Project proposal for Machine Learning

Dataset: Wine quality dataset [1].
Topic: Predicting the wine quality using Random Forest approach.
Team members: Javier de Jesus Flores Herrera, Nejada Karriqi

Most of the people like to drink wine while they enjoy their food. This has made the wine industry a very important industry nowadays and people are buying wines expecting a good quality, since it is impossible for a person to infer the quality of the wine just by looking at the bottle and we cannot taste it before buying it. We would like to offer another solution to the wine lovers by building a regression model using random forest (RF) algorithm to meet the goal: The identification of wine quality. For this approach, we will be doing some exploratory analysis to understand the data we are working with and based on this analysis we would proceed with the data preprocessing. If necessary, we might consider as part of our project, using a outlier detection algorithm in order to identify the excellent and poor wines.

For this model we decided to choose this dataset because it contains enough observations, which make it easy to create a model and be able to split the data into two subsets: training and test for the model validation. It is composed by two subsets: red and white wine variants. The red wine contains 1599 observations and the white wine 4898 observations. Both wines have 12 attributes which the input values can differ from each wine type. For each wine there are different attributes that play important role to its quality. We will identify the most important ones and we will focus on them when creating the model in order to have a pleasant result.

Here are some papers we are going to use to support our project development:

[1] Paulo Cortez, Antonio Cerdeira, Fernando Almeida, Telmo Matos, and Jos Reis. Modeling wine preferences by data mining from physico-chemical properties. Decision Support Systems, 47(4):547 – 553, 2009.Smart Business Networks: Concepts and Empirical Evidence.

[2] Andrew T. Jebb, Scott Parrigon, and Sang Eun Woo. Exploratory data analysis as a foundation of inductive research. Human Resource Management Review, 27(2):265 – 276, 2017.