# Problem Identification - Project Charter

Group 8

Ahmed Khair, Jiazu Zhang & Mengjia Kira Wei

Thursday, September 7, 2023

Dataset: <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>

**1 Overview**

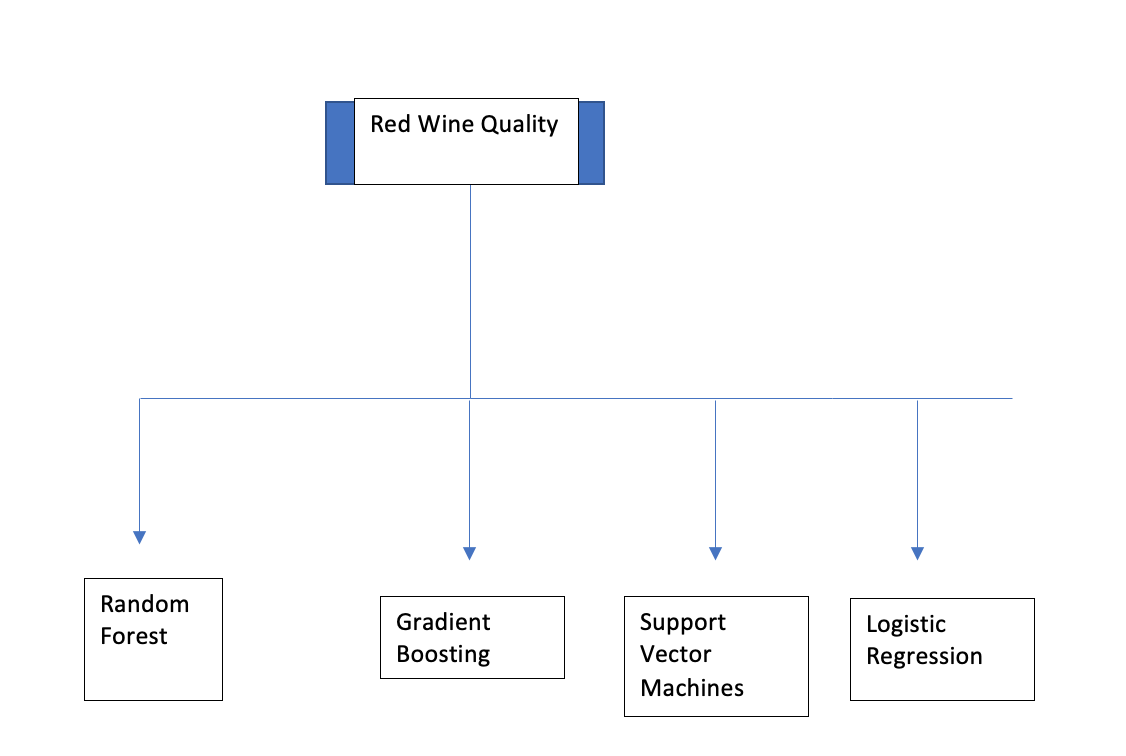
Wine tasting and quality assessment are intriguing subjects due to the intricacies involved in evaluating wine. Factors such as human preferences, grape types, and region all contribute to the complexity of this problem. This research focuses on wine quality. Understanding and quantifying wine quality is vital for the wine industry's growth and competitiveness, our solution will assist wine companies and investors in making strategic decisions. Both classification and regression techniques will be employed to classify and predict wine quality scores.

**1.1 Project Background and Description**

The sample dataset that will be used in this research is the wine dataset from the UCI Machine Learning Repository (Wine Quality, 2009). The dataset contains physicochemical and sensory data of red and white variants of the Portuguese "Vinho Verde" wine. According to TPN/Lusa (2022), Vinho Verde wine is one of the most popular portuguese wines that get exported to around 104 markets around the world. This will provide us with a good variety of both red and white wine data. The study will test the physicochemical variables' effect on wine quality and build a model to predict wine quality based on those variables. The solution can also be applied to other wine types.

**1.2 Project Scope**

By using the wine quality prediction dataset, we will utilize generative methods such as gradient boosting, random forest, support vector machine, and logistic regression. These generative methods will be trained using the training data and will be evaluated on the test dataset in terms of different matrices such as accuracy, precision, recall & F1 score. Linear regression will be used as a baseline model and based on the performance of all the models, we will select the best-performing models and try to enhance their performance by doing fine-tuning and resampling techniques. Thus, the model will be trained again using resampled data and the best set of parameters to check the updated performance of the model.



The deliverable of this project is to assess the quality of wine using attributes provided in the dataset. Firstly, we will clean the data and conduct any relevant pre-processing to ensure the data is ready for exploratory data analysis, then we split the data and train the model. After training the model, we will evaluate the generative methods on the test data and select the best-performing models in order to improve their performance as previously mentioned. Ultimately, the end goal of the project is for it to be deployed either using Heroku or Docker.

In terms of the project, non-generative methods would be considered out of scope as well as predicting the quality using features that are not present in the dataset such as customer preferences, brand names, etc.

## **1.3**. S**uccess criteria and constraints**

1.3.1. The success of the ML model is evaluated based on two criteria. One is the model accuracy. In classification models, the accuracy is measured with precision, recall and F1 scores. In regression models, MSE, R-squared, p-value, AIC & BIC are essential accuracy metrics to evaluate the success of modeling. The second criteria is whether the model makes sense from the food science and wine industry’s perspective, i.e., the features selected as predictive variables should be meaningful and recognized by the industry. The deployed model should also have enough Generalizability, meaning that the model’s accuracy should also be high when applied to the population where the sample is drawn or other sampled wines. Above all, the model should be able to help wine businesses to succeed in terms of wine inventory selection from the whole-seller, shelf displays and promotions.

1.3.2. There are three main constraints of this project. Given that the target variable is numeric (3-8), the subjectivity of bin size for the classification would largely impact the accuracy score, i.e., dividing the quality scores into 2 bins (good and bad) would be easier to achieve high accuracy than dividing into 5 bins. There are no standard criteria about quality scores categorization, adhering to the most popular industry and business requirement is the rule. Secondly, the dataset includes data only up to 2009, which might be missing recent wine quality measures and trends. Thirdly, the end users of the deployment might not have access to all the wine features used in the model, then the accuracy of the prediction would be impacted as a result. Third, this dataset is limited to Portuguese wines, hence the model’s result might only be applicable to this specific wine type.

References

*Record year for Vinho Verde*. The Portugal News. (2022, January 25). <https://www.theportugalnews.com/news/2022-01-25/record-year-for-vinho-verde/64832>

*Wine quality*. UCI Machine Learning Repository. (2009). <https://archive.ics.uci.edu/dataset/186/wine+quality>