# Project Proposal: Predicting Wine Quality

## Data Description

Our project's data source is the Wine Quality dataset from the UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/dataset/186/wine+quality). The dataset includes two subsets: red wine (1,599 observations) and white wine (4,898 observations), totaling 6,497 samples. The data consist of physicochemical measurements collected between 2004 & 2007 from Portuguese "Vinho Verde" wines & released publicly in 2009. Wine samples were obtained from a commercial producer, with laboratory technicians performing standardized chemical analyses (e.g., titration, spectrophotometry) to measure physicochemical attributes, & then certified wine experts assigned sensory quality scores (0-10). The population of interest is commercially produced red and white Vinho Verde wines from northwestern Portugal. The dataset contains 11 continuous physicochemical variables (e.g., acidity levels, chlorides, residual sugar, alcohol content, sulfur dioxide levels, pH, density, sulphates) & one ordinal target variable, wine quality score. The predictor variables (X) are the 11 physicochemical attributes & the target variable (Y) is the expert-assigned wine quality rating. Our primary predictors of interest are alcohol content, acidity measures, sulfur dioxide levels, & interaction effects among chemical properties.

## Hypothesis of Interest

We hypothesize that wines with higher alcohol content tend to receive higher quality scores. We also expect that at least one interaction between physicochemical variables significantly influences wine quality. Lastly, we hypothesize that nonlinear models will outperform linear models in prediction accuracy. These hypotheses form testable claims that align with our objectives of variable importance, interaction effects, & model performance differences.

## Model Training and Evaluation Plan

We plan to train linear regression, regularized models (LASSO and Ridge). We will use a train-validation-test split or k-fold cross-validation as well as hyperparameter tuning to ensure accuracy & generalizability. We will evaluate models using MSE & R squared, selecting the best model based on test-set performance to avoid overfitting. We will assess variable importance by examining model coefficients for linear models, permutation importance for regularized models, & feature importance & SHAP values for tree-based models. We will interpret model estimates by examining coefficient magnitudes & direction & analyzing feature importance for non-linear models. We will include XGBoost, a gradient boosting method not explicitly covered in class.

## Evaluating Hypotheses

To evaluate the alcohol-quality hypothesis, we will examine the sign and significance of the alcohol coefficient in linear & regularized models & review alcohol's feature importance ranking and SHAP values in nonlinear models.

To assess interaction effects, we will inspect potential interaction terms in linear models and leverage XGBoost to detect nonlinear interactions, using SHAP values and feature importance charts to identify meaningful combined effects between predictors.

If nonlinear models yield significantly lower MSE and higher R squared than linear models, this will support our hypothesis that nonlinear relationships improve wine quality prediction.

https://www.sciencedirect.com/science/article/pii/S0167923609001377?via%3Dihub