**DATA AND INFORMATION QUALITY PROJECT REPORT**

PROJECT ID: 29

PROJECT NUMBER: 1

ASSIGNED DATASETS: abalone, users

STUDENTS:

Gabriele Ginestroni (10687747)

Tommaso Capacci (10654230)

ASSIGNED TASK: Clustering

1. **SETUP CHOICES**

**Chosen ML algorithms:**

Since the given datasets presented different characteristics we decided to use different algorithms for each of them.

In particular, for the “abalone” dataset we used:

* **K-Prototypes**: is a clustering method used to cluster mixed-type data (i.e. data that contains both categorical and numerical variables). This algorithm combines the K-Means algorithm for numerical variables and the K-Modes algorithm for categorical variables.

It works by first initializing the centroids for each cluster using either random sampling or a user-specified initialization method. Then, for each iteration, it assigns each data point to the cluster with the closest centroid. Next, it calculates the new centroid for each cluster by taking the mean of the numerical variables and the mode of the categorical variables of the data points assigned to that cluster. Finally, it repeats these steps until the centroids no longer change or a maximum number of iterations is reached.

* **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** is a density-based clustering algorithm that groups data points that are closely packed together (i.e. dense regions of data points) and separates data points that are sparsely located (i.e. less dense regions of data points).

DBSCAN works by defining a neighborhood around each data point, and then identifying clusters based on the density of data points in these neighborhoods. The algorithm takes as input two parameters: eps and minPts.: the eps parameter defines the radius of the neighborhood around each data point, and minPts defines the minimum number of data points required to form a dense region.

For the “users” one, instead, we opted for:

* **K-modes:** is a clustering algorithm that is similar to the k-means algorithm but is used for categorical data rather than numerical data. Instead of calculating the mean of a cluster, it calculates the mode, which is the most common value in a cluster.
* **Spectral clustering**: is a clustering algorithm that can be used to find clusters in non-linearly separable data. It works by first constructing a similarity matrix for the data, which encodes the similarity between each pair of data points. The algorithm then constructs the Laplacian matrix, which is a transformation of the similarity matrix. The Laplacian matrix encodes the relationships between data points in terms of their similarity and proximity to one another. Next, the algorithm finds the eigenvectors of the Laplacian matrix corresponding to its K lowest non-zero eigenvalues, where K is the number of clusters. These eigenvectors are used as the coordinates of the data points in a new, low-dimensional space. Finally, the data points in this new space are clustered using k-means algorithm.

**Chosen ML performance evaluation metrics:**

In order to analyse all the relevant aspects we decided to use 2 different types of metrics: the first one gave us general information about how good the results of the application of our ML algorithms were (in fact we also used them to select the values of their parameters), the second one instead allowed us to compare the results of the same algorithms applied on the imputed datasets and on the original ones.

Inside the first category we chose to use:

* **Silhouette score**
* **Calinski-Harabasz score**

While for the second one:

* **Adjusted Mutual Information (AMI)**
* **Completeness score**

**Imputation techniques selected:**

As for the ML algorithms, different datasets required different imputation techniques due to the nature of their features.

“Abalone” dataset:

**Standard**) Single imputation using the mode of features

**Advanced**) Multiple Imputation by Chained Equations (MICE) with a KNN classifier

“Users” dataset:

**Standard**) sklearn SimpleImputer to infer the mode

**Advanced**) KNN classifier

2. PIPELINE IMPLEMENTATION

a. Description of the steps you performed

3. RESULTS

a. Description of the main results obtained

b. ML performance comparison between the imputation/outlier detection techniques you have implemented