**Proposal: Predicting Wine Quality with Machine Learning and Feature Selection**

### Introduction

This project is a collaborative effort by three graduate students, **Darius, Weilun, and Santosh**, as part of a new team that is partnering with ABC Wine Company to assist in optimizing their selection and production of high-quality wines with machine learning. Our goal is to help the company improve its wine-making process, enhance quality control, and increase competitiveness in the industry through data-driven insights.

California is a well-known winery area. Napa has more than 500 wineries, Sonoma has more than 400 wineries, and Paso Robles has more than 200. Running a winery is very difficult in California. To help wineries and winemakers make wine, we use machine learning methodology to predict the quality of wine. Our models can help wineries to have more customers and become more competitive.

Using machine learning models, we aim to classify wines into different quality categories, allowing the company to make more informed decisions regarding production, ingredient adjustments, and marketing strategies. Rather than merely training a classifier, we will take a deeper approach by exploring feature importance and simplifying the model to determine if a smaller subset of features can still yield strong predictive performance. The data will give the company a cost-effective method to assess wine quality without requiring extensive chemical analysis.

### Business Impact

1. **Improved Wine Quality:** The company can adjust production methods to enhance taste and consistency by identifying key physicochemical properties influencing wine quality.

2. **Cost Reduction:** Reducing reliance on expensive laboratory testing using a data-driven model to predict wine quality efficiently.

3. **Competitive Advantage:** Providing the winery with an analytical approach to optimize quality control, allowing them to produce better wines and stand out in the market.

4. **Better Decision-Making:** Offering actionable insights to improve ingredient selection, fermentation techniques, and production efficiency.

### Problem Statement

The goal of this project is to classify wines into three categories: **Low quality (0-5), Medium quality (6-7), and High quality (8-10)** based on their chemical composition. Using the Wine Quality Dataset from the UCI Machine Learning Repository, we will build machine learning models to predict wine quality while investigating which features are most crucial for making accurate predictions. Our goal is to assist ABC Wine Co. in selecting the highest quality wines efficiently and effectively, reducing reliance on manual testing or costly lab tests. Additionally, these insights can help the company **fine-tune its wine production process** to improve overall quality and market competitiveness.

### Approach

1. **Exploratory Data Analysis (EDA):**

1. **Heatmap:** To visualize the correlation between each feature and wine quality, identifying redundant or highly correlated features.

2. **Scatter Plots:** To analyze the relationship between individual physicochemical f features and wine quality scores.

3. **Histograms:** To examine the distribution of each feature and determine if transformations (e.g., normalization, log scaling) are needed.

4. **Quality Count Plot:** To check if the dataset is imbalanced, ensuring appropriate resampling techniques are applied.

5. **Distribution Plots:** To visualize the probability density of each feature, helping us determine trends and outliers in the data.

2. **Data Preprocessing:**

1. Handle class imbalances through resampling techniques (oversampling/under sampling).

2. Standardize numerical features for algorithms sensitive to feature scaling.

3. **Feature Selection:**

1. Use **correlation analysis** to remove low-correlated features.

2. Train models on **all features** and **a reduced feature subset** to compare performance.

4. **Model Training & Evaluation:**

1. Train various classification models:

1. **Baseline Models:** Logistic Regression, Decision Tree

2. **Advanced Models:** Random Forest, Support Vector Machine (SVM), Convolutional Neural Networks (CNN)

2. Evaluate models using **accuracy score and F1-score** to measure overall performance and balance between precision and recall.

3. Interpret results using **SHAP values (Shapley Additive Explanations)** to understand f feature importance and how different physicochemical properties impact predictions. SHAP visually represents how much each feature contributes to the final prediction, making the model more interpretable for decision-making.

4. Generate **feature importance plots** to help the winery understand which attributes influence wine quality the most.

5. **Comparing Simpler Models:**

1. Train a classifier using **only the top 3-5 most important features**.

2. Compare its performance with models using all features.

3. If the simpler model performs well, this suggests a cost-effective way for the wine company to assess quality using fewer tests.

### Expected Insights

1. Identifying which chemical properties have the highest impact on wine quality.

2. Determining if a reduced feature set can yield comparable performance to models using all features.

3. Providing actionable insights for the wine company to optimize its selection, production, and marketing processes.

4. Gaining hands-on experience in machine learning through model development, feature engineering, and performance evaluation.

### Challenges & Considerations

1. Handling class imbalances since high-quality wines are underrepresented.

2. Addressing multicollinearity between physicochemical features.

3. **Ensuring model interpretability:** Since machine learning models can be complex, we will use SHAP values to explain how the model makes predictions, allowing the winery to trust and act on the results.

4. **Appropriateness of CNN for tabular data:** Traditionally, CNNs are used for image processing; we explore whether they can be adapted for structured data representation to improve feature extraction and classification accuracy.

5. As first-time machine learning students, balancing model complexity with our ability to interpret the results effectively.

This study will develop a predictive model and offer insights into the most critical factors affecting wine quality. We aim to provide a practical, data-driven approach to wine quality assessment by investigating the importance of features and testing simpler models. Additionally, this project will be a foundational learning experience that enhances our understanding of AI-driven data analysis while delivering valuable recommendations to the wine company.

### References

1. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). **Wine Quality Dataset**. *UCI Machine Learning Repository*. Retrieved from https://archive.ics.uci.edu/ml/datasets/wine+quality

2. Lundberg, S. M., & Lee, S. I. (2017). **A unified approach to interpreting model predictions**. *Advances in Neural Information Processing Systems, 30*.

3. Guyon, I., & Elisseeff, A. (2003). **An introduction to variable and feature selection**. *Journal of Machine Learning Research, 3*, 1157-1182.

4. Breiman, L. (2001). **Random forests**. *Machine Learning, 45*(1), 5–32.

5. Tibshirani, R. (1996). **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.