Predicting Wine Quality Using Binary Classification

By: Miguel Franco

Repo:
https://github.com/MFranco2000/Supervised-Learning-Final-Project-Predicting-Wine

Problem: Assessing Wine Quality

Consumer:

- Can a model help a casual wine consumer make an informed assessment of a wine’s quality?

Business:

- Could a model help a business predict if a wine they are developing will be liked by consumers?

Goal:

- Develop a model to predict whether a wine is good based on its properties.

Data Used

Wine Quality Dataset

Donated by: Paulo Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis

UC Irvine Machine Learning Repository

https://archive.ics.uci.edu/dataset/186/wine+quality

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Wine Quality [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.

Wine Quality Dataset

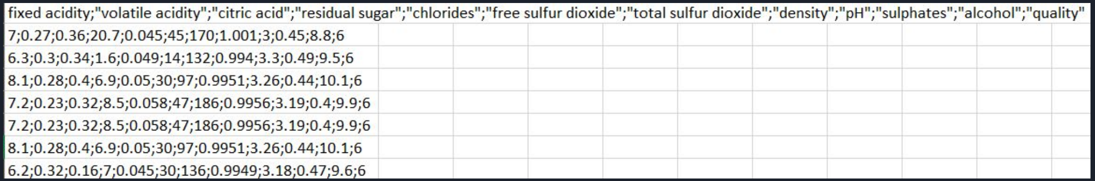
Two datasets are included, related to red and white “Vinho Verde” wine samples, from the north of Portugal.

Input variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol

Output variable: quality (score between 0 and 10)

Red Wine Number of Entrees: 1599

White Wine Number of Entrees: 4898



```
fixed acidity;"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide";"density";"pH";"sulphates";"alcohol";"quality"
7;0.27;0.36;20.7;0.045;45;170;1.001;3;0.45;8.8;6
6.3;0.3;0.34;1.6;0.049;14;132;0.994;3.3;0.49;9.5;6
8.1;0.28;0.4;6.9;0.05;30;97;0.9951;3.26;0.44;10.1;6
7.2;0.23;0.32;8.5;0.058;47;186;0.9956;3.19;0.4;9.9;6
7.2;0.23;0.32;8.5;0.058;47;186;0.9956;3.19;0.4;9.9;6
8.1;0.28;0.4;6.9;0.05;30;97;0.9951;3.26;0.44;10.1;6
6.2;0.32;0.16;7;0.045;30;136;0.9949;3.18;0.47;9.6;6
```

Methods and Approach

Binary classification approach (average and above: 6+ on the quality scale)

Models:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Classifier (SVC)

Logistic Regression Performance

Best Cross-Validation Accuracy: 0.7440801066113867

Tuned Logistic Regression Report

	precision	recall	f1-score	support
False	0.61	0.59	0.60	451
True	0.78	0.80	0.79	849
accuracy			0.72	1300
macro avg	0.70	0.69	0.69	1300
weighted avg	0.72	0.72	0.72	1300

Decision Tree Performance

Best Cross-Validation Accuracy: 0.7750603390834383

Tuned Decision Tree Report

	precision	recall	f1-score	support
False	0.62	0.67	0.64	451
True	0.82	0.79	0.80	849
accuracy			0.74	1300
macro avg	0.72	0.73	0.72	1300
weighted avg	0.75	0.74	0.75	1300

Random Forest Performance

Best Cross-Validation Accuracy: 0.8248983860220627

Tuned Random Forest Report

	precision	recall	f1-score	support
False	0.76	0.75	0.75	451
True	0.87	0.87	0.87	849
accuracy			0.83	1300
macro avg	0.81	0.81	0.81	1300
weighted avg	0.83	0.83	0.83	1300

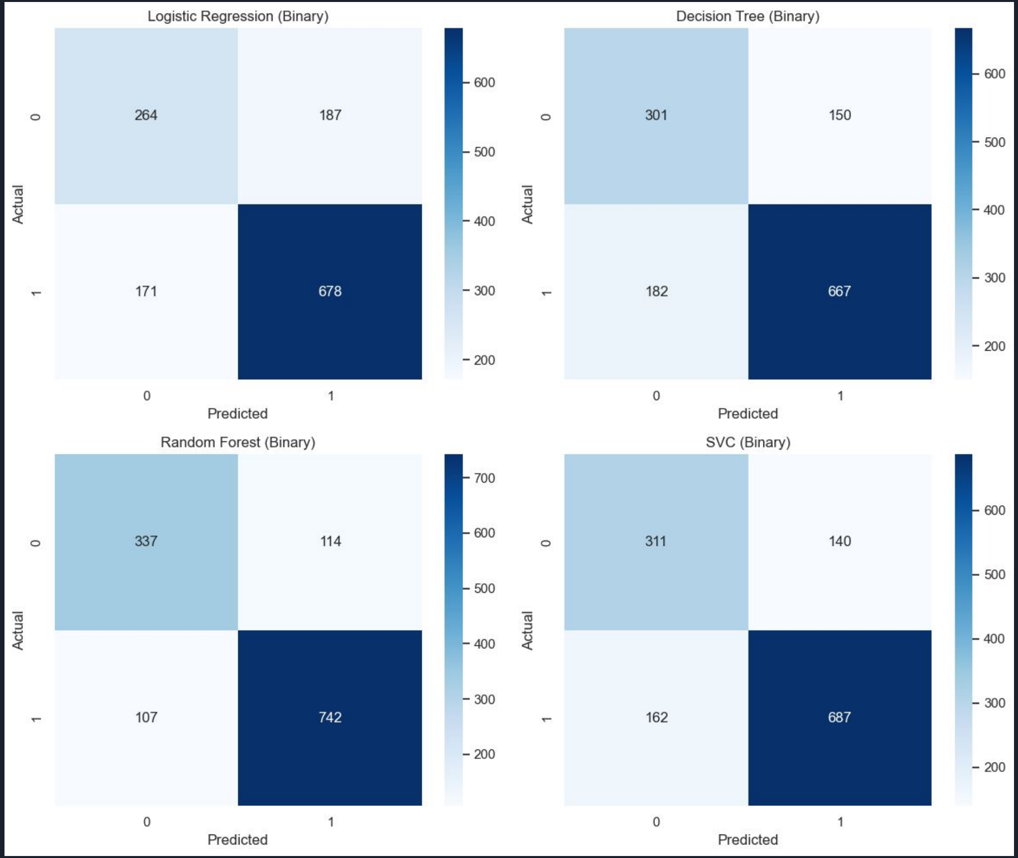
SVC

Best Cross-Validation Accuracy: 0.7802589398089879

Tuned SVC Report

	precision	recall	f1-score	support
False	0.61	0.69	0.67	451
True	0.83	0.81	0.82	849
accuracy			0.77	1300
macro avg	0.74	0.75	0.75	1300
weighted avg	0.77	0.77	0.77	1300

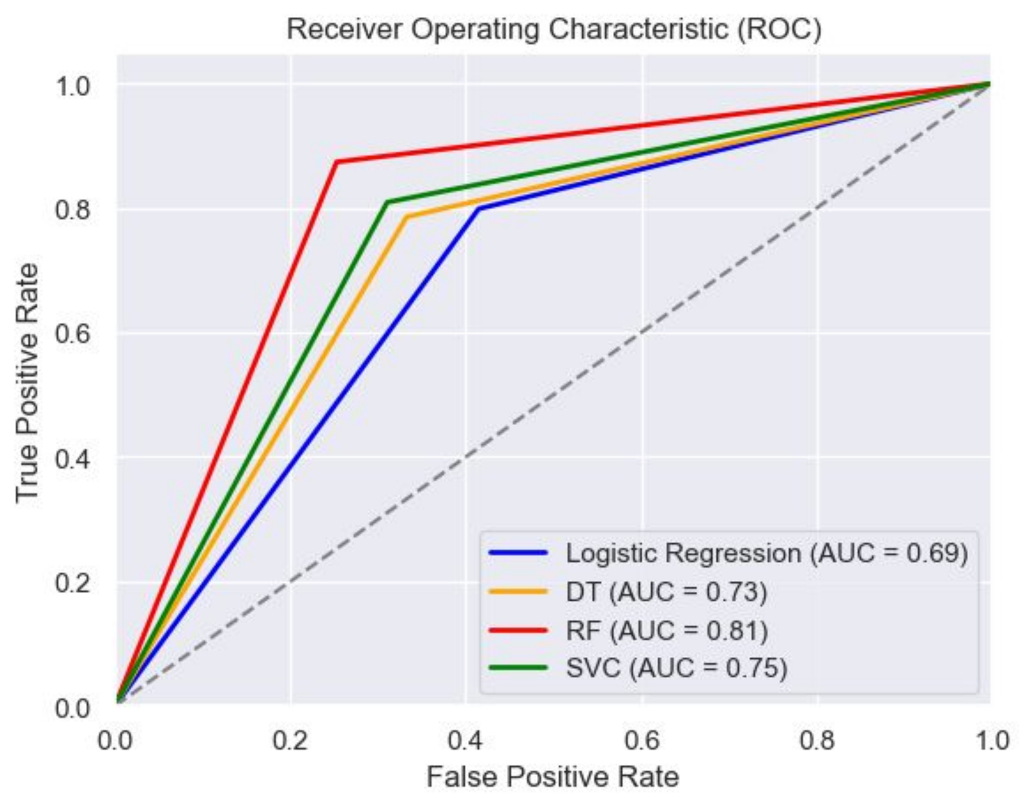
Results - Confusion Matrix



Random Forest correctly predicts the most True Positives and True Negatives

Followed by SVC Model

Results - ROC Curve



Best-performing model:

Random Forest

AUC Score: 0.81

Summary and Conclusion

Goal:

- Develop a model to predict whether a wine is good based on its properties.

Data:

- Wine Quality Dataset

Best Performing Model:

- Tuned Random Forest Model