Data And Information Quality Project Report

- **PROJECT ID**: 21
- **PROJECT NUMBER**: 1
- **ASSIGNED DATASET**: USERS
- **STUDENT**: PASQUALE CASTIGLIONE 10657816
- **ASSIGNED TASK**: CLUSTERING

Contents

1 SETUP CHOICES

1.1 Clustering

Clustering was performed used two different methods: K-Modes and K-Means.

- **K-Modes** Because of the categorical nature of the data, *K-Modes* was used. In order to select the number of clusters, *elbow* method analysis was performed.
- **K-Means** The second clustering techniques used was K-Means using Jaccard as distance measure.

1.2 Chosen ML performance evaluation metrics

1.3 Imputation

1.3.1 Simple Imputation

Simple imputation was performed by propagating valid values to the cells with missing values using *fillna* method from *Pandas*. Firstly values were propagated forward then, to avoid null values in the first row, values were propagated backward also. This method showed very good result with the original data, but applying it to a shuffle version of the dataset showed worse results.

1.3.2 Advanced Imputation

K-Nearest Neighbors was used as the advanced technique used to impute missing value. In order to apply this method, data were firstly encoded as one-hot numeric arrays, then *KNNImputer* from *sklearn* was used to fit and transform the dirty dataset. This method turned out to be robust to shuffling.

2 PIPELINE IMPLEMENTATION

2.1 Description of the steps you performed

- **Data Exploration:** In this phase histograms of the datasets and heatmaps of the datasets with missing values were plotted.
- **Imputation:** In this phase imputation was performed. In particular for the advanced imputation, features were encoded as one hot vector using the *get_dummies()* method from *pandas*

3 RESULTS

3.1 Imputation

To assess the quality of the imputation, accuracy was computed as the number of rows that were imputed as the original over all the rows.

3.1.1 Simple Imputation Accuracy

	CT	CU	LT	TC
50%	0.92	0.91	0.9	1.00
60%	0.92	0.93	0.93	1.00
70%	0.94	0.97	0.93	0.99
80%	0.96	0.95	0.97	1.00
90%	0.97	0.99	0.99	1.00

Table 1: Simple Imputation

	CT	CU	LT	TC
50%	0.69	0.81	0.67	0.61
60%	0.71	0.83	0.76	0.69
70%	0.83	0.90	0.81	0.78
80%	0.89	0.93	0.87	0.83
90%	0.91	0.98	0.94	0.91

Table 2: Simple Imputation Shuffled

3.1.2 Advanced Imputation Accuracy

	CT	CU	LT	TC
50%	0.76	0.90	0.79	0.80
60%	0.84	0.92	0.85	0.87
70%	0.88	0.96	0.92	0.91
80%	0.94	0.97	0.95	0.94
90%	0.97	1.00	0.99	0.98

Table 3: KNN Imputation

	CT	CU	LT	TC
50%	0.79	0.91	0.78	0.78
60%	0.83	0.91	0.86	0.83
70%	0.88	0.94	0.91	0.91
80%	0.96	0.96	0.94	0.95
90%	0.98	0.99	0.99	1.00

Table 4: KNN Imputation Shuffled

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