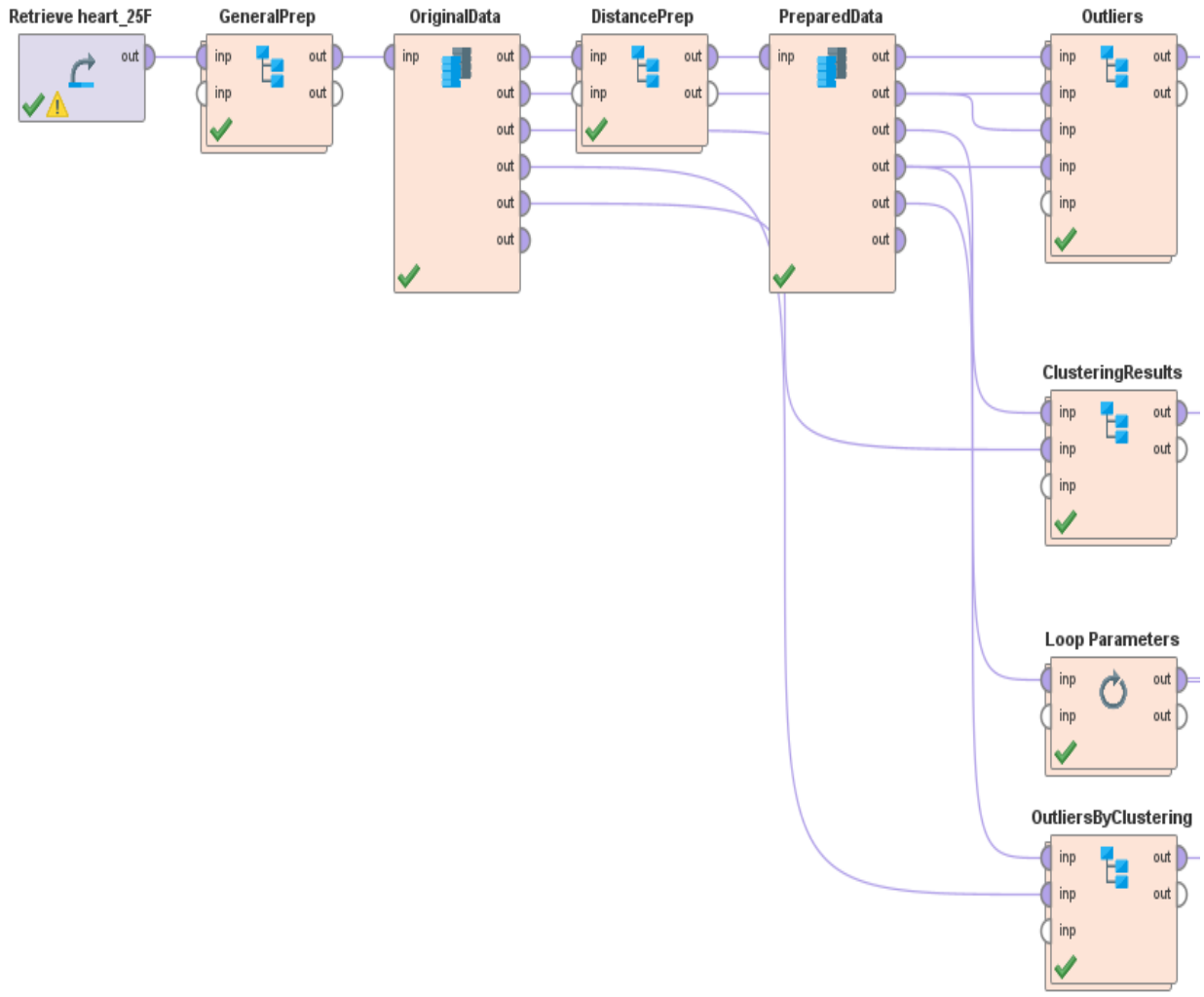
CST8502 – Assignment 2
Outlier Detection & Clustering - RapidMiner

Due Date: Check Brightspace for due dates.

Introduction

The goal of this assignment is to cluster the Heart Failure dataset using kMeans and find outliers using "Detect Outlier (LOF)", "Detect Outlier (Distances)" and by clustering approach. Every operator should be named like <<firstname>>_<<operator>> (Ex: Anu_Normalize). As you are not creating tables in this assignment, there will be specific marks for following the naming conventions.

The general template of the full process must look like this:



Once you get outliers from outlier detection methods and clustering approaches, you should join them to see the instances flagged as outliers by all approaches (the final join and filtering is not shown in the above template). You must have subprocesses named GeneralPrep and DistancePrep for data preparation and subprocesses named Outliers, ClusteringResults, LoopParameters and OutliersByClustering for various modeling.

Data Preparation

1. Load the data into RapidMiner.
2. As part of GeneralPrep, remove any duplicates, generate an ID, set the correct data types for each attribute and set the role if you have any special attributes (like class etc.). Now, we have the original data ready to apply model-specific preparation steps. Multiply it so that we can use the original data to join with various results.
3. DistancePrep: We will be using distance-based methods for clustering and outlier detection. So, make sure to prepare your data accordingly. Numerical columns should be normalized and nominal columns should be one-hot encoded (**must** use Nominal to Numerical operator).
4. Include the list of attributes that you have normalized and the list of attributes that you have one-hot encoded.
5. Multiple the prepared data so that we can use it for various analyses.
6. Take a screenshot of the current process and subprocesses – GeneralPrep and DistancePrep and paste them in the answer document.

Outlier Detection

7. Do outlier detection using LOF. Use "Generate Outlier Flag" operator to convert outlier score to an outlier flag. When you generate a flag, you can keep the contamination factor as 5% (0.05). Now, join this result with the original data to see original attributes.
8. Now, Use Detect Outliers (distances) operator to detect outliers. This will create a new column named outlier. Set the number of outliers as 45, which is around 5% of the total number of instances. Now, join with the original data to see original attributes.
9. Rename outlier columns as LOF_Outlier and Distances_Outlier. As we cannot keep more than one column with the role "Outlier", change the role of Distances_outlier to "Interpretation"
10. Join both results and filter those instances flagged as outliers by both methods. Take a screenshot of the filtered instances and paste it in the answer document.
11. Take a screenshot of the subprocess and paste it in the answer document.

Clustering

12. Run kMeans with k=8 and join the clustered results with the original attributes.
13. Find the patterns of each cluster (why those instances are clustered together). If you have a Cluster Model Visualizer operator, use it to get more details about the patterns in each cluster. If not, do a manual analysis.
14. Include your interpretations for each cluster in the answer document.
15. Take a screenshot of the subprocess and paste it in the answer document.

Loop Parameters

16. Use "Loop parameters" operator to run kMeans multiple times. Once you get the result, from the "plot view" tab, plot the elbow diagram and paste it in the answer document.
17. Take a screenshot of the subprocess and paste it in the answer document.

Outlier Detection by Clustering

18. Redo clustering with k=20 to see outliers.
19. Filter those clusters with less than 10 instances in it. Now, filter the instances in those small clusters. You must use aggregate, filter examples and join operators for this task. Paste a screenshot of the filtered instances in the answer document.
20. Paste a screenshot of the subprocess and paste it in the answer document.

Common Outliers

21. Join the results from Outliers and OutliersByClustering subprocesses to get the outliers flagged by all three approaches.
22. Take a screenshot of the outliers and paste it in the answer document.
23. Interpret why they are outliers and include your reasoning in the answer document
24. Paste the screenshot of the process where you do the final join and filtering.
25. Now, take a screenshot of the entire 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. Your lab will not be graded if you miss the rmp file OR the answer document.
3. Don't zip files. Zipped files will not be graded.