**I**nstituto **S**uperior de **E**ngenharia de **L**isboa

# Mestrado em Engenharia Informática e Computadores

# Semestre de Verão 2023/2024

**Laboratório 3**

Mineração de Dados em Larga Escala

**Trabalho realizado por:**

Pedro Carvalho nº47113 G01

Nuno Bartolomeu nº47233 G01

Índice

[Part II. Orange 1](#_Toc163463919)

[1) Orange environment and analysis of existing exemples 1](#_Toc163463920)

[2) Feature ranking and selection 2](#_Toc163463921)

[3) Feature reduction with principal component analysis and discretization 4](#_Toc163463922)

[Part III. R Studio 6](#_Toc163463923)

[1) Feature Selection 6](#_Toc163463924)

[2) Feature Reduction 9](#_Toc163463925)

[3) Feature Discretization 10](#_Toc163463926)

# Visualize the dataset

1. Use the sparklyr sdf schema function to check the schema of the df variable.
2. Check the content of the SPARK data frame df, using the head function.

Uma imagem com texto, captura de ecrã

Descrição gerada automaticamente

1. Use the stopifnot function to guarantee that the number of columns and rows in df is correct. To achieve this goal, apply the nrow and ncol functions (or the equivalent in Sparklyr), and compare the values with the ones in Table 1.

Ao usar a função stopifnot esta irá interromper a execução do código e um erro será gerado. O código continuou a correr indicando que o número de colunas(545 atributos) e o número de linhas(2190 instâncias está garantido.

Uma imagem com Tipo de letra, texto, Gráficos, captura de ecrã

Descrição gerada automaticamente

# Feature Selection

1. Use the magrittr’s pipe operator, %>% and the select function to reduce df features to the features in the indexes 1, 2, 5, 6, 9, 10, 11, 14, 16, 17, 19, 21, 24, 25, 26, 31, 32, 33, 34, 35, 41, 44, 49, 50, 54. Store the resulting SPARK Dataframe in the df.sel variable. Notice that the first feature is the dependent variable, named CLASS.
2. Use the head function to overview the resulting dataset.

# Use of generic sampling techniques

1. Apply the sparklyr sdf random split function to produce two datasets: one for training (2/3) and other for testing (1/3). Use the seed value 123, for this and all random functions from this point forward.
2. Use the R table function to determine the number of instances for each class in both datasets. Explain why this function cannot be used directly on df.train and df.test.

Não pode ser diretamente usado nas variáveis df.train e df.test pois a função table espera um objeto R local e essas variáveis são objetos distribuídos do Spark DataFrame que residem em um cluster do Spark.