CST8390 - Lab 1
Data exploration and integration with Weka - Iris dataset

Due Date: must be demoed within the **first two weeks** of classes.

Introduction

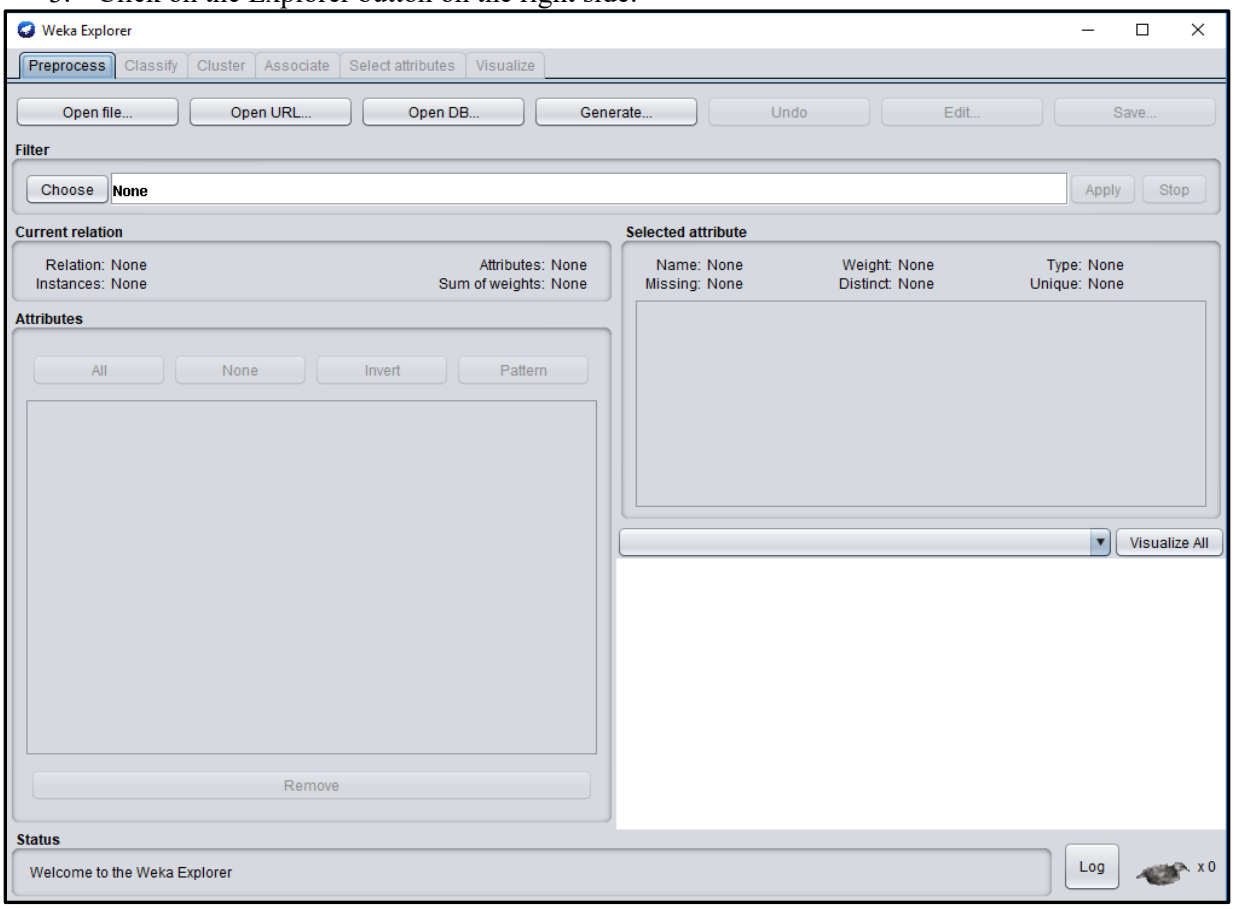
The goal of this lab is to install and familiarize with Weka.

Part 1

1. Download and install Weka. You can find it here:
   http://www.cs.waikato.ac.nz/ml/weka/downloading.html

2. Open Weka and have a look at the interface. It is an open-source project written in Java from the University of Waikato.

3. Click on the Explorer button on the right side:



4. Check different tabs to familiarize with the tool.

5. Weka comes with a number of small datasets. Those files are located at C:\Program Files\Weka-3-8 (If it is installed at this location. Or else, search for Weka-3-8 to find the installation location). In this folder, there is a subfolder named 'data'. Open that folder to see all files that comes with Weka.

6. For easy access, copy the folder 'data' and paste it in your 'Documents' folder.

7. In this lab, we will work with the dataset Iris. To open Iris dataset, click on 'Open file' in the 'Preprocess tab'. From your 'data' folder, select iris.arff and hit open.

8. To know more about the iris dataset, open iris.arff in notepad++ or in a similar tool and read the comments.

9. Click on visualize tab to see various 2D visualizations of the dataset.

    a. Click on some graphs to see more details about it.
    b. In any of the graph, click one 'x' to see details about that data record.

10. Fill this table:

Flower Type	Count

11. Fill this table:

Attribute	Minimum	Maximum	Mean	StdDev

Part 2

12. Download EmployeesSalary.csv file from Brightspace.
13. Open EmployeesSalary.csv in excel and explore it.
14. Read https://www.cs.waikato.ac.nz/ml/weka/arff.html to find the expectations of an ARFF file.
15. Identify the attributes of the data. Record the attributes and the type of attribute for the data.
16. Closely analyse data. In excel, do the required modifications to match with the requirements for an ARFF file. (Hint: Check the requirements if the data has spaces in it.)
17. Open the file in notepad++. Add the required section headers and corresponding information in the file. Save the file as EmployeesSalary.arff. This involves creating the @relation line, one @attribute line per attribute, and @data to signify the start of data. It is good to add comments at the top of the file describing where you obtained this data set, explanation about your attributes etc. A comment in the ARFF format starts with the percent character % and continues until the end of the line.

18. Open the ARFF file as you did in lab 1 (by selecting ‘Open file’ in the ‘Preprocess tab’.). You may run into errors as you load your ARFF file. If so, check the requirements to troubleshoot your problem.

19. Which are the four important attributes that are relevant for data analysis?

    a.

    b.

    c.

    d.

20. For the nominal attributes from Question 19, fill in the following table:

Attribute Name:		**Attribute Name:**	
Label	**Count**	**Label**	**Count**

Attribute Name:	
Label	**Count**

21. Analyze your data to see any anomalies. List the identified anomalies below. Write why you think those records are anomalies in the following format:

Id	first_name	last_name	email	Address	Country	Branch	Currency	Salary	Reason

In order to get the credit for this lab:

1. Show the Iris file in Weka during demo.
2. Show the EmployeesSalary file in Weka during demo.
3. Fill the answers in Lab1_Answers.doc and upload it to Brightspace.
4. Show the answer document during demo.

Both demo during lab hours and submission in Brightspace are required to get credits for the lab.