Predicting Wine Quality Using Machine Learning

Noah Marc Paulo Amante
Lehann Enzo Galang
Elijem Timothy Jaso

CSCI 111 - Introduction to AI

Introduction

To predict whether a wine is good or not depending on its chemical properties

- Supervised classification problem
- Comparison of two different models
  - Logistics Regression
  - Random Forest
- Dataset
  - https://archive.ics.uci.edu/dataset/186/wine+quality

Dataset

- 1,599 samples of red wine
- 11 different chemical features
- Target variable : Quality (0-10)

- Shaped to multi-class quality scores into a binary classification
- 0 = Average or Below (quality < 7)
- 1 = Good Wine (quality >= 7)

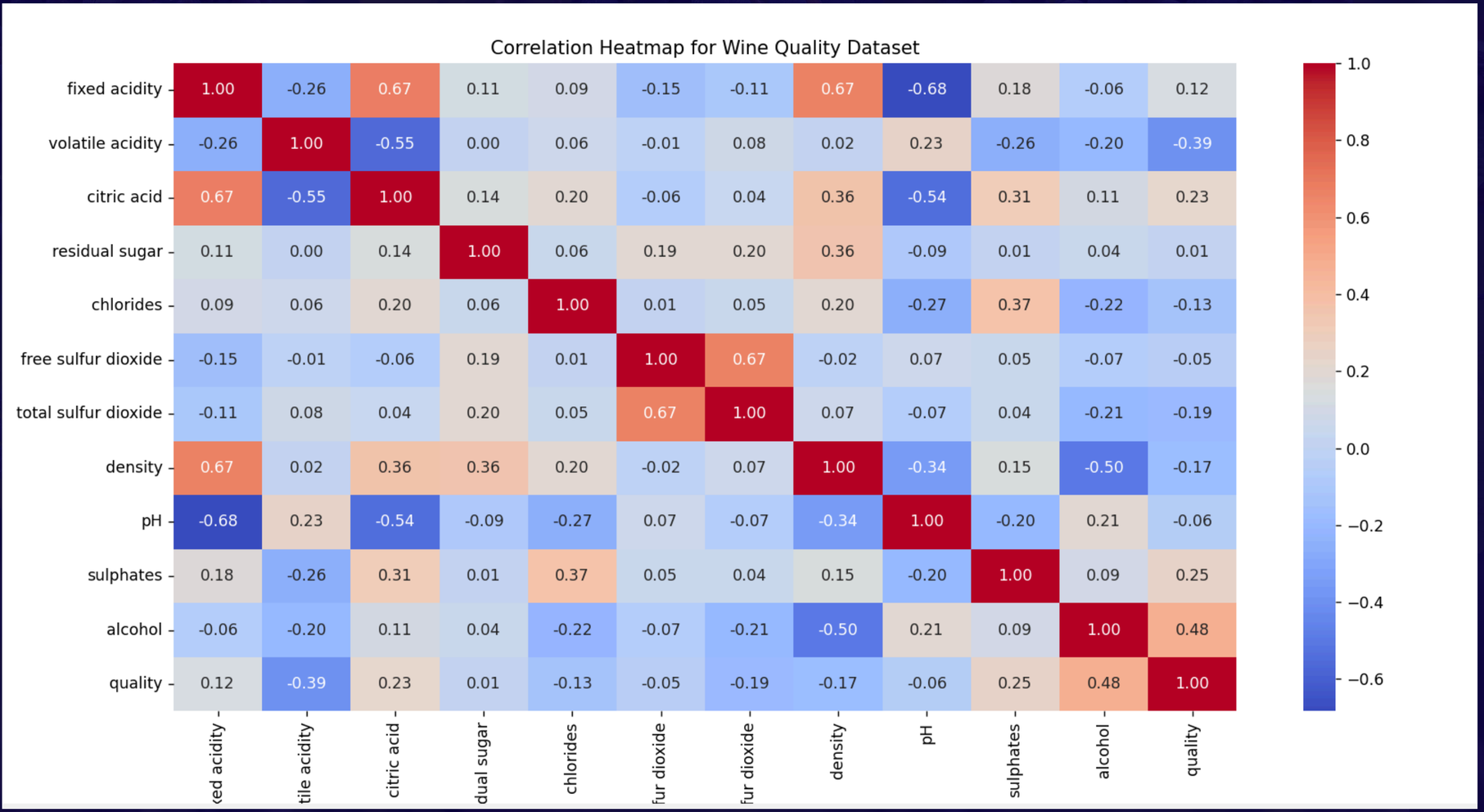
Data Preparation

- Checked for missing values (none)
- Feature scaling with Standard Scaler
- 80/20 train-test split

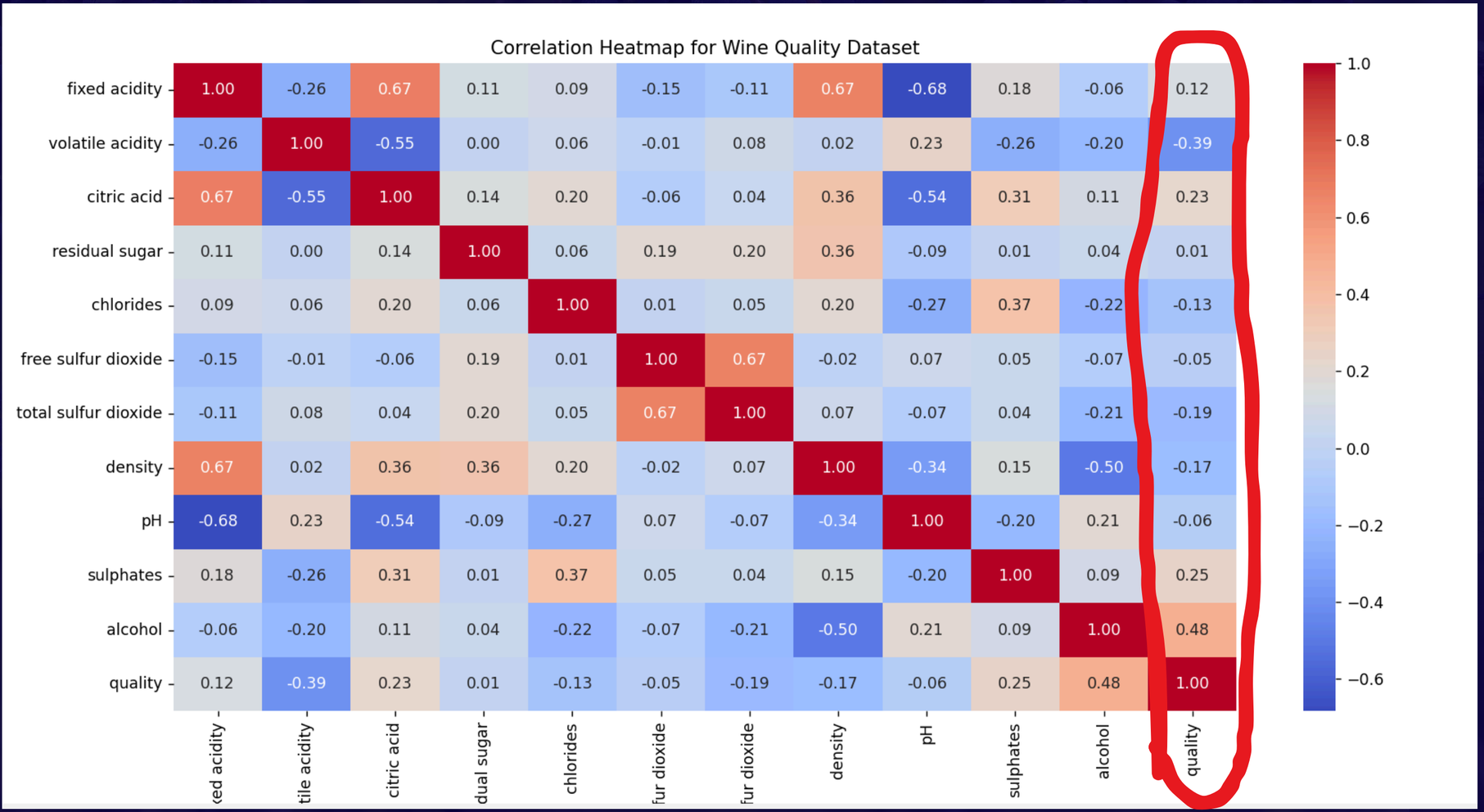
	x	y
Train	(1279, 11)	(1279)
Test	(320, 11)	(320)

Exploratory Data Analysis

Correlation Heatmap



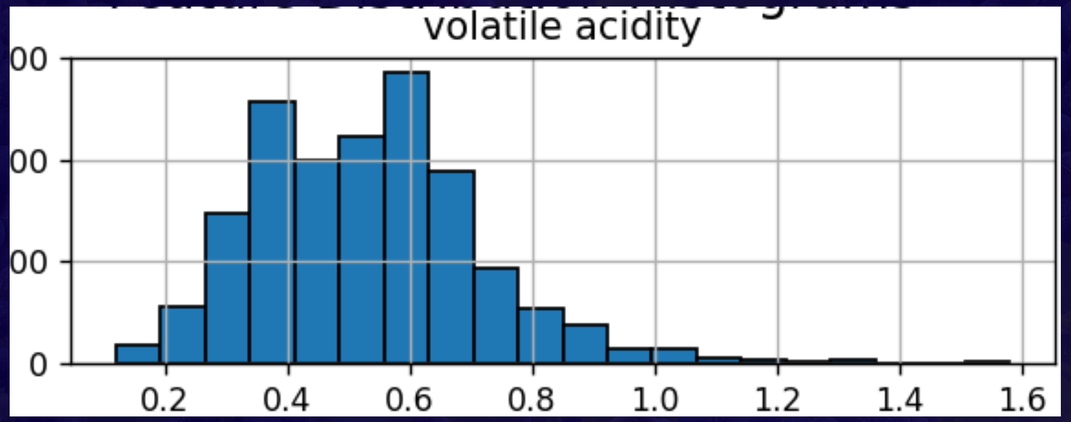
Correlation Heatmap

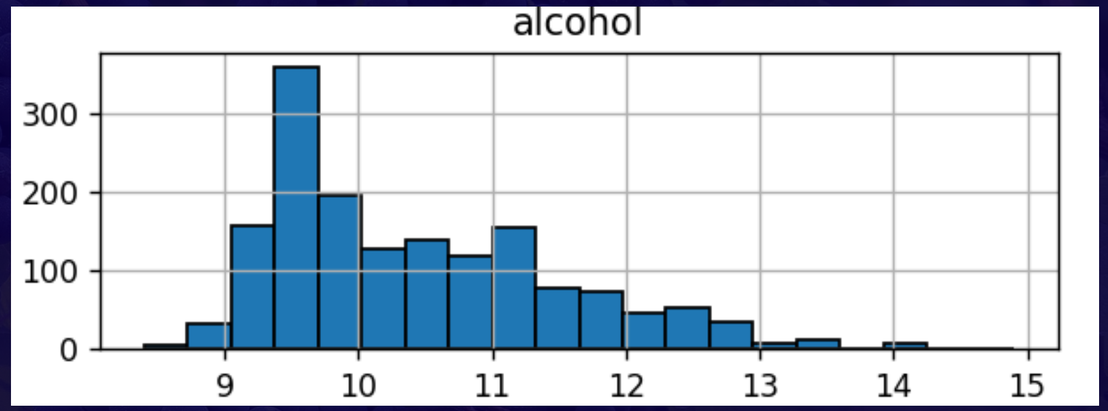


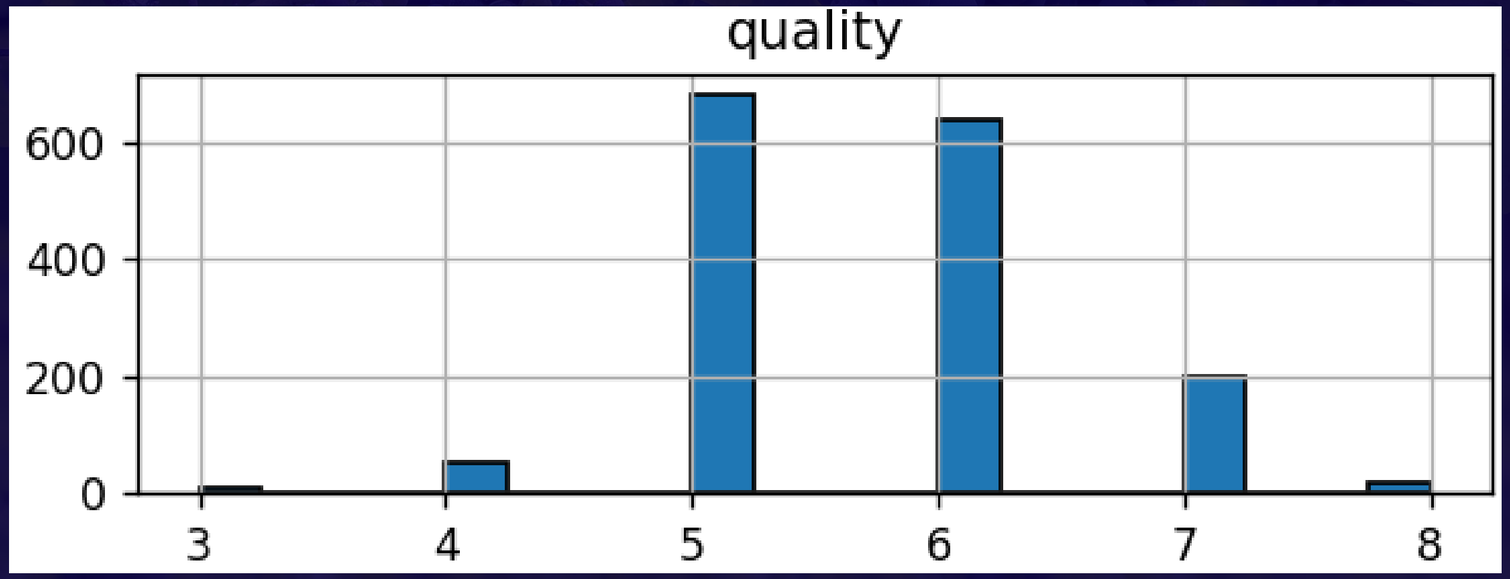
Correlation Heatmap

- Alcohol (0.48): Strongest positive correlation to quality
- Sulphates (0.25) & Citric Acid (0.23): Moderate positive correlation to quality
- Volatile Acidity (-0.38): Strongest negative correlation to quality
- Citric Acid and Sulphates have a moderate to high negative correlation with Volatile acidity

Feature Distribution







Models Used

Model A: Logistic Regression

- Assumes a straight-line relationship between features and the probability of a class
- Works best when the data is linearly separable

Model B: Random Forest

- Nonlinear, tree-based ensemble model
- Makes predictions by combining the output of many decision trees
- Can capture complex interactions between features

Classification Report

Model A: Logistic Regression

	precision	recall	f1-score	support
0	0.89	0.97	0.92	273
1	0.59	0.28	0.38	47
accuracy			0.87	320
macro avg	0.74	0.62	0.65	320
weighted avg	0.84	0.87	0.84	320

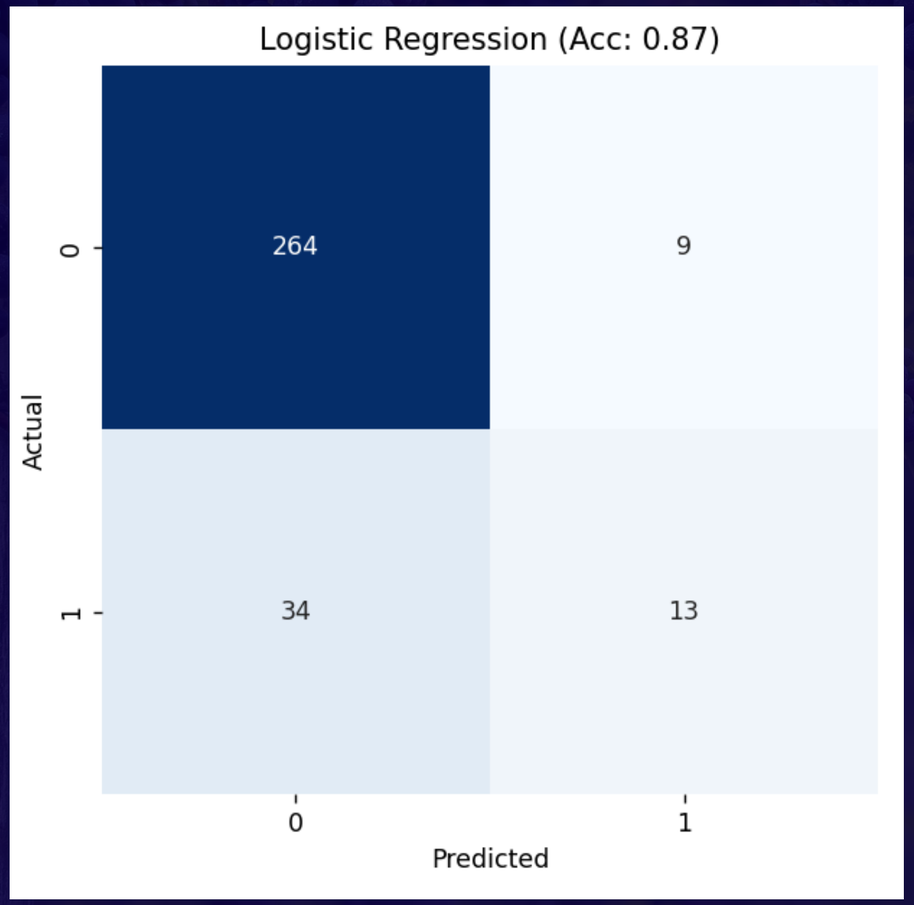
Classification Report

Model B: Random Forest

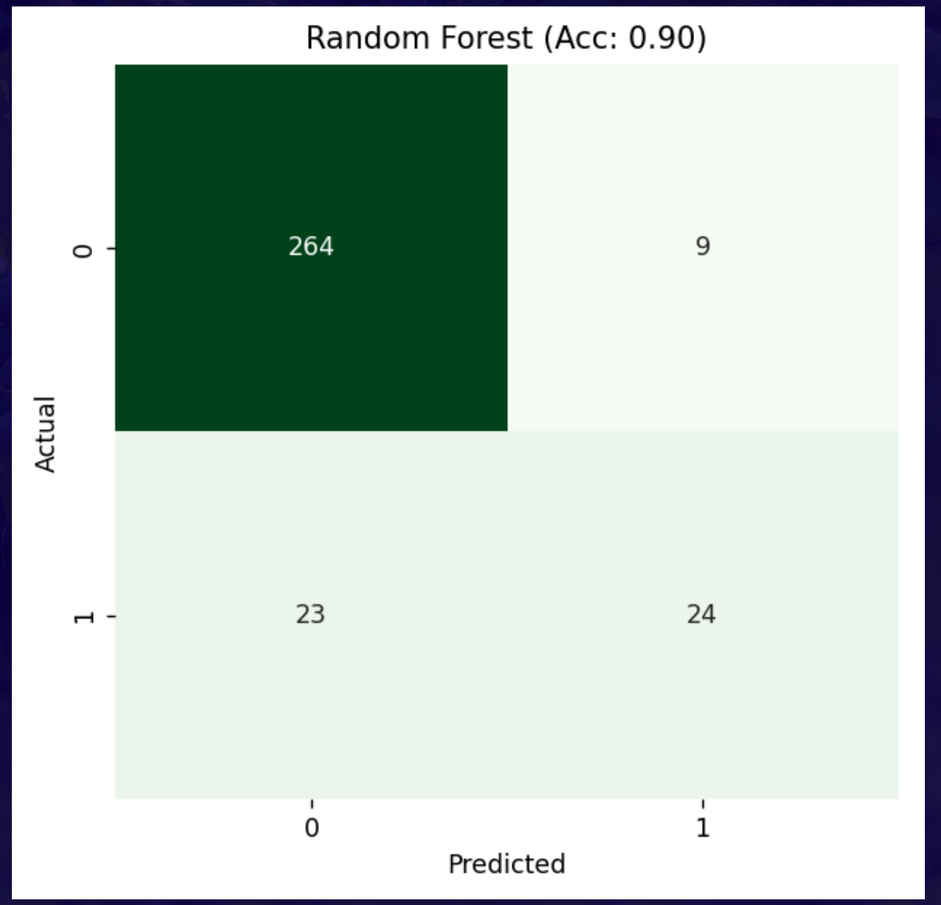
	precision	recall	f1-score	support
0	0.92	0.97	0.94	273
1	0.73	0.51	0.60	47
accuracy			0.90	320
macro avg	0.82	0.74	0.77	320
weighted avg	0.89	0.90	0.89	320

Confusion Matrix

Model A: Logistic Regression



Model B: Random Forest



Results

Model A: Logistic Regression

- 87% overall accuracy
- High recall and f1-score for detecting bad wines
- Poor performance in detecting ‘good’ wines

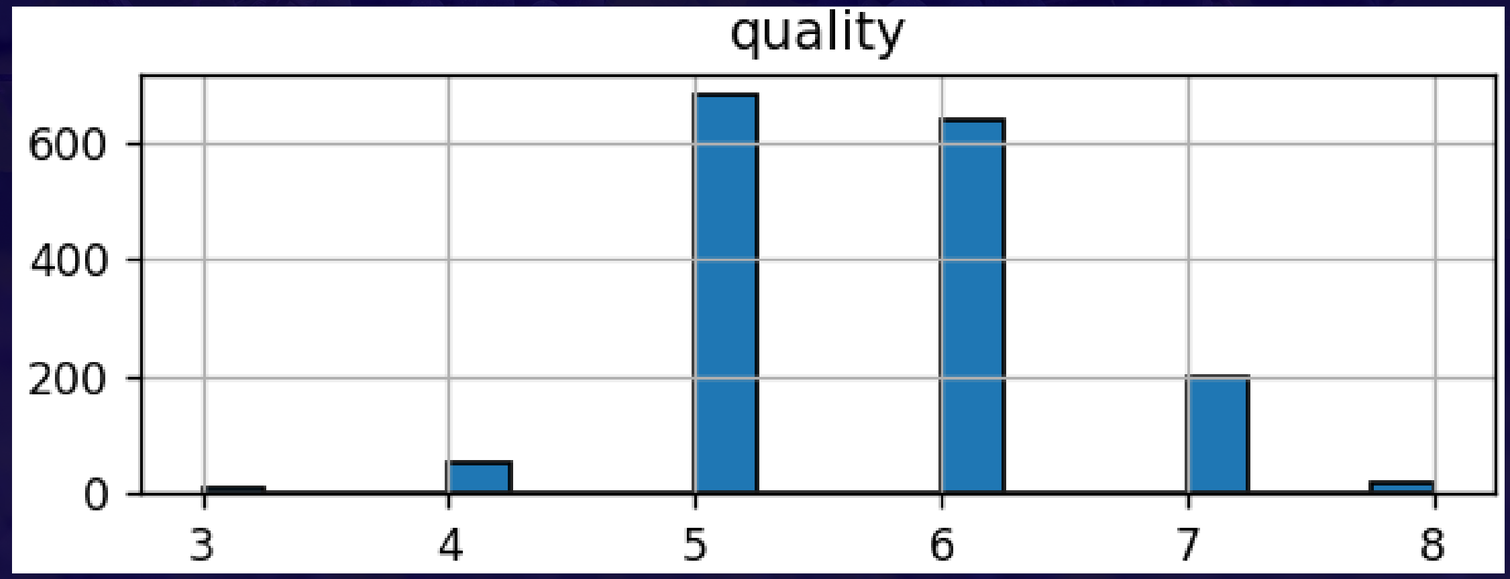
Model B: Random Forest

- Higher accuracy at 90 %
- High precision with negatives
- Much higher precision with positives compared to logistic regression
- Although still relatively low recall and f1-score at .51 and .60, respectively

Conclusions

Overall insights: (will organize tomorrow morning)

- Lower scores in predicting good quality wines are due to the dataset
- Majority of wines are in the 5-6 quality range
- So the model is biased towards predicting bad wines



Conclusions

Overall insights:

- Random Forest is likely the better-performing model
- Chemical features most strongly linked to quality:
- Alcohol ↑
- Volatile acidity ↓
- Sulphates & citric acid ↑
- Binary classification approach was effective

Recommendations

- Using a larger dataset
- Experimenting with other types of wine (white wine)
- Having a more evenly spread out dataset
- Trying multi-class prediction instead

Thank you