Assignment 1 Report

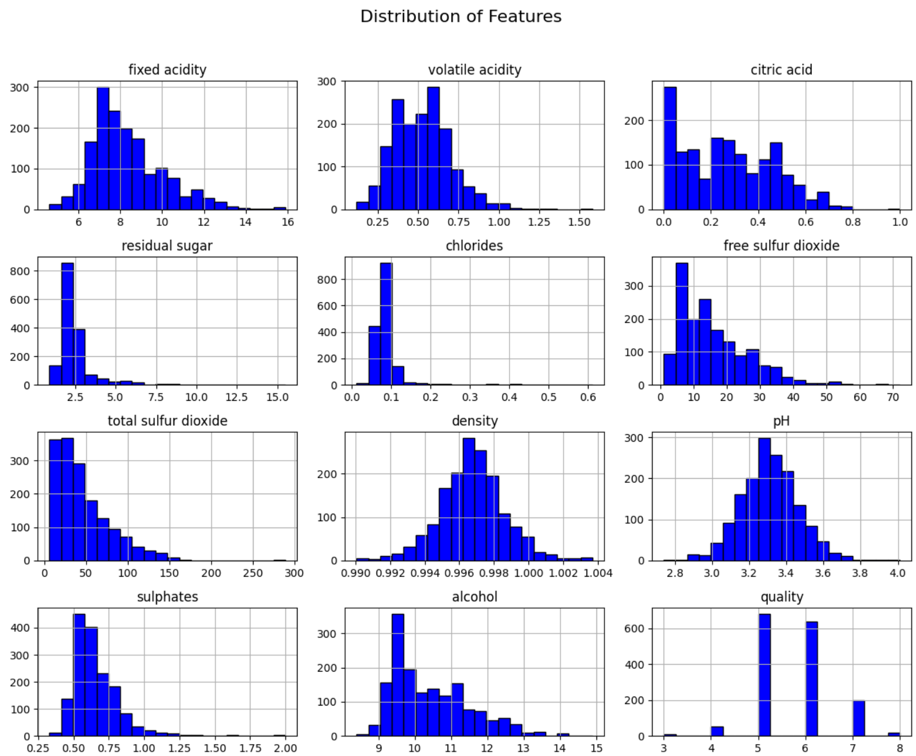
1. Obtaining Dataset

In this assignment, I used the **Wine Quality** datasets from **UCI Dataset**.

The table below contains all the attribute for the dataset:

Variable Name	Role	Type
fixed_acidity	Feature	Continuous
volatile_acidity	Feature	Continuous
citric_acid	Feature	Continuous
residual_sugar	Feature	Continuous
chlorides	Feature	Continuous
free_sulfur_dioxide	Feature	Continuous
total_sulfur_dioxide	Feature	Continuous
density	Feature	Continuous
pH	Feature	Continuous
sulphates	Feature	Continuous
alcohol	Feature	Continuous
quality	Target	Integer

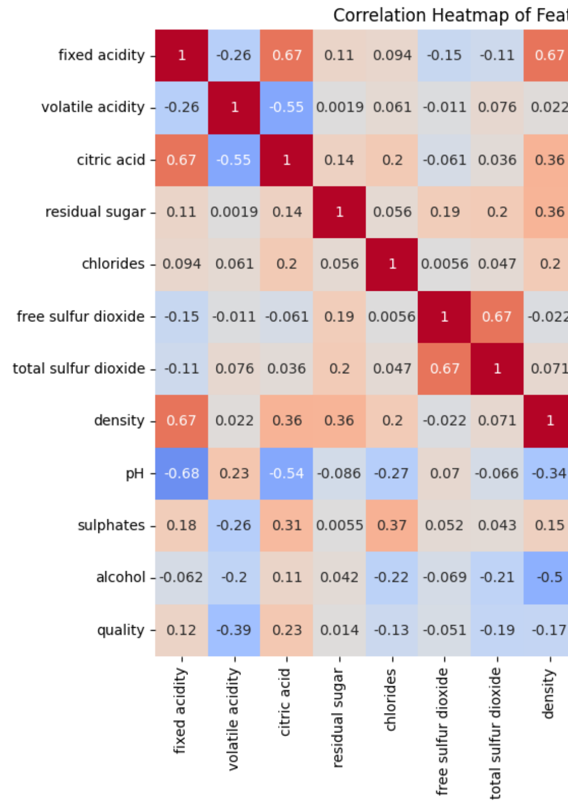
Below is the distribution of our dataset for all features.



We notice that only Density and pH are normally distributed. From the look of distribution, we need to collect more data for abnormal features, as it skews to the left.

2. Preprocessing Dataset

The data in the dataset do have **missing value**. However, there exist some outlier for each features. In this assignment, I decided to keep the outlier since it does not affects our model.



We have the correlation heatmap between all features as below:

From the correlation heatmap, we can extract the correlation between our target, **quality** and other features:

Feature	Correlation
fixed acidity	0.124052
volatile acidity	-0.390558
citric acid	0.226373
residual sugar	0.013732
chlorides	-0.128907
free sulfur dioxide	-0.050656
total sulfur dioxide	-0.185100
density	-0.174919
pH	-0.057731
sulphates	0.251397
alcohol	0.476166

We noticed that features such as **residual sugar**, **free sulfur dioxide** and **pH** do not play a huge role for our model so we can exclude the features from our model.

3. Regression Model

3.1. SGDRegressor

We use **Standard Scaler** to normalize our dataset in order to have better performance. We divide the datasets into **80/20** split for training and testing.

After testing the SGDRegressor with fixed paramter of **max iterations: 100_000**, **tolerance: 1e-3**, we have an **R2** score of around **0.35**.

I initially thought that the low score might be due to the outlier data in the dataset. I tried removing the outlier from the dataset. However, the **R2** score seems to be worse than the one without removed outliers.

After than I tried hyperparameter tuning using **Grid Search** from sklearn, with the parameter as below: - max-interations: [1000, 100_000, 1_000,000] - tolerance: [1e-3, 1e-4, 1e-5, 1e-6] - learning_rate; [constant, optimal, invscaling, adaptive]

Below is the table that shows the R2 for each set hyperparameters combination:

Alpha	Learning Rate	Max Iteration	Tolerance	R^2 Mean	R^2 Std
0.0001	constant	1000	0.001	0.2967	0.0128
0.0001	constant	1000	0.0001	0.3229	0.0070
0.0001	constant	1000	1e-05	0.2725	0.0435
0.0001	constant	2000	0.001	0.2892	0.0252
0.0001	constant	2000	0.0001	0.2965	0.0201
0.0001	constant	2000	1e-05	0.3083	0.0188
0.0001	constant	3000	0.001	0.2968	0.0282
0.0001	constant	3000	0.0001	0.3051	0.0229
0.0001	constant	3000	1e-05	0.2971	0.0233
0.0001	optimal	1000	0.001	-90173210627838425694208.0000	165394206094869238120448.0000
0.0001	optimal	1000	0.0001	-780824247723995307180032.0000	1441206628650898065195008.0000
0.0001	optimal	1000	1e-05	-538290081469203578888.0000	977464197749896806400.0000
0.0001	optimal	2000	0.001	-1097054547387706769408.0000	1730631422335407882240.0000
0.0001	optimal	2000	0.0001	-6258277578128301752320.0000	10526606428246999302144.0000
0.0001	optimal	2000	1e-05	-55299555802975064555520.0000	68127160860636130312192.0000
0.0001	optimal	3000	0.001	-809560119702566679347200.0000	1567210898727617293189120.0000
0.0001	optimal	3000	0.0001	-18523175183136791199744.0000	22309428505095648051200.0000
0.0001	optimal	3000	1e-05	-964813168567264695287808.0000	636723143316226231500800.0000
0.0001	invscaling	1000	0.001	0.3241	0.0108
0.0001	invscaling	1000	0.0001	0.3261	0.0109
0.0001	invscaling	1000	1e-05	0.3242	0.0117
0.0001	invscaling	2000	0.001	0.3243	0.0111
0.0001	invscaling	2000	0.0001	0.3245	0.0131
0.0001	invscaling	2000	1e-05	0.3232	0.0144
0.0001	invscaling	3000	0.001	0.3245	0.0104
0.0001	invscaling	3000	0.0001	0.3243	0.0136
0.0001					

invscaling | 3000 | 1e-05 | 0.3249 | 0.0123 | | 0.0001 | adaptive | 1000 | 0.001 | 0.3251 | 0.0113 | | 0.0001 | adaptive | 1000 | 0.0001 | 0.3246 | 0.0128 | | 0.0001 | adaptive | 1000 | 1e-05 | 0.3245 | 0.0130 | | 0.0001 | adaptive | 2000 | 0.001 | 0.3247 | 0.0128 | | 0.0001 | adaptive | 2000 | 0.0001 | 0.3249 | 0.0125 | | 0.0001 | adaptive | 2000 | 1e-05 | 0.3247 | 0.0124 | | 0.0001 | adaptive | 3000 | 0.001 | 0.3244 | 0.0129 | | 0.0001 | adaptive | 3000 | 0.0001 | 0.3247 | 0.0126 | | 0.0001 | adaptive | 3000 | 1e-05 | 0.3246 | 0.0127 | | 0.001 | constant | 1000 | 0.001 | 0.2767 | 0.0236 | | 0.001 | constant | 1000 | 0.0001 | 0.3082 | 0.0295 | | 0.001 | constant | 1000 | 1e-05 | 0.2783 | 0.0401 | | 0.001 | constant | 2000 | 0.001 | 0.3076 | 0.0212 | | 0.001 | constant | 2000 | 0.0001 | 0.3038 | 0.0331 | | 0.001 | constant | 2000 | 1e-05 | 0.3021 | 0.0281 | | 0.001 | constant | 3000 | 0.001 | 0.2619 | 0.0514 | | 0.001 | constant | 3000 | 0.0001 | 0.3002 | 0.0135 | | 0.001 | constant | 3000 | 1e-05 | 0.3135 | 0.0147 |

After performing the **Grid Search** on **SGDRegressor**, we have the following best hyperparamter: **{'learning_rate': 'optimal', 'max_iter': 3000, 'tol': 0.0001}** with an **R2 score** of 0.32

3.2. Ordinary Least Square(OLS)

I used the **OLS** model from **statsmodel** package. The OLS model gave us a similar **R2** score compared to the **SGDRegressor**.

After running the model, we obtained the following data: | Variable | Coefficient | Std Err | t | P>|t| | [0.025 | 0.975] | | :———————————- | :————— | :———— | :——- | :—- | :——- | :——— | | const | 4.2632 | 0.458 | 9.303 | 0.000 | 3.364 | 5.162 | | volatile_acidity | -1.0383 | 0.114 | -9.114 | 0.000 | -1.262 | -0.815 | | chlorides | -1.8379 | 0.432 | -4.256 | 0.000 | -2.685 | -0.991 | | total_sulfur_dioxide | -0.0023 | 0.001 | -4.046 | 0.000 | -0.003 | -0.001 | | pH | -0.4467 | 0.132 | -3.376 | 0.001 | -0.706 | -0.187 | | sulphates | 0.8565 | 0.120 | 7.143 | 0.000 | 0.621 | 1.092 | | alcohol | 0.2977 | 0.019 | 15.418 | 0.000 | 0.260 | 0.336 |

Metric	Value
Dep. Variable	quality
R-squared	0.348
Adj. R-squared	0.345
F-statistic	113.2
Prob (F-statistic)	1.56e-114
Log-Likelihood	-1276.1
No. Observations	1279
AIC	2566
Df Residuals	1272
BIC	2602
Df Model	6
Covariance Type	nonrobust
Omnibus	19.066

Metric	Value
Durbin-Watson	1.933
Prob(Omnibus)	0.000
Jarque-Bera (JB)	24.800
Skew	-0.188
Prob(JB)	4.12e-06
Kurtosis	3.570
Cond. No.	1.61e+03

From the result itself, we noticed the R2 value is 0.348, indicating that about 34.8% of the variability in the target variable (quality) is explained by the model. We have a high F-statistic indicating that all the features that we choose are significantly relevant to the model itself. As for the feature itself, all feature seems to have a coefficient that impacts the quality in some ways, as an improvement, we can remove the total_sulfur_dioxide as it have the lowest coefficient out of all features, indicating that it is less relevant to the model.

4. Conclusions

In my conclusion for the low **R2** score, I believe that **Linear Regression** model is not suitable for this problem because the data that we have is not close to linear, causing our regression results to deviate from the ground truth.