TAA 3

1 Anomaly detection

1. Preparing the dataset

   (a) Load the `gapminder_data_graphs.csv` dataset. Filter to keep only the year 2015. Remove textual columns.

   (b) Apply an anomaly detection using `IsolationForest` from sklearn.

   (c) Plot the distribution of anomaly scores (`decision_function`)

   (d) Plot the most anomalous countries

   (e) To understand why these countries are considered anomalies, print for each variable the difference from the mean. Using a relative distance (scale by variance) can be more informative. Interpret. Use also a dimensionality reduction technique (e.g., PCA) to visualize the relative position of those anomalous nodes.

   (f) Do the same process using PCA as anomaly detection method. You can use `pca.inverse_transform`

   (g) Do the same process using Gaussian Mixture (`score_samples`).

2 Imbalance

(a) Load the `cars_synth_clean.csv` dataset

(b) Create a new target variable called `15k+` , which is true if a car is worth more than 15 000. Remove the price column

(c) Check that there is class imbalance

(d) Using `RandomForestClassifier` perform a classification to predict this target variable.

(e) Compute the Accuracy, the ROC_AUC. Plot the confusion matrix

(f) Using `RandomUnderSampler` from `imblearn` library, perform under-sampling. If you prefer, you can do it manually.

(g) Check the new scores for accuracy, ROC_AUC. Plot the confusion matrix

(h) Explain the difference: what is this model better and worst at doing ? How can it explains the score differences ?

3 Feature selection

(a) In the `cars_synth_clean.csv` dataset, use a `clustermap` from `seaborn` library to visualize the correlations between variables (`df.corr`)

(b) If we were using a threshold of 0.5, what variable should we keep?

4 Going Further

2. SMOTE and SMOTER

   (a) Use SMOTE strategy from library `imblearn` to do class imbalance correction and compare the results

(b) When trying to predict directly the price, you are confronted to the same problem of data imbalance. Plot the distribution of this target value to observe this imbalance. Train a model to predict directly the price. To observe the poor predictions on rare values, you can draw a scatterplot with a relation between target value and average errors.

(c) Search a solution to perform SMOTER. Observe how the performance is affected.