Assignment 2: Data classification and feature analysis using WEKA machine learning toolbox

1. Introduction.

The current assignment focuses on the classification of data and analysis of the effects of different features on the classification tasks. It examines the use of combination of features, feature reduction and the differences resulting in the use features with different ranking scores.

2. Start-up instructions

The tasks related to the presented assignment will be conducted using the WEKA machine learning toolbox. WEKA is a freely distributed toolbox, containing a wide variety of classification and ranking algorithms applicable in tasks related to the classification and analysis of data.

Once the program is installed and started the following window will appear:

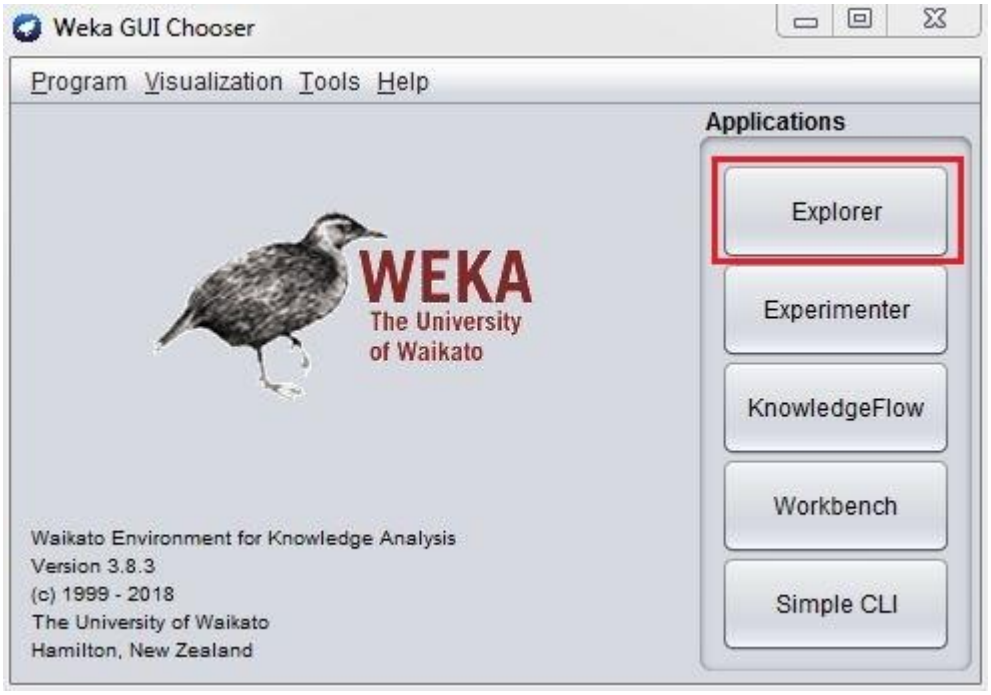

Figure 1: WEKA's main menu window

After the "Explorer" menu is selected the Explorer window in opened (Figure 2).

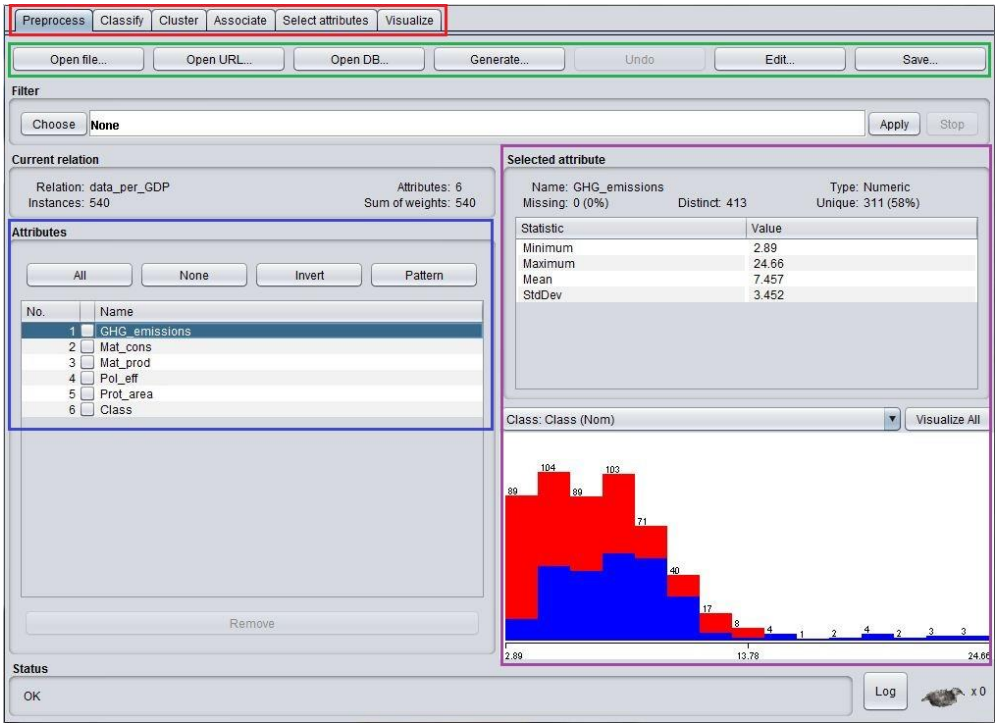

Figure 2: Explorer window

The main elements of the Explorer window are the:

- **Explorer navigation tabs** (Figure 2, marked with red): The tabs are used to select windows for different types of tasks such as classification and feature ranking.
- **Data operations** (Figure 2, marked with green): They are used for data loading and editing. Files are opened using the "Open file…" button.
- **Attributes window** (Figure 2, marked with blue): The attribute window provides information about the features contained in the opened file. It allows each feature group be selected, removed or included in the tests.
- **Selected attributes window** (Figure 2, marked with purple): The Selected attribute window provides statistical information about the features, selected in the Attributes window. It also presents graphical information about the number of elements of each class, the range in which they vary and their distribution.

The classification of the data can be performed after data is loaded and the "Classify" tab is selected from the Explorer navigation tabs. The selection of the "Classify" tab opens the following window:

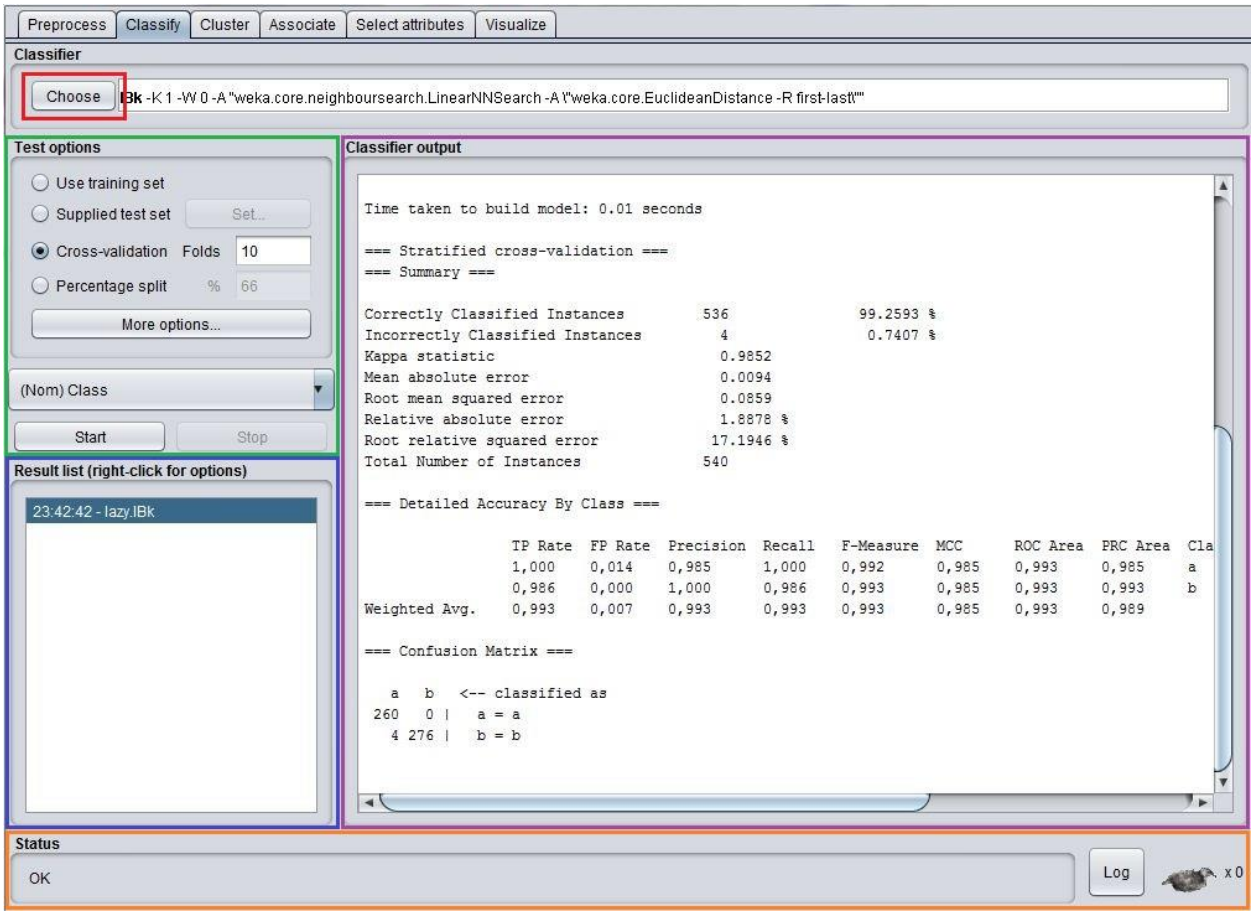

Figure 3: "Classify" window

When the "Classify" window is opened a classification algorithm must be selected. This is done by pressing the "Choose" button in the Classifier menu (Figure 3, marked in red). After the button is pressed a drop-down menu appears, in which different classifiers are grouped in folders, in accordance with their type.

After the classifier is chosen the parameters of the experiment can be set from the "Test options" section (Figure 3, marked with green) and the experiment can be initiated from the "Start" button. Information about the settings and the results is given in the "Classifiers output" window (Figure 3, marked with purple), while information about the previously conducted can be accessed from the Results list menu (Figure 3, marked with blue). Finally, information about the status of the experiment and the current test iteration is provided in the Status field positioned in the bottom of the windows (Figure 3, marked with orange).

3. Tasks

For the presented task the WEKA machine learning toolbox is required. The toolbox can be downloaded completely free from the following address: https://waikato.github.io/weka-wiki/downloading_weka/

In addition to the current assignment a dataset has been compiled – it is included in the csv file EU_GDP.csv. The file contains information about the 27 member states of the EU regarding the following indicators:

- **Greenhouse gas (GHG) emissions**: Carbon dioxide (CO2) emissions measured in Tonnes/capita;
- **Air pollution effects**: Mortality measured per 1 000 000 inhabitants;
- **Protected areas**: Terrestrial protected areas measured in % of the total land area of the country;
- **Material consumption**: Measured in total tonnes/capita;
- **Material productivity**: Total produced non-energy materials measured in dollars/kilogram;

The used data is sourced from the Organization for Economic Co-operation and Development (OECD - https://data.oecd.org/environment.htm) and covers the time period 2000 – 2019. The compiled indicators are assigned classes according to the nominal GDP of each country: class "a" denotes countries with GDP of over 30 000$ while class "b" denotes countries with GDP lower than 30 000$.

Task 1:

Start WEKA, open the "Explorer" menu and load the EU_GDP.csv file. After the file is loaded select the "Classify" menu and classify the data using the following classifiers (*note*: *the main folder under which the classifier is listed is also provided*):

- trees/J48;
- lazy/IBk;
- bayes/NaiveBayes;
- functions/MultilayerPerceptron;
- functions/SMO

Write down the obtained results in Table 1.

Table 1. Classification results

Classifier	J48	IBk	NB	MLP	SMO
Result					

Task 2:

Load the EU_GDP.csv file in the "Explorer" menu. After the file is loaded select the "Select attributes" menu and choose the "*ReliefFAttributeEval*" algorithm in the Attribute Evaluator section. Run the algorithm, which will rank the 5 indicators according to their relevance to the designated classes. List the ranking in Table 2.

Table 2. Feature ranking

Rank	Indicator
#1	
#2	
#3	
#4	
#5	

Task 3:

Based on the ranking obtained on in Task 2 repeat the classification process while removing the feature with the lowest score (from the "Explorer" menu). Repeat the task until only 1 feature is left. Write down the obtained results in the Table 3 and observe how the reduction of the features affects the classification accuracy:

Table 3. Classification results for data with reduced feature number

Dataset	J48	IBk	NB	MLP	SMO
4 Features					
3 Features					
2 Features					
1 Feature					

Using the obtained results and the results obtained it Task 1 calculate the accuracy drop-off (difference) in the results after each feature reduction. Write down the results in Table 4.

Table 4. Difference in classification results when datasets with 1 and 5 features are used

Result difference	J48	IBk	NB	MLP	SMO
4F drop-off					
3F drop-off					
2F drop-off					
1F drop-off					

Task 4:

Repeat the experiment using only the lowest ranking feature and compare the results to the case where only the highest ranking feature is used for classification. Write the results down in table 5.

Table 5. Difference in classification results when a low ranked and high ranked feature is used

Setup	J48	IBk	NB	MLP	SMO
Rank #1 feature					
Rank #5 feature					
Result difference					

Task 5:

Based on the observations on the obtained results formulate a conclusion on the effects of feature reduction and feature selection.

Answer:

4. Guidance

If you require guidance, consultation, suggestions or have difficulty completing the assignment please contact assist. Prof. Firgan Feradov – Technical University of Varna, Bulgaria at firgan.feradov@tu-varna.bg