**DDS DS 6306 Project 2: Wine Quality**

**A computer on a table with a vineyard and wine glass

Description automatically generated**

**Introduction**

Wine quality is a multifaceted attribute influenced by a myriad of factors, including chemical properties, production techniques, and environmental conditions. The goal of this data science project is twofold:

1. **Objective A**: Build a predictive model that minimizes the Mean Absolute Error (MAE) on the test set.
2. **Objective B**: Analyze the data to uncover key factors that contribute to high-quality wine, providing actionable insights for vintners to enhance their production processes.

**Objective A: Building a Predictive Model**

The primary objective is to develop a robust predictive model that accurately forecasts wine quality based on various features. We aim to achieve the lowest possible Mean Absolute Error (MAE) on the test set, ensuring high precision in our predictions.

**Mean Absolute Error (MAE)** is a commonly used metric for evaluating regression models. It measures the average magnitude of errors in a set of predictions, without considering their direction. The formula for MAE is:

A math equations and numbers

Description automatically generated with medium confidence

Minimizing MAE ensures that our model predictions are as close as possible to the actual values, providing a reliable tool for assessing wine quality.

The student with the lowest MAE on the test set will win an extra 5 points on the project and Bragging Rights! ☺

**Objective B: Mining Data for Insights**

Beyond prediction, this project aims to delve into the dataset to identify the key determinants of high-quality wine. By analyzing the relationships between various chemical properties and wine quality, we can provide valuable insights to vintners, potentially leading to improvements in wine production.

Key factors to investigate may include:

* **Acidity Levels**: Both fixed and volatile acidity could influence the taste and preservation of wine.
* **Sugar Content**: Residual sugar levels may affect the sweetness and fermentation process.
* **pH Level**: The pH of wine can impact its color, stability, and overall balance.
* **Alcohol Content**: Higher alcohol levels might correlate with perceived wine quality.

By examining these and other variables, we aim to identify patterns and correlations that distinguish high-quality wines from lower-quality counterparts. These insights can then inform production techniques, helping vintners optimize their processes to produce superior wines.

**Feature Details:**

Input variables (based on physicochemical tests):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 – alcohol

12 – Type

13 - Location

Output variable (based on sensory data):

14 - quality (score between 0 and 10)

The features above include objective tests (e.g. PH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

**Audience:**

You have been contracted by [Robert Renzoni](https://robertrenzonivineyards.com/) Vineyards in Riverside California. All presentations are geared towards Head Vintner Olivia Bue.

**Rubric:**

Objective A: 30%

* Models that use concepts like feature engineering and imputation of missing values will receive higher scores.
* Note: all missing values in the dataset can be considered to be missing at random.

Objective B: 30%

* Identify the “top” features related to the quality of the wine. “Top” is subjective here, it could be one it could be 10, it is up to you to identify and communicate the features that are practically important.
* Of course it is not enough to simply identify the features. Explain how and the extent to which they are important.

Video Presentation: 30%

* Polished Opening
* Avoid extensive reading from slides
* Labeled Plots
* Faded Bullet points
* Voice Inflection
* Consistent format
* No spelling or grammatical errors
* Polished exit / conclusion slide
* Staying with the max time

Peer Review Submission: 10%

**Dates:**

Final Project Deadline: Sunday, December 15th at 11:59pm CST

**Deliverables:**

7 min recorded presentation with the link posted on this Google Doc:

<https://docs.google.com/document/d/1iTB-ccaY03L40rSBgCY8l4J5JISUU21hEaq829BL0nQ/edit?usp=sharing>

Predicted quality for wines test set. This should be in CSV format and should have two columns: ID and quality.

RShiny Dashboard that provides two-way plots of quality versus the other features.

Rmarkdown that describes your analysis.

GitHub Repo with all files that pertain to the deliverables above.

Peer Review

**Groups:**

This is a group project. Communication is critically important to the success of the project. Please reach out to your partner immediately and work out a schedule works for you both. ☺

Here are groups:

1. **Zainab Anwar** (zanwar@mail.smu.edu) & **Blake Armstrong** (armstrongb@mail.smu.edu)
2. **Arman Azhand** (aazhand@mail.smu.edu) & **Eric Graham** (ericgraham@mail.smu.edu)
3. **Ammar Shafi** (ashafi@mail.smu.edu) & **Megan Huy** (mhuy@mail.smu.edu)
4. **Ekam Brar** (ebrar@mail.smu.edu) & **Kyle Evans** (kyleevans@mail.smu.edu)
5. **Gerden Clark** (gerdenc@mail.smu.edu), **Vanessa Nkongolo** (vnkongolo@mail.smu.edu), & **Solomon Mathew** (stmathew@mail.smu.edu)
6. **Isiah Perine** (iperine@mail.smu.edu) & **Katherine Nguyen** (nguyenkh@mail.smu.edu)
7. **Bryan Pruneda** (bpruneda@mail.smu.edu) & **Brittany Blackmon** (brittanyblackmon@mail.smu.edu)
8. **Tommy Mai** (tommymai@mail.smu.edu) & **Miguel Zavala** (mzvalagamero@mail.smu.edu)