P.PORTO ESCOLA SUPERIOR DE TECNOLOGIA E GESTÃO	Tipo de Prova Assignment 1	Ano letivo 2018/2019	Data
	Curso Mestrado em Engenharia Informática		Hora
	Unidade Curricular Tecnologias Escaláveis para Análise de Dados		Duração

A. Introduction

This assignment is meant to all Erasmus students enrolled in the Curricular Unit Tecnologias Escaláveis para Análise de Dados, of the Mestrado em Engenharia Informática.

This assignment has a weight of 50% on the final score, and a minimum score of 8.0.

B. Enunciado

The student should choose a dataset from an online source such as Kaggle, import it into a MongoDB instance and implement the proposed exercises using the MongoBD's Aggregation Framework

1) Lookup Collection – Create a new collection that contains the unique values of an enumeration in the original dataset, together with a unique numeric code. Replace, in the original collection, each enumeration value with the corresponding code, that acts as a foreign key in the relational model
2) Oversampling – Implement the oversampling technique in an attribute of your choosing containing an enumeration, so that all values are equally represented in the dataset
3) Undersampling - Implement the undersampling technique in an attribute of your choosing containing an enumeration, so that all values are equally represented in the dataset
4) Discretizing – Discretize a numeric or date attribute according to criteria defined by you
5) Probabilistic Analysis – Choose a nominal attribute. Calculate the probability of a document belonging to each of the attribute's value.
6) Tf-idf – Calculate the term frequency–inverse document frequency, a statistical measure that measures the importance of a given word in a set of documents.
7) Index – Create an index of the words that exist in an attribute, containing the id's of the documents in which the words appear
8) K-fold Cross Validation – Generate k different disjoint datasets, each with approximately N/K instances (N is the size of the dataset)
9) Normalization – Create a new variable on the dataset that results from normalizing an existing numeric variable
10) Remove noise – Remove instances containing outlier values in a given attribute
11) Fill missing values– Fill the missing values in an attribute of the dataset with the average value (in the case of a numeric variable) or with the most frequent value (in the case of a enumeration)
12) Pivot Table – Create a collection that represents a Pivot Table (e.g. sum, mean, count) for two variables of your choice.