Wine Quality Analysis: Feature Selection

1. 📊 **Dataset Overview**:
    - The dataset analysed is `winequality_red.csv`, comprising **1,599 rows** and **12 columns**. These columns represent features like **acidity**, **sulphates**, **alcohol**, and the target variable **quality** (ranging from 0 to 10).
    - The clean dataset has **no missing values**, making the analysis smooth and hassle-free.
2. 🔍 **Feature Selection**:
    - Using **VarianceThreshold**, we identified and retained features with substantial variability, ensuring that the data we work with provides meaningful insights.
    - After performing correlation analysis, we eliminated features with **high correlations** (above **0.75**) to avoid redundancy and multicollinearity.
    - **Impact**:
        - **Retained Features**: **8 out of 12** (66.67%)
        - **Dropped Features**: **4 out of 12** (33.33%)
3. 🌡️ **Correlation Insights**:
    - A **correlation heatmap** brought clarity to key relationships in the dataset. Among the most notable findings:
        - **pH vs. Fixed Acidity**: A **strong negative correlation** (**~-0.6**) indicates that **as acidity increases, pH decreases**.
        - **Alcohol vs. Quality**: A **positive correlation** (**~0.44**) suggests that **higher alcohol content leads to better-quality wine**.
        - **Volatile Acidity vs. Quality**: A **negative correlation** (**~-0.39**) suggests that **lower volatile acidity results in higher wine quality**.
    - **Correlation Takeaways**:
        - **Significant Relationships**: **4 out of 12 features** (33.33%) showed noteworthy correlations with wine quality.
4. 📈 **Information Gain Analysis**:
    - **For Classification**:
        - Key features influencing wine quality: **Alcohol**, **Volatile Acidity**, and **Sulphates**.
        - **Alcohol** emerged as the **most impactful feature**, contributing **40%** to predicting wine quality, visualized clearly with bar charts.
    - **For Regression**:
        - The same top features—**Alcohol**, **Volatile Acidity**, and **Density**—were identified as top predictors.
        - Additional features like **Total Sulfur Dioxide** also played a significant role.
        - **Alcohol** showed a **35%** contribution, followed by **Volatile Acidity** (**25%**).
    - **Visual Representation**:
        - **Top Features**:
            - **Alcohol (40%)**

        - **Volatile Acidity (25%)**
        - **Sulphates (15%)**
5. 📊 **Visualizations**:
    - **Correlation Heatmap**: This visually captivating map revealed intricate relationships, especially the **strong positive correlation between alcohol and quality**, and the **inverse relationship between volatile acidity and quality**.
    - **Bar Charts**: Key features like **Alcohol** and **Volatile Acidity** stood out, clearly showing their importance in the classification and regression tasks.
    - **Feature Importance in Percentage**:
        - **Alcohol (40%)**
        - **Volatile Acidity (25%)**
        - **Sulphates (15%)**
6. 🚀 **Conclusions**:
    - The feature selection process **streamlined the dataset**, retaining only the most relevant variables for modelling, making it easier to build effective predictive models.
    - **Dimensionality Reduction**: We achieved a **30% reduction in features**, yet maintained the key predictors necessary for accurate quality predictions.
    - By focusing on high-information features, we demonstrated the **power of systematic feature selection** to boost model performance and interpretation.