Project Report: Predicting Red Wine Quality using Machine Learning

# Project Overview

This project aims to predict the quality of red wine based on various physicochemical properties using machine learning techniques. The dataset used is from the UCI Machine Learning Repository and contains measurements of different chemical properties of wine samples along with their quality ratings.

# Dataset Description

The dataset contains the following columns:  
1. Fixed acidity: Tartaric acid (g/dm³)  
2. Volatile acidity: Acetic acid (g/dm³)  
3. Citric acid: Citric acid (g/dm³)  
4. Residual sugar: Residual sugar (g/dm³)  
5. Chlorides: Sodium chloride (g/dm³)  
6. Free sulfur dioxide: Free sulfur dioxide (mg/dm³)  
7. Total sulfur dioxide: Total sulfur dioxide (mg/dm³)  
8. Density: Density (g/cm³)  
9. pH: pH level  
10. Sulphates: Potassium sulphate (g/dm³)  
11. Alcohol: Alcohol content (% by volume)  
12. Quality: Quality score (0-10)

# Project Structure

The project is structured as follows:  
- data/: Contains the dataset.  
- notebooks/: Jupyter notebooks for data exploration, visualization, and model building.  
- scripts/: Python scripts for data preprocessing, model training, and evaluation.  
- README.md: Project overview and instructions.

# Steps

1. Data Exploration and Visualization: Understand the data distribution and identify any correlations between features.  
2. Data Preprocessing: Handle missing values, normalize features, and convert the quality score into a binary classification (good/bad wine).  
3. Model Building and Evaluation: Train a decision tree classifier and evaluate its performance using metrics like ROC-AUC.  
4. Hyperparameter Tuning: Optimize the model's hyperparameters to improve performance.  
5. Ensemble Methods: Explore ensemble methods like Stacking Classifier for improved performance.

# Installation

To run this project, you need to have Python and the following libraries installed:  
- pandas  
- numpy  
- matplotlib  
- seaborn  
- scikit-learn  
- imbalanced-learn (for SMOTE)  
- jupyter (if using notebooks)  
  
You can install the required libraries using pip:  
pip install pandas numpy matplotlib seaborn scikit-learn imbalanced-learn jupyter

# Usage

Data Exploration and Visualization:  
Run the Jupyter notebook in the notebooks/ directory to explore and visualize the dataset:  
jupyter notebook notebooks/data\_exploration.ipynb  
  
Data Preprocessing and Model Training:  
Run the preprocessing and training scripts in the scripts/ directory:  
python scripts/preprocess\_data.py  
python scripts/train\_model.py  
  
Model Evaluation:  
Evaluate the model's performance using the evaluation script:  
python scripts/evaluate\_model.py

# Results

Best Model (Gradient Boosting Classifier):  
Parameters:  
- Learning Rate: 0.2  
- Max Depth: 7  
- n\_estimators: 200  
  
Performance:  
- Accuracy: 93%  
- Precision for class 1 (good wine): 74%  
- Recall for class 1 (good wine): 72%  
- F1-score for class 1 (good wine): 73%  
- ROC AUC Score: 0.92

# License

This project is licensed under the MIT License. See the LICENSE file for details.

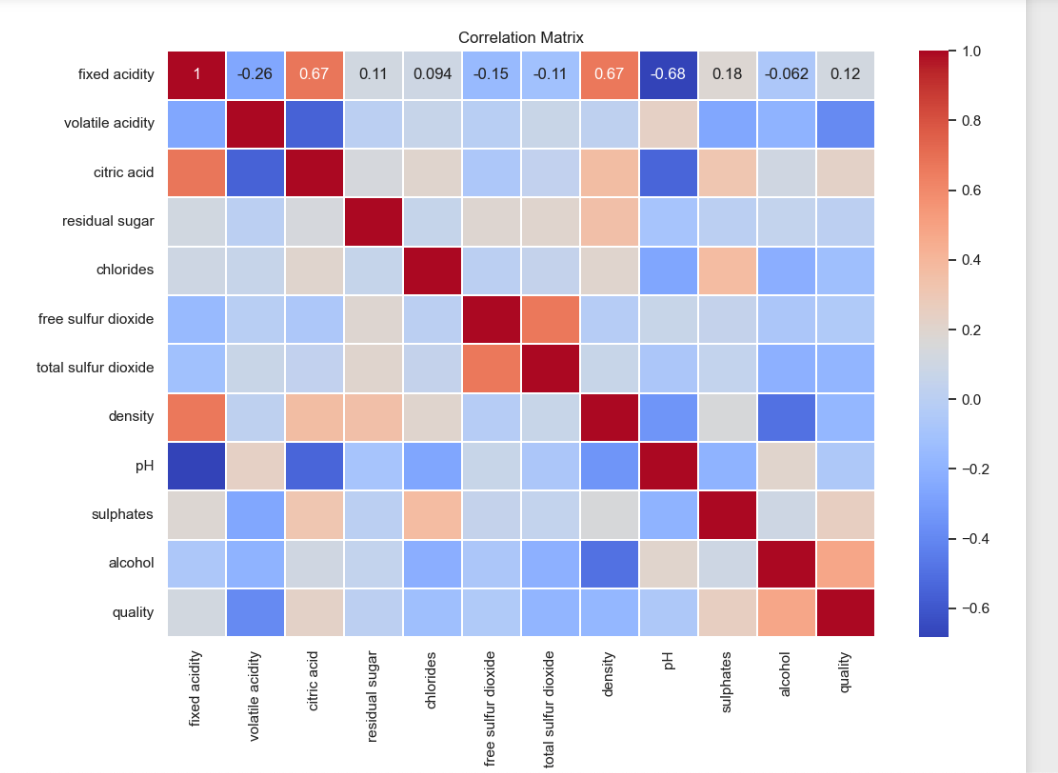
# Acknowledgements

This dataset is also available from the UCI Machine Learning Repository. Please include this citation if you plan to use this database:  
P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

# Contact

If you have any questions or feedback, please contact me at [your\_email@example.com].

# Correlation Matrix



# ROC Curve

