Script

A: Predicting wine quality is a significant topic in the wine industry, given that traditional evaluation processes such as sensory tests based on human perception, can be subjective and time-consuming,

E: Leveraging data-driven approaches, our group aims to identify the most influential features that determine quality of wine and to develop a classification model for accurate quality prediction.

Y: Our data set is sourced from Kaggle and contains attributes such as alcohol content, chlorides, residual sugar, and other chemical properties.

A: By utilizing machine learning and data analysis, we aim to gain insights that can enhance quality prediction, helping producers maintain high standards through objective metrics. Understanding the key factors that contribute to wine quality could optimize production costs, inform strategic marketing decisions, and introduce consistency. We believe our work holds potential for broader applications in other areas of product quality assessment.

(Our problem definition for this dataset is, what are the features that are most important in predicting wine quality rating and can we develop a wine classification model based off these properties?)

A: In our analysis of the wine quality dataset, we aimed to understand the relationships between various chemical properties and the quality ratings of wine.

A: During data preparation and cleaning, we encountered values with excessive decimal places such as 0.7200000000000001 (7.2 * 10^-15) for the chlorides feature. These values were cleaned by rounding to three decimal places to ensure consistency. We verified the data types, which were all doubles or integers, to ensure the appropriate formatting throughout the dataset.

Y: Using a correlation heatmap, we observed most features showed weak correlations with wine quality. This suggests that no single attribute is a dominant predictor of quality on its own, and that the nature of wine quality evaluation is multifactorial. The strongest observed correlation to wine quality is alcohol content, with a value of 0.48; and the second strongest correlation to wine quality is volatile acidity with a value of –0.41. This suggests that higher alcohol content tends to be associated with better quality ratings; the lower volatile acidity level is, the higher wine quality tends to be. Though, we cannot say with certainty that these are the sole predicting factors.

E: Through visualizations such as histograms, box plots, and violin plots, we noticed that many features displayed significant skewness and contained many outliers. We decided not to remove them because these outliers might hold valuable information, and their removal could oversimplify the data and potentially lead to the loss of critical insights.

A: To address skewness and the presence of outliers, we applied a new scaling technique log transformations to specific features such as chlorides and residual sugar. We also experimented with robust scaling on the chlorides feature to minimize the influence of extreme values. However, all of these methods did not yield significant improvement in our analysis. This outcome suggests that handling outliers and skewed distributions remains a challenge in our dataset. Future work could involve experimenting with alternative approaches for outlier detection and transformation, or incorporating models specifically designed to handle outliers more effectively.

E: To explore predictive modeling, we initially applied a linear regression approach. However, this method proved inadequate because our target variable, quality, is ordinal and ranges from 3-8 as whole numbers. The relationship between predictors and the target is non-linear, exhibiting significant horizontal clustering and leading to a poor fit. The lack of linearity indicated that a linear regression model was ill-suited for capturing the complexities of the data. We then shifted our focus to decision trees and random forest classifiers, which are more effective for capturing non-linear relationships and handling complex data interactions.

Y: Our primary objective was to identify the features that most significantly impact wine quality ratings and develop a reliable classification model based on these properties. While linear regression failed to capture these, decision trees provided insights into feature importance by segmenting data based on feature values, revealing which attributes (e.g., alcohol content, acidity levels) were most influential in predicting wine quality. We also used a new technique, random forest classifiers also helped our model's effectiveness by reducing overfitting and improving accuracy through the combination of multiple decision trees.

E: For the decision tree classifier, our accuracy was fairly low, with an average value of 0.5 to 0.6 for both training and testing data. For example, in the categories fixed acidity, volatile acidity, citric acid, and residual sugar, the train accuracy was about 0.55, a test accuracy of about 0.51, and a mean squared error of about 0.764. The results were fairly similar across the other categories as well. To combat the high mean squared error, we decided to use a new technique, random forest classifiers, to hopefully improve our predictive accuracy. Running it in the same categories mentioned before, we improved our accuracy slightly, with a test accuracy of about 0.53 and a mean squared error value of about 0.656.

A: We evaluated the decision tree models' performance using a confusion matrix, which highlighted areas of misclassification and provided a clearer view of model strengths and weaknesses. Building on this, we utilized random forest classifiers, which improved prediction accuracy by combining multiple decision trees, to offer a more reliable predictive framework.

Y: With the data analysis and machine learning tools and techniques we applied, we determined that alcohol content exhibited the strongest correlation with wine quality. Most other features demonstrated minimal to no significant correlation with quality ratings, which may be attributed to various factors such as dataset limitations or complex interactions not captured in our models.

E: We developed multivariate classification models that grouped features based on chemical composition similarities. For example, we created models where density, pH, and alcohol were grouped; another model included fixed acidity, volatile acidity, citric acid, and residual sugar; while chlorides, free sulfur dioxide, total sulfur dioxide, and sulphates formed a separate group. This categorization aimed to capture feature interactions more effectively, especially since individual features within these groups often displayed limited correlation with wine quality.

A: One key insight from our findings is that many individual features appear to have little predictive power in isolation, suggesting that wine quality may be determined by complex interactions not reflected in our dataset. We acknowledge that correlation does not imply causation, but our analysis indicates that factors beyond the chemical attributes we examined could play a significant role in

determining wine quality. This realization points to the need for future models that incorporate sensory data, expert ratings, and potentially other external variables to improve predictive accuracy.

Y: Moreover, our data was highly skewed, likely impacting the performance of our models and the interpretability of our results. This skewness suggests that more effective data preparation methods, such as normalization or outlier management, could potentially lead to improved model accuracy and reliability in future studies.

E: In conclusion, while our models did not fully solve the original problem of predicting wine quality with high accuracy, they provided valuable insights into the limitations of purely chemical-based predictions.

Moving forward, incorporating more diverse data sources and employing more rigorous data preparation techniques will be critical to achieving better predictive performance.