Project Title

Wine Quality Predictions Analysis

1. Title

- Title: **Wine Quality Predictions Analysis**
- Subtitle:
- Name: Faleye Doyin Opeyemi
- Date: 24-08-2025

2. Dataset Overview

The dataset contains physicochemical attributes of wine samples, along with a quality score (typically ranging from 0 to 10). Key features include:

Feature	Description
Fixed Acidity	Tartaric acid concentration
Volatile Acidity	Acetic acid concentration
Citric Acid	Citric acid content
Residual Sugar	Sugar remaining after fermentation
Chlorides	Salt content
Free Sulfur Dioxide	SO_2 not bound to other compounds
Total Sulfur Dioxide	Total SO_2 content
Density	Mass per unit volume
pH	Acidity level
Sulphates	Sulfate concentration
Alcohol	Alcohol percentage

Feature	Description
Quality	Target variable (wine quality score)

3. Inspect Dataset

- **df.info() – checking for the information of the data**
- **df.describe() – Summary Statistics of the data**

4. Exploratory Data Analysis (EDA)

- **df.isnull()**.sum() – checking for missing values
- **Correlation heatmaps** to identify relationships between features
- **Pairplots** to explore feature interactions
- **Boxplots** to compare feature distributions across quality levels

5. Data Preprocessing

- Dropping unwanted column
- df.shape – checking for how many rows and columns in the dataset
- create another column
- Standardized the column
- Train_Test_Split

4. Models Training and Evaluation Used

🔹 **Random Forest Classifier (RFC)**

- **Type**: Ensemble method using decision trees
- **Strengths**: Handles non-linear relationships, reduces overfitting

- **Accuracy**: **89%**
- **Challenge**: Can be computationally intensive with large datasets

🔹 **Support Vector Classifier (SVC)**

- **Type**: Margin-based classifier
- **Strengths**: Effective in high-dimensional spaces
- **Accuracy**: **86%**
- **Challenge**: Sensitive to parameter tuning and scaling

🔹 **Stochastic Gradient Descent (SGD)**

- **Type**: Linear classifier optimized via gradient descent
- **Strengths**: Fast and scalable for large datasets
- **Accuracy**: **83%**
- **Challenge**: Requires careful feature scaling and tuning

5. Data Visualization & Libraries

🔧 **Libraries Used**

- **Pandas**: Data manipulation
- **NumPy**: Numerical operations
- **Scikit-learn**: Model building and evaluation
- **Matplotlib & Seaborn**: Visualization

📈 **Visualization Techniques**

- **Correlation heatmaps** to identify relationships between features
- **Boxplots** to compare feature distributions across quality levels
- **Pairplots** to explore feature interactions

📊 **Interpretation**

Three machine learning models were employed to classify wine quality:

- **Random Forest Classifier (RFC)**: Achieved the highest accuracy at **88%**
- **Support Vector Classifier (SVC)**: Delivered an accuracy of **86%**
- **Stochastic Gradient Descent (SGD)**: Reached an accuracy of **81%**

Among these, the **Random Forest Classifier** demonstrated superior performance in terms of predictive accuracy and robustness. Given its ensemble nature and ability to handle feature interactions effectively, **RFC is recommended as the primary model for wine quality prediction**.

6. ✅ Recommendation

- **Use Random Forest Classifier** for model deployment due to its superior accuracy and interpretability.
- **Explore advanced ensemble models** like **XGBoost**, **CatBoost**, or **LightGBM** for further performance gains.
- **Feature engineering** and **hyperparameter tuning** can significantly improve model accuracy.
- **Consider dimensionality reduction** (e.g., PCA) if scaling to larger datasets.