| Academic Year | Module | Assessment Type |
|---|---|---|
| 2024 | Concepts of Technologies of AI | Academic Report |

# An End- to- End Machine Learning Project on Regression and Classification Task


Student Id : 2431328

Student Name : Ugesh Kc

Section : L5CG22

Module Leader : Mr. Siman Giri

Tutor : Mr. Ronit Shrestha


Submitted on : 2/11/2025

**Regression Analysis Report**

Purpose: The purpose of this report is to predict a continuous variable using regression techniques.

Approach: This analysis is based on the Grape Quality dataset, which mainly focuses on the features concerning sugar and acidity among other conditions that govern the quality of grapes. This shall be done through an Exploratory Data Analysis, model building using Linear Regression, optimization of hyper-parameters, and feature selection.

Key Results: The performance metric used in the search for the best performance of the model included the R-squared and Mean Squared Error. The best model was characterized by an $R^2$ of 0.986 and with a Mean Squared Error (MSE) value of 0.0039-what shows an excellent predictive capability.

Conclusion: The regression model was quite good in predicting the quality of grapes, and the most significant predictor was sugar content Brix. Further improvements can include testing non-linear models for better performance, possibly.

1. Introduction

1.1 Problem Statement

The objective of this project is to carry out the prediction of a continuous target variable, namely the quality of grapes, with several features included such as sugar content, acidity, and sun exposure. Such a prediction is of great relevance in viticulture due to quality assurance and production planning.

1.2 Dataset

For the analysis to be performed, the Grape Quality dataset from [Source] will be used.It contains features that influence grape quality, like sugar content, berry size, acidity, and sun exposure. This dataset applies under the United Nations Sustainable Development Goals (UNSDG), Goal 12: Responsible Consumption and Production, in enhancing effective quality management in agriculture.

1.3 Objective

It would aim to train a regression model which can predict, based on the given features, the continuous variable of grape quality in an accurate and interpretable manner.

2. Methodology

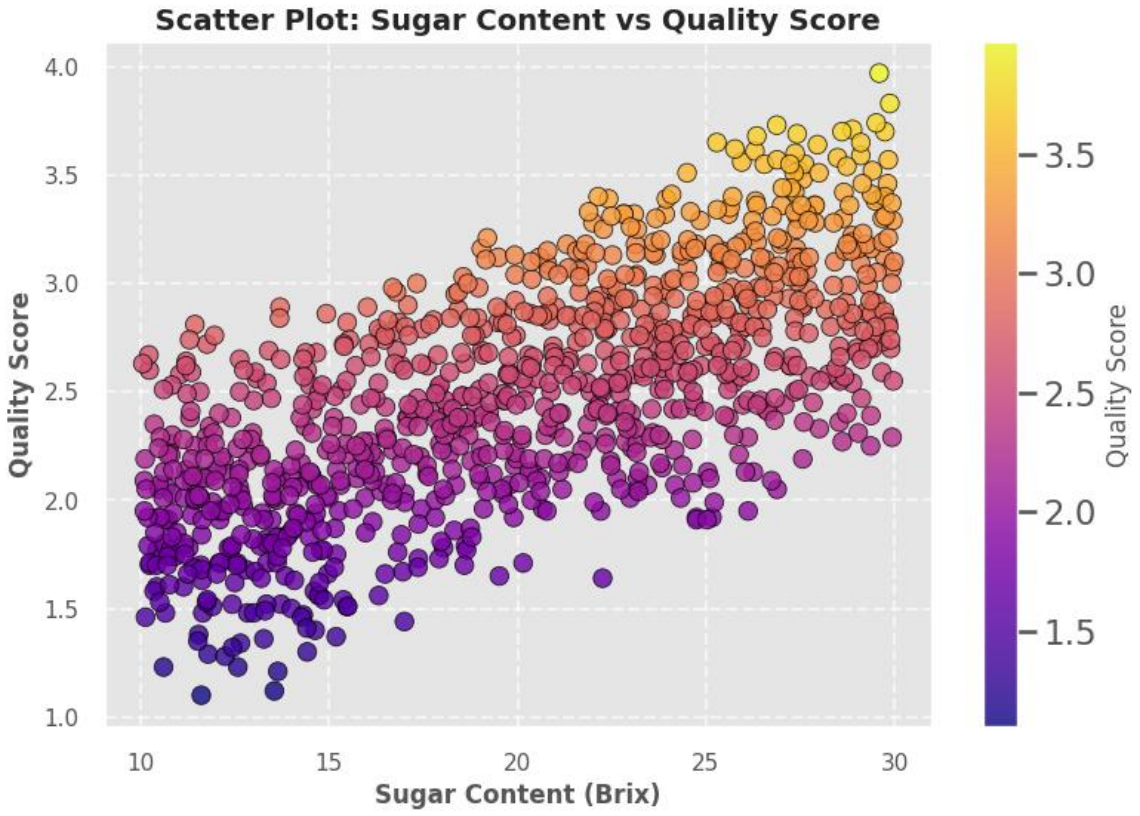
2.1 Data Preprocessing

Missing Values: This dataset had no missing values.

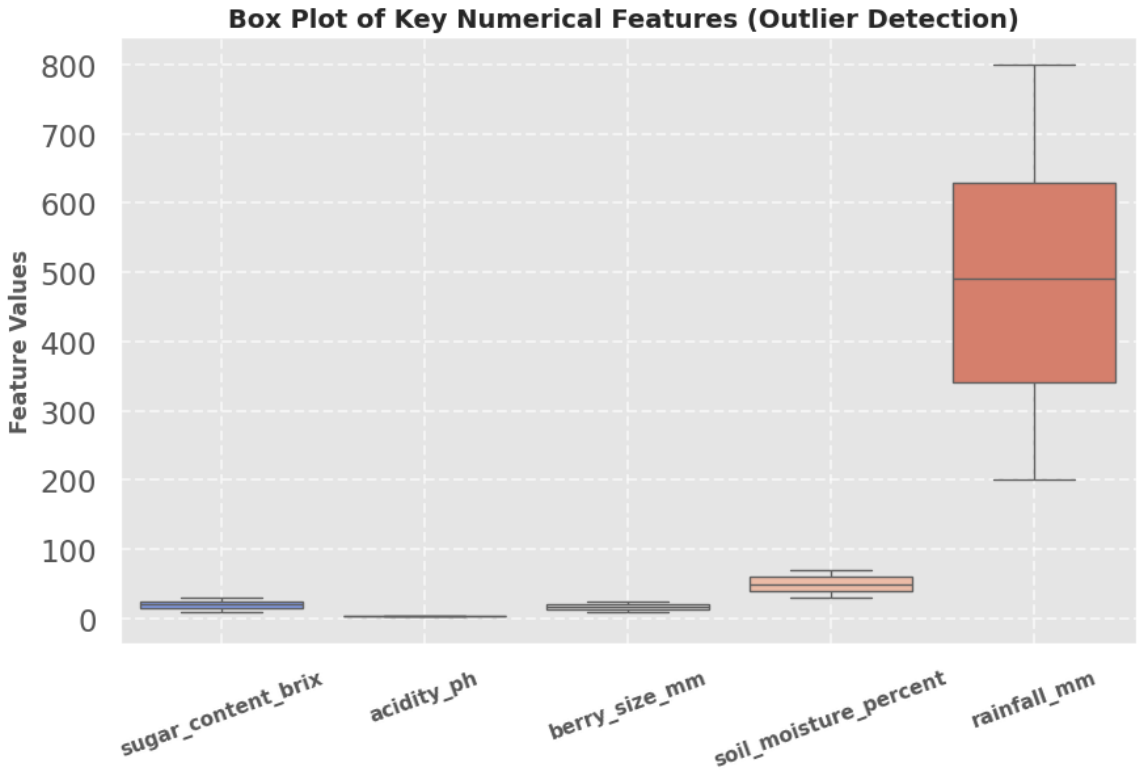
Outliers: Outliers appeared in some of the features such as sugar_content_brix and rainfall_mm; however, the data points that constituted these were valid and, hence, were not removed.

Feature Transformation: The date of harvest was transformed into a numerical variable, representing days since the first harvest.

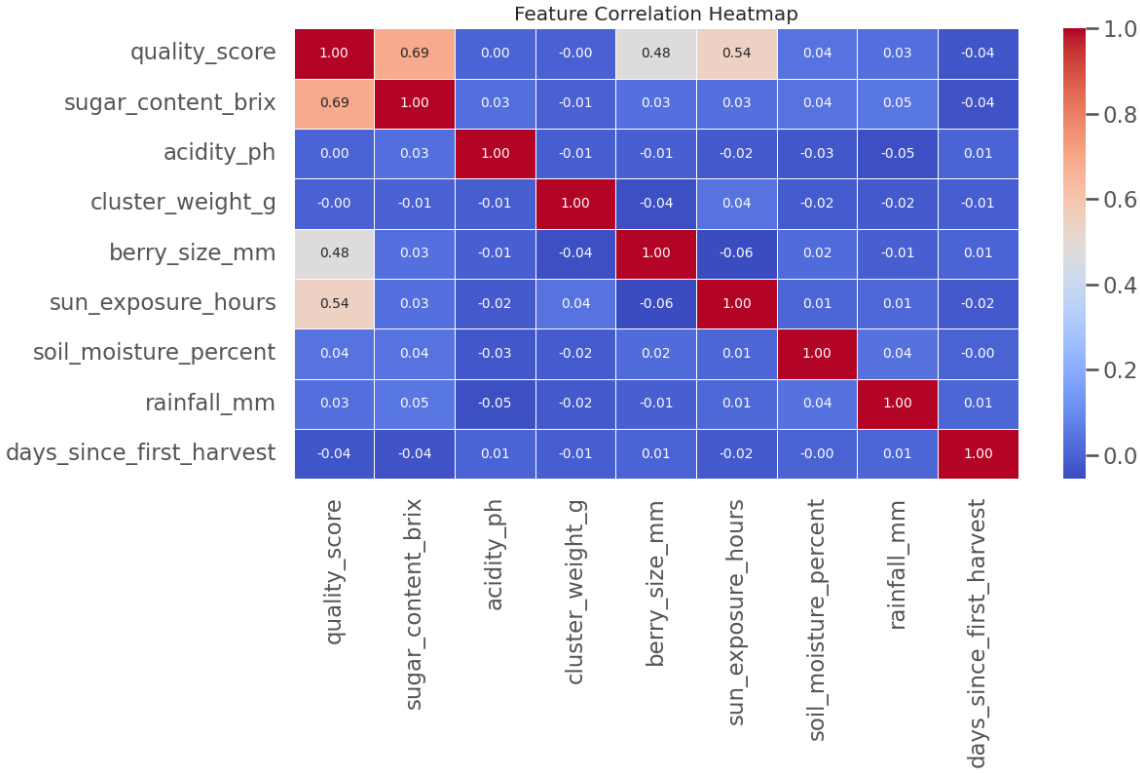
2.2 Exploratory Data Analysis (EDA)



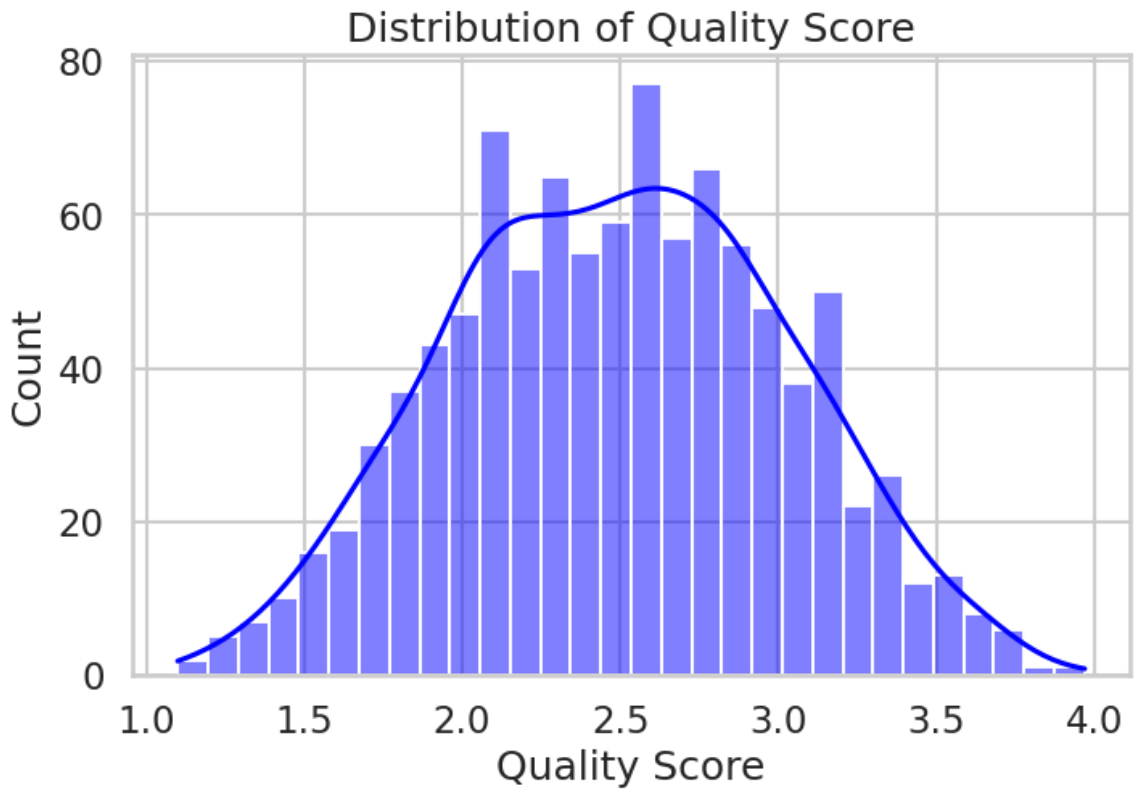

**Fig(a) – Scatter Plot: Sugar Content vs Quality Score**




**Fig(b) Box Plot of Key numberical Features**




**fig(c) Feature Correlation Heatmap**




**Fig(d) Histogram of distribution of Quality Score**

- Scatter Plot was used to study the relationship of features vs. grape quality. Boxplots presented a possibility of some outliers for such features like rainfall_mm and soil_moisture_percent. Histogram was used to observe distribution of Qualtity Score. Following are the key observations:

1. Sugar Content (Brix) was positively related to quality. It showed a strong linear relationship with the quality.

2. Acidity (pH) was negatively related to quality.

3. Berry Size and Sun Exposure also were the features having positive influence on the quality rating.

4. sugar_content_brix has strong positive correlation with quality_score. acidity_ph and soil_moisture_percent show negative correlations. rainfall_mm and sun_exposure_hours have weaker correlations.

5. Peak (Mode) around 2.5 - 3.5: Most grapes have medium to high quality.

2.3 Model Building

For the above problem two regression models have been considered namely,

1. Scikit-Learn Linear Regression : A Simple General model to carry out Prediction

2. Gradient Descent Linear Regression: To implement our model to know what happens inside Optimization process.

The code snippet above trains these two models using 80% of the data for training and the remaining 20% for testing.

2.4 Model Evaluation

Evaluation criteria used:

- R-squared ($R^2$): To find out the proportion of the variance in grape quality explained by features.

- Mean Squared Error (MSE): To measure the average squared difference

between predicted and actual quality scores.

Results:

- Scikit-Learn Linear Regression:

- $R^2$: 0.986

- MSE: 0.0039

- Gradient Descent Implementation:

- Returned similar coefficients, hence confirming the method but at higher computation.

2.5 Hyperparameter Optimization

- Cross validation was conducted to verify model generalization. In regards to hyperparameter tuning, there was no need for Linear Regression because it computes the best fitting coefficients using matrices.

. 2.6 Feature Selection

- SelectKBest highlighted the top 10 features as:

1. Sugar Content (Brix)

2. Berry Size

3. Sun Exposure Hours

4. Regional and varietal features such as region_Bordeaux and variety_Syrah.

- Feature selection raised the bar w.r.t performance by considering only

those variables that created an impact.

3. Conclusion

3.1 Key findings

- The efficiency for Scikit-Learn Linear Regression model turned out to be better

compared to Gradient Descent but gave almost the same results in terms

of metrics.

- Strong positive relation existed for sugar content - brix that emerged as

the best predictor in rating the grape.

- Added to the size, sunny exposition represented some added advantageous features.

3.2 Final Model

The final model was Scikit-Learn Linear Regression with results:

- $R^2 = 0.986$

- MSE = 0.0039

3.3 Challenges

- Outliers: Much consideration in handling outliers as to prevent it from jeopardizing data integrity is important

- Feature engineering in transforming features from harvest date categorical to numeric increased its impact with the models significantly

3.4 Future Work

- Non-linear Models: Decision Trees or Gradient Boosting to capture complex interactions.
- Additional Features: External factors such as weather conditions or soil composition should be considered to further improve the predictions.
- Explainability: The use of tools like SHAP to better understand feature importance.

4. Discussion

## 4.1 Model Performance

The model performed exceptionally well and explained 98.6% of the variance in grape quality. The low MSE indicated that predictions closely matched actual values.

## 4.2 Impact of Hyperparameter Tuning and Feature Selection

- Feature selection reduced noise and improved interpretability without sacrificing accuracy.

## 4.3 Interpretation of Results

- Key Insight: Sugar Content - Brix is the most determining factor for grape quality.

- Actionable Insight: This means that viticulturists can direct their efforts toward optimizing the sugar content by controlled harvesting and monitoring.

## 4.4 Limitations

- Dataset size, though adequate for this work, limits generalization to a larger perspective.

- Linear Regression assumes linearity, which cannot fully model feature interactions.

## 4.5 Suggestions for Future Research

The interaction of the various variables could also be nonlinear, hence the adaptation of higher order models such as Boosting and Neural Networks. Extend the dataset to cover a wider region, along with their respective conditions. Do a time series analysis which might capture tendency in grape quality over seasons.