# Introduction

Clustering is a very well-known unsupervised machine learning technique that has been used for a long time, through different algorithms, it allocates data points that are somehow similar to each other into distinct groups. It has a wide range of applications in different domains and multiple algorithms based on different metrics that originate different clusters. However, grouping data into meaningful clusters can sometimes be quite challenging, from selecting the most suitable algorithm to the ongoing increase of the data’s complexity itself, the clusters found may not produce the best results.

On the other hand, when dealing with classification problems, supervised learning, not every sample from each class behaves the same way, creating a sort of subgroups among those classes. Finding these groups can reveal to be tricky as the classification models don’t consider them directly and labelling the samples correctly, prior to the application of a model is very costly to achieve.

This thesis explores a different approach, Supervised Clustering with SHAP values, to identify clusters and different behaviour samples among labelled classes and compare it with the state of art approaches to access if it performs better. Through the use of an explainable artificial intelligence (XAI) tool, Shapley Additive exPlanations (SHAP), as a pre-processing step

# Methodology

## Pre-Processing techniques

The supervised clustering with SHAP values approach requires pre-processing procedures so the clustering algorithms can be applied efficiently. In this section, it is carefully explained the multiple steps of the approach as well as the pre-processing techniques detailed in the order they were implemented.

Two different approaches are taken, at first, the normal unsupervised clustering is conducted, followed by the supervised clustering with SHAP values technique. Both approaches require pre-processing procedures to be able to apply the clustering algorithms efficiently. In this section, all the pre-processing and clustering behind the approaches are carefully explained in the order in which they were implemented.

### Feature Selection

The quantity of features varies according to the domains and models analysed, often datasets come with a reasonable large number of features, including features that are considered noise and end up being irrelevant to the machine learning models. Feature selection focus on only retaining the relevant features to be used by the models.

Among the feature selection techniques, the Pearson correlation coefficient was the one chosen. It measures the strength of the linear relationship between the features assigning a correlation value to each pair of features. A threshold value is selected, eliminating all the features with an absolute correlation above the threshold settled.

where is the number of features of the correspondent dataset, is the number of samples of the dataset, are the feature and the feature respectively, correspond to the sample of the feature , same reasoning is applied for , corresponds to the sample mean for the respective feature. is the correlation between each feature pair and can take a value between -1 and 1, where the larger the absolute value is, the stronger is the relation between the two features.

The Pearson correlation coefficient was selected since it is an efficient technique to eliminate features that don’t give any new contribution to the models and it aligns with the SHAP values methodology, that will be explained further bellow in this section, as they can’t deal perfectly with correlated features.

### One-Hot encoding

### Standardization of the features

The features of datasets usually measure different units and often come in different scales, for some machine learning estimators, it is required to apply standardization techniques to the datasets so that each feature equally contributes to the output of the model.

The inbuilt class StandardScaler of Scikit-learn Python library was used for this purpose.

for each sample , it calculates a standard score where is the samples mean and is the samples standard deviation.

### SHAP Values

SHAP values are an adaptation of the Shapley values to the machine learning framework, instead of having games and players, it is focused on models and features. For each feature, it measures the contribution to the model output by doing a weighted summation average of all possible features subsets and calculating the marginal contribution of that feature to the model.

The SHAP values are a local method, which means that for every row/instance in the dataset, it calculates the SHAP values for the feature , given a model and a vector of features to be explained;

where is a subset of the features used in the model and the total set of the features in , having as all the possible subsets excluding the feature , and corresponds to the predictions of the model , given a set of features , with and without the feature , respectively, marginalized over the features that are not included in that set .

### Dimensionality Reduction

With big amounts of features to deal with, some machine learning algorithms struggle in processing all that information, making it harder to find a solution in a feasible amount of time (curse of dimensionality). On the other hand, it is hard for humans to interpret graphs above 3 dimensions, so to produce meaningful visualisations of data and for machine learning algorithms to be efficient, it is necessary to reduce the dimension of the features to a lower dimensional space.

This technique plays an important role when it comes to clustering algorithms as it allows the algorithm to overcome the curse of dimensionality as it allows to visualize the data in two or three dimension graphs that humans can extract information off.

Essentially, the Dimensionality reduction algorithms fall in two categories; matrix factorization or neighbour graphs. The first category tends to preserve more the global structure of the data while the second category prioritizes the preservation of the local structure of the data. Therefore, these algorithms play a trade-off of trying to keep the maximum structure of the data while reducing its dimension.

Uniform Manifold Approximation and projection (UMAP) was the algorithm chosen for this approach. This algorithm belongs to the neighbour graphs category and can separate and identify clusters on higher dimensions much better than algorithms like Principal Component Analysis (PCA) that belong to the matrix factorization category.

UMAP also tends to outperform t-distributed stochastic neighbour embedding (t-SNE) when it comes to scalability, achieving results much faster, and also on preserving the global structure of the data, which is very important when it comes to clustering analysis.

Leland, John and James (2018), the authors of UMAP, carefully explain all the mathematical assumptions and algorithms behind it.