Project Proposal

Guilherme Peralta – up201704277

Pedro Duarte Lopes – up201706106

Project Title: Vinho Verde’s Quality Classification Using Machine Learning

Data Set: Samples from 1599 red wines and 4898 white wines[1];

- Attributes:
  - **Input variables** (based on physicochemical tests):
    - fixed acidity;
    - volatile acidity;
    - citric acid;
    - residual sugar;
    - chlorides;
    - free sulfur dioxide;
    - total sulfur dioxide;
    - density;
    - pH;
    - sulphates;
    - alcohol.
  - **Output variable** (based on sensory data):
    - quality (score between 0 and 10 according to a median of at least 3 evaluations made by wine experts).

Project Idea: The main goal is to create a model which is able to predict the quality of a wine based on its physicochemical properties and sensorial properties, evaluated by a panel of experts. The relation between the two properties could mean an enhancement in the wine quality all over the sector.

In order to achieve this goal, the following steps will be taken:

- Data pre-processing:
  - Normalization of the data;
- Separation of the train and the test data from the data set:
  - Check if the distribution of the data is equal between train and test data;
- Feature evaluation regarding data dispersion (ex: average, variance, etc);
- Choosing a model to relate the features and the output;
- Training the model using the train data from the data set;
- Evaluate the model using the test data from the data set;
- Compare the model with previous results from the chosen articles.

Required Software: Python using Scikit-learn, NumPy and Pandas libraries.

Papers to read:

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., &amp; Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553. doi:10.1016/j.dss.2009.05.016

[2] Nebot, Àngela & Escobet, Antoni. (2015). Modeling Wine Preferences from Physicochemical Properties using Fuzzy Techniques. 501-507. 10.5220/0005551905010507.