

Decision Tree and Neural Networks on Red-Wine Quality Dataset

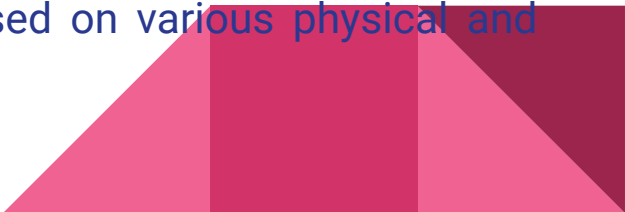
Sriya Gorrepati
Hari Chandana Kotnani

Introduction

- Our project aims to analyze the factors that influence the quality of red wine.
- To ensure quality control and provide consumers with accurate information, proper wine traceability is crucial
- Our analysis aims to empower winemakers to focus on these essential features to achieve the best possible quality of wine.



Dataset

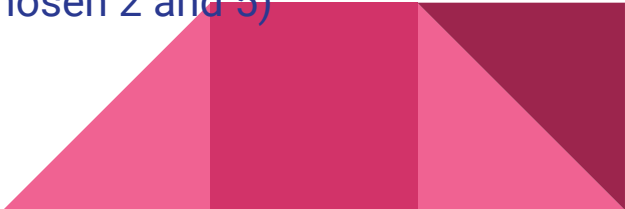
- The red wine quality dataset, chosen from Kaggle.
 - comprises 1599 samples of red wine. Each sample is described by 12 attributes, including Fixed acidity, Volatile acidity, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Density, pH, Sulfates, Alcohol, and Quality.
 - Quality is the target attribute.
 - suitable for classification and regression tasks related to wine production, marketing, and quality control, and can provide valuable insights into the factors that influence wine quality.
 - Decision trees can be used to identify the most important attributes.
 - Implemented neural networks to predict wine quality based on various physical and chemical properties.
- 

PreProcessing

- binarized the data by converting values below 5 to '0', considering values below 5 as 'bad' and values above 5 to '1', considering values above 5 as 'good'.
- For our neural network model, we applied the StandardScaler preprocessing step to the data by fitting it to the training data and then transforming both the training and test data.



Decision Tree Analysis

- Calculated the correlation between each pair of features
 - Top 5 features with the highest correlation with target variable: Alcohol, Volatile acidity, Sulfates, Citric acid, and Total sulfur dioxide
 - Two approaches: Multi-class and Binary classification
 - In the multi-class approach, we treated each quality value from 3 to 8 as a separate class.
 - In binary classification, we treated quality values above 5 as one class and below or equal to 5 as another class.
 - GridSearchCV to find best parameters for the model
 - Visualized decision trees with different max_depth value(chosen 2 and 5)
- 

Decision Tree Analysis: Feature importance

- We used 'feature_importances_' to identify the choices the algorithm made in which attributes to split on.
- Gini importance method is used to calculate feature importance.
- Evaluated using performance metrics and compared for both approaches.



Approach 1 Results : Decision tree- multi-class

Max Depth 2:

Accuracy: 0.59375

alcohol: 0.854680408049223
volatile acidity: 0.0
sulphates: 0.14531959195077704
citric acid: 0.0
total sulfur dioxide: 0.0

	precision	recall	f1-score	support
3	0.00	0.00	0.00	1
4	0.00	0.00	0.00	13
5	0.63	0.78	0.69	140
6	0.55	0.60	0.58	134
7	0.00	0.00	0.00	30
8	0.00	0.00	0.00	2
accuracy			0.59	320
macro avg	0.20	0.23	0.21	320
weighted avg	0.51	0.59	0.55	320

Decision tree-multi-class

Max Depth 5:

		precision	recall	f1-score	support
	3	0.00	0.00	0.00	1
alcohol: 0.4701451258266833	4	0.00	0.00	0.00	13
volatile acidity: 0.10130562881723514	5	0.62	0.74	0.68	140
sulphates: 0.20798219808634727	6	0.60	0.54	0.57	134
citric acid: 0.018271224911649526	7	0.41	0.40	0.41	30
total sulfur dioxide: 0.2022958223580848	8	0.00	0.00	0.00	2
	accuracy			0.59	320
	macro avg	0.27	0.28	0.28	320
	weighted avg	0.56	0.59	0.57	320

Decision tree- multi-class-using hyper parameters

Using hyper parameters

```
Best parameters:  {'criterion':  
'entropy', 'max_depth': 5,  
'max_leaf_nodes': None,  
'min_samples_leaf': 5}
```

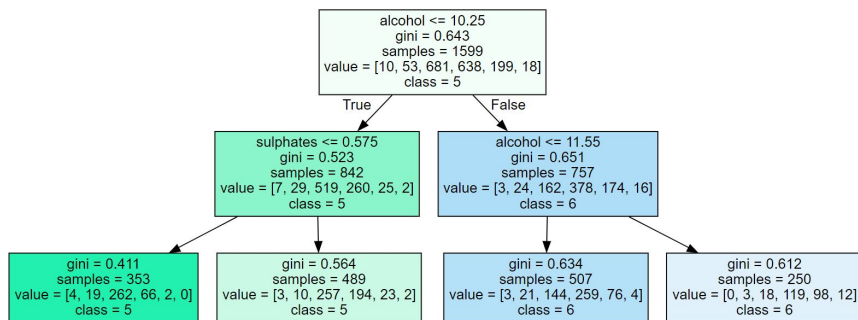
Accuracy: 0.59375

	precision	recall	f1-score	support
3	0.00	0.00	0.00	1
4	0.00	0.00	0.00	13
5	0.66	0.68	0.67	140
6	0.57	0.59	0.58	134
7	0.41	0.53	0.46	30
8	0.00	0.00	0.00	2
accuracy			0.59	320
macro avg	0.27	0.30	0.29	320
weighted avg	0.57	0.59	0.58	320

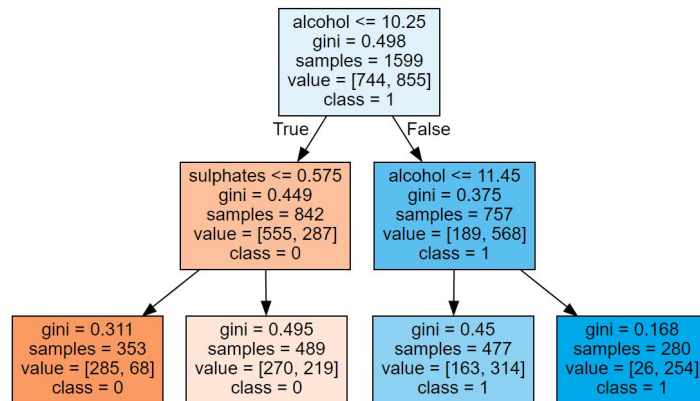
Decision tree:

Max Depth 5: Due to larger size trees, uploaded in Git for both approaches

Max Depth 2: Multi-class



Max Depth 2: Binary class



Approach 2 Results : Decision tree- Binary class

Max Depth 2: Feature importance

```
alcohol: 0.8534468516082325  
volatile acidity: 0.0  
sulphates: 0.146553148391767  
citric acid: 0.0  
total sulfur dioxide: 0.0
```

- Performance:
- Accuracy: 0.70625

Approach 2 Results : Decision tree- Binary class

Max Depth 5: Feature importance:

```
alcohol: 0.5091813094225128  
volatile acidity: 0.09609707716101944  
sulphates: 0.22966563969990947  
citric acid: 0.02057261994318562  
total sulfur dioxide: 0.14448335377337274
```

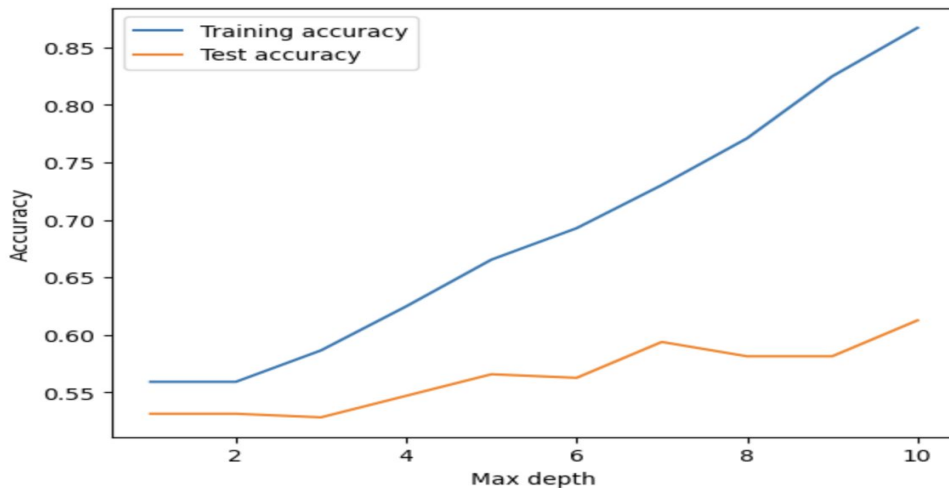
- Performance:
- Accuracy: 0.746875

Comparison: Results

- Based on the results, it can be concluded that the binary classification approach outperformed the multi-class classification approach for this particular dataset.
- The binary classification accuracy is significantly higher than the multi-class accuracy.
- Also, adding more hyperparameters through grid search did not significantly improve the accuracy of the model

Effect of Max depth on bias and variance

- Increase the max_depth more leads to overfitting, resulting in a slightly lower test accuracy and also increase in more difference between train and test accuracy.



Neural Networks


- The dataset divides quality into two classes, and we chose to use all features without selecting any best features .
- Visualize the trained MLP classifiers for best accuracy
- Implemented Grid search t choose hidden layer sizes
- Selected best three models

```
Configuration 1 :  
Parameters: {'hidden_layer_sizes': (30, 30)}
```

```
Configuration 2 :  
Parameters: {'hidden_layer_sizes': (30, 30, 15)}
```

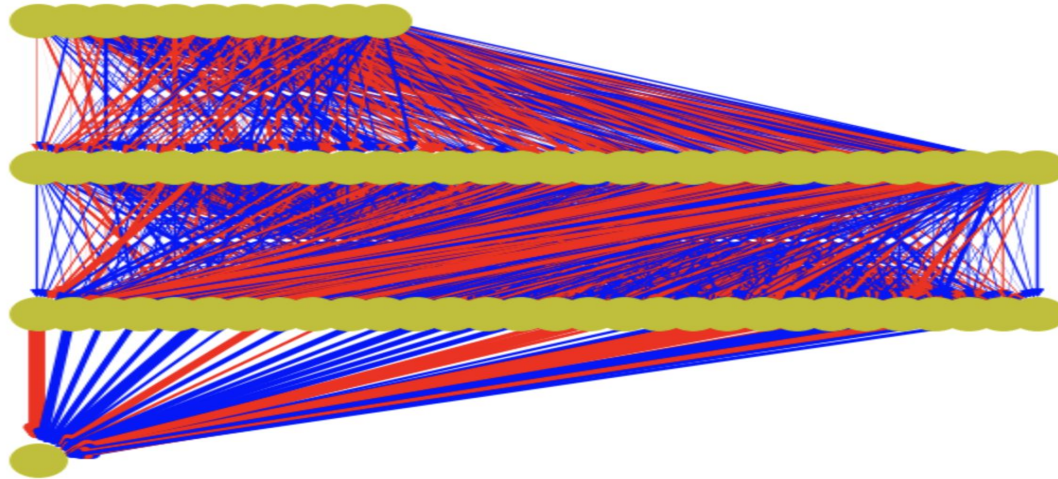
```
Configuration 3 :  
Parameters: {'hidden_layer_sizes': (15, 15, 15)}
```

```
MLPClassifier(hidden_layer_sizes=(30, 30), max_iter=10000)
```



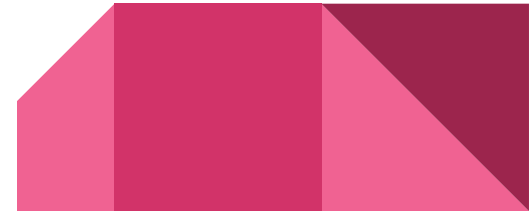
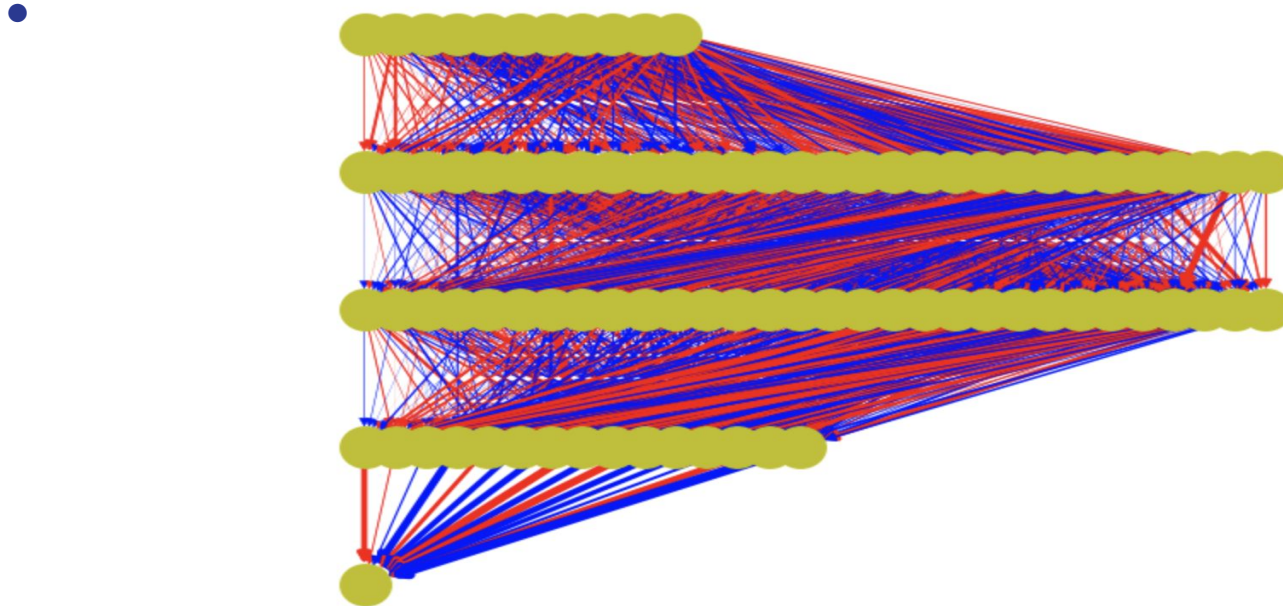
Architecture 1:

- Hidden layer sizes as (30, 30), 2 hidden layers, each with 30 neurons.



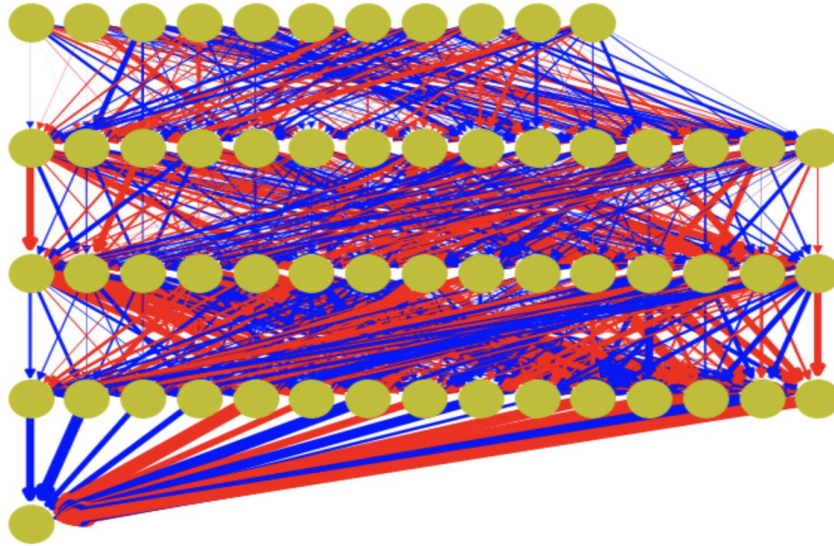
Architecture 2:

- Hidden layer sizes as (30, 30, 15), 3 hidden layers, first and second containing 30 neurons and 15 neurons in third layer.



Architecture 3:

- Hidden layer sizes as (15, 15, 15) ,3 hidden layers, each containing 15 neurons.



Parameters comparison

					precision	recall	f1-score	support						precision	recall	f1-score	support		
					precision	recall	f1-score	support	0	0.75	0.65	0.69	184	0	0.74	0.72	0.73	184	
									1	0.73	0.81	0.77	216	1	0.76	0.78	0.77	216	
0	0.68	0.71	0.69	184															
1	0.74	0.71	0.73	216															
					accuracy				0.74	400				accuracy				0.75	400
					macro avg	0.74	0.73	0.73	400				macro avg	0.75	0.75	0.75	400		
					weighted avg	0.74	0.74	0.74	400				weighted avg	0.75	0.75	0.75	400		

(30,30) Layers

(30,30,15) Layers

(15,15,15) Layers

- layers (15,15,15) is having highest accuracy and f1 score.

Conclusion:

- Alcohol, Volatile acidity, Sulfates, Citric acid, and Total sulfur dioxide are the most important features for wine quality
- Decision trees and neural networks provide valuable insights for wine production and quality control
- Proper data preprocessing is crucial for accurate results

Thank you