Project Proposal

Problem Statement and Motivation:

Predict the quality of red and white wines based on certain features and understand how these features affect each other.

The conclusions from this capstone analysis are meant to inform, enrich, entertain, and inspire people interested in wines. The analysis will help answer questions like what does 'structure' of a wine imply/mean, the 'science' behind wine aromas, or what is "freshness" in wines. The project will give us a deeper, more scientific understanding of the structure of wines which might help get a more accurate and scientific insight into suitable foods/cheeses pairings with a particular type of wine, since it all depends on taste and flavor of the wine and food in question.

The conclusions from this analysis might also help companies like the Wine Enthusiast which is a prominent multichannel marketer of wines, addressing the wholesale, retail, and consumer-direct markets of wine, which seek to provide answers to similar questions. This analysis will also cater to a more independent search for anyone looking to buy a wine which suits his taste and budget.

The clientele interested in this analysis will be anyone interested in wines, who wants to know more about what makes a wine taste sweet/sour/bitter, what makes a good quality wine, and the differences between red and white wines, so they can make a well informed decision when buying/suggesting wines.

Data:

Datasets for red and white wines will be obtained from Kaggle.
https://www.kaggle.com/rajyellow46/wine-quality

Description of dataset:

The data includes information about red and white vinho verde wine samples, from the north of Portugal, with their 11 features: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol.

Method:

1. **EDA and visualization:**

The dataset will be checked for missing values and will be cleaned if needed.

Exploratory Data Analysis will be performed on red and white wine quality datasets.

Correlation heatmaps will be drawn to visualize different features for both the wine types. Features with significant correlation will be plotted in the form of boxplots to investigate the exact relationship between them.

For ease of visualization of boxplots, the wine quality which ranges from 0 to 10, will be divided into three types: high (8 and above), medium (5-8) and low (<= 5).

Scatter plots may be plotted to further visualize the different features of wines.

2. **Statistical Testing**

After EDA, the dataset features will be tested for statistical significance to check if there are strong correlations between pairs of independent variables or between an independent and a dependent variable in the dataset.

Appropriate methods and conclusions will be drawn in order to build a Machine Learning model which predicts the wine quality.

3. Machine Learning

Methods for building the ML model could be a Decision Tree classifier, Random Forest, Support Vector Classifier, KNN or regression.

One or more of these methods will be applied to the dataset to check which model performs the best in terms of accuracy, precision, and sensitivity.

Deliverables:

Deliverable include source code, a report and presentation slides.