**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 our first experience with machine learning. Coming from diverse academic and professional backgrounds, we are eager to apply theoretical knowledge to a practical dataset while deepening our understanding of machine learning techniques. Various chemical properties influence the quality of wine, and accurately predicting wine quality can help winemakers maintain consistency and improve production processes.

In this project, we will leverage machine learning techniques to classify wines into different quality categories based on their physicochemical attributes. However, 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. This will allow us to understand how machine learning models work and which features are most relevant in assessing wine quality.

### **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. Since this is our first exposure to machine learning, we aim to focus not just on achieving high model accuracy but also on interpretability and feature importance.

### **Approach**

1. **Exploratory Data Analysis (EDA):**
   * **Heatmap:** To visualize the correlation between each feature and wine quality, identifying redundant or highly correlated features.
   * **Scatter Plots:** To analyze the relationship between individual physicochemical features and wine quality scores.
   * **Histograms:** To examine the distribution of each feature and determine if transformations (e.g., normalization, log scaling) are needed.
   * **Quality Count Plot:** To check if the dataset is imbalanced, ensuring appropriate resampling techniques are applied.
   * **Distribution Plots:** To visualize the probability density of each feature, helping us determine trends and outliers in the data.
2. **Data Preprocessing:**
   * Convert the quality score (0-10) into categorical labels (Low, Medium, High).
   * Handle class imbalances through resampling techniques (oversampling/undersampling).
   * Standardize numerical features for algorithms sensitive to feature scaling.
3. **Feature Selection:**
   * Use **correlation analysis** to remove highly correlated features.
   * Apply **Random Forest feature importance** and **L1-regularized logistic regression** to identify the most influential features.
   * Train models on **all features** and **a reduced feature subset** to compare performance.
4. **Model Training & Evaluation:**
   * Train various classification models:
     + **Baseline Models:** Logistic Regression, Decision Tree
     + **Advanced Models:** Random Forest, XGBoost
   * Evaluate models using **accuracy, F1-score, and ROC-AUC**.
   * Interpret results using **SHAP values** and feature importance plots.
5. **Comparing Simpler Models:**
   * Train a classifier using **only the top 3-5 most important features**.
   * Compare its performance with models using all features.
   * If the simpler model performs well, this suggests a cost-effective way to assess wine quality using fewer tests.

### **Expected Insights**

* Identifying which chemical properties have the highest impact on wine quality.
* Determining if a reduced feature set can yield comparable performance to models using all features.
* Providing actionable insights for winemakers to optimize their testing processes.
* Gaining hands-on experience in machine learning through model development, feature engineering, and performance evaluation.

### **Challenges & Considerations**

* Handling class imbalances since high-quality wines may be underrepresented.
* Addressing multicollinearity between physicochemical features.
* Ensuring the interpretability of models to provide meaningful insights for winemakers.
* As first-time machine learning students, balancing model complexity with our ability to interpret the results effectively.

### **References**

* 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
* Lundberg, S. M., & Lee, S. I. (2017). **A unified approach to interpreting model predictions**. *Advances in Neural Information Processing Systems, 30*.
* Guyon, I., & Elisseeff, A. (2003). **An introduction to variable and feature selection**. *Journal of Machine Learning Research, 3*, 1157-1182.
* Breiman, L. (2001). **Random forests**. *Machine Learning, 45*(1), 5–32.
* Tibshirani, R. (1996). **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.

This study will develop a predictive model and offer meaningful 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, as we navigate our first machine learning course, this project will serve as a foundational learning experience that enhances our understanding of AI-driven data analysis.