

Name: Daniella Omenogor

Report on Exploratory Data Analysis and Modelling of the Wine Quality Dataset using Regression Algorithms.

The exercise given was to perform exploratory data analysis, data modelling and evaluation, on the given dataset: **wine_quality-red.csv**. The wine quality dataset describes the amount of various chemicals present in different types of red wine and their effect on the quality of the wine. It contains eleven(11) input(independent) variables, and one(1) dependent(target) variable which is represented in numerical form.

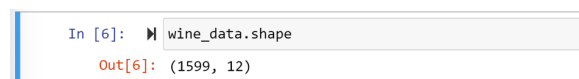
The tasks performed in this exercise are:

- Exploratory Data Analysis.
- Data modelling with linear regression, ridge and lasso regression, to predict wine quality on a scale of 0 to 10.
- Compute the performance of the aforementioned models and compare.
- Check the assumptions of linear regression.

Data Loading

The prerequisite for completing the above tasks was to load the data. Upon loading, it was observed that the given dataset consists of:

- 12 features: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality
- 1599 observations



```
In [6]: wine_data.shape
Out[6]: (1599, 12)
```

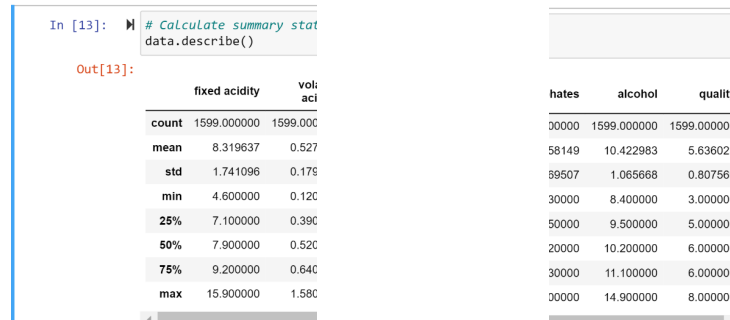
Exploratory Data Analysis.

After loading the data, the next step was to perform exploratory data analysis (EDA) on the dataset. EDA is the process of analysing and summarising the main characteristics of a dataset in order to better understand the data and uncover patterns and relationships between the variables. In this step, I observed the distribution of the data features, summary statistics, relationship(correlation) between features, balance or imbalance in the dataset, and so on.

I began the EDA process by examining the dataset's features and summary statistics. As earlier stated, the dataset consists of 1599 observations and 12 features. The quality feature, which is the dependent or target variable, ranges from 0 to 10 and is represented in numerical form.

Some key insights derived from carrying out EDA include:

- The average wine quality in the dataset is 5.636, with the wine with the least quality in the dataset having a quality score of 3 and the wine with the highest quality having a quality score of 8.



```
In [13]: # Calculate summary statistics
data.describe()
```

Out[13]:

	fixed acidity	volatile acidity
count	1599.000000	1599.000000
mean	8.319637	0.527
std	1.741096	0.176
min	4.600000	0.120
25%	7.100000	0.390
50%	7.900000	0.520
75%	9.200000	0.640
max	15.900000	1.580

	total acidity	alcohol	quality
count	1599.000000	1599.000000	1599.000000
mean	10.422983	5.636023	5.636023
std	1.065668	0.807569	0.807569
min	8.400000	3.000000	3.000000
25%	9.500000	5.000000	5.000000
50%	10.200000	6.000000	6.000000
75%	11.100000	6.000000	6.000000
max	14.900000	8.000000	8.000000

- There are no missing values in the dataset.
- I also plotted the correlation between each feature and the target variable and found that most of the features are not strongly correlated with the target variable(wine quality). This was discovered by using a heatmap to visualise the correlation between respective features in the dataset. Particularly, the features with the highest correlation with the target variable are; alcohol(0.45 correlation) and volatile acidity(-0.39 correlation). The lack of strong correlation might affect how well the models would be able to predict wine quality as the features would not be very good or reliable determinants of the wine quality.
- The wine samples with high quality(8) had the highest alcohol content. However, wine samples with lower quality scores also have fairly high alcohol content (greater than or equal to 10).
- I also examined the balance or imbalance of the dataset by visualising the distribution of the wine quality counts as well as the percentage counts of wine quality scores. I found that the dataset is imbalanced, with about 80% of the wine samples having a quality score of 5 or 6. This could also potentially impact the performance of the models.
- Finally, I examined the distribution of the features in the dataset. I found that while a number of the features were approximately normally distributed, a good number of them were either left or right skewed such as; citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, and alcohol.

Data modelling.

LINEAR REGRESSION:

The linear regression model was built using all the 11 input features. The model achieved an MAE of 0.47, MSE of 0.38, RMSE of 0.62 and an R2 score of 0.328. The R-squared(R2) value of 0.328 indicates that only about 32.8% of the variance in the target variable(wine quality) is explained by the model. This suggests that the model is not a good fit for the data.

RIDGE REGRESSION:

The ridge regression model was also built using all the 11 input features. A value of alpha=0.01 was used for regularisation. The model achieved an MAE of 0.47, MSE of 0.38, RMSE of 0.62 and an R2 score of 0.331. The R-squared(R2) value of 0.331 indicates that only about 33.1% of the variance in the target variable(wine quality) is explained by the model, which is similar to the linear regression model. This depicts that regularisation did not significantly improve the performance of the model.

LASSO REGRESSION:

The lasso regression model was once again built using all the 11 input features. A value of alpha=0.01 was also used for regularisation. The model had an MAE of 0.51, MSE of 0.42, RMSE of 0.65 and an R2 score of 0.26.

Models Performance and Comparison.

```
In [122]: performance = {
            'Algorithm': ['Linear Regression', 'Ridge Regression', 'Lasso Regression'],
            'MAE': [mae_linear_reg, mae_ridge_reg, mae_lasso_reg],
            'MSE': [mse_linear_reg, mse_ridge_reg, mse_lasso_reg],
            'RMSE': [rmse_linear_reg, rmse_ridge_reg, rmse_lasso_reg],
            'R2': [r2_linear_reg, r2_ridge_reg, r2_lasso_reg]
          }

performance_df = pd.DataFrame(performance)
print(performance_df)
```

	Algorithm	MAE	MSE	RMSE	R2
0	Linear Regression	0.469633	0.384471	0.620057	0.328389
1	Ridge Regression	0.468638	0.382631	0.618572	0.331604
2	Lasso Regression	0.511414	0.423894	0.651071	0.259524

The results suggest that both Linear Regression and Ridge Regression models have similar performance and outperformed the Lasso Regression model. This is because the Lasso model had the highest MAE, MSE and RMSE values, indicating that the model fitted the data the least. However, the difference in performance between Linear and Ridge Regression models is marginal. The R2 score indicates that only about 33% of the variability in the dependent variable can be explained by the independent variables in the models. Overall, the Linear Regression model and Ridge Regression model perform better than the Lasso Regression model.

Finally, based on the R-squared values obtained, it can be concluded that none of the regression models performed well in predicting wine quality. This may be due to the fact that the dataset is highly imbalanced, with most of the wine samples having a quality score of 5 or 6. As a result, the models may have been biased towards these scores, resulting in poor performance.

Checking Assumptions of Linear Regression.

Recall that there are four(4) main assumptions of linear regression which are;

- Linearity: There is a linear relationship between the independent variable(s) and the dependent variable.
- Independence: The observations in the dataset must be independent of each other.
- Homoscedasticity: The variance of the errors (residuals) should be constant across all values of the independent variable(s).
- Normality: The errors (residuals) should be normally distributed.

The aforementioned assumptions were verified using data visualisation methods. The results of verifying the above assumptions can be found in the notebook attached (*Red-Wine-Dataset_EDA-and-Modelling.ipynb*).