**Executive Summary: Wine Quality Prediction Project**

**Project Overview**

This project focuses on predicting wine quality scores using advanced statistical and machine learning tools. By leveraging the **WineQT dataset**, we employed **Exploratory Data Analysis (EDA)**, **Random Forest (RF)**, and **Support Vector Machines (SVM)** to forecast wine quality based on various physicochemical attributes. This analysis aims to provide actionable insights into the factors influencing wine quality and develop a robust predictive model for future wine quality scores over the next two years.

**Objectives**

1. **Understand** the current status of wine quality through data analysis and visualization.
2. **Identify key predictors** influencing wine quality, such as alcohol content, volatile acidity, and sulphates.
3. **Build and compare models** (Random Forest and SVM) to predict wine quality scores accurately.
4. **Simulate and forecast** wine quality for 190 new data points.
5. Present findings through impactful visualizations using **Tableau** and detailed statistical reporting.

**Dataset**

The dataset consists of **1,143 records** with **13 features**:

* **Features**: Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol.
* **Target Variable**: Quality (score ranging from 3 to 8).
* **ID Column**: Row identifier for each record.

**Exploratory Data Analysis (EDA)**

1. **Wine Quality Distribution**:
   * Most wines have a quality score of **5 or 6**.
   * Higher quality wines (7–8) are less frequent but associated with specific trends in features.
2. **Feature Insights**:
   * **Alcohol Content**: Higher alcohol levels correlate strongly with higher quality scores.
   * **Volatile Acidity**: Lower values of volatile acidity are associated with better-quality wines.
   * **Sulphates**: Positive correlation observed with wine quality, but outliers exist.
3. **Correlation Analysis**:
   * Significant correlations were identified between **alcohol**, **volatile acidity**, and **sulphates** with the target variable (quality).
   * A heatmap visualization highlighted the importance of these features for predictive modeling.

**Model Development and Comparison**

Two machine learning models were developed and compared:

**1. Random Forest (RF)**

* **Methodology**: 10-fold cross-validation with **1,000 trees**.
* **Accuracy**: High predictive accuracy with robust handling of feature importance.
* **Feature Importance**: Alcohol, volatile acidity, and sulphates were identified as the most critical predictors.

**2. Support Vector Machine (SVM)**

* **Methodology**: Radial kernel with hyperparameters **cost = 10** and **gamma = 0.1**.
* **Accuracy**: Competitive performance but slightly less accurate for extreme quality scores.
* **Performance Metrics (check R file for result)**

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **Kappa** |
|  |  |  |

**Key Insight: Random Forest outperformed SVM in accuracy and consistency, making it the preferred model for wine quality prediction. (check R file result)**

**Future Predictions**

**To simulate future scenarios, 190 new data points were generated using randomized feature values within observed ranges. The Random Forest model was applied to predict the wine quality for these new observations. Predictions were saved and exported for further analysis.**

**Visualizations**

**The project findings were communicated through a series of visualizations:**

1. **Wine Quality Distribution: Bar plots showing the concentration of quality scores.**
2. **Model Performance Comparison: Accuracy and Kappa metrics for Random Forest and SVM.**
3. **Feature Relationships:**
   * **Alcohol vs. Quality (boxplots).**
   * **Volatile Acidity Trend (line graph showing decline with higher quality).**
   * **Sulphates vs. Quality (scatter plots).**
4. **Prediction Comparisons: RF vs SVM predictions against actual quality scores.**

**These visualizations were created using Tableau and R libraries such as ggplot2 and corrplot.**

**Key Insights and Recommendations**

1. **Alcohol Content is the most significant predictor of wine quality, followed by volatile acidity and sulphates.**
2. **Lower levels of volatile acidity are critical for improving wine quality.**
3. **Random Forest provides a robust, reliable method for predicting wine quality and should be prioritized for deployment in production.**
4. **Future wine quality predictions for new data points indicate an upward trend with optimized alcohol content and acidity management.**

**Conclusion**

**This project successfully combined data analysis, predictive modeling, and visual storytelling to understand and forecast wine quality. By identifying key predictors and leveraging the superior performance of the Random Forest model, we achieved actionable results that can guide winemakers to improve wine quality. The insights and predictions presented in this project provide a strong foundation for future studies and real-world applications in the wine industry.**