Wine Quality

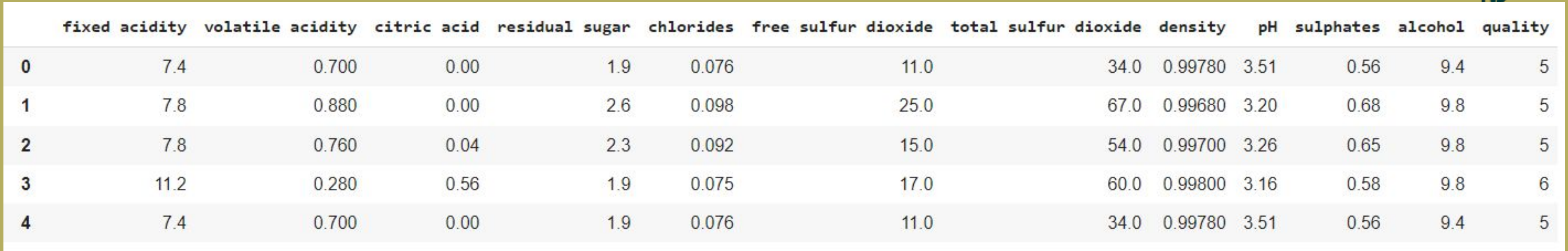
Dante Lee, Leyang Li, Ethan Koran, Jose Pantaleon

Introduction

- In the past, determining wine quality has always been a subjective test.
- Our aim is to transform this subjective matter into an objective one by considering certain chemical properties.
- Moreover, we also hope to determine the correlation between quality wine and its specific chemical properties. Allowing us to determine which factors are most important.
- Several groups have pursued classification models to increase objectivity, utilizing different methods and algorithms.
- Our contribution lies in adopting an altered algorithm while building upon the foundations laid by these previous models.

Data Set

- The dataset provided are specifically focused on the red and white variations of the renowned Portuguese "Vinho Verde" wine.
- The input variables are based on the following physicochemical tests: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol
- The dataset also includes a quality rating (dependent variable) for each wine sample, ranging from 0 to 10. This rating was obtained through sensory data provided by experts who assessed the quality of the wines.



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5

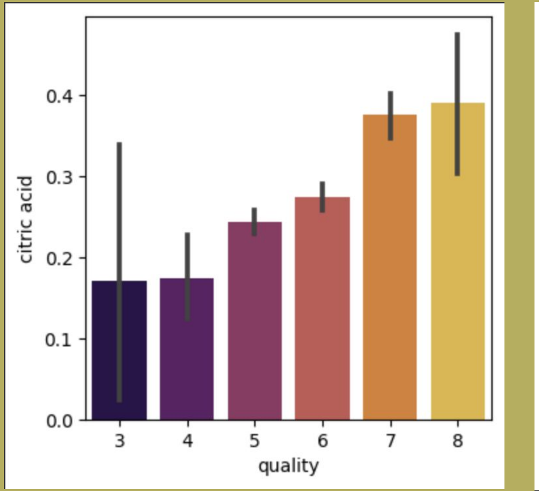
First 5 rows of the dataset

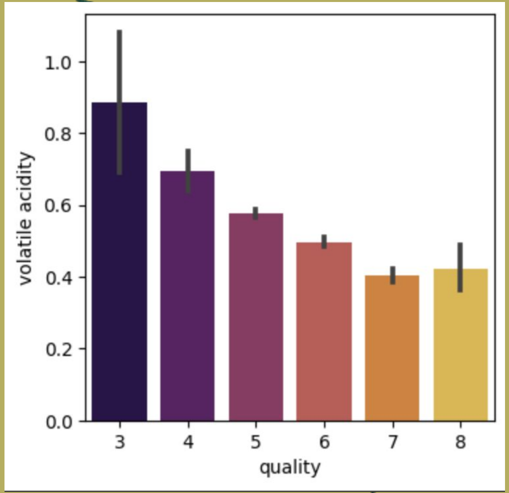
Visualizing the Data

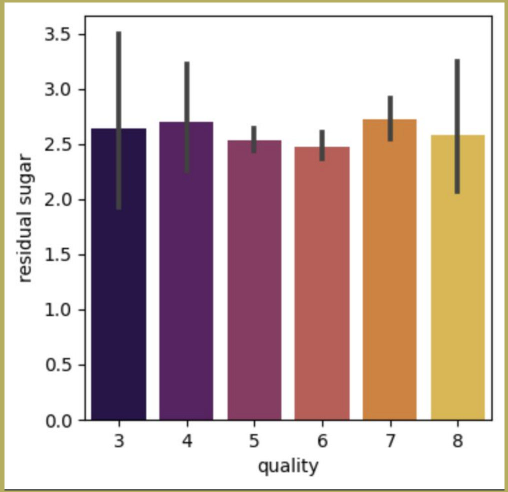
To visualize our data, we first created distribution bar graphs of each chemical property vs the quality grade given to each wine in the data set

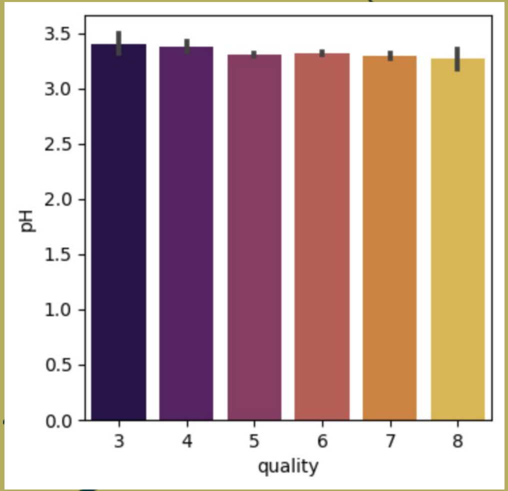
This showed us that there is a correlation between quality for some of the properties of wine we looked at, while other chemical properties had no significant effect on the wine's quality

Conclusion: regularization in our model could be helpful to eliminate these unnecessary variables









Methodology: Logistic Regression

Logistic regression is suitable for categorical or binary classification tasks.

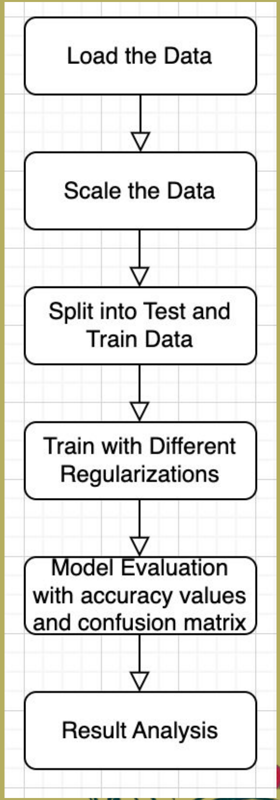
Scale the data and split it into test (20%) and train (80%) data.

Use regularization to avoid overfitting.

Compare the performance of models with L1, L2, or Elastic net regularization in the same train and test data.

Calculate the accuracy by dividing correct prediction number by total prediction number.

Visualize the result using confusion matrix and analyze it.



Load the Data

Scale the Data

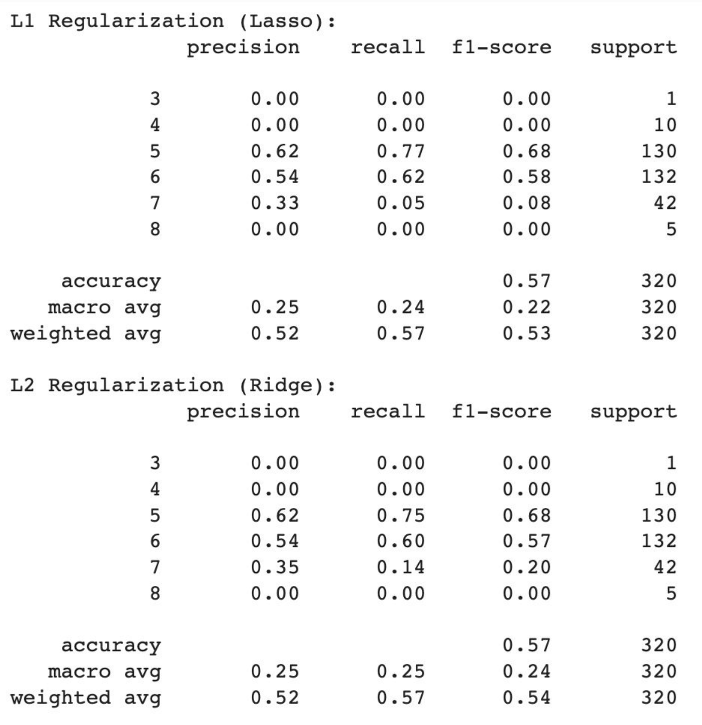
Split into Test and Train Data

Train with Different Regularizations

Model Evaluation with accuracy values and confusion matrix

Result Analysis

L1 & L2 Regularization

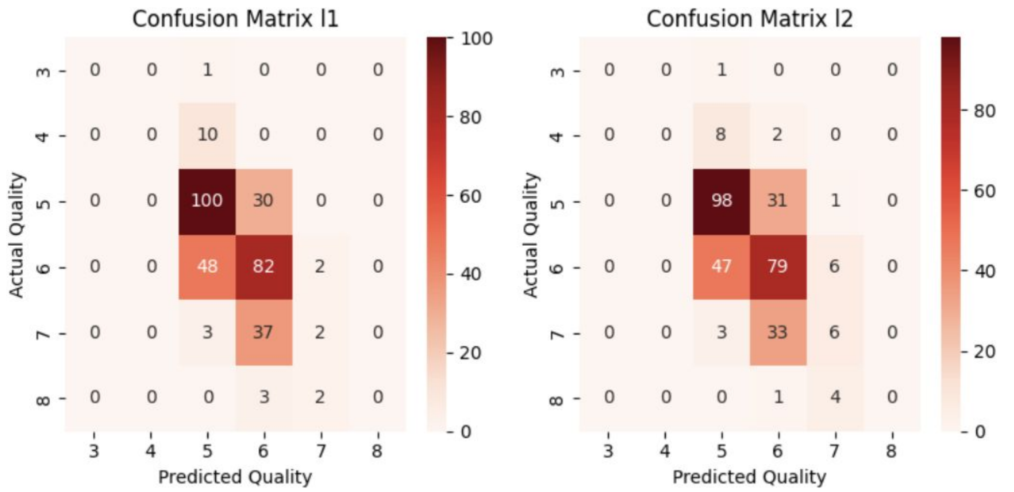


```
L1 Regularization (Lasso):
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.00      0.00      0.00        10
           5       0.62      0.77      0.68       130
           6       0.54      0.62      0.58       132
           7       0.33      0.05      0.08        42
           8       0.00      0.00      0.00         5

    accuracy                           0.57       320
   macro avg       0.25      0.24      0.22       320
weighted avg       0.52      0.57      0.53       320

L2 Regularization (Ridge):
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.00      0.00      0.00        10
           5       0.62      0.75      0.68       130
           6       0.54      0.60      0.57       132
           7       0.35      0.14      0.20        42
           8       0.00      0.00      0.00         5

    accuracy                           0.57       320
   macro avg       0.25      0.25      0.24       320
weighted avg       0.52      0.57      0.54       320
```

L1 and L2 model reports

L1 solver: liblinear, L2 solver: lbfgs

Note that although the accuracy lies around 57%, it is 97-98% with tolerance scope of 1.



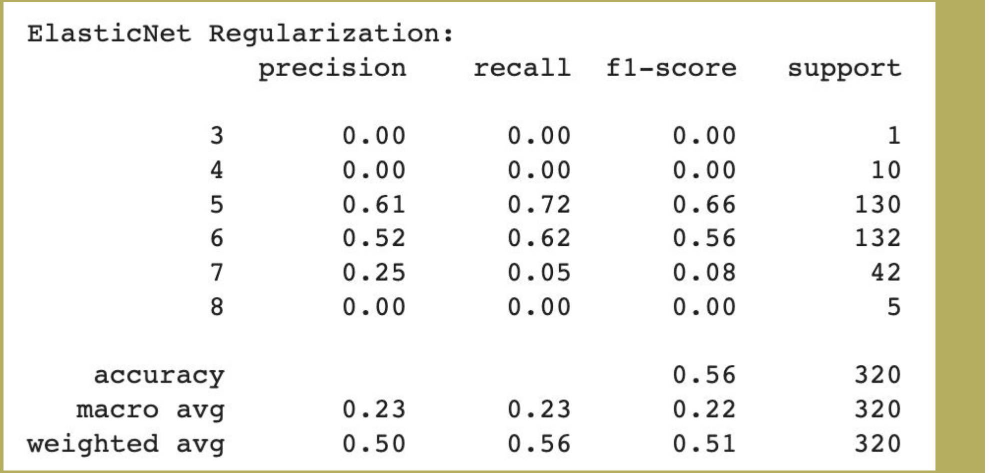
L1 and L2 confusion matrices

Elastic Net Regression

Solver: saga, L1 ratio: 0.5

Accuracy: 56%,

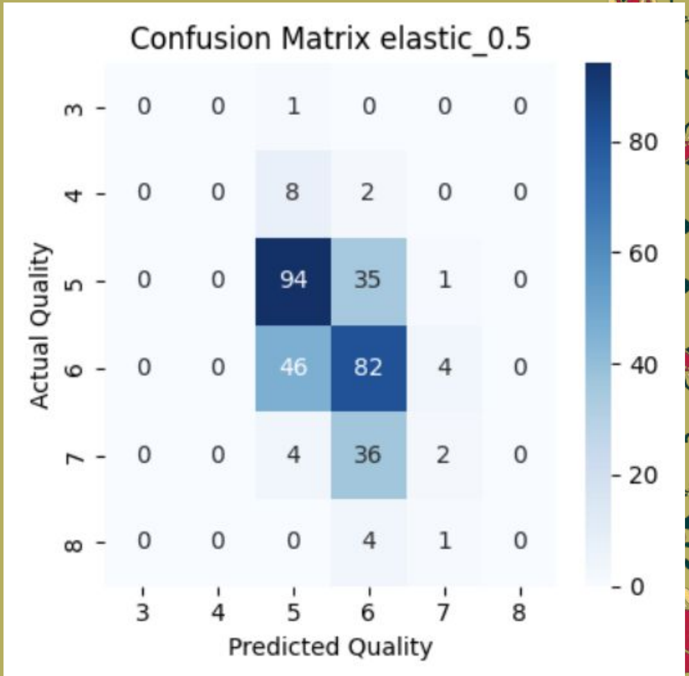
Accuracy with tolerance scope of 1: 96%.



```
ElasticNet Regularization:
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.00      0.00      0.00        10
           5       0.61      0.72      0.66       130
           6       0.52      0.62      0.56       132
           7       0.25      0.05      0.08        42
           8       0.00      0.00      0.00         5

    accuracy                           0.56       320
   macro avg       0.23      0.23      0.22       320
weighted avg       0.50      0.56      0.51       320
```

Elastic net model report

Elastic net confusion matrix

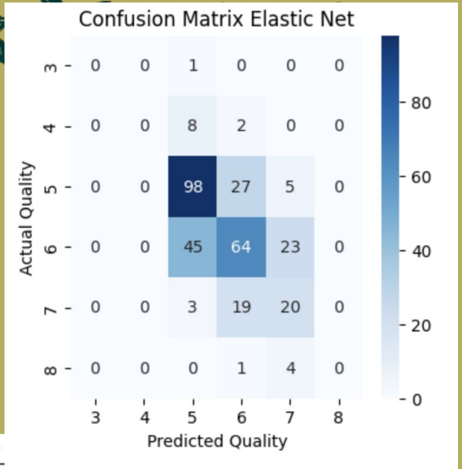


	3	4	5	6	7	8
3	0	0	1	0	0	0
4	0	0	8	2	0	0
5	0	0	94	35	1	0
6	0	0	46	82	4	0
7	0	0	4	36	2	0
8	0	0	0	4	1	0

SGDClassifier

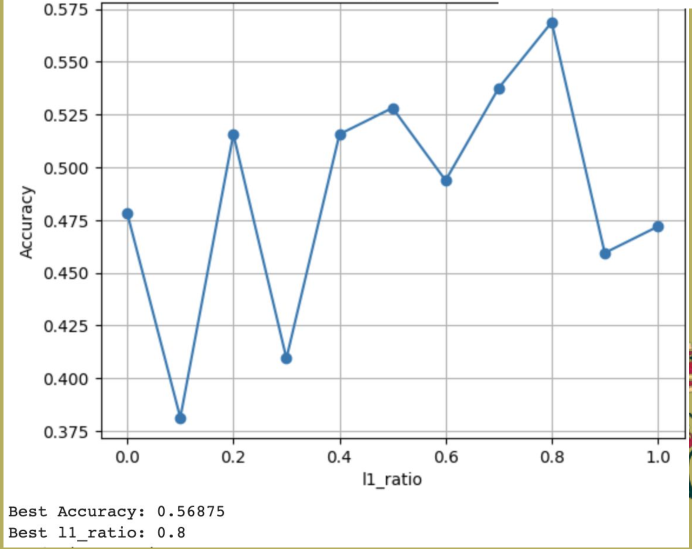
Stochastic gradient descent (SGD) learning: iteratively updates the model weights by taking small steps in the direction of the negative gradient of the loss function with respect to the parameters.

The stochastic aspect refers to the fact that it randomly samples a subset of the training data (also known as a mini-batch) in each iteration, making it efficient for large datasets.


Confusion Matrix Elastic Net

Actual Quality \ Predicted Quality	3	4	5	6	7	8
3	0	0	1	0	0	0
4	0	0	8	2	0	0
5	0	0	98	27	5	0
6	0	0	45	64	23	0
7	0	0	3	19	20	0
8	0	0	0	1	4	0

We trained the model eleven times, with elastic net regularization of different L1 ratios. The graph of accuracy vs. L1 ratio is shown, and the best result (accuracy: 56.9%) is presented. The fluctuation is due to the randomness of SGD.

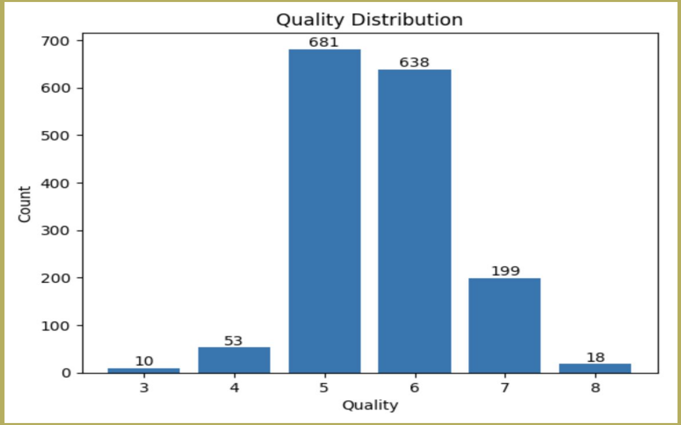


```
Best Accuracy: 0.56875
Best l1_ratio: 0.8
```

Conclusion of Categorical Classification

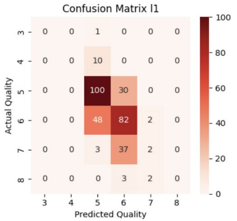
In terms of accuracy, L1 (Lasso regression 57.5%) outperforms L2 (Ridge regression 57.2%) and Elastic net regression (56.9% maximum).

Confusion matrices show models tend to overestimate wines of quality 3 and 4 and underestimate wines of quality 7 and 8, probably due to the lack of corresponding data in dataset.



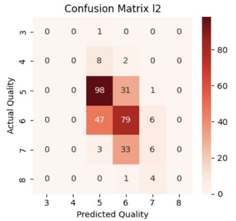
Quality distribution in data set

Notice that the best model of SGD Classifier performs better in quality 7. Probably the randomly generated subsets happen to favor quality 7 prediction.



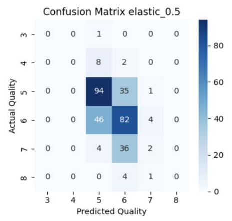
Confusion Matrix l1

	3	4	5	6	7	8
3	0	0	1	0	0	0
4	0	0	10	0	0	0
5	0	0	100	30	0	0
6	0	0	48	82	2	0
7	0	0	3	37	2	0
8	0	0	0	3	2	0



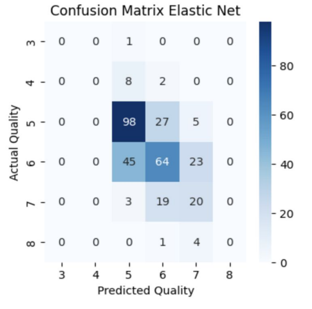
Confusion Matrix l2

	3	4	5	6	7	8
3	0	0	1	0	0	0
4	0	0	8	2	0	0
5	0	0	98	31	1	0
6	0	0	47	79	6	0
7	0	0	3	33	6	0
8	0	0	0	1	4	0



Confusion Matrix elastic_0.5

	3	4	5	6	7	8
3	0	0	1	0	0	0
4	0	0	8	2	0	0
5	0	0	94	35	1	0
6	0	0	46	82	4	0
7	0	0	4	36	2	0
8	0	0	0	4	1	0



Confusion Matrix Elastic Net

	3	4	5	6	7	8
3	0	0	1	0	0	0
4	0	0	8	2	0	0
5	0	0	98	27	5	0
6	0	0	45	64	23	0
7	0	0	3	19	20	0
8	0	0	0	1	4	0

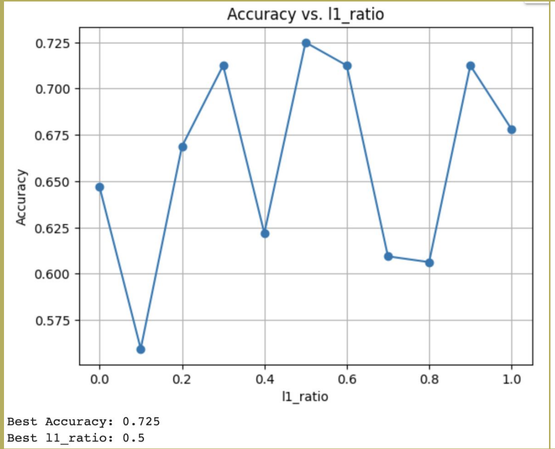
Binary Classification

Used the same approaches in binary classification with 6 as threshold.

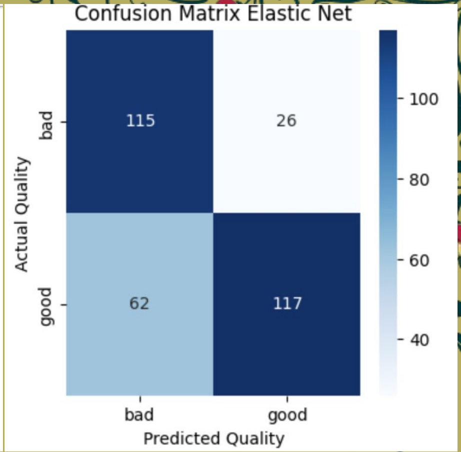
L2 (74.1%) outperforms L1 (73.8%) and Elastic net regression (72.5% maximum),

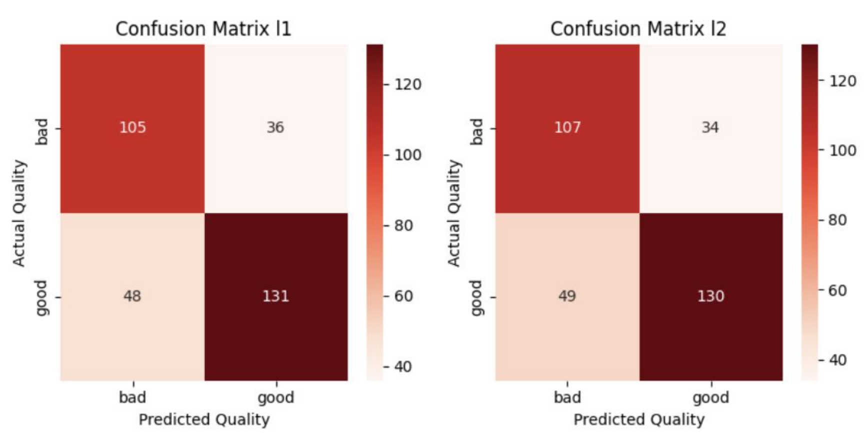
Accuracy:

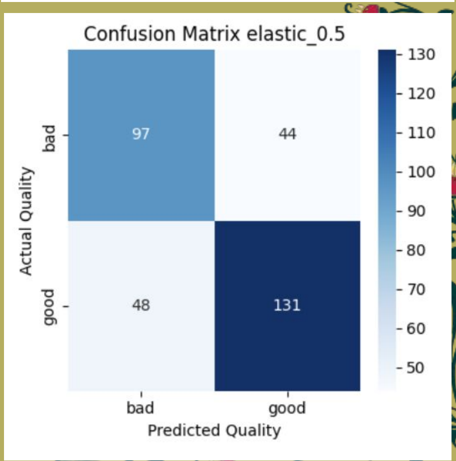
L1:	73.8%
L2:	74.1%
Elastic:	71.3%
SGD Elastic:	72.5%



```
Best Accuracy: 0.725
Best l1_ratio: 0.5
```





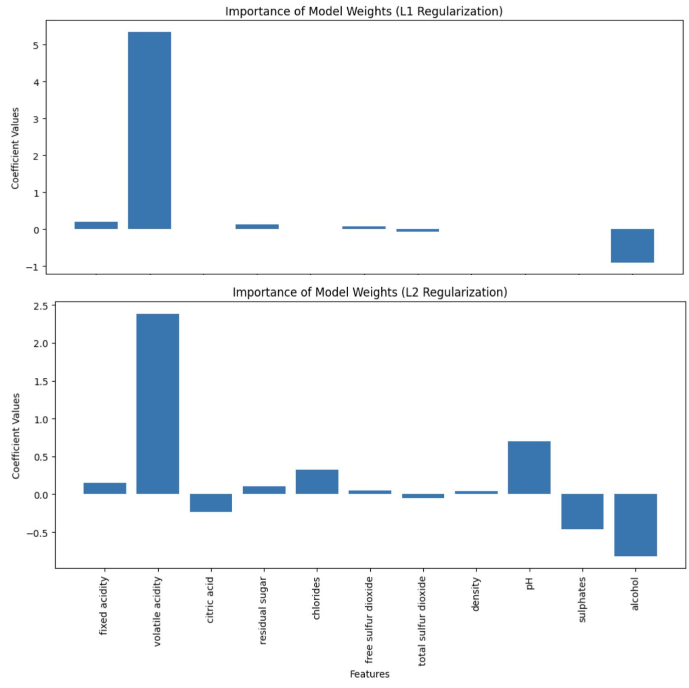


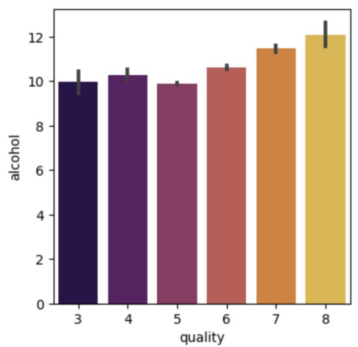
Conclusion

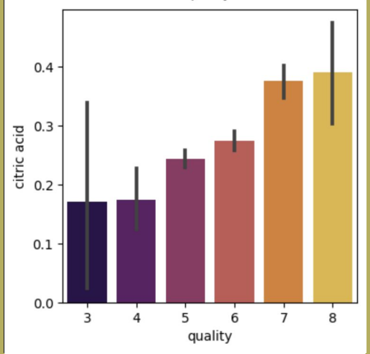
Binary classification (L2, 74.1%) achieves a significant higher accuracy than categorical classification (L1, 57.5%). However, categorical classification has more reference value with tolerance scope of 1 (L1, 97.8%).

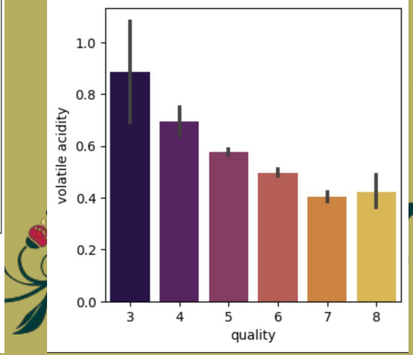
According to the weights of L1 categorical model, volatile acidity is considered most important, followed by alcohol.

This matches the data set visualization as there is a significant correlation between volatile acidity and quality. Although a correlation is also present in citric acid, there are overlaps, especially between quality 5 and 6. Therefore, it's reasonable for model not considering citric acid.









Live Demonstration

Thank You.