Data mining: concepts and algorithms

Exam – Data mining

Name: Surname: \_

Student ID: \_ Dataset: \_

# Objective

Exploit data mining algorithms to analyze a real dataset using the RapidMiner machine learning tool.

# Dataset

Download the zip <http://dbdmg.polito.it/wordpress/wp-content/uploads/2020/01/DatasetsUCI.zip> and select **the assigned dataset**. The dataset has a class attribute that will be used during the first part of the exam as label for the classification problem. If needed, a brief description of each dataset is available in the zip folder together with the dataset. The examination test is composed of two parts: the first one focuses on the classification problem and the second one the clustering problem.

# Part I: Classification problem

Analysts want to predict the class of new data, according to the already classified data. To this purpose, you must exploit two different classification algorithms: a decision tree (Decision Tree) and a distance-based classifier (K-NN) to create classification models based on the given dataset and give an answer to a set of questions in order to understand what is the best classification algorithm, in terms of accuracy, for the given problem/dataset.

To evaluate classification performance, different configuration settings have to be tested and compared with each other. A 10-fold Stratified Cross-Validation process must be used to validate classifier accuracy.

# Questions

Answer to the following questions:

1. Learn a Decision Tree using the whole dataset as training data and the default configuration setting for algorithm Decision Tree.
   * Report the tree structure (Take a screenshot and place it below).
   * Which attribute is deemed to be the most discriminative one for service class prediction? Why?

**Answers**:

1. Use a 10-fold Stratified Cross-Validation approach to validate the accuracy of the generated classification model.
   * What is the impact of the **minimal leaf size** parameter on the average accuracy achieved by the generated Decision Tree? Why?
   * Compare the confusion matrices achieved using different parameter settings (keep the default configuration for all the other parameters).
   * Report the accuracy achieved for at least three values of the **minimal leaf size** parameter.

**Answers**:

1. Considerer the K-Nearest Neighbor (K-NN) classification algorithm and use a 10-fold Stratified Cross-Validation approach to validate the accuracy of the generated classification models.
   * What is the impact of parameter **k** on the classifier performance (i.e., accuracy)? Why?
   * Compare the confusion matrices and the accuracy achieved using different values of **k**. Report the accuracy achieved for at least three values of **k**.
   * Does K-NN perform on average better or worse than Decision Tree on the analyzed data?

**Answers**:

# Part II: Clustering problem Goal

Analysts want to split the available data objects in groups. To this purpose, you must exploit two different clustering algorithms: a k-Means clustering algorithm (**K-Means**) and a density-based algorithm (**DBScan**) and select the most appropriate one for the given problem. Consider **only numerical attributes** for the clustering task.

# Questions

Answer to the following questions:

1. Apply the **k-Means** algorithm to cluster the given data and analyze the impact of parameter **k** (number of generated cluster) on the generated clusters.

More specifically, perform an empirical analysis by using the average within cluster distance measure to evaluate the impact of the value of **k** on the quality (in terms of Cluster Cohesion) of the generated clusters.

* + What is the impact of k on the cluster cohesion (average within cluster distance)? Which **k** would you choose as best solution? Why?

**Answers**:

1. Consider the **DBScan algorithm** (a density based algorithm) and compare its performance with that of the **k-Means** algorithm.
   * Report an example of clustering result (Take a screenshot and place it below).
   * What is the impact of parameter **min points** on the performance of DBScan?
   * Strongly increase/decrease the **epsilon parameter** (keep the default configuration for all the other parameters). What is the impact of the **epsilon parameter** onthe noise cluster?Why?
   * Does **DBScan** perform better than **k-Means** (in terms of average within cluster distance)?

**Answers**: