

Data Science Lab1

Team nr: Insert here	Student 1: Antero Morgado	IST nr: 1119213
	Student 2: David Ferreira	IST nr: 1107077
	Student 3: José Fernandes	IST nr: 1103727
	Student 4: Olha Buts	IST nr: 1116276

CLASSIFICATION

1 DATA PROFILING

May be used to describe any useful observation about the data, and that was used in the current project. An example is the use of any domain knowledge to process the data or evaluate the results. **Shall not exceed 200 characters.**

Data Dimensionality

Shall contain all relevant information and charts respecting to the data dimensionality perspective, such as the number of records and number of dimensions, and their impact on the following analysis. **Shall not exceed 500 characters.**

Figure 1: Nr Records x Nr variables for dataset 1 (left) and dataset 2 (right)

Figure 2: Nr variables per type for dataset 1 (left) and dataset 2 (right)

Figure 3: Nr missing values for dataset 1 (left) and dataset 2 (right)

Data Distribution

Shall contain all relevant information and charts respecting to the data distribution perspective, such as each variable distribution, type, domain and range. May be used to describe any useful observation about the data, and that was used in the current project. **Shall not exceed 500 characters.**

Figure 4: Global boxplots dataset 1 (left) and dataset 2 (right)

Figure 5: Single variables boxplots for dataset 1

Figure 6: Single variables boxplots for dataset 2

Figure 7: Histograms for dataset 1

Figure 8: Histograms for dataset 2

Figure 9: Outliers study dataset 1

Figure 10: Outliers study dataset 2

Figure 11: Class distribution for dataset 1

Figure 12: Class distribution for dataset 2

Data Granularity

Shall contain all relevant information and charts respecting to the data granularity perspective, such as the impact of different granularities considered for each variable. May present additional taxonomies if needed. **Shall not exceed 500 characters.**

Figure 13: Granularity analysis for dataset 1

Figure 14: Granularity analysis for dataset 2

Data Sparsity

Shall contain all relevant information and charts respecting to the data sparsity perspective, such as domain coverage and correlation among variables. **Shall not exceed 500 characters.**

Figure 15: Sparsity analysis for dataset 1

Figure 16: Sparsity analysis for dataset 2

Figure 17: Correlation analysis for dataset 1

Figure 18: Correlation analysis for dataset 2