MAS369-220251110-assignment1

***Abstract*—In this project the effectiveness of linear and polynomial regression models are explored for predicting wine quality based on physicochemical properties. A linear regression model from scratch using gradient descent optimisation is implemented to complete the task and feature engineering is performed to enhance the model performance. The results indicate that while linear regression provides a foundational understanding, polynomial regression with appropriately chosen degrees significantly improves prediction accuracy.These findings highlight the importance of model selection and feature transformation in developing reliable predictive models for wine quality assessment.**

I. Introduction

Wine quality assessment traditionally relies on expert tasters, introducing subjectivity and scalability challenges in industrial production. Accurate prediction of wine quality using machine learning models can streamline quality control processes and enhance consistency. This paper argues that implementing polynomial regression models, which capture non-linear relationships in the data, provides a superior balance between accuracy and complexity for practical wine quality prediction.

The task begin with processing and exploring the wine dataset to understand underlying patterns. Then a linear regression model using gradient descent method from scratch is implemented to evaluate the performance. To recognising the limitations of linear models in capturing complex relationships, a polynomial regression with varying degrees is introduced into the project. Through systematic experimentation and analysis, this project demonstrate how polynomial regression improves prediction accuracy, which supports the argument for its suitability in wine quality prediction tasks.

II. Data Processing and Exploration

A. Data Description

The dataset used in this study is the UCI Red Wine Quality dataset, comprising 1,599 observations of red wine samples. Each sample includes 11 physicochemical properties, such as **fixed acidity**, **volatile acidity**, **citric acid**, **residual sugar**, **chlorides**, **free sulfur dioxide**, **total sulfur dioxide**, **density**, **pH**, **sulphates**, and **alcohol** (1).

The target variable is **quality**, rated on a scale from 0 to 10, representing the wine's overall quality as determined by expert tasters.

B. Data Cleaning and Preprocessing

Handling Missing and Anomalous Data:

Initial exploration showed no missing values in the dataset. However, we examined the possibility of anomalous zero values in features where zero may not be plausible. Specifically, we investigated the **citric acid** feature:

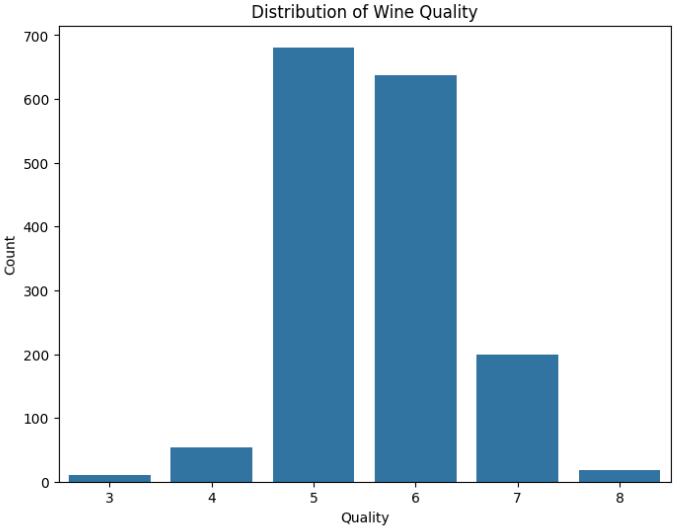


Fig. 1: Distribution of Wine Quality

```
df[’citric acid’].isnull().sum()
```

Upon analysis, we found that some wines legitimately have zero citric acid, as it is sometimes added to enhance freshness but is not always present. Therefore, we retained these zero values, acknowledging their validity. We also observed that there's a huge difference in the mean value for each variables, some of them are between 0 and 1 and some of them have the mean values around 10, and for the total sulfur dioxide it has the mean of 46.5. For each of the standards we try to see (mean, std, min, 25%, 50% , 75% and max), their value of total sulfur dioxide is much more bigger than the other variables.

C. Statistical Analysis and Visualization

Summary Statistics:

The summary statistics is computed to understand the distribution of each feature:

```
df.describe().T
```

Key Observations:

- **Alcohol** content has a mean of approximately 10.42%, ranging from 8.4% to 14.9%.
- **Quality** scores are concentrated between 3 and 8, with a mean of 5.64.

Visualizations:

The wine quality scores are concentrated around the middle of the scale. As shown in Figure 1, wines rated 5 and 6 comprise approximately 68% of the samples, with quality 5 accounting for 42.6% and quality 6 for 24.9%. There are significantly fewer wines at the extremes; high-quality wines (rated 7 and above) represent about 13%, while low-quality

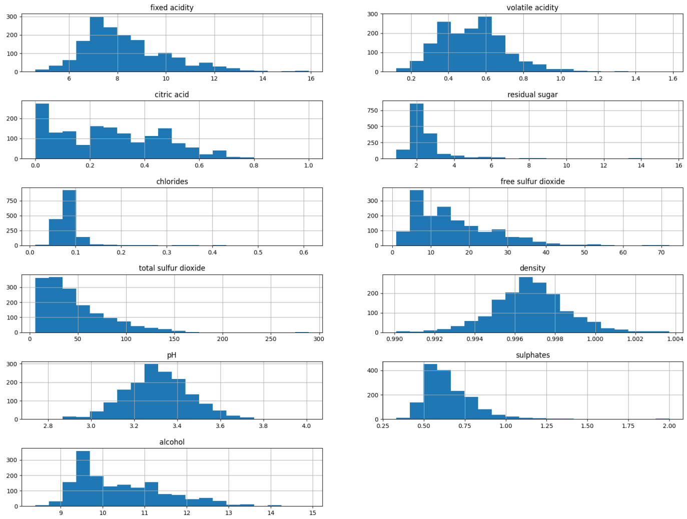


Fig. 2: Histograms of Features

wines (rated 4 and below) make up less than 9%. This central tendency indicates an imbalance toward average-quality wines, which may affect the model's ability to predict rare high or low-quality ratings due to limited training examples.

Fig. 2 reveals distinct distribution patterns for each variable. Notably, several features exhibit right-skewed distributions. For example, volatile acidity and chlorides show a concentration of lower values with a long tail towards higher values, indicating that most wines have low levels of these components, while a few samples have unusually high concentrations. This skewness suggests the presence of outliers and non-normal distributions, which may impact the modelling process. Features like residual sugar and free sulfur dioxide also display right-skewness, though less pronounced. In contrast, pH and density have more symmetrical distributions centred around their mean values, suggesting a normal distribution. Understanding these distribution characteristics is crucial for selecting appropriate preprocessing techniques, such as scaling or transformation, to improve model performance and accuracy.

Fig. 3 visually depicts the relationships between pairs of variables in the dataset, highlighting potential correlations and patterns. Notably, the plots reveal that alcohol has a positive relationship with wine quality; as the alcohol content increases, the quality rating tends to improve. Conversely, volatile acidity exhibits a negative relationship with quality, indicating that higher levels of volatile acidity are associated with lower quality wines. These observations suggest that alcohol content and volatile acidity are significant predictors of wine quality.

Fig. 4 quantitatively confirms these relationships by displaying the correlation coefficients between variables. It shows that alcohol has the highest positive correlation with quality (approximately 0.48), while volatile acidity has a substantial negative correlation (approximately -0.39). Additionally, sulphates and citric acid also have moderate positive correlations with quality. The heat map further reveals correlations among predictor variables themselves, such as the strong positive correlation between fixed acidity and citric acid, and the negative correlation between alcohol and density. These insights are crucial for feature selection and modelling, as they help identify the most influential variables and account for multicollinearity among predictors.

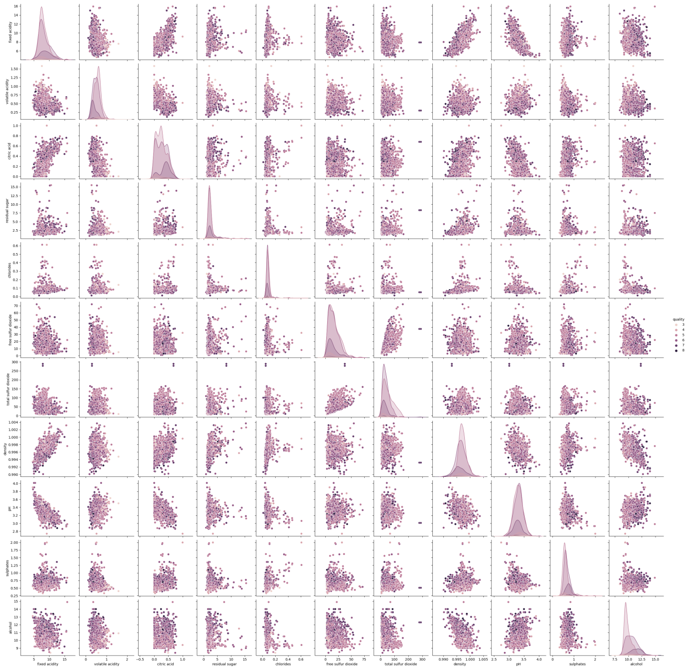


Fig. 3: Pair Plots of Features

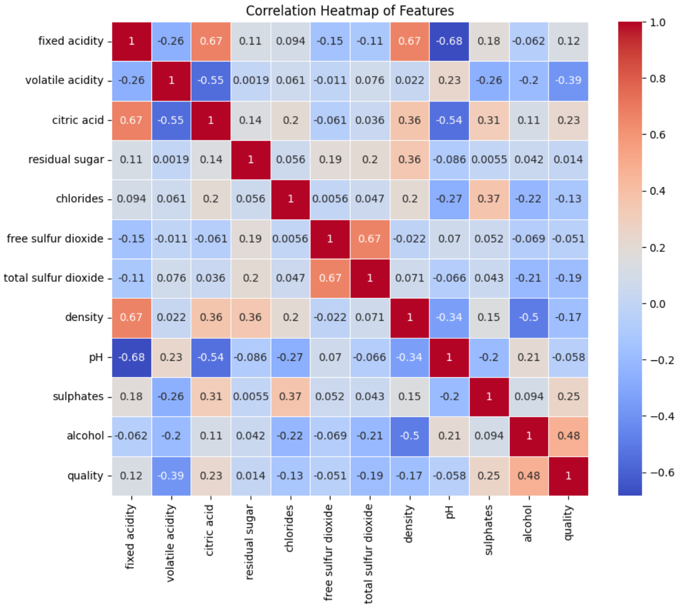
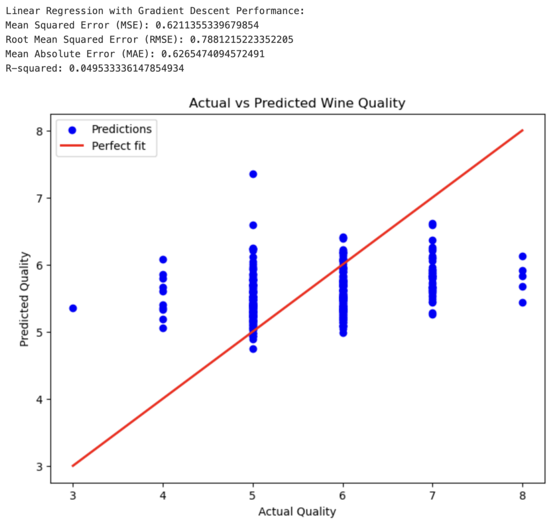


Fig. 4: Correlation Heatmap of Features

III. Linear Regression Implementation and Analysis

A. *Implementation of Linear Regression from Scratch*

We developed a custom linear regression model using gradient descent optimisation. The model minimises the mean squared error between predicted and actual quality scores by iteratively updating weights and bias.



```
Linear Regression with Gradient Descent Performance:
Mean Squared Error (MSE): 0.6211355339679854
Root Mean Squared Error (RMSE): 0.7881215223352205
Mean Absolute Error (MAE): 0.6265474094572491
R-squared: 0.049533336147854934
```

Fig. 5: Actual vs Predicted Wine Quality (Linear Regression)

Additionally, we used all 11 features as input variables, scaled using Min-Max scaling to ensure efficient convergence during optimisation.

B. Model Training and Evaluation

Training Process:

- **Data Split:** 80% training, 20% testing.
- **Scaling:** Features scaled to [0, 1] range.
- **Parameters:** Learning rate of 0.01, 1,000 iterations.

Evaluation Metrics:

- **Mean Squared Error (MSE):** 0.621
- **Root Mean Squared Error (RMSE):** 0.788
- **Mean Absolute Error (MAE):** 0.627
- **R-squared:** 0.050

The linear regression model trained using gradient descent resulted in an R-squared value of 0.050, indicating that the model explains only 5% of the variance in wine quality. The Mean Squared Error (MSE) was 0.621, with a Root Mean Squared Error (RMSE) of 0.788 and a Mean Absolute Error (MAE) of 0.627. These metrics suggest that the model's predictive performance is relatively poor. The low R-squared value signifies that the linear model fails to capture the underlying patterns between the physicochemical properties and wine quality effectively. This underperformance is visually evident in Fig. 5, where the scatter plot of actual versus predicted quality scores shows a wide dispersion of points and lacks a clear linear relationship. The predicted values do not align closely with the actual values, demonstrating the model's limited ability to generalise. These results imply that a simple linear model may be insufficient for predicting wine quality and that more complex models or feature transformations are necessary to improve accuracy.

IV. Feature Engineering

A. Correlation-Based Feature Selection

To enhance model performance, we performed feature selection based on correlation with the target variable:

```
corr_target = df.corr()['quality'].sort_values(ascending=False)
selected_features = corr_target[abs(corr_target) > 0.1].index
```

Selected Features:

- **Positive Correlation:**
  - *Alcohol* (0.476)
  - *Sulphates* (0.251)
  - *Citric acid* (0.226)
  - *Fixed acidity* (0.124)
- **Negative Correlation:**
  - *Volatile acidity* (-0.391)
  - *Total sulfur dioxide* (-0.185)
  - *Density* (-0.175)
  - *Chlorides* (-0.129)

B. Model Retraining and Evaluation

Using the selected features, we retrained the linear regression model.

Evaluation Metrics:

- **MSE:** 0.621 (unchanged)
- **RMSE:** 0.788
- **MAE:** 0.627
- **R-squared:** 0.050

Analysis:

Despite feature selection, the model's performance did not improve. This suggests that linear relationships captured by the model are insufficient to explain the variability in wine quality.

V. Polynomial Regression

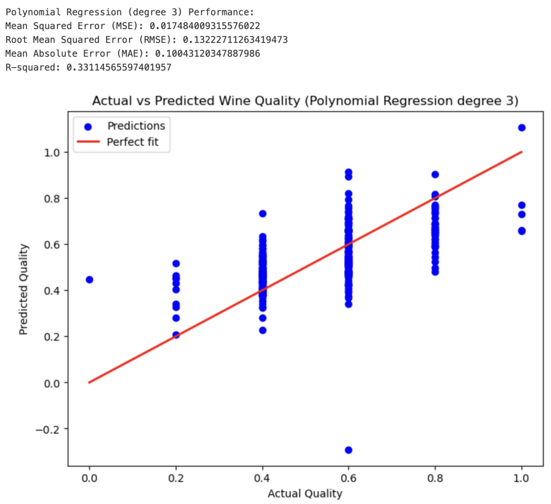
A. Implementation of Polynomial Regression

To capture non-linear relationships, we implemented polynomial regression with degrees 2 and 3, the codes of which are attached in the Jupyter notebook file.

B. Model Evaluation and Comparison

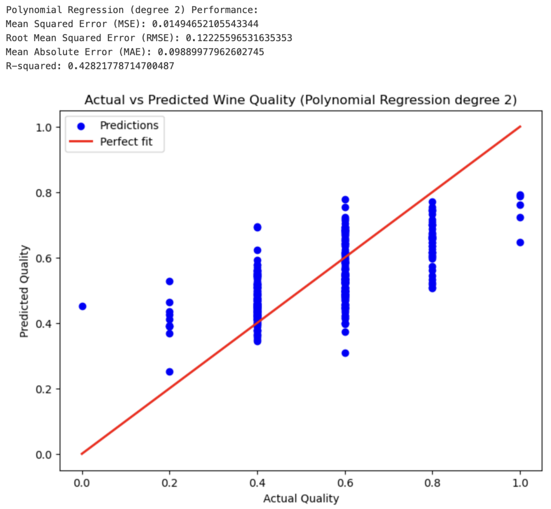
Degree 2 Polynomial Regression:

- **MSE:** 0.0149
- **RMSE:** 0.122
- **MAE:** 0.099
- **R-squared:** 0.428



```
Polynomial Regression (degree 3) Performance:
Mean Squared Error (MSE): 0.017484009315576022
Root Mean Squared Error (RMSE): 0.13222711263419473
Mean Absolute Error (MAE): 0.10043120347887986
R-squared: 0.33114565597401957
```

Fig. 7: Actual vs Predicted Wine Quality (Polynomial Regression Degree 3)



```
Polynomial Regression (degree 2) Performance:
Mean Squared Error (MSE): 0.01494652105543344
Root Mean Squared Error (RMSE): 0.12225596531635353
Mean Absolute Error (MAE): 0.09889977962602745
R-squared: 0.42821778714700487
```

Fig. 6: Actual vs Predicted Wine Quality (Polynomial Regression Degree 2)

Degree 3 Polynomial Regression:

- **MSE:** 0.0175
- **RMSE:** 0.132
- **MAE:** 0.100
- **R-squared:** 0.331

Analysis:

The degree 2 polynomial regression significantly improved the R-squared value to 0.428, indicating that the model now explains approximately 43% of the variance in wine quality. However, increasing the degree to 3 did not yield better results and slightly decreased performance.

VI. Discussions and Conclusions

The transition from linear to polynomial regression markedly improved model performance, affirming our argument that capturing non-linear relationships is crucial in wine quality prediction. The degree 2 polynomial model, in particular, strikes an optimal balance between complexity and accuracy. By introducing non-linear terms, the model better captures the intricate relationships between the physicochemical properties and wine quality, leading to a substantial increase in the R-squared value from 0.050 in the linear model to 0.428.

However, the use of polynomial regression introduces certain trade-offs. The increase in model complexity can lead to a higher risk of overfitting, especially with higher-degree polynomials. While the degree 2 model improved performance without significant overfitting, the degree 3 model showed diminished returns and a slight decrease in R-squared value, suggesting that additional complexity did not translate into better predictive power. This underscores the importance of careful degree selection to balance model fit and generalizability.

Computational efficiency is another consideration. The expansion of features through polynomial terms increases computational demands. In our case, the degree 2 polynomial features increased the number of input variables but remained manageable for the dataset size. However, higher degrees or larger datasets might require more computational resources, which could be a limitation in practical applications.

Interpretability is also affected when moving from linear to polynomial models. Linear regression offers straightforward interpretations of the relationship between each predictor and the target variable. In contrast, polynomial regression introduces interaction and squared terms, making it more challenging to discern the impact of individual features on wine quality. This reduced interpretability can be a drawback in contexts where understanding the influence of specific variables is important for decision-making.

Despite the improvements achieved with polynomial regression, the model explains less than half of the variance in wine quality, indicating that other factors not captured in the dataset may significantly influence quality assessments. Factors such as grape variety, vineyard practices, fermentation processes, and sensory attributes could contribute to wine quality but are not included in the physicochemical measurements. This limitation suggests that incorporating additional relevant features or exploring more advanced modeling techniques might further enhance predictive performance.

In summary, while polynomial regression significantly enhances the predictive capability over a simple linear model by capturing non-linear relationships, it is essential to consider the associated trade-offs in complexity, computational efficiency, and interpretability. Future work could focus on integrating more comprehensive data and employing regularization techniques or alternative machine learning algorithms to address these challenges and improve the robustness of wine quality predictions.

References

[1] “UCI Red Wine Dataset.” [Online]. Available: https://archive.ics.uci.edu/dataset/186/wine+quality