CST8390 - Lab 4 (Part 1)
Clustering

Due Date: Check Brightspace for due dates.

Introduction

The goal of this lab is to cluster the Salary file using kMeans in **RapidMiner**.

Steps:

1. Load the data into RapidMiner. Make sure that you have the right data types. If not, do the required data type conversions.
2. Generate a new attribute named SalaryOrig to store the original salary. Once we normalize salary, we cannot see the original salary.
3. Make a copy of the file. This will be used when you join the results.
4. Set Id as the Id column.
5. Select the relevant attributes.
6. Do all data preparation steps. We will be using a distance-based method for clustering. So, make sure to prepare your data accordingly.
7. Do clustering using kMeans. In the lecture, we saw that the optimal k for this dataset is 5. So, run clustering with k=5. Take a screenshot of the clustering model and paste it in the answer document.
8. Redo clustering with k=10 to see outliers.
9. Write your results to a csv file. Interpret your results and filter those clusters with less than 5 instances in it. Take a screenshot of the filtered instances and paste it in the answer document.
10. Export your process and save it at your preferred location. This will be saved as an rmp file. Also, take a screenshot of the process and paste it in the answer document.

In order to get grades,

1. For the demo, you should be ready with your rmp file in RapidMiner.
2. Submit the rmp file and the answer document to Brightspace.