea z-score

$ python norm.py '..\house-prices.csv' 'norm\_multi\_column\_min\_max.csv' LotArea,MSSubClass z-score

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

* Command line:

**python operations.py <input> --opt <OPT (\*)> --column1 <first\_column> --column2 <second\_column> --res <result\_column>**

* Example:

$ python operations.py '..\house-prices.csv' -opt \* --column1 LotFrontage –column2 LotArea –res MultOfLotFrontageAndArea

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated with medium confidenceTable

Description automatically generated

(From *house-prices.csv*)

# 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.