Data And Information Quality Project

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

Contents

1 SETUP CHOICES

1.1 Chosen ML algorithms

1.2 Chosen ML performance evaluation metrics

1.3 Imputation techniques selected

2 PIPELINE IMPLEMENTATION

2.1 Description of the steps you performed

3 RESULTS

3.1 Description of the main results obtained

3.1.1 Simple Imputation

	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

3.1.2 Simple Imputation Shuffled

	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

3.1.3 KNN Imputation

	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

3.1.4 KNN Imputation Shuffled

	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

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