Practice 2A: Feature Preprocessing

Supervised Learning
Wine dataset

1. Load the dataset provided on iCorsi (`wine_red.csv` and `wine_white.csv`). You can find them in the datasets folder.

2. There are two files, one for red wines and the other for white wines. Merge the datasets eventually resulting in a single dataframe.

3. How big is the final dataset you are working with?

4. How does the dataset looks like? You can visualize first five and last five rows of the dataframe.

5. Try to graphically visualize the dataset.

   (a) There is a column named `TARGET` (that represents the *quality* of the wine), visualize the distribution of its values (you can use `countplot` from `seaborn`).

   (b) Visualize the histogram or bar plot of all the features in the dataframe.

   (c) Visualize the joint scatter plot of two variables that you guess are important in determining the quality of the wine (later we will use methodologies to find important features). You can use `joint plot` from `seaborn`.

   (d) Visualize the Probability Density Function (PDF) of each feature the dataset, you can use `displot` from seaborn.

   (e) Visualize the pair plot matrix to understand the features pairwise relationship.

   (f) Experiment on other kinds of plots.

6. What kind of features are there in the dataset? Analyse the type of columns in the dataframe.

7. Are there any outliers in the data? How do you detect and handle them? Reason on your choices.

8. Are there any missing values in the data? How do you detect and handle them? Reason on your choices.

9. Use KNNImputer from scikit-learn to impute the categorical variable.

10. Write half a page report about data exploration and data quality handling (outliers and missing values) and submit it on the dedicated section of the iCorsi page.