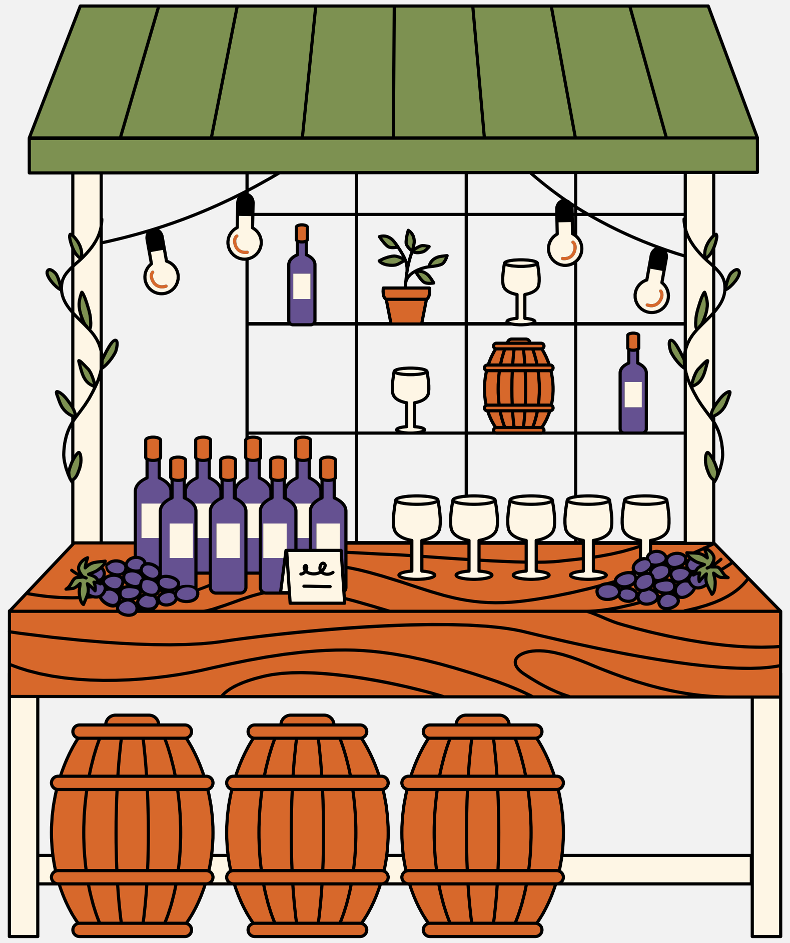
FINAL PROJECT
RED WINE QUALITY PREDICTION

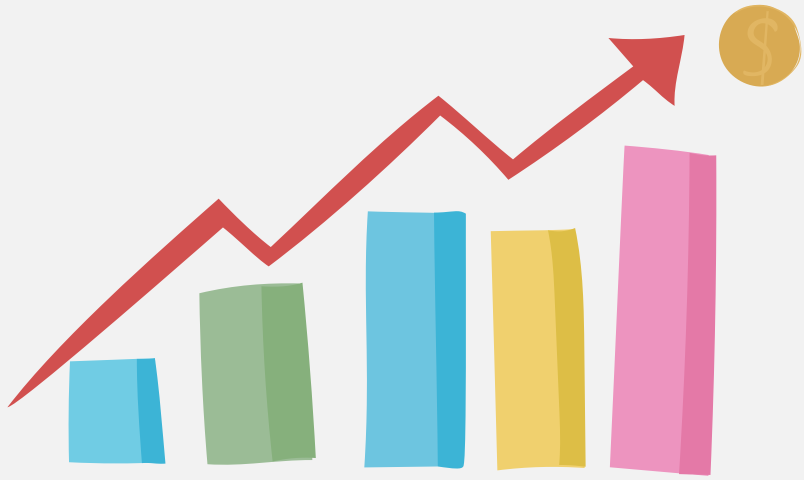


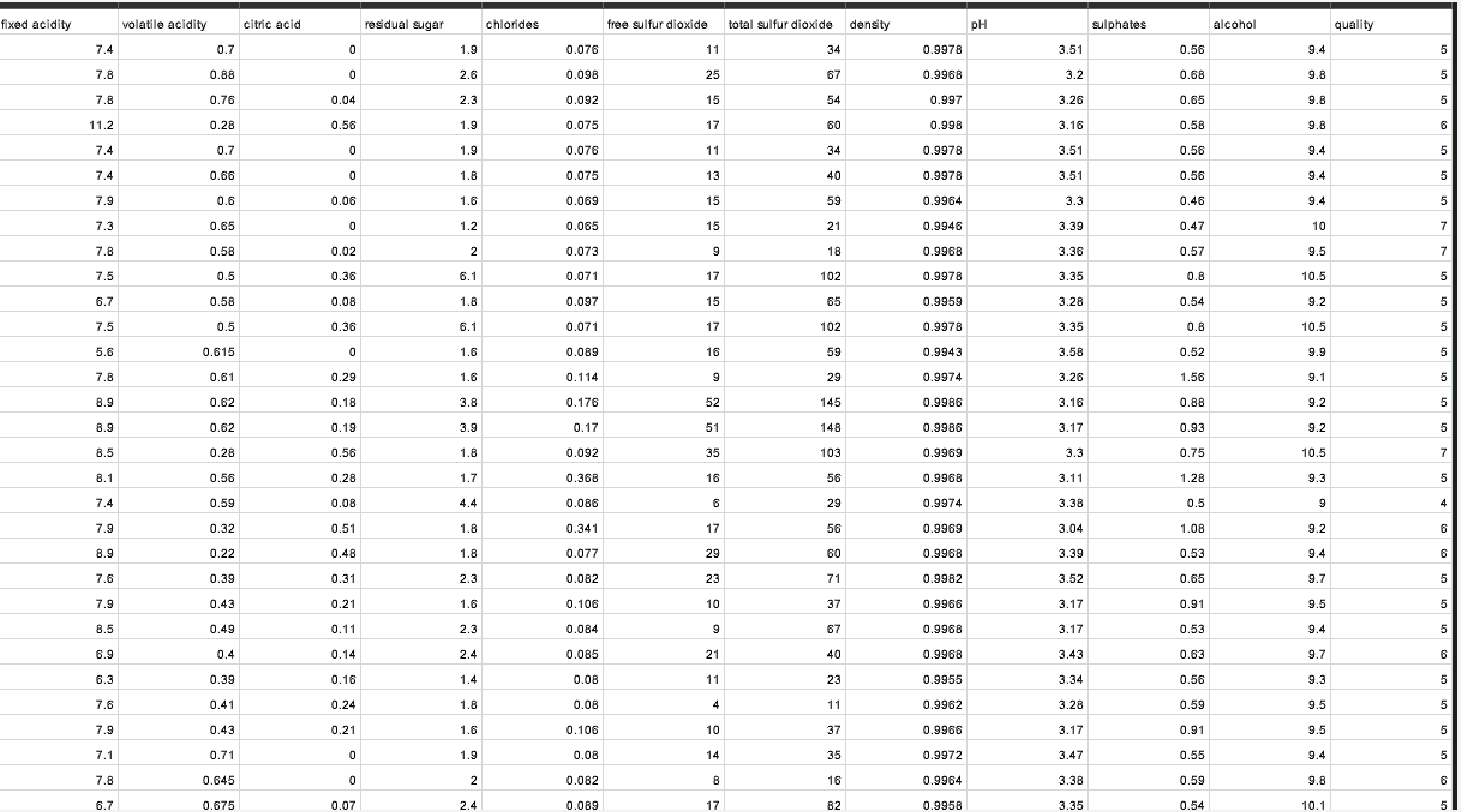
Presented by: Mya Ei Win

INTRODUCTION

- Dataset: Red Wine Quality (UCI Machine Learning Repository)
- Goal: Predict red wine quality score (regression) and high-quality label (classification)
- Used models: Linear, Logistic, MLP, Gradient Boosting

RED WINE DATASET





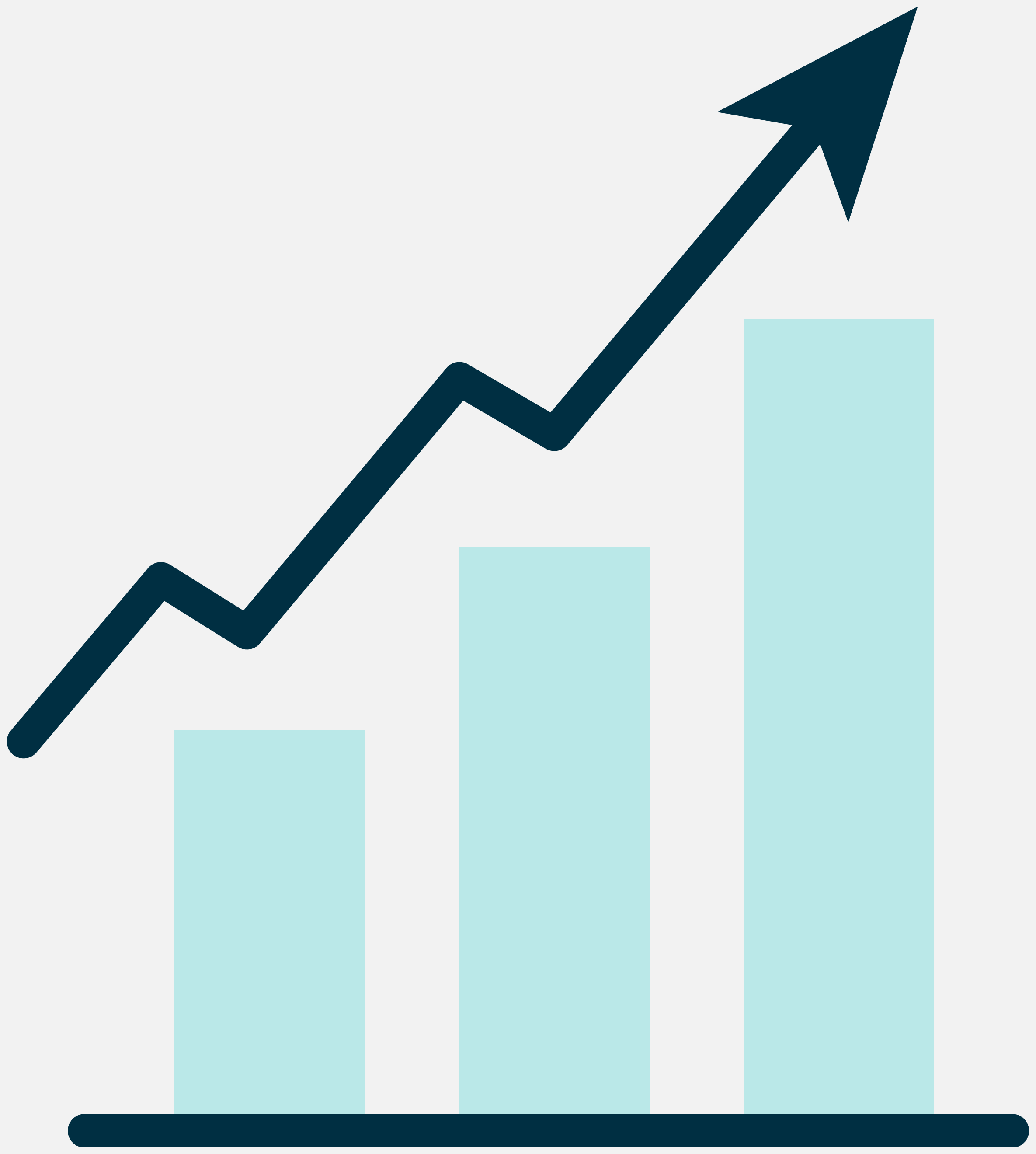
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7
8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5
7.4	0.59	0.08	4.4	0.086	6	29	0.9974	3.38	0.5	9	4
7.9	0.32	0.51	1.8	0.341	17	56	0.9969	3.04	1.08	9.2	6
8.9	0.22	0.48	1.8	0.077	29	60	0.9968	3.39	0.53	9.4	6
7.6	0.39	0.31	2.3	0.082	23	71	0.9982	3.52	0.65	9.7	5
7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5
8.5	0.49	0.11	2.3	0.084	9	67	0.9968	3.17	0.53	9.4	5
6.9	0.4	0.14	2.4	0.085	21	40	0.9968	3.43	0.63	9.7	6
6.3	0.39	0.16	1.4	0.08	11	23	0.9955	3.34	0.56	9.3	5
7.6	0.41	0.24	1.8	0.08	4	11	0.9962	3.28	0.59	9.5	5
7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5
7.1	0.71	0	1.9	0.08	14	35	0.9972	3.47	0.55	9.4	5
7.8	0.645	0	2	0.082	8	16	0.9964	3.38	0.59	9.8	6
6.7	0.675	0.07	2.4	0.089	17	82	0.9958	3.35	0.54	10.1	5

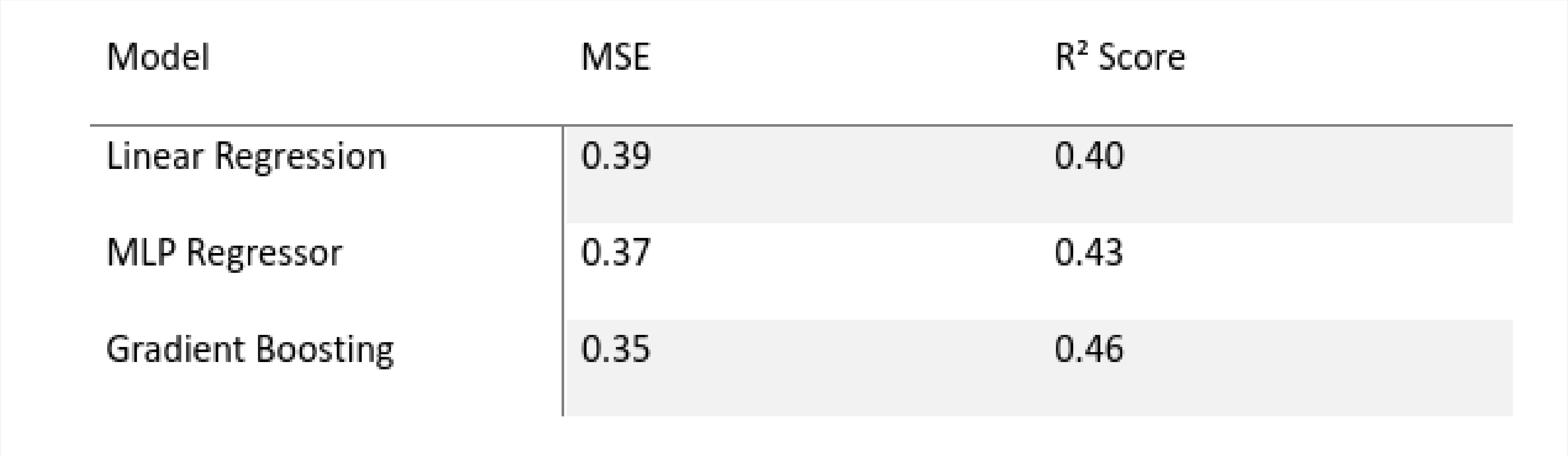
DATA PREPARATION

- Loaded dataset from CSV
- Checked for and removed missing values
- Separated features and target (quality)
- Split into training and test sets (80/20)

MODEL USED

- Linear Regression & Logistic Regression (baseline models)
- MLP Regressor & Classifier (neural network, tuned)
- Gradient Boosting Regressor & Classifier (tree-based, tuned)





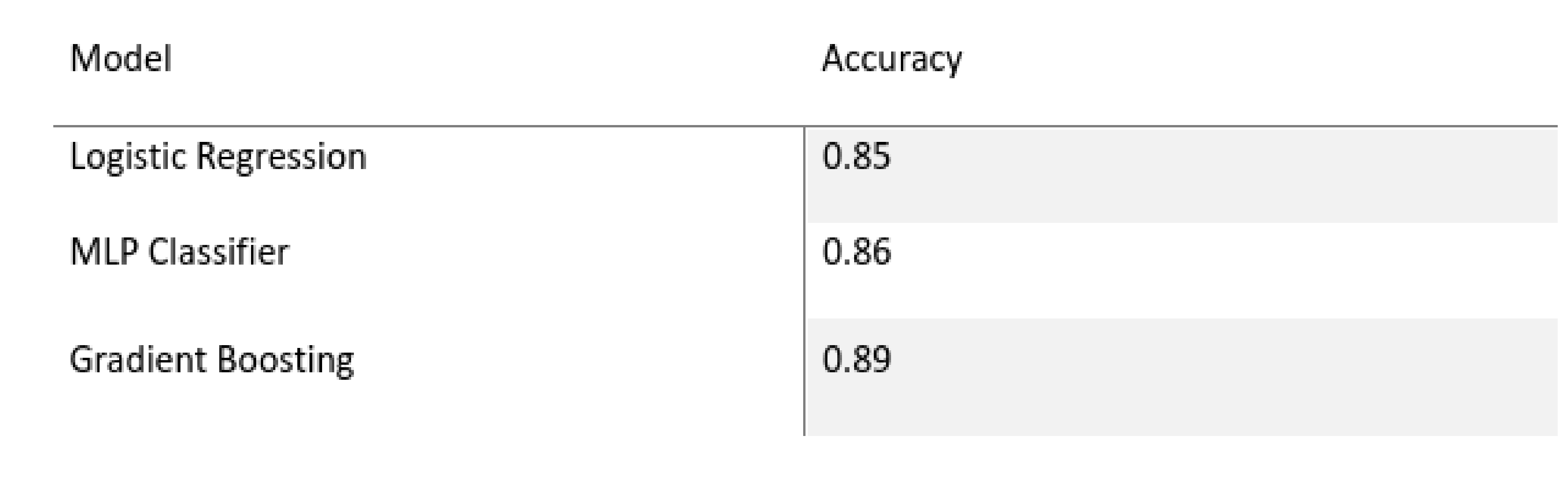
Model	MSE	R² Score
Linear Regression	0.39	0.40
MLP Regressor	0.37	0.43
Gradient Boosting	0.35	0.46

REGRESSION RESULTS

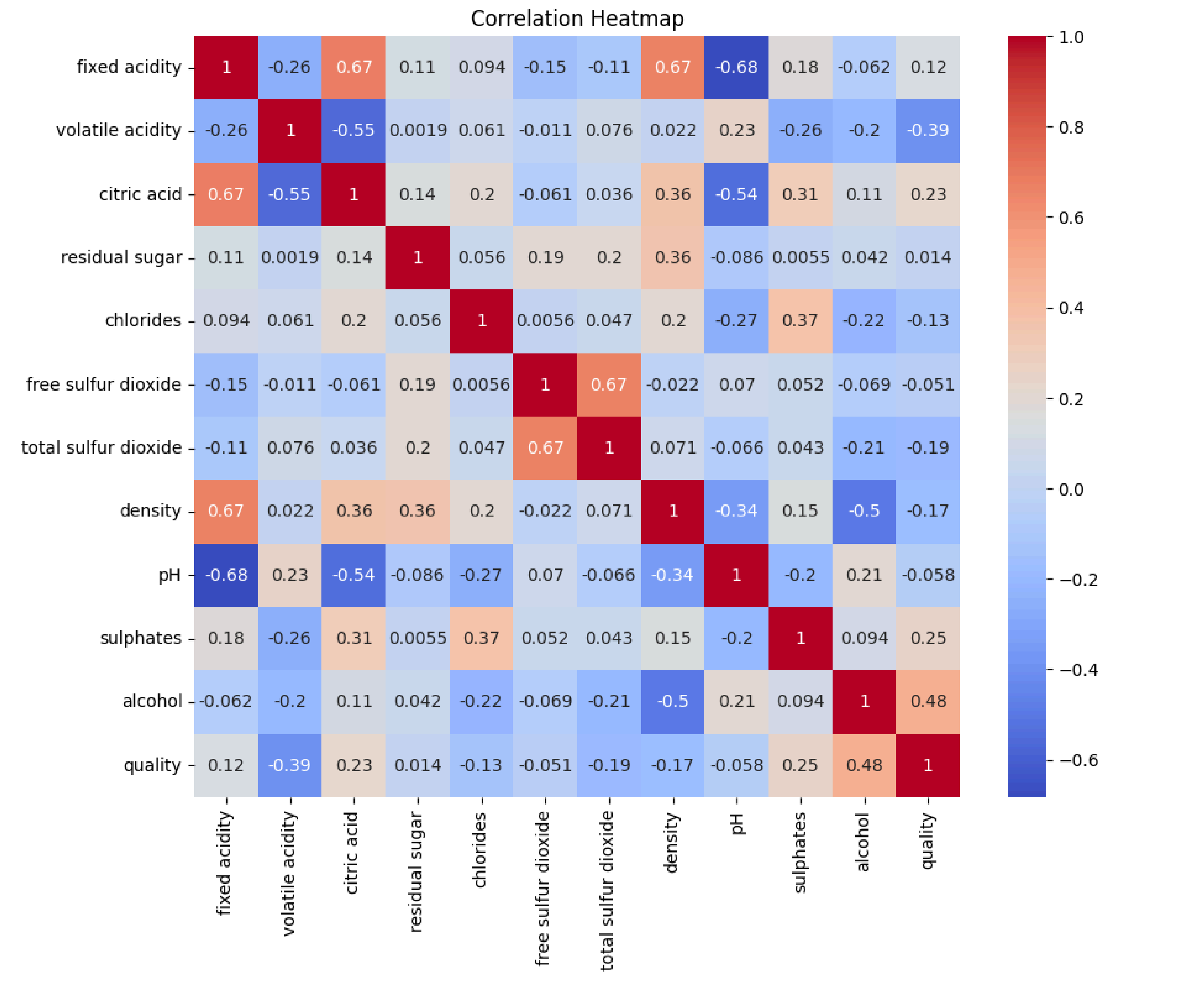
- Gradient Boosting Regressor performed best:
- - Linear Regression: MSE = 0.390, R^2 = 0.403
- - MLP Regressor: MSE = 0.372, R^2 = 0.431
- - Gradient Boosting: MSE = 0.353, R^2 = 0.460
- Gradient Boosting handled non-linear relationships

CLASSIFICATION RESULTS

- Gradient Boosting Classifier achieved highest accuracy:
- - Logistic Regression: Accuracy = 0.85
- - MLP Classifier: Accuracy = 0.86
- - Gradient Boosting: Accuracy = 0.89



Model	Accuracy
Logistic Regression	0.85
MLP Classifier	0.86
Gradient Boosting	0.89



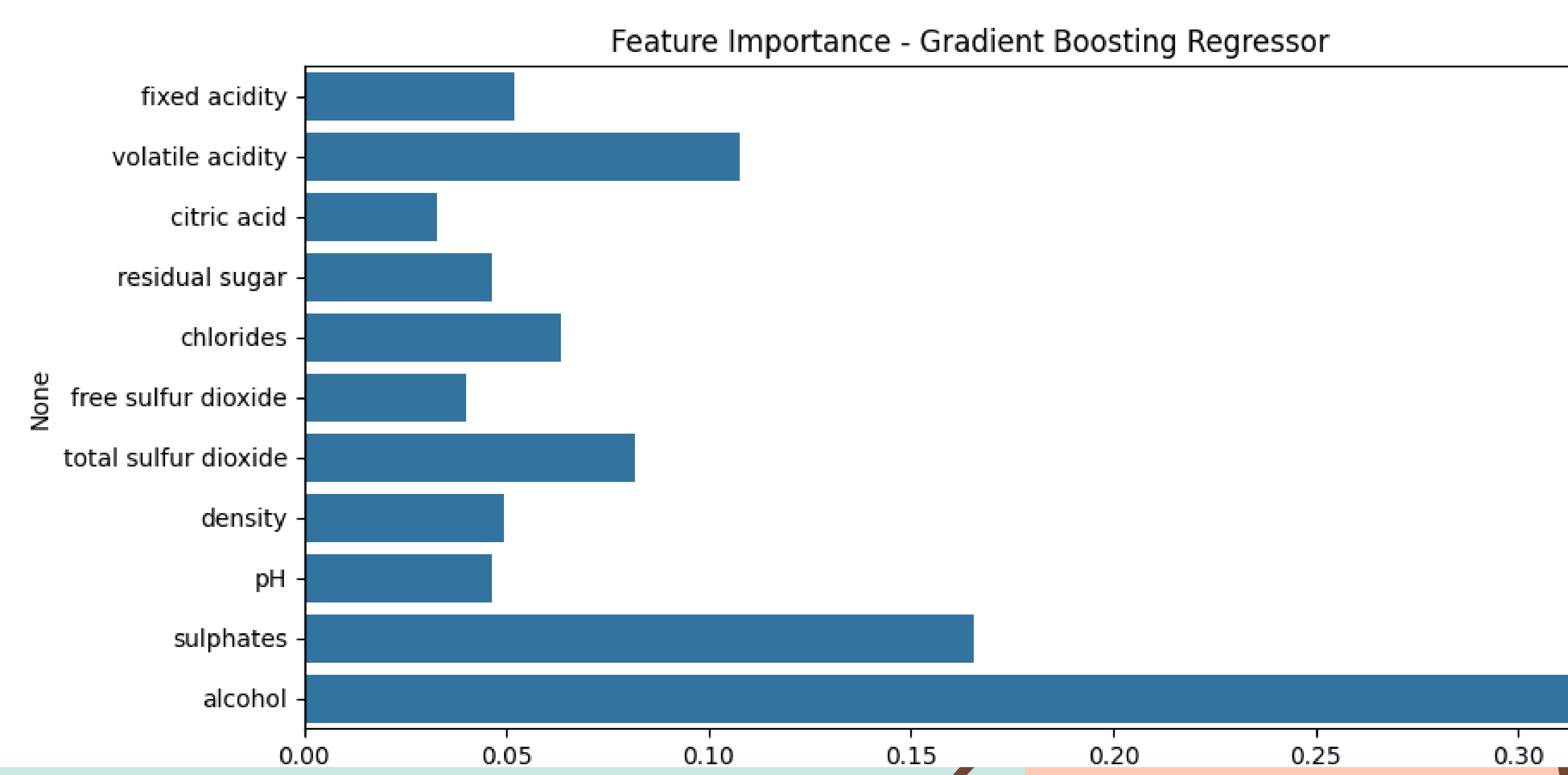
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.26	0.67	0.11	0.094	-0.15	-0.11	0.67	-0.68	0.18	-0.062	0.12
volatile acidity	-0.26	1	-0.55	0.0019	0.061	-0.011	0.076	0.022	0.23	-0.26	-0.2	-0.39
citric acid	0.67	-0.55	1	0.14	0.2	-0.061	0.036	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.11	0.0019	0.14	1	0.056	0.19	0.2	0.36	-0.086	0.0055	0.042	0.014
chlorides	0.094	0.061	0.2	0.056	1	0.0056	0.047	0.2	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	-0.15	-0.011	-0.061	0.19	0.0056	1	0.67	-0.022	0.07	0.052	-0.069	-0.051
total sulfur dioxide	-0.11	0.076	0.036	0.2	0.047	0.67	1	0.071	-0.066	0.043	-0.21	-0.19
density	0.67	0.022	0.36	0.36	0.2	-0.022	0.071	1	-0.34	0.15	-0.5	-0.17
pH	-0.68	0.23	-0.54	-0.086	-0.27	0.07	-0.066	-0.34	1	-0.2	0.21	-0.058
sulphates	0.18	-0.26	0.31	0.0055	0.37	0.052	0.043	0.15	-0.2	1	0.094	0.25
alcohol	-0.062	-0.2	0.11	0.042	-0.22	-0.069	-0.21	-0.5	0.21	0.094	1	0.48
quality	0.12	-0.39	0.23	0.014	-0.13	-0.051	-0.19	-0.17	-0.058	0.25	0.48	1

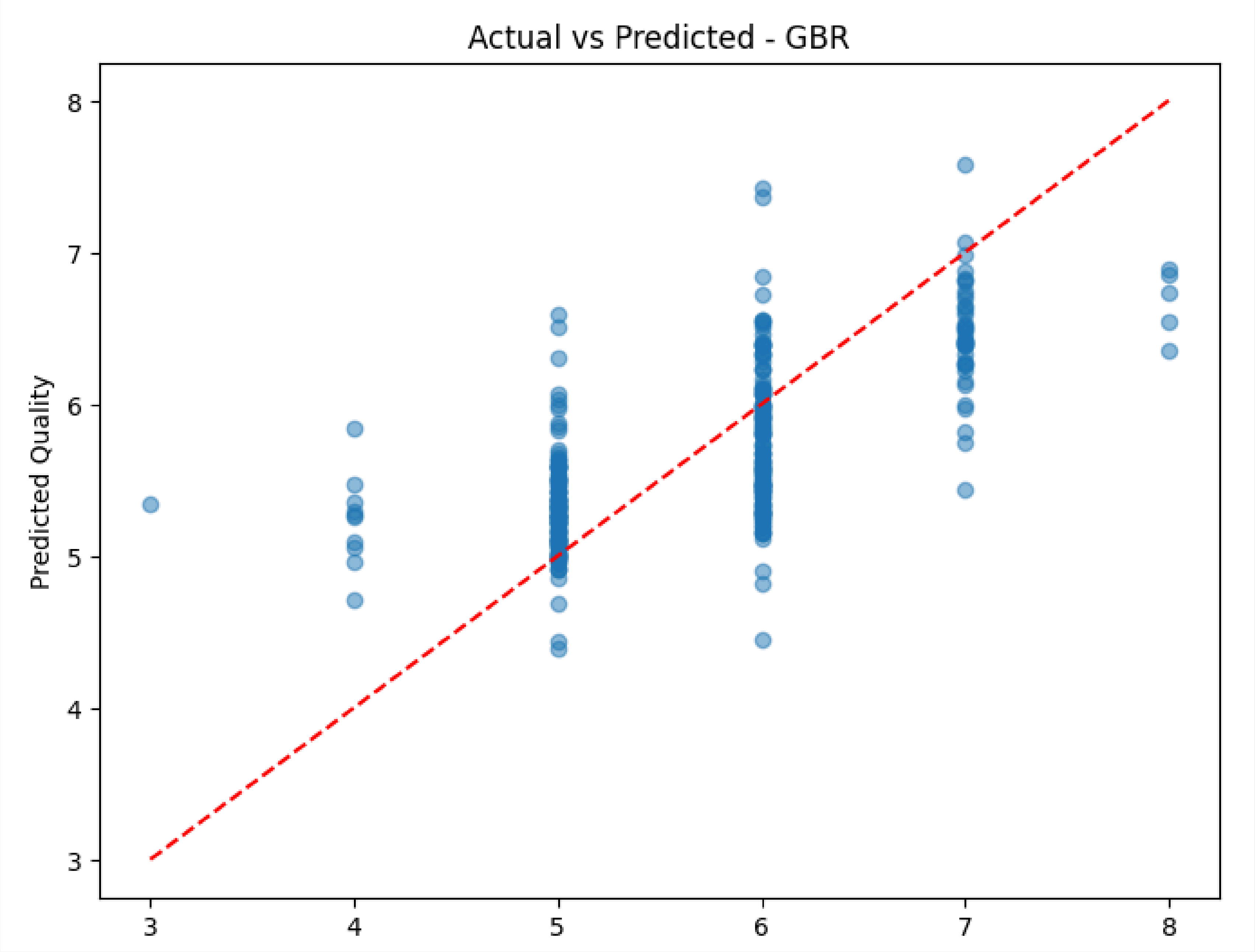
KEY VISUALS USED

- Correlation Heatmap: Alcohol had strongest positive correlation

KEY VISUALS USED

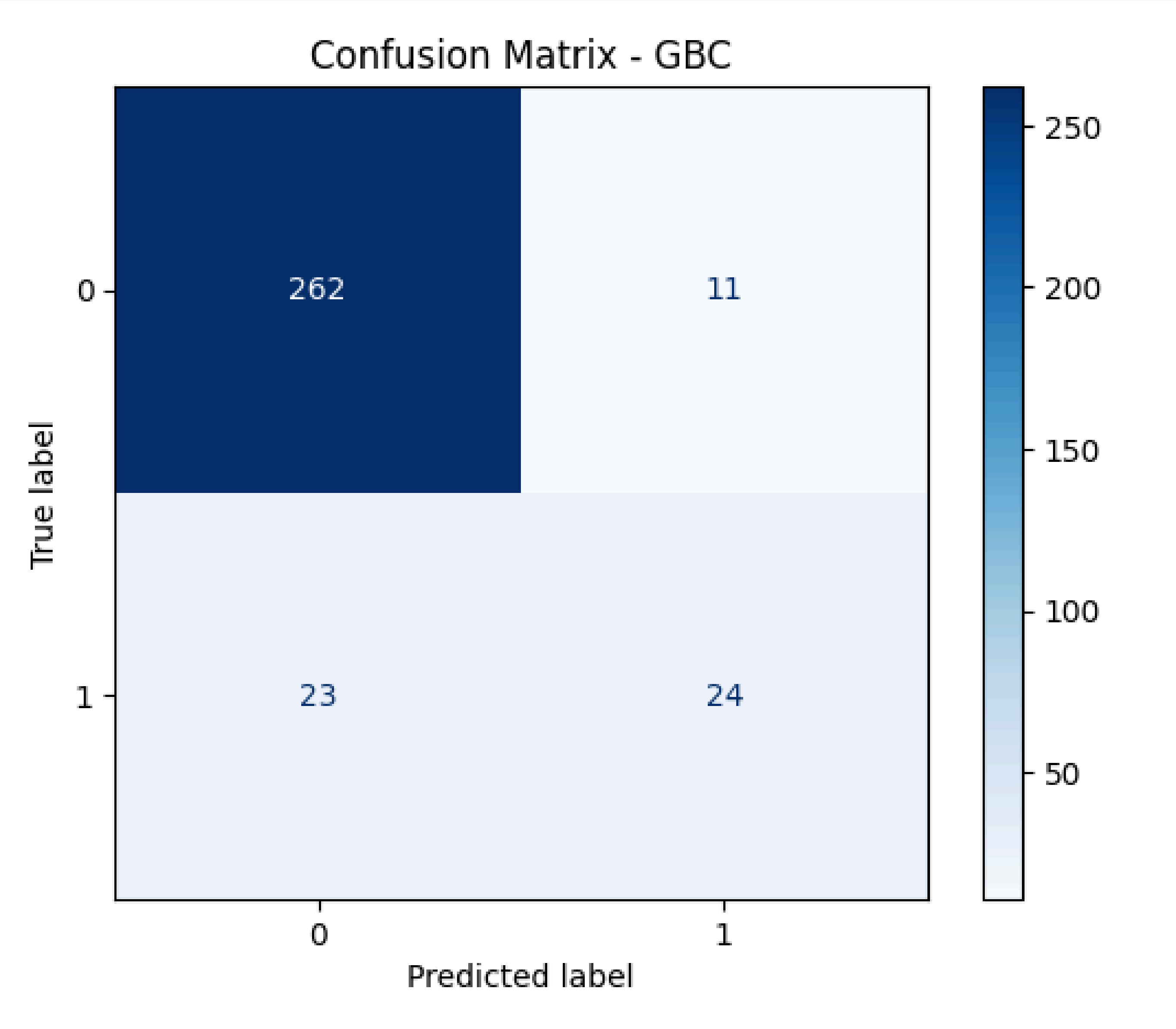
- Feature Importance (Gradient Boosting): Alcohol, sulphates most important





KEY VISUALS USED

- Actual vs Predicted Plot (GBR): Points close to ideal line

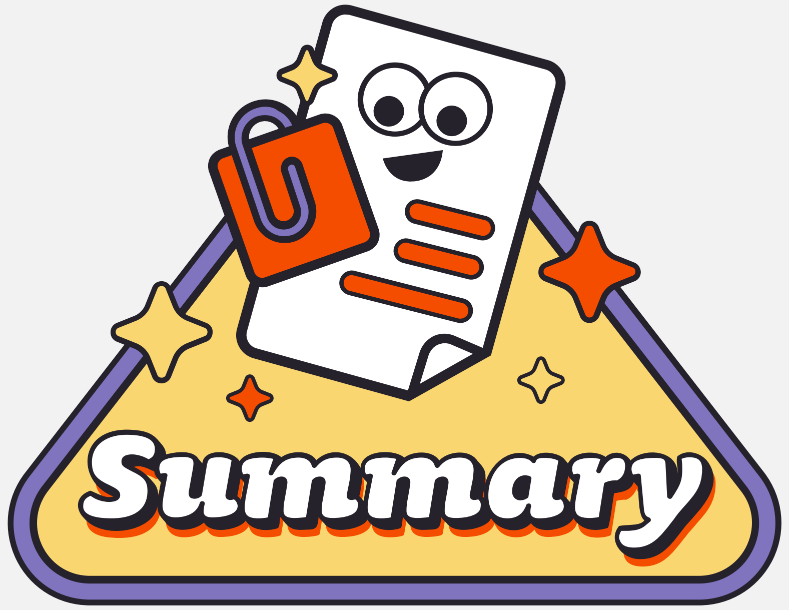


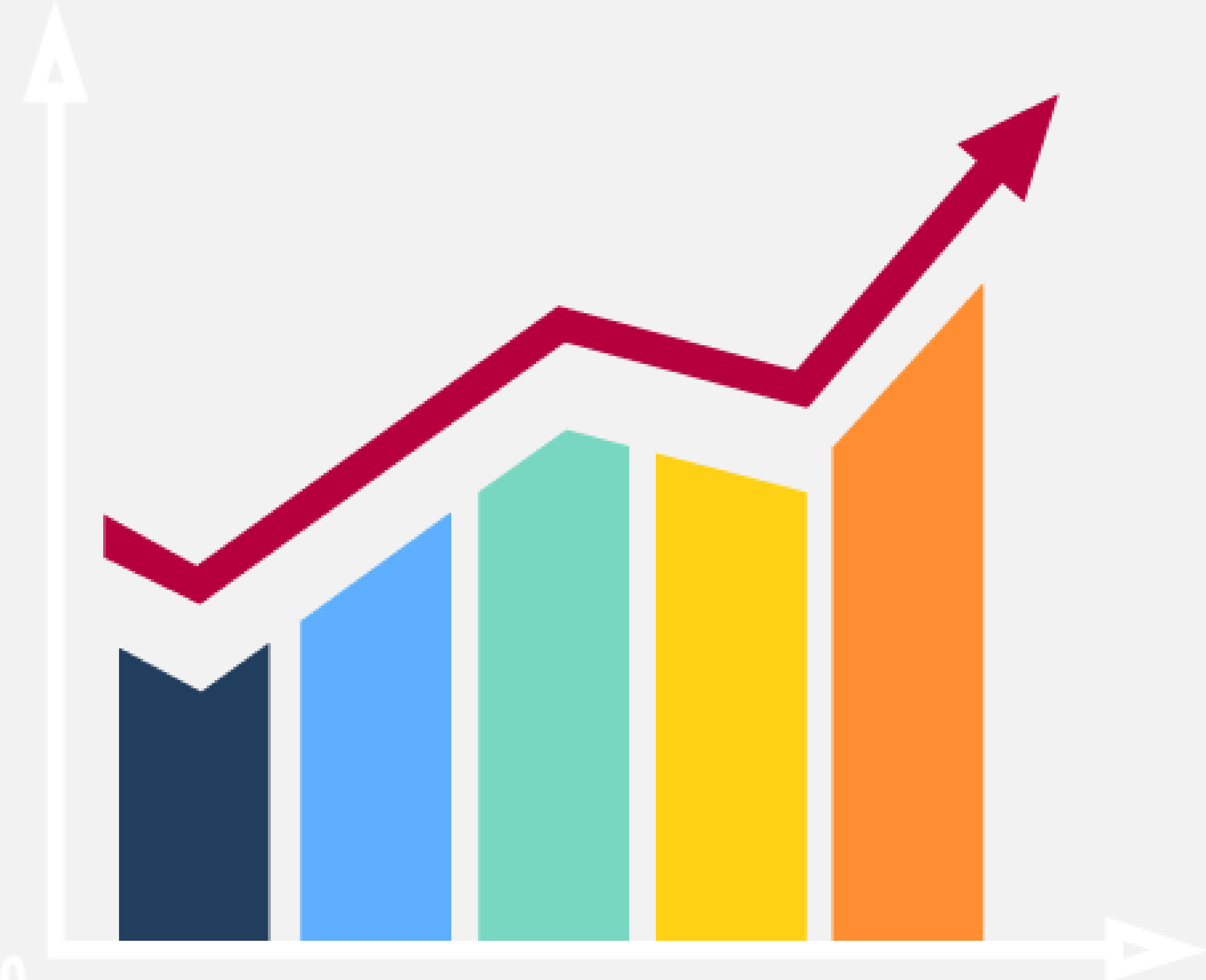
KEY VISUALS USED

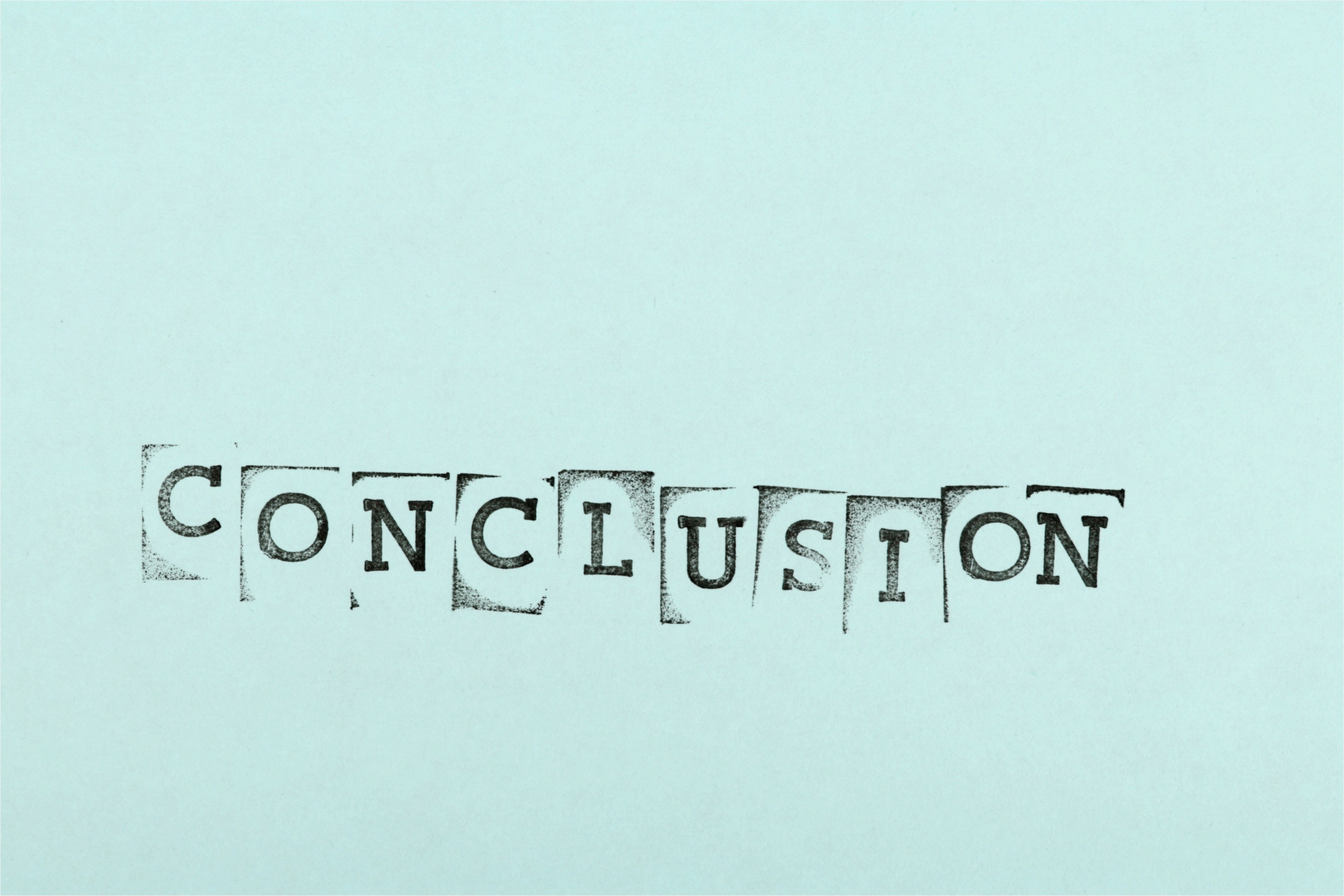
- This matrix shows the model is really strong at identifying non-high-quality wines. It correctly labeled 262 of them.
- For high-quality wines, it caught 24 out of 47.
- So, while overall performance is good, there’s still room to improve that side.

SUMMARY

- Gradient Boosting performed best overall in both tasks
- MLP improved with tuning, but slower to train
- Linear and Logistic models were good baselines







CONCLUSION & FUTURE WORK

- Learned how different models perform on the same dataset
- Gradient Boosting was most effective
- Future: handle class imbalance, tune tree depth, use full cross-validation

REFERENCE

- https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/data
- https://www.canva.com/