ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# Predicting Red Wine Quality

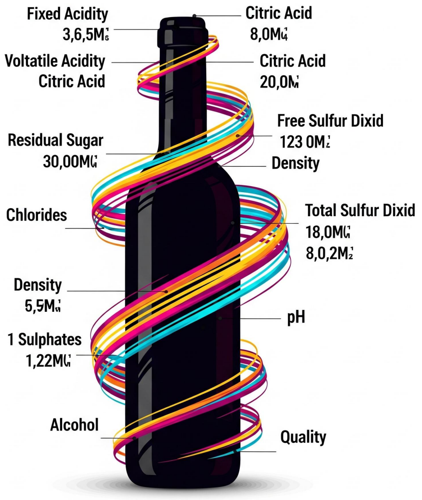
**By : Keith Barnes**

**Course:** Matteo Francia - Machine Learning and Data Mining (Module 2) - A.Y. 2024/25

Machine Learning Course Project

# PROJECT OVERVIEW

**Problem:** Can we develop a machine learning model to classify red wines as "good" or "bad" based on their physicochemical properties?

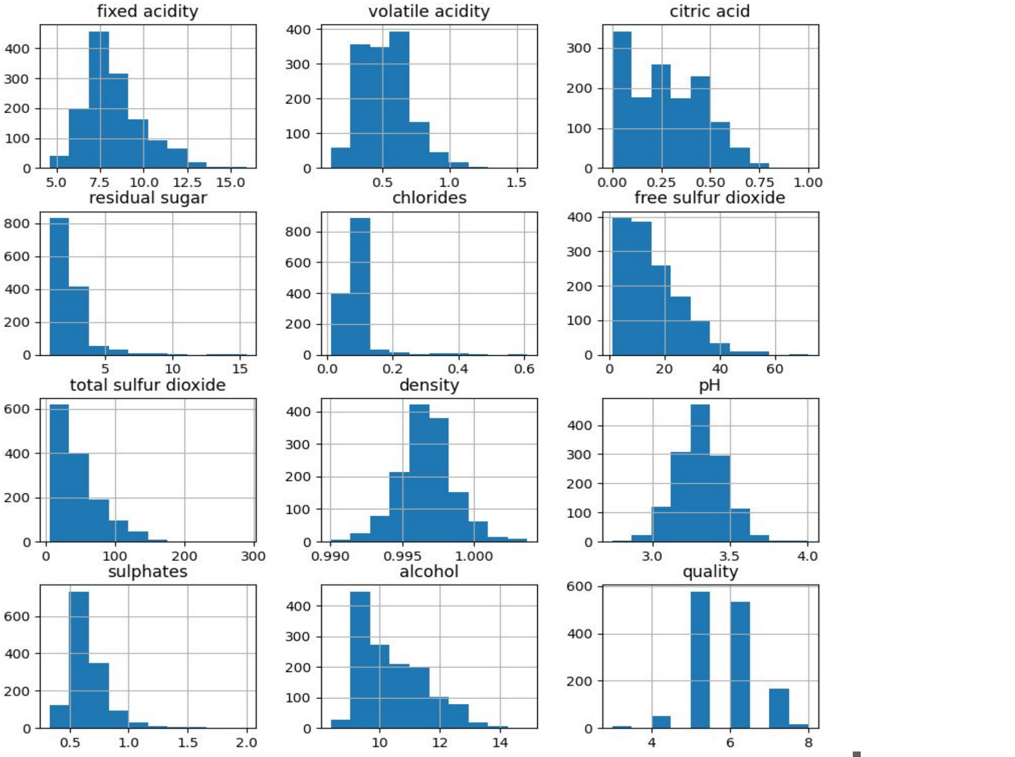


- Dataset: Red Wine Quality (Cortez et al., 2009)
- Features: 11 physicochemical attributes of wine
- Target: Quality score from 1 to 10 - dataset 3  to 8
- Class Imbalance: majority of the wines are in the range of 5 - 6

# Exploratory Data Analysis (EDA)

- **No Null values**
- **240 duplicates removed**

- Histogram of Featurest to check distribution
  - Target Class imbalance (majority 5/6) - SMOTE
  - Skewed distribution



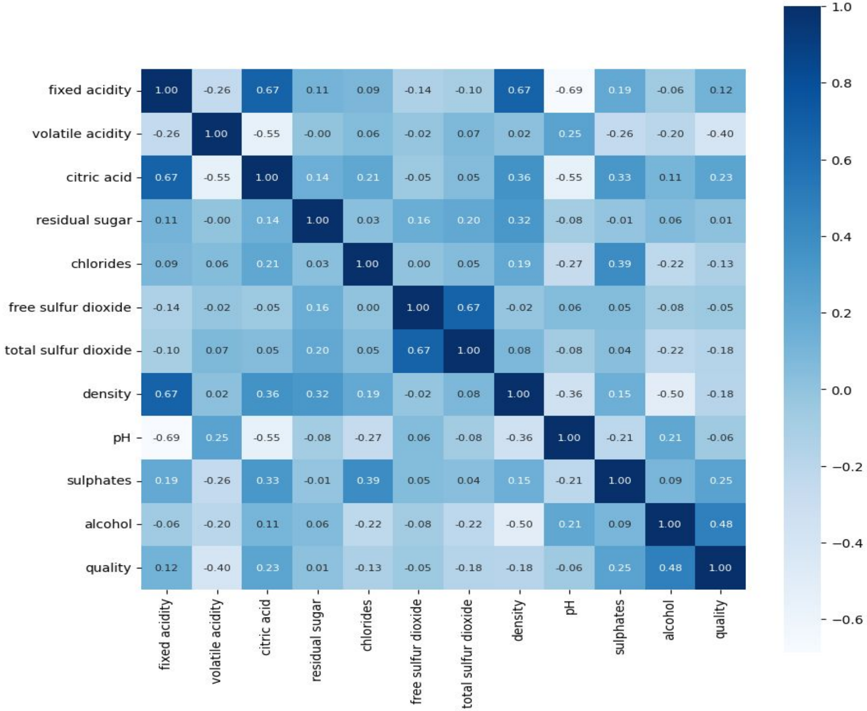
# Pearson Correlation: Heatmap

With target:

- Alcohol -> 0.48
- Volatile Acidity -> 0.4 (-ve)
- Weak - sulphates/ citric acid -0.25/0.23
- No relation - residual sugar, pH

Amongst Feature:

- Fixed acidity & Citric Acid - 0.67
- Free Sulphur Dioxide & Total Sulphur Dioxide - 0.67



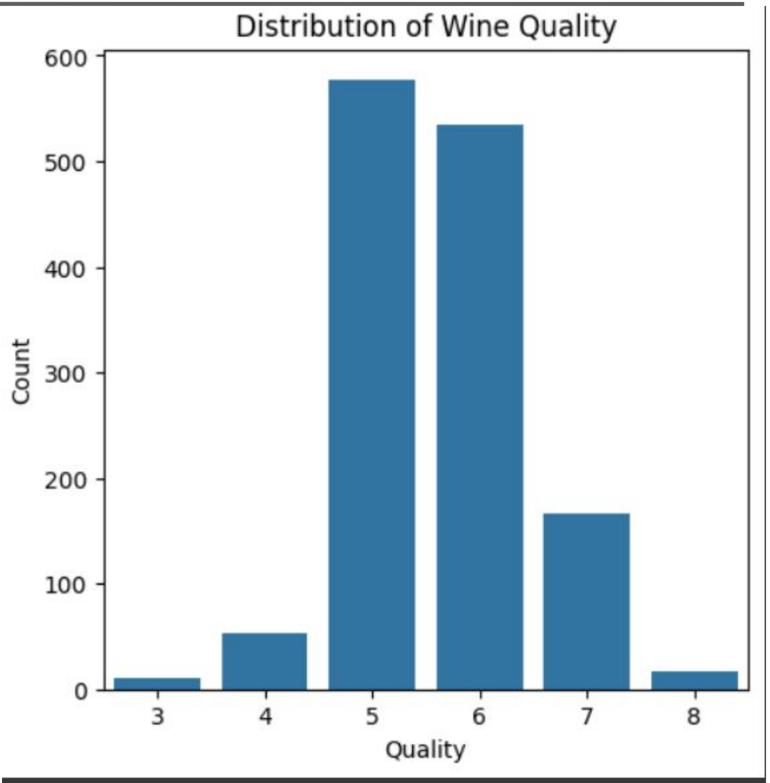
| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| fixed acidity | 1.00 | -0.26 | 0.67 | 0.11 | 0.09 | -0.14 | -0.10 | 0.67 | -0.69 | 0.19 | -0.06 | 0.12 |
| volatile acidity | -0.26 | 1.00 | -0.55 | -0.00 | 0.06 | -0.02 | 0.07 | 0.02 | 0.25 | -0.26 | -0.20 | -0.40 |
| citric acid | 0.67 | -0.55 | 1.00 | 0.14 | 0.21 | -0.05 | 0.05 | 0.36 | -0.55 | 0.33 | 0.11 | 0.23 |
| residual sugar | 0.11 | -0.00 | 0.14 | 1.00 | 0.03 | 0.16 | 0.20 | 0.32 | -0.08 | -0.01 | 0.06 | 0.01 |
| chlorides | 0.09 | 0.06 | 0.21 | 0.03 | 1.00 | 0.00 | 0.05 | 0.19 | -0.27 | 0.39 | -0.22 | -0.13 |
| free sulfur dioxide | -0.14 | -0.02 | -0.05 | 0.16 | 0.00 | 1.00 | 0.67 | -0.02 | 0.06 | 0.05 | -0.08 | -0.05 |
| total sulfur dioxide | -0.10 | 0.07 | 0.05 | 0.20 | 0.05 | 0.67 | 1.00 | 0.08 | -0.08 | 0.04 | -0.22 | -0.18 |
| density | 0.67 | 0.02 | 0.36 | 0.32 | 0.19 | -0.02 | 0.08 | 1.00 | -0.36 | 0.15 | -0.50 | -0.18 |
| pH | -0.69 | 0.25 | -0.55 | -0.08 | -0.27 | 0.06 | -0.08 | -0.36 | 1.00 | -0.21 | 0.21 | -0.06 |
| sulphates | 0.19 | -0.26 | 0.33 | -0.01 | 0.39 | 0.05 | 0.04 | 0.15 | -0.21 | 1.00 | 0.09 | 0.25 |
| alcohol | -0.06 | -0.20 | 0.11 | 0.06 | -0.22 | -0.08 | -0.22 | -0.50 | 0.21 | 0.09 | 1.00 | 0.48 |
| quality | 0.12 | -0.40 | 0.23 | 0.01 | -0.13 | -0.05 | -0.18 | -0.18 | -0.06 | 0.25 | 0.48 | 1.00 |

# Target Binarization

Quality >=7 is “Good” else bad

Imbalanced dataset ~13% good samples -SMOTE

Pros: Meaningful segregation, identifies truly good wines



# Data Preparation

Train-Test split:

- No information leak during training

Log Transformation:

- To address the right skew of features

Standardization:

- Learn the scaling rules from training set
- Transform both the sets using the same

Smote: Synthetic Minority Over-sampling technique

- Generate Synthetic samples for minority Class
- Avoid Bias by models

# Model Training - Tree based Classifiers

**Reasoning for Tree-Based Models:**

- These models are generally robust to multicollinearity and  skewed feature distributions and are effective at capturing non-linear relationships.
- The most popular when it comes to classification problems

**Models Trained:**

1. **Decision Tree:** A baseline model to understand the basic feature splits.
2. **Random Forest :** An ensemble method to reduce overfitting and improve generalization.
3. **AdaBoost :** A boosting algorithm that iteratively focuses on misclassified samples.
4. **Gradient Boosting:** Another powerful boosting method that builds trees sequentially to correct errors.

**Evaluation Metric:**

**F1-Score** was the primary metric due to the class imbalance. It provides a better measure of success than accuracy by balancing precision and recall.

# Modelling phase 2 - Distance Based

**Potential Problems:**

- The performance of Distance based models might be hindered by **high dimensionality** and **correlated features.**
- Note that we handled skew during log transformation.

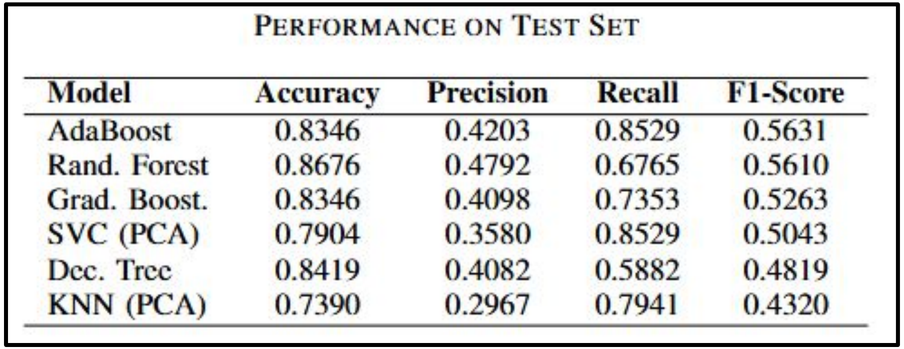
**Principal Component Analysis (PCA):**

- PCA was applied to the standardized training data - **Dimensionality reduction** and to combine correlated features into new, **uncorrelated components**.
- Components were retained to explain **90% of the variance**.

**Models Trained on PCA-transformed data:**

1. **k-Nearest Neighbors (KNN):** A distance-based algorithm that often benefits from a reduced and less noisy feature space.
2. **Support Vector Machine (SVM):** Also sensitive to feature scale and dimensionality, making it a good candidate for PCA transformed data.

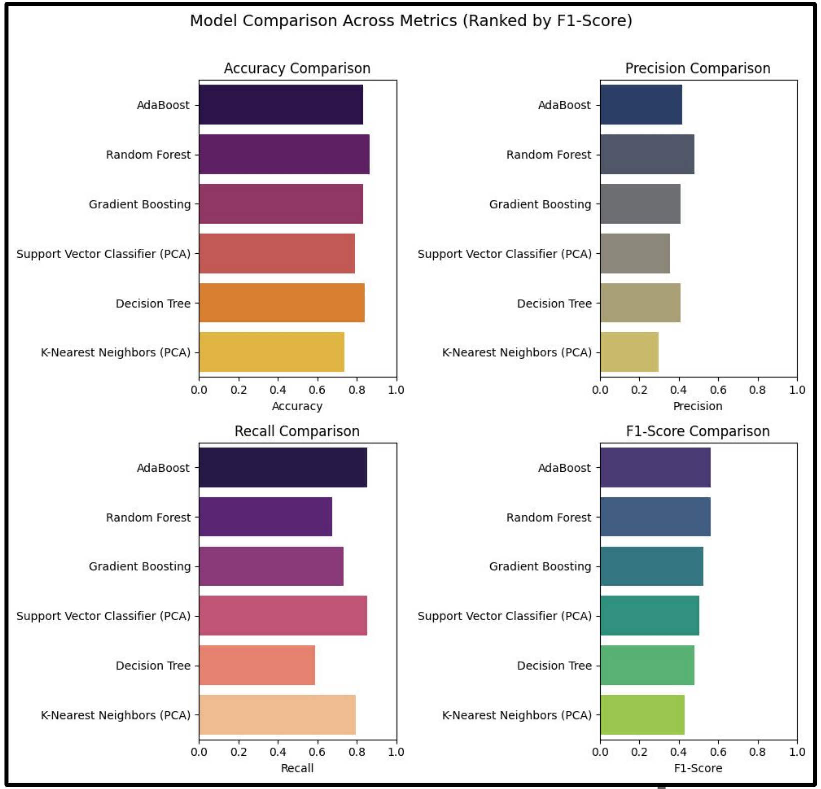
# Combined results



PERFORMANCE ON TEST SET

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| AdaBoost | 0.8346 | 0.4203 | 0.8529 | 0.5631 |
| Rand. Forest | 0.8676 | 0.4792 | 0.6765 | 0.5610 |
| Grad. Boost. | 0.8346 | 0.4098 | 0.7353 | 0.5263 |
| SVC (PCA) | 0.7904 | 0.3580 | 0.8529 | 0.5043 |
| Dec. Tree | 0.8419 | 0.4082 | 0.5882 | 0.4819 |
| KNN (PCA) | 0.7390 | 0.2967 | 0.7941 | 0.4320 |

- **Ranked by F1-Score:** Primary metric for imbalanced data.
- **High Accuracy:** Masks Precision/Recall trade-off for the minority class.
- **Ensemble Methods Dominated:** AdaBoost strongest overall; Random Forest achieved best Precision.
- **Distance-Based Classifiers:** Mixed results and inferior to ensembles.



# Conclusions

- The main learning from the dataset was how to handle class imbalance
- Ensemble models provided the **most balanced performance** for this classification task when evaluated on F1 Scores
- SMOTE Likely enabled **higher Recall**, but potentially contributed to lower Precision by broadening decision boundaries.
- A consistent Precision/Recall tradeoff was observed, suggesting inherent dataset challenges.
- This suggests that while the physicochemical features are informative, they do not provide a perfectly clear boundary to separate "good" from "bad" wines, reflecting the subjective nature of wine tasting.

.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# THANK YOU!