# **Wine Quality Prediction**

Statement of Work

Submitted by –

Vasundara Chandre Gowda

100800734

Contents

[**Wine Quality Prediction** 1](#_Toc55588198)

[Submitted by 1](#_Toc55588199)

[Problem Statement 3](#_Toc55588200)

[Scope of the project 3](#_Toc55588201)

[Dataset 3](#_Toc55588202)

[Independent variables 3](#_Toc55588203)

[Dependent variables 3](#_Toc55588204)

[Dataset Source 4](#_Toc55588205)

[Assumptions 4](#_Toc55588206)

[Steps to perform the classification 4](#_Toc55588207)

[Working on new data 5](#_Toc55588208)

## Problem Statement

A wine has a lot properties which determines the quality of wine. However, for a layman, it is difficult to judge the quality of wine based on parameters such as acidity, sugar etc. Canada’s wine industry creates about $6.8 billion economic impact and is a significant driver to Canadian economy. More than 31000 jobs in Canada are related to wine industry. Canadians enjoy more than 1 billion glasses of wine, therefore, as a consumer it is important to know the quality of wine. Since, it is hard to know what makes a good wine, therefore the quality can be judged by applying Machine Learning algorithm to various factors.

## Scope of the project

The scope is limited to predictions only of wine quality and no other beverage. Once the best model is selected and it performs well on the testing data, the final model and the code will be the deliverable.

## Dataset

For this problem, I will be using Wine Quality Dataset from Kaggle, that will help use to predict whether the quality of wine is good or not. The variables for predicting the wine quality are-

### Independent variables

* Fixed acidity
* Volatile acidity
* Citric acid
* Residual sugar
* Chlorides
* Free Sulphur dioxide
* Total Sulphur Dioxide
* Density
* Ph
* Sulphates

### Dependent variables

* Quality (1-10)

The dataset has about 1600 rows and 12 columns.

Since it will be hard for regression as well as classification algorithm to predict 10 different outputs just with a small amount of data, therefore we will manipulate the dataset and change the output variables by changing the output into categorical variables as follows-

* The score between 1 and 4 will be modified to **poor.**
* The score between 5 and 7 will be modified to **normal**.
* The score between 8 and 10 will be modified to **good**.

Therefore, with the three unique outputs rather than 10, it will be easy for and algorithm to predict the quality of the wine.

## Dataset Source

The source of the dataset is-

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

## Assumptions

It is assumed that the dataset is from authentic sources. Therefore, no changes will be made to the original values of the dataset. There seems to be the sufficient variables in the dataset, therefore we assume that there is no default assumptions to be made.

## Steps to perform the classification

The steps that will be followed to perform the classification are-

1. Import the dataset
2. Clean the dataset. This will include removing the null values, removing the columns that are not useful for the prediction.
3. Data Visualization. In order to get the general idea about the data like the distribution of individual columns as well as determining the relationship between two variables
4. Preprocess the dataset. Preprocessing will include converting the string values into numeric format and converting the data which does not follow any ranking into One hot encoding. Also, preprocessing will include removal of outliers and scaling the data.
5. Model Fitting. Various types of classification models will be fitted. Each model will have certain parameters, and best parameters for each model will be selected through Grid Search Cross validation. Some of the models are-
   * Logistic Regression
   * Decision Tress
   * Random Forests
   * Support Vector Machines
   * Neural Networks
6. Model Selection. This includes selecting the best model out of all the models trained. We can achieve this through various metrics and separating the validation data. After the best model is selected, the next step will be moving towards model evaluation.
7. After the model has been trained on out training data, the model will be tested on testing data, which is separated from the training data. The model will be evaluated on various metrices such as accuracy, precision, recall, and f1-score. If the model is performing well on testing data, the model will be deployed on production, otherwise, one must rework on the data.

## Working on new data

After the model is deployed on production, we have to make the prediction on new data, that is the parameters received for the wine quality. The new data will be preprocessed in the same was as it was done on training data. The model that we have trained and tested will be saved on a particular location as a pickle file. The preprocessed data will be passed in the model and the user will receive the output as prediction from the model, which will be the quality of the wine.