CST8390 Assignment 3

Due: March 18, 2023, at 11:59 PM Sharp!!!

(Late submissions will not be accepted)

Goal: The goal of this lab is to explore and analyze Glass dataset and perform clustering using kMeans and farthestFirst, and outlier detection using Local Outlier Factor and Isolation Forest.

Steps:

- Dataset – Glass
  - https://archive.ics.uci.edu/ml/datasets/Glass+Identification
  - This dataset has 7 classes, out of which class 6 is a minority class that has 9 instances. Those 9 instances should be marked as outliers, while all other instances are inliers. For outlier detection, you need to create a column named Outlier and mark 'yes' for outliers and 'no' for all other attributes. To avoid confusion, you can keep two processed files – one for clustering and the other for outlier detection.

Data Understanding

You must include a brief description (10 sentences) about the 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. 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.

Data Preparation

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. Tabulate statistics and counts (whichever apply) for each attribute. Provide that information in **one** table. Perform all required CRISP steps. Specify which all filters you applied and the corresponding reason. Now, 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 outliers, clusters, separable classes, inseparable classes etc.). You need to compare the attributes on your x and y axes and their impact on the class attribute.

Modeling

Clustering**:** Now perform clustering using k-Means for different 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 or until you see an elbow in the elbow graph). Highlight the row with the best k. You must create a **single** table with results. Scanned images and different tables are NOT acceptable. Next, perform clustering using farthestFirst method and tabulate the results.

Outlier detection**:** Make sure that Outlier column was added in the earlier stage. If not, add it now as instructed before. Perform Outlier Detection using Local Outlier Factor method (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 in the report. Find how many of the actual outliers are predicted as outliers. If the result is not close enough, repeat the process with only selected attributes. Give **detailed** explanation on your findings.

Now, perform Outlier Detection using Isolation Forest method on the 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. Find how many of the actual outliers are predicted as outliers.

Discussion of Results

Outlier Detection Results: Combine results from LOF and ISF by creating an excel file named combinedResults_datasetname and find the ensemble results. Paste the screenshot of final results (as we did in Outlier detection lab). Also, include the excel sheet in the zipped folder.

Outlier Detection and Clustering: Provide a discussion on comparison of clustering results and outlier detection results. When we did the clustering task of the employeeSalary file, those outliers were clustered into separate clusters when we increased the number of clusters. For this dataset also, check whether outliers are grouped as separate clusters. If not, increase the number of clusters to see whether they gets grouped into separate clusters.

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 is important. Every step/question should be answered with explanation. The assignment should look like a professional report. You should have a cover page, table of content, table of pictures, and report should have sections like Introduction, Data Understanding, Data Preparation, Modeling, Discussion of results, Conclusion, and References. There will be **negative** marks if you miss explanation for any of the steps. Also, deductions will be applied if the report is not professional.