Lab 3
Data Preparation

GOAL

To prepare the data according to the issues identified in the previous KDD task.

WORK TO BE DONE

To prepare the data, following the needs identified before, with respect to:

Encoding.
Missing value imputation.
Outliers treatment.
Scaling.
Balancing.
Feature selection and generation.

Per each preparation task:

1. Pick the dataset and split it into two datasets: one for training and another for testing.
2. For each preparation task, choose two different approaches and apply them to the training dataset. Then apply the adequate transformation to the testing set, when necessary.
3. Train a KNN and a Naïve Bayes model for each resulting dataset.
4. Compare the performance of both Naïve Bayes and both KNN models and select the dataset that yields the best performance.
5. Pick this dataset to proceed to the next preparation task.

REPORT TO DELIVER

PDF file only with charts and tables. No analysis or justification is needed at this point.
Suggested charts per dataset, and per preparation task:

Naïve Bayes and KNN performance per each approach applied.
Identification of the best dataset and corresponding approach (best approach).
Naïve Bayes and KNN confusion matrices for the best approach.

Good work!!!