CST8390 Assignment 3

Due: July 25, 2021 at 11:59 PM Sharp!!!

(Late submissions will not be accepted)

Goal: The goal of this lab is to explore and analyze one dataset from the given list and perform clustering using kMeans and farthestFirst and outlier detection using Local Outlier Factor and Isolation Forest.

Steps:

1. Select **one** dataset from the list below:
   - Dataset 1 – Glass
     - https://archive.ics.uci.edu/ml/datasets/Glass+Identification
     - http://odds.cs.stonybrook.edu/glass-data/
     - This dataset contains attributes regarding several glass types (multi-class). Here, class 6 is a clear minority class, as such points of class 6 should be marked as outliers, while all other points are inliers. For outlier detection, you need to create a column named Outlier and mark class 6 instances as Yes and all other attributes as No. After this, remove class attribute.
   - Dataset 2 – Lymphography
     - https://archive.ics.uci.edu/ml/datasets/Lymphography
     - http://odds.cs.stonybrook.edu/lympho/
     - It is a multi-class dataset having four classes, but two of them are quite small (2 and 4 data records). Therefore, those two small classes should be merged and considered as outliers compared to other two large classes (81 and 61 data records). For outlier detection, you need to create a column named Outlier and mark instances of smaller classes as Yes and all other attributes as No. After this, remove class attribute.

2. You have to include a brief description (10 sentences) about the selected dataset. From the papers given with the dataset, you may be able to find the performance of some clustering and outlier detection methods applied on those datasets. If so, include that also in the description. **(Marks: 4)**

3. Thoroughly analyze your data to have a clear understanding of your data and their attributes and types. Tabulate attributes, its description (if available), and its data types.

   (Marks: 3)

4. Load your file to Weka. Double check the type of your attributes in Weka. If they are not as expected, apply filters to convert them to the right types. Make sure that your class attribute (i.e., Outlier attribute) is nominal.

5. Tabulate statistics and counts (whichever apply) for each attribute. Provide that information in **one** table. **(Marks: 2)**

6. Perform preprocessing, data cleaning, remove duplicates, handle missing information etc. Specify which all filters you applied and the corresponding reason. **(Marks: 3)**

7. Once you are done with data preparation, navigate to Visualize tab to visualize your data. Include 3 interesting charts in your submission. You need to specify how those charts are interesting (you may have clusters, or classes are separable, or classes have too much of overlapping etc.). You need to compare the attributes on your x and y axes and their impact on the class attribute. **(Marks: 3)**

8. Now perform clustering using k-Means for different k, i.e., numClusters (which makes sense for your dataset) and tabulate those results. (Hint: if you have 3 class labels, then 3 and above may be a good value for k. You need to run with at least 5 different values of k). Highlight the row with the best k. You have to create a **single** table with results. Scanned images and different tables are not acceptable. *(select "Classes to clusters evaluation" and select class attribute from the drop down list).* **(Marks: 3)**

9. Next, perform clustering using farthestFirst method and tabulate the results (like step 8). **(Marks: 3)**

10. Now, prepare dataset to do outlier detection. Based on the class attribute, you have to create a new column named "Outlier". Once "Outlier" column is created, remove class column. Based on the description of the dataset, type "Yes" for outlier instances and "No" for other instances. Perform Outlier Detection using Local Outlier Factor method (For LOF, perform it with 10-fold cross validation). Open "Visualize classifier errors" and save the file as datasetName_LOF.arff. Open datasetName_LOF.arff and select predicted Outlier in the attributes list. Get a screenshot and paste it here. **(Marks: 3)**

11. Open Edit window for datasetName_LOF.arff file and sort the data based on Predicted Outlier column. Find how many of the actual outliers are predicted as outliers. **(Marks: 2)**

12. If the result is not close enough, repeat steps 10 and 11 with only selected attributes. Give **detailed** explanation on your findings. **(Marks: 5)**

13. Perform Outlier Detection using Isolation Forest method on the original dataset. Open "Visualize classifier errors" and save the file as datasetName_ISF.arff. Open datasetName_ISF.arff and select predicted Outlier in the attributes list. Get a screenshot and paste it here. **(Marks: 3)**

14. Open Edit window for datasetName_ISF.arff file and sort the data based on Predicted Outlier. Find how many of the actual outliers are predicted as outliers. **(Marks: 2)**

15. Combine results from LOF and ISF by creating an excel file named combinedResults_datasetname with the actual attributes, PredictedOutlier from LOF, PredictedOutlier from ISF, and a new column named ensemble results (2 if the instance is classified as an outlier by both methods, 1 if by only one of the methods and 0 if it is not an outlier). Provide a screenshot of the opened file. **(Marks: 4)**

Submission Details:

This is a partner assignment. Assignment should have a cover page with the name (Last name, first name of both students) and student numbers. Create a zipped folder named LastNameFirstStud_FirstNameFirstStud_LastNameSecondStud_FirstNameSecondStud.zip with the report, datasetName_LOF.arff, datasetName_ISF.arff and combinedResults_datasetname.xls, and model files of LOF, ISF, kMeans and FarthestFirst. *There will be mark deduction if folder name doesn't match with the requirements.* Upload the zipped folder to Brightspace.

Marks:

This assignment will have a total of 40 marks. Each step (with and without marks) is important. There will be **negative** marks if you miss explanation for any of the steps. Every step/question should be answered with explanation. The assignment should look like a professional report.