CST8390 - Lab 4
k Nearest Neighbor (kNN)

Due Date: Week 4 in corresponding lab sessions

Introduction

The goal of this lab is to perform classification on wine dataset using **kNN**.

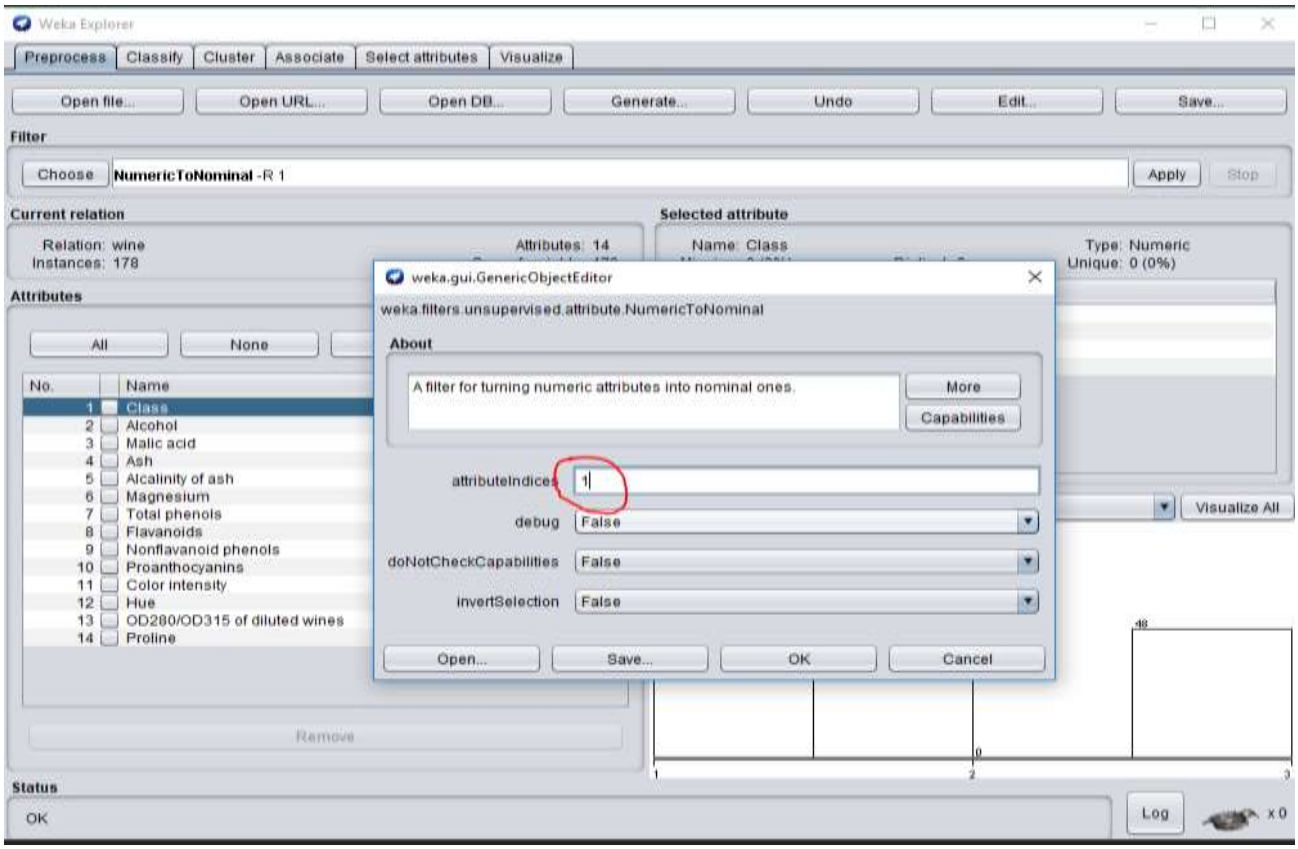
Steps:

1. Get "Wine" dataset from https://archive.ics.uci.edu/ml/datasets.html (or https://archive.ics.uci.edu/ml/index.php) and save it as a **CSV** file. (**Data** is in Wine.data and info is in data.names). Add **attribute** names as the first row in the csv file. (For every row, first value is the class, remaining values are various attributes)

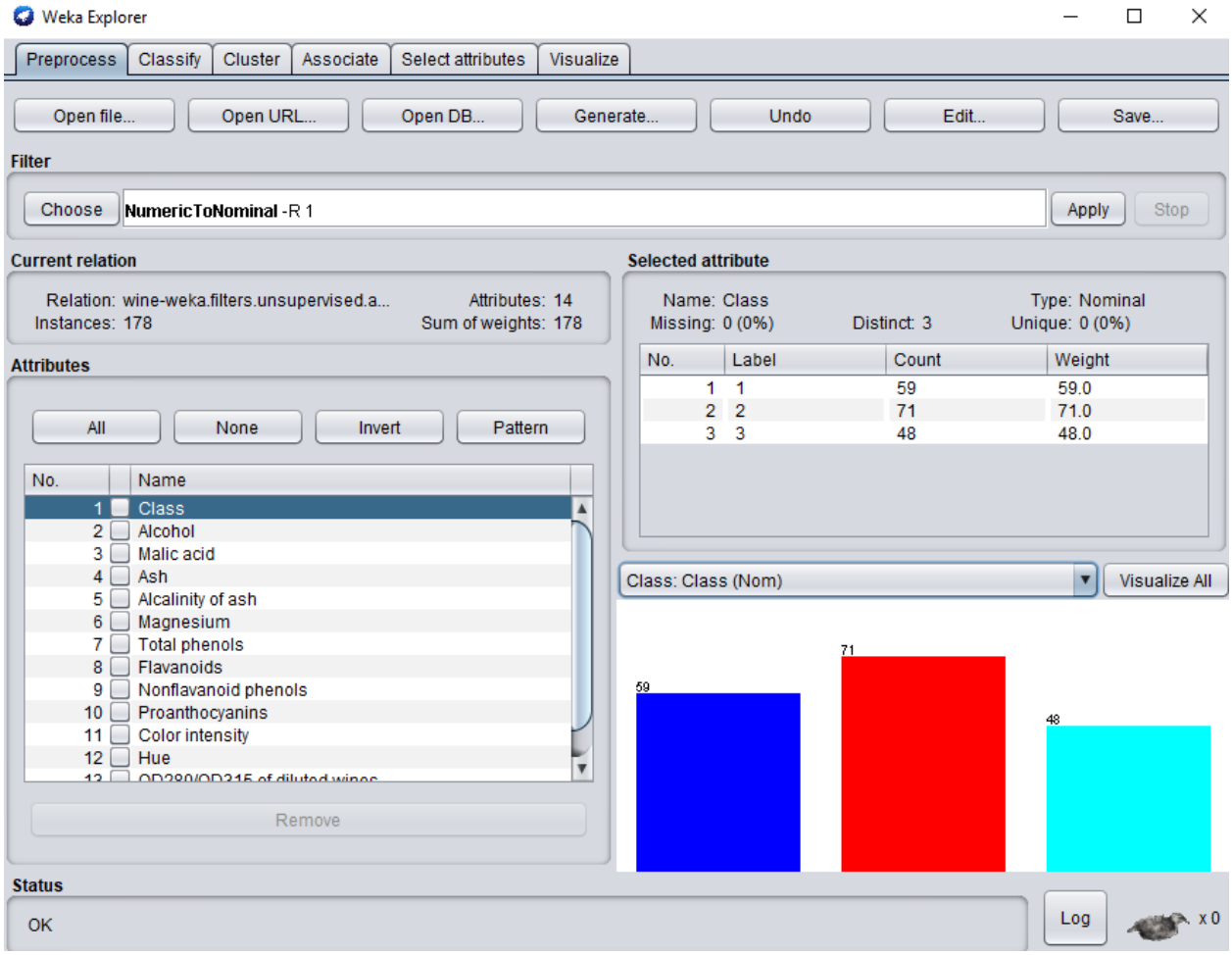
2. Explore and learn about the **relevance** of various attributes of the dataset

3. Load the file to **Weka**.

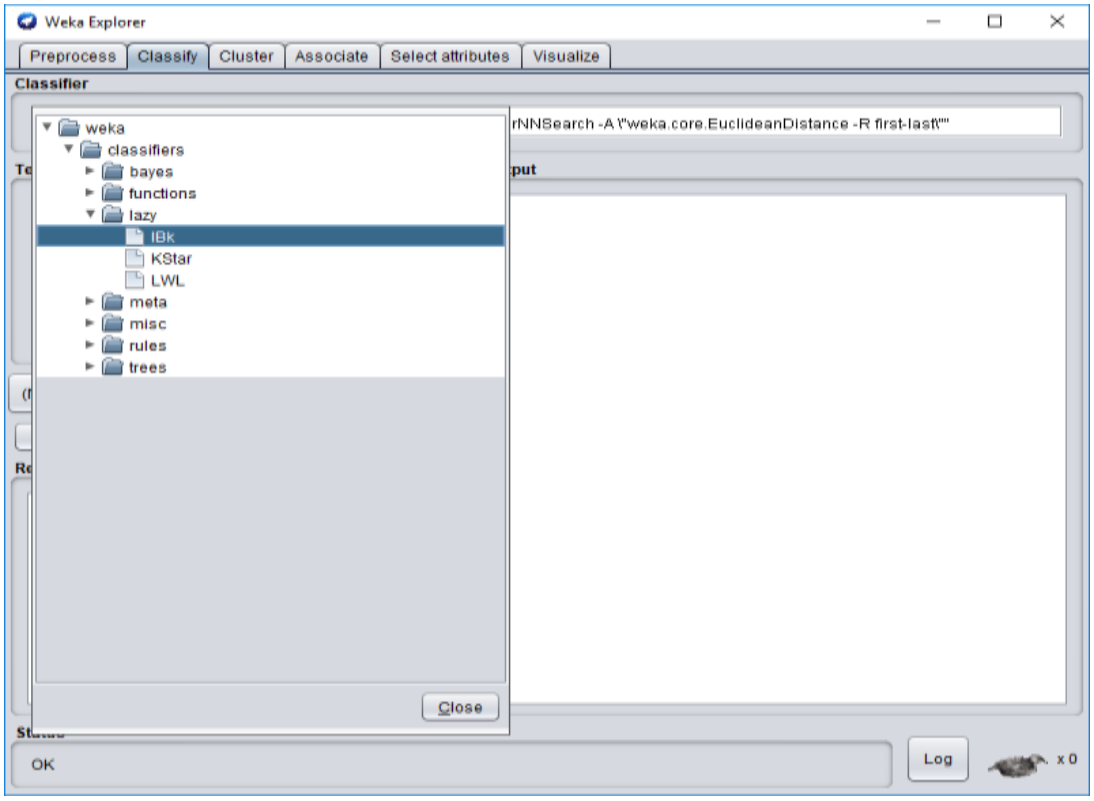
4. Check how various attributes are converted in **Weka**. Class is considered as **numeric** instead of **nominal**. Apply **filter** NumericToNominal to convert class datatype to nominal. When you apply filter, you need to specify the index of the attribute you need to apply the filter.



Now, you should see like this:

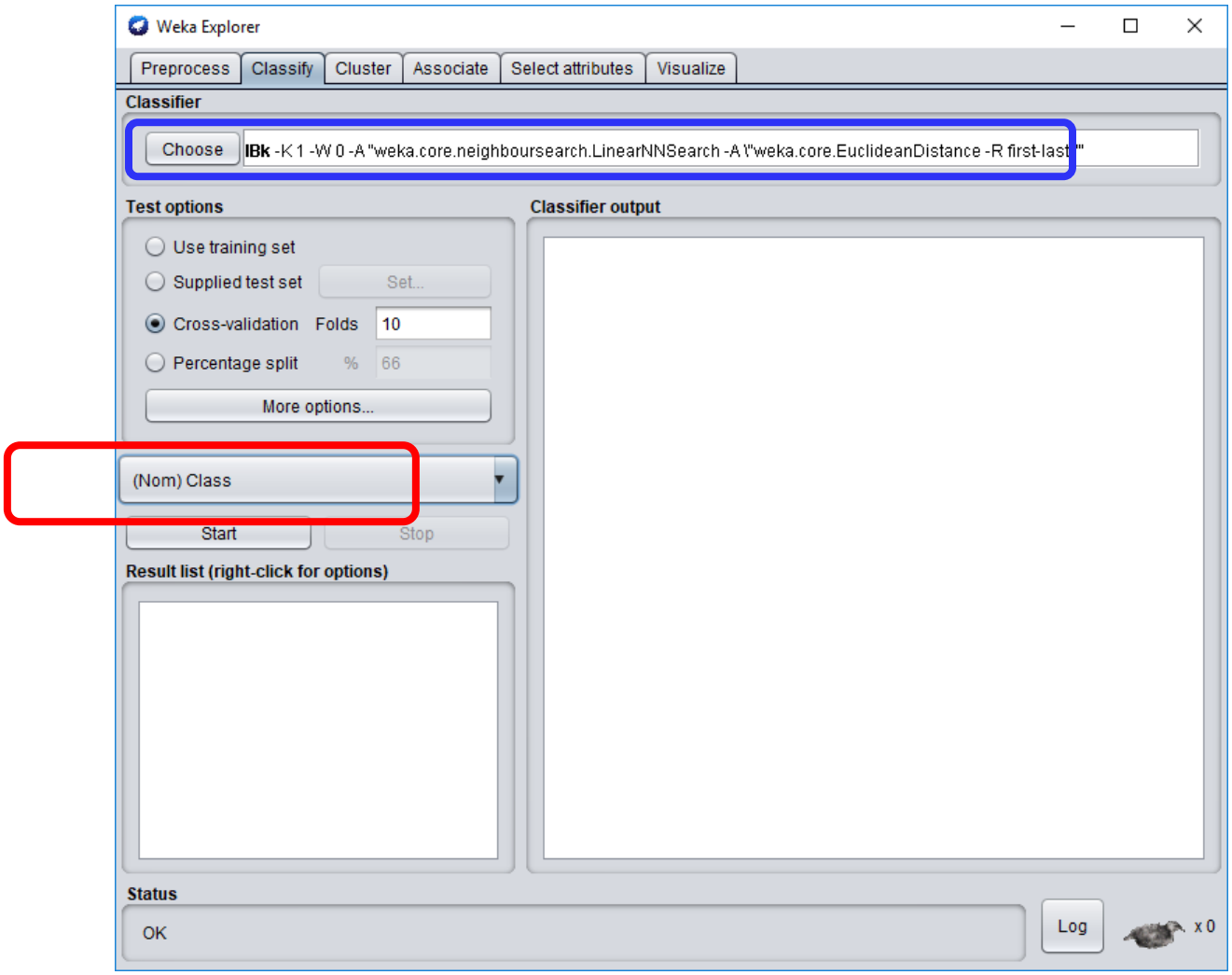


5. Now, we need to perform classification using kNN method. For that, click on "Classify" tab. For this lab, we use kNN. For that, choose IBk which is **Instance Based k Nearest Neighbors** from Lazy in the tree view.



6. As mentioned earlier, our first attribute is the class label. We need to set that now in the classify panel. (Marked in red below)

   IBk –K 1 –W 0 –A "weka.core.neighboursearch.LinearNNSearch …." This is the **parameter list** for the algorithm (Marked in blue). Click on this text to set the value of k. Set k as 3. Close the window. Now, set the cross-validation to 10 Folds if it's not already there. Now click "Start" to run the algorithm.



7. There should be a lot of text in the right-hand side of the window with the results of the algorithm. Find the line that says "Correctly classified instances".

   a. What is the **percentage** of correctly classified items?
   b. What are the **True Positive (TP)** rates of each class?
   c. Look at the **confusion matrix**, which class is incorrectly classified?

8. Now click on the "Choose" button to modify the number of neighbours that are used in the **kNN** search to 5.

   a. What is the **percentage** of correctly classified instances? _____

   b. What are the **True Positive (TP)** rates of each class? _____

   c. Look at the **confusion matrix**, which classes are incorrectly classified? _____

9. Run the algorithm **several times**, always increasing the value of N by two, and always an **odd number**: 1, 7, 9, 11, 13. Each of your tests will be in the window of the lower left. Fill in the following table.

K	percentage of correctly classified instances
1	
7	
9	
11	
13	

Which class is being mis-classified? _____

10. Repeat **step 9** with "Percentage Split" of 70. Fill in the following table.

K	percentage of correctly classified instances
1	
3	
5	
7	
9	
11	
13	

REMEMBER:

Show your answers to the lab professor when you are done (in Weka and document).

This lab has 5 marks so ensure that you have all your answers filled in.

FOR YOUR ANALYSIS:

What is the purpose of "confusion matrix"? What is its importance?

Explain with your own words the kNN method.

Ottawa, Jan 2020.