CST8390 - Lab 3
K Nearest Neighbor (kNN)

Due Date: Check Brightspace for due dates.

Introduction

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

Steps:

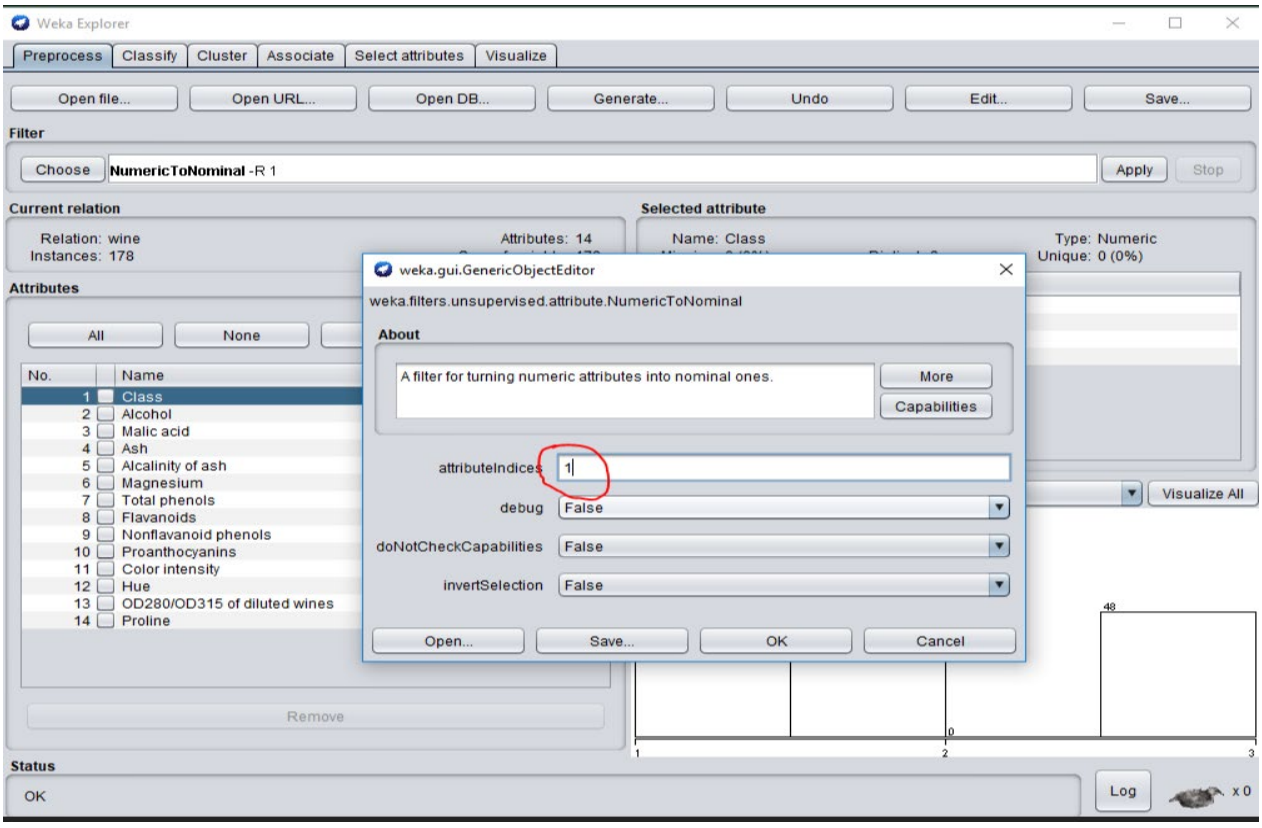
1. Open the word document named Lab3_Answers.doc.

2. Get "Wine" dataset from https://archive.ics.uci.edu/ml/datasets.php (this link will show you a list of datasets, you need to search for wine dataset in this page) and save the content of the file 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)

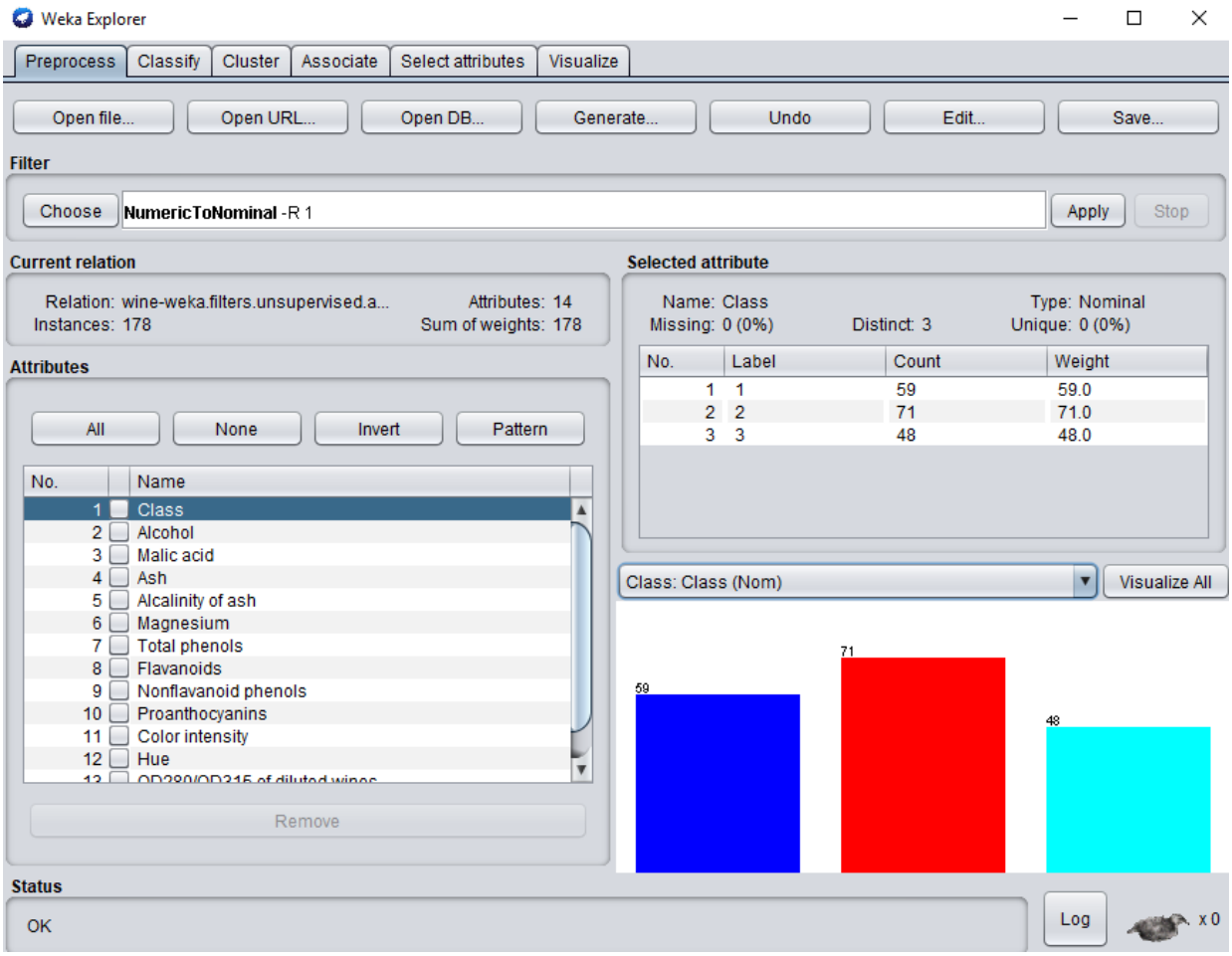
3. Explore and learn about the relevance of various attributes of the dataset.

4. Load the file to Weka.

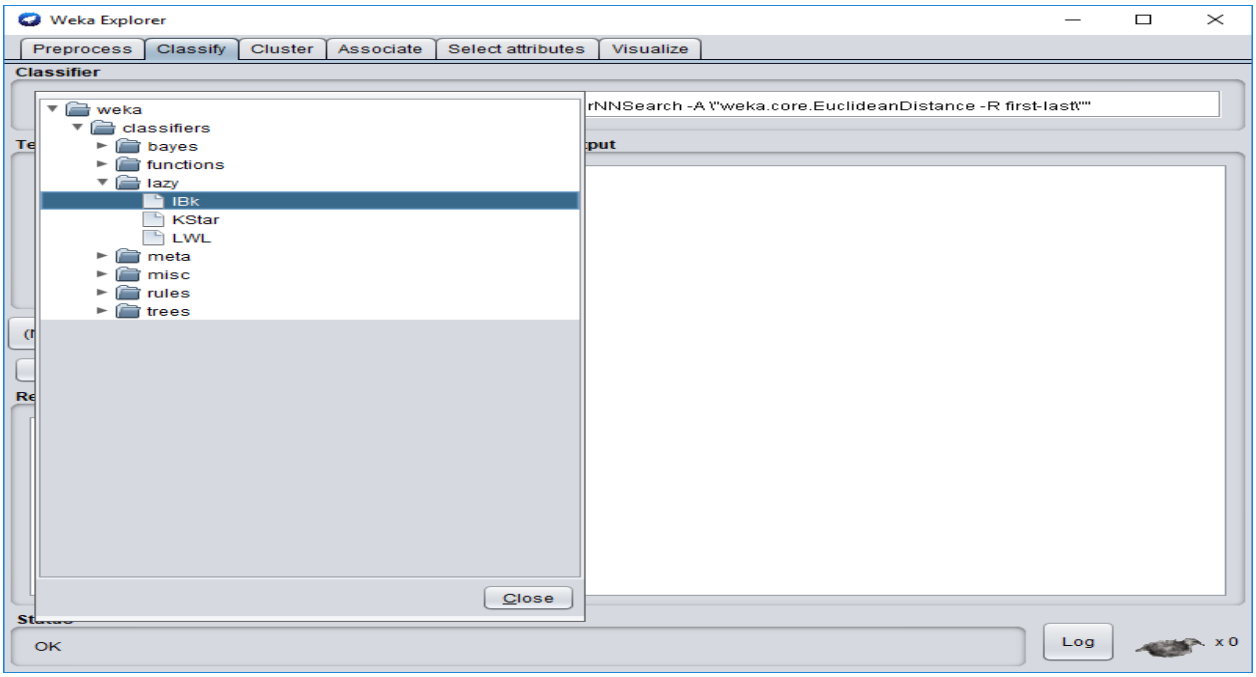
5. Check how various attributes are converted in Weka. Class is considered as numeric instead of nominal. Apply filter NumericToNominal to convert class datatype to nominal. When you apply filer, you need to specify the index of the attribute you need to apply the filter. Make sure that you are converting only the attributes to be converted. If you didn't specify attribute index, by default, all numeric attributes will be converted.



Now, you should see like this:

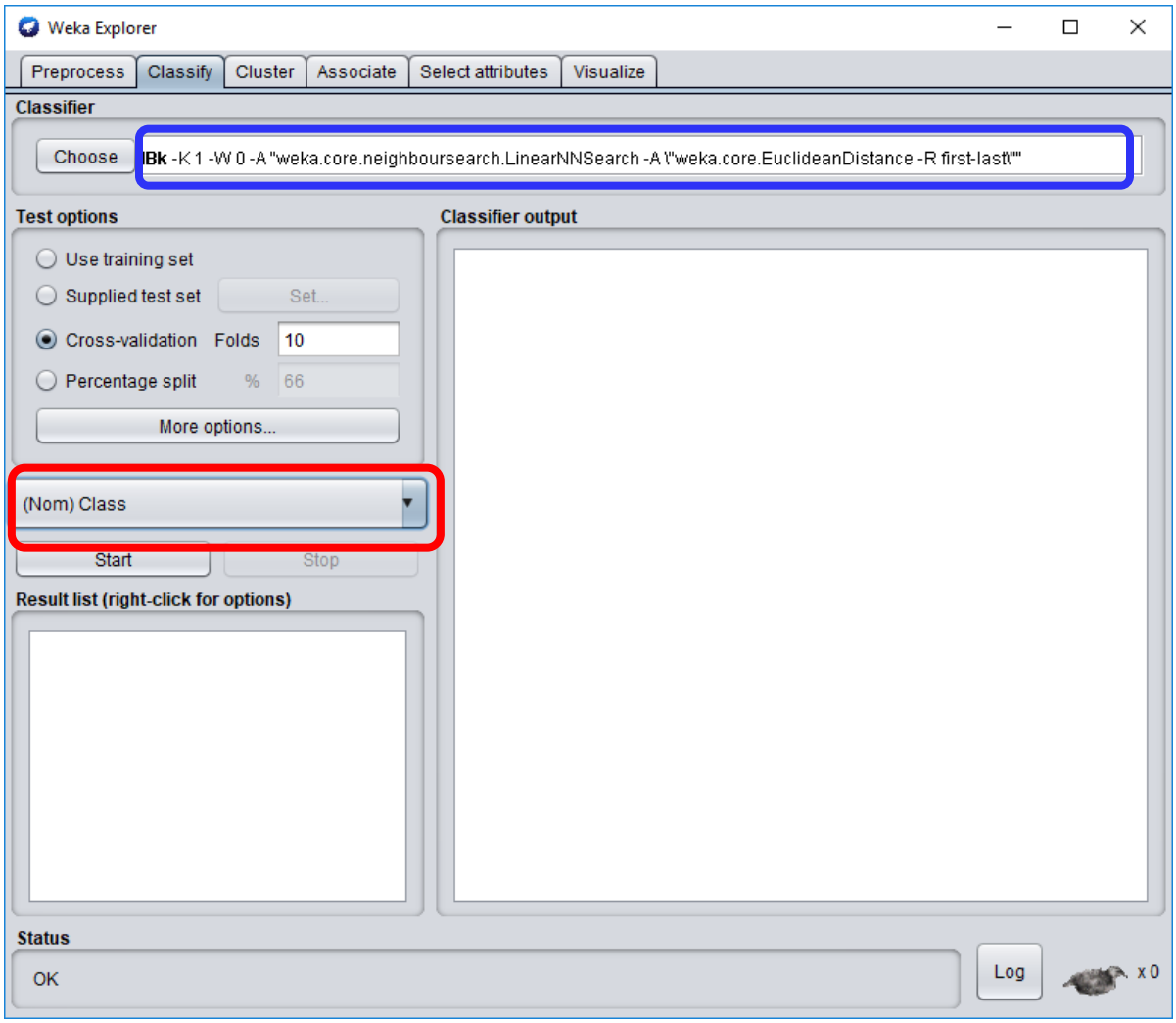


6. Now, we need to perform classification using kNN method. For that, click on “Classify” tab. For that, choose IBk which is Instance Based k Nearest Neighbors from Lazy in the tree view.



7. 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 1. Close the window. Now, set the cross-validation to 10 Folds if it's not already there. Now click "Start" to run the algorithm.



8. 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?

9. Now click on the "Choose" button to modify the number of neighbours that are used in the kNN search to 3.

   d. What is the **percentage** of correctly classified instances? _____
   e. What are the True Positive (TP) rates of **each** class? _____
   f. Look at the confusion matrix, which classes are incorrectly classified? _____

10. Run the algorithm several times, always increasing the value of N by two, and always an odd number: 5, 7, 9. Each of your tests will be in the window of the lower left. Fill in the table given in the answer document.

11. Repeat step 9 with "Percentage Split" of 70. Fill in the table in the answer document.

12. Take one instance as a test instance and show the calculations in Excel to find the class of that instance by applying 5NN. You need to explain the process AND include the screenshot of the excel file (of the final stage where you make the prediction of the class of the test instance) in the Answer document.

In order to get credit for this lab,

1. You should be ready to show these 10 executions in Weka and the excel sheet during demo.
2. Your answer document should have answers for questions 8 - 12.
3. Upload the answer document to Brightspace.

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

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