Homework 2

Data Analysis and Classification 2019-2020
Logistic Regression

winequality-red.csv dataset (iCorsi)
winequality-white.csv dataset (iCorsi)
research paper: Cortez et. al. [2009] (iCorsi)

This homework has to be developed on a Jupyter Notebook. Each question needs to have at least a Code Cell (implementation) and a Markdown Cell (explanation and/or answer). The notebook developed, named as **<surname_homework_2>.ipynb** has to be sent via email at michela.papandrea@supsi.ch by sunday 3.11.2019.

This homework is based on the wine quality classification research paper and dataset shared on the course page.

What to do

The homework consists in building a classification model which is able to predict the quality of the wine (binary classification) based on its physicochemical values.

Approach

The idea is to apply different approaches and evaluate them, in terms of Accuracy, Precision-per class and Recall-per class. You can apply the hold-out validation methodology, selecting randomly 20% of a dataset for testing. Per each approach, plot also the Confusion Matrix and make some reasoning over it.

1. Apply a Logistic Regression algorithm on the 2 datasets, separately. Create a new target class binning the quality in 2 sets:for example, if quality>=5 => target_quality=high, if quality<5 => target_quality=low. Build a model for the red wine dataset, and a model for the white wine dataset.

2. Apply Logistic Regression algorithm on the complete dataset, considering together red and white wines, and adding a feature column named "wine_type", whose elements can assume the values {'red', 'white'}.

3. Apply a Linear Regression algorithm, considering the quality measure learned as a continuous value. When evaluating the prediction, make an approximation on the quality value predicted by the linear regression model to retrieve a target_quality (with the same procedure asin point 1). Evaluate the target_quality prediction on the binned real_quality.