|  |  |
| --- | --- |
| **Q1: Classification** | |
| **Features Ranking** |  |
| Measure used for ranking is mutual\_info\_classif from the library sklearn.feature\_selection. It computes the mutual information between each feature and the target using the mutual\_info\_classif() function from scikit-learn. Mutual information measures the amount of information shared by a feature and the target, so it provides a score of how useful a feature is for predicting the target.  feature mutual\_info  10 alcohol 0.086082  9 sulphates 0.057628  1 volatile acidity 0.050628  2 citric acid 0.035945  0 fixed acidity 0.033534  7 density 0.031057  4 chlorides 0.030113  6 total sulfur dioxide 0.022118  8 pH 0.016291  5 free sulfur dioxide 0.009932  3 residual sugar 0.002430 | |
| **Best Parameter values, for Model** | **Result: F1, macro-precision, macro-recall** |
| **Neural Net:**  Hidden nodes/ layers: (100,50)  batch-size=500 | **Neural Net (in code nn2):**  F1=0.702  macro-prec=0.738  macro-rec=0.679 |
| **Decision Tree**  Dept=5 | **Decision Tree (in code dtc2):**  F1=0.699  macro-prec=0.697  macro-rec=0.701 |
| *Some conclusions for Q1: one-paragraph of critical thinking of the results of classification.*  *Overall, the results showed that the Decision Tree with max\_depth=5 performed better that the other one with max\_depth=3 that we experimented with, achieving an F1 score of around 0.70. Among the Neural Networks, the one with two hidden layers of sizes 100 and 50 performed the best, achieving an F1 slightly above 0.70. These results suggest that a Neural Network with multiple hidden layers is a little bit better suited to this dataset than a Decision Tree with a deeper maximum depth.* | |

**Report**

**Alkiviadis Kariotis 241735**

|  |  |
| --- | --- |
| **Q2: Clustering Report** | |
| **K-Means** | **10 clusters, Silhouette-coefficient = 0.39, inertia = 86577.81** |
| *Some conclusions for Q2: one-paragraph of critical thinking of the results of classification.*  *The results of the k-Means clustering on the wine quality dataset showed that the best number of clusters to use depended on the evaluation metric used. When evaluating based on inertia, which measures the sum of squared distances between each data point and its cluster centroid, the best result was obtained with k=10 clusters. On the other hand, when evaluating based on silhouette score, which measures the similarity of each data point to its own cluster compared to other clusters, the best result was obtained with k=2 clusters. Therefore, there is no single best number of clusters for this dataset, and the optimal number depends on the evaluation metric and the specific goals of the analysis. But the most optimal choice from the choices that we explored is to go with k=10.* | |