A00268744: Kajal Singh

Data Science

Data Mining Analysis Report

Table of Contents

[1. Regression Analysis 2](#_Toc153537411)

[1.1 Business Understanding 2](#_Toc153537412)

[1.2 Data Understanding & Preparation 2](#_Toc153537413)

[1.2.1 Dataset Source 2](#_Toc153537414)

[1.2.2 Exploratory Data Analysis (EDA) 2](#_Toc153537415)

[1.2.3 Data Pre-processing 3](#_Toc153537416)

[1.3 Modelling 3](#_Toc153537417)

[1.3.1 Linear Regression 3](#_Toc153537418)

[1.3.2 Polynomial Regression 4](#_Toc153537419)

[1.4 Evaluation 4](#_Toc153537420)

[1.4.1 Linear Regression Results 4](#_Toc153537421)

[1.4.2 Polynomial Regression Results 5](#_Toc153537422)

[1.4.3 Visualizations 5](#_Toc153537423)

[2. Decision Tree Analysis 8](#_Toc153537424)

[2.1 Business Understanding 8](#_Toc153537425)

[2.2 Data Understanding & Preparation 8](#_Toc153537426)

[2.2.1 Dataset Download 8](#_Toc153537427)

[2.2.2 Exploratory Data Analysis (EDA) 8](#_Toc153537428)

[2.2.3 Data Preprocessing 8](#_Toc153537429)

[2.3 Modelling 9](#_Toc153537430)

[2.3.1 Decision Tree Classifier 9](#_Toc153537431)

[2.3.2 Decision Tree Visualization 9](#_Toc153537432)

[2.3.3 Visualizing Decision Boundaries 10](#_Toc153537433)

[2.4 Evaluation 10](#_Toc153537435)

[2.4.1 Decision Tree Results 10](#_Toc153537436)

[3. kNN Analysis 12](#_Toc153537437)

[3.1 Business Understanding 12](#_Toc153537438)

[3.2 Data Understanding & Preparation 12](#_Toc153537439)

[3.2.1 Dataset 12](#_Toc153537440)

[3.2.2 Data Preparation 12](#_Toc153537441)

[3.2.3 Additional Features 12](#_Toc153537442)

[3.3 Modelling 12](#_Toc153537443)

[3.3.1 kNN Classifier 12](#_Toc153537444)

[3.3.2 Random Forest Classifier 12](#_Toc153537445)

[3.4 Evaluation 12](#_Toc153537446)

[3.4.1 kNN Results 12](#_Toc153537447)

[3.4.2 Random Forest Analysis 14](#_Toc153537448)

[3.4.3 Added Evaluation Metrics 15](#_Toc153537449)

[3.4.3 Advanced Visualization - Decision Boundaries 16](#_Toc153537450)

[Conclusion 17](#_Toc153537451)

[References 17](#_Toc153537452)

# 1. Regression Analysis

## 1.1 Business Understanding

The primary objective is to understand and predict the quality of white wine based on various features. This analysis aims to provide insights into the factors influencing wine quality and develop predictive models to assist in quality assessment.

## 1.2 Data Understanding & Preparation

### 1.2.1 Dataset Source

The dataset used is the [White Wine Quality dataset](https://archive.ics.uci.edu/dataset/186/wine+quality.zip), (1) consisting of features such as acidity levels, alcohol content, and other chemical properties.

### 1.2.2 Exploratory Data Analysis (EDA)

Performed EDA to understand the distribution of features and the quality of white wines.

Visualizations, including a histogram of wine quality, were used to gain initial insights into the dataset. The visualization below shows the distribution of wine quality.

A diagram of a quality of wine

Description automatically generated

*Fig1: Distribution of White Wine Quality*

### 1.2.3 Data Pre-processing

Checked for missing values and No missing values were found in the dataset, ensuring its completeness.

- Explored summary statistics.

- Split the data into features and target variables.

- Split the data into training and testing sets.

## 1.3 Modelling

### 1.3.1 Linear Regression

Implemented a linear regression model and evaluated its performance on the test set.

The model was trained on the training set and tested on the unseen data to predict wine quality.

A screen shot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

### 1.3.2 Polynomial Regression

Implemented a polynomial regression model with a degree of 2 and evaluated its performance on the test set. The model was trained and tested similarly to the linear regression model.

A computer screen with text

Description automatically generated

## 1.4 Evaluation

### 1.4.1 Linear Regression Results

The linear regression model's performance was evaluated using Mean Squared Error (MSE) and R-squared metrics. The evaluation aimed to understand the model's accuracy and predictive capabilities.

- Mean Squared Error: 0.55

- R-squared: 0.27

A screenshot of a computer program

Description automatically generated

### 1.4.2 Polynomial Regression Results

The polynomial regression model's performance was assessed using the same metrics to compare its predictive power with the linear regression model.

- Mean Squared Error: 0.50

- R-squared: 0.32

A computer screen shot of a program

Description automatically generated

### 1.4.3 Visualizations

Visualizations were created to compare actual and predicted values for both linear and polynomial regression models. The focus was on the relationship between alcohol content and wine quality.

#### 1.4.3.1 Linear Regression 2D Visualization

A graph with red lines and blue dots

Description automatically generated

*Fig2: Linear Regression 2D Visualization*

#### 1.4.3.2 Polynomial Regression 2D Visualization

A graph with red and blue dots

Description automatically generated

*Fig3: Polynomial Regression 2D Visualization*

#### 1.4.3.3 Linear Regression 3D Visualization

A diagram of a wine quality

Description automatically generated with medium confidence

*Fig4: Linear Regression 3D Visualization*

#### 1.4.3.4 Polynomial Regression 3D Visualization (All Features)

A diagram of a graph

Description automatically generated with medium confidence

*Fig5: Polynomial Regression 3D Visualization*

# 2. Decision Tree Analysis

## 2.1 Business Understanding

The objective is to classify wines into different classes based on their chemical properties. This information aids in quality control and production management.

## 2.2 Data Understanding & Preparation

### 2.2.1 Dataset Download

I used the [Wine Classification dataset](https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data) (2)

### 2.2.2 Exploratory Data Analysis (EDA)

We explored the dataset, checked for missing values, and visualized the distribution of wine classes.

A bar chart with blue rectangular bars

Description automatically generated with medium confidence

*Fig6: Distribution of Wine Classes*

### 2.2.3 Data Preprocessing

- Identified and processed categorical and numerical features.

- Split the data into training and testing sets.

- Applied Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance.

A computer screen shot of a number of text

Description automatically generated with medium confidence

## 2.3 Modelling

### 2.3.1 Decision Tree Classifier

Implemented a Decision Tree classifier and visualized the resulting tree.

A black background with white text

Description automatically generated

### 2.3.2 Decision Tree Visualization

The Decision Tree visualization provides insights into feature importance and decision-making.

A diagram of different colored squares

Description automatically generated

*Fig7: Decision Tree Visualization*

### 2.3.3 Visualizing Decision Boundaries

Visualized decision boundaries using the **mlxtend** library for both the original and resampled datasets.

### A diagram of a bar graph Description automatically generated with medium confidence

*Fig8: Decision Boundaries Visualization*

## 2.4 Evaluation

### 2.4.1 Decision Tree Results

- Accuracy: 0.94

- Classification Report and Confusion Matrix

A computer screen shot of code

Description automatically generated

A screenshot of a computer screen

Description automatically generated

*Fig9: Decision Tree Accuracy and Classification Report*

# 3. kNN Analysis

## 3.1 Business Understanding

Similar to the Decision Tree analysis, the goal is to classify wines into different classes using the kNN algorithm.

## 3.2 Data Understanding & Preparation

### 3.2.1 Dataset

We used the same [Wine Classification dataset](https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data) (2)

### 3.2.2 Data Preparation

- Loaded and explored the dataset.

- Checked for missing values.

- Split the data into training and testing sets.

- Addressed class imbalance using Synthetic Minority Over-sampling Technique (SMOTE).

- Standardized features.

### 3.2.3 Additional Features

Added some additional features to enhance the kNN analysis:

- **Feature Scaling**: Standardized features to ensure they have the same scale.

- **Hyperparameter Tuning**: Used GridSearchCV to find the optimal values for k and other parameters.

- **Cross-Validation**: Employed cross-validation to assess the model's generalization performance.

## 3.3 Modelling

### 3.3.1 kNN Classifier

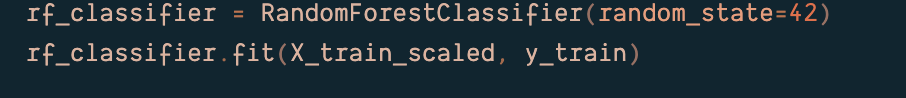
We implemented a kNN classifier with k=5.

A close up of a computer screen

Description automatically generated

### 3.3.2 Random Forest Classifier

Implemented a Random Forest classifier and trained it on the pre-processed data.



## 3.4 Evaluation

### 3.4.1 kNN Results

- Accuracy: 0.75

- Classification Report and Confusion Matrix

A screenshot of a computer screen

Description automatically generated

A blue squares with white text

Description automatically generated

*Fig10: kNN Accuracy, Classification Report and Confusion Matrix*

### 3.4.2 Random Forest Analysis

This section is the Random Forest analysis, including model performance metrics such as accuracy, classification report, and confusion matrix.

A screenshot of a computer screen

Description automatically generated

*Fig11: Random forest Accuracy and Classification Report*

A graph of different types of substances

Description automatically generated

*Fig12: A bar chart showing feature importance*

### 3.4.3 Added Evaluation Metrics

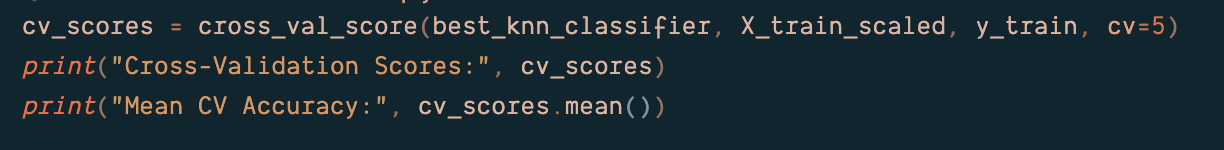
I’ve incorporated additional evaluation metrics to provide a comprehensive view of the kNN model's performance:

* **Hyperparameter Tuning Results:** Displayed the best hyperparameters obtained from the grid search.

A screen shot of a computer code

Description automatically generated

* **Cross-Validation Scores:** Examined the mean accuracy from cross-validation.



A screenshot of a graph

Description automatically generated

*Fig13: A bar chart showing cross-validation scores*

### 3.4.3 Advanced Visualization - Decision Boundaries

This section focuses on visualizing decision boundaries in a 2D space, providing insights into how the kNN classifier separates different classes based on the selected features. Adjust the feature indices as needed for your analysis.

A computer screen shot of a program code

Description automatically generated

A map of different colored areas

Description automatically generated

*Fig14: Decision Boundaries Visualization (KNN)*

# Conclusion

In this comprehensive data mining analysis, we delved into the prediction of white wine quality and the classification of wines based on chemical properties. Here are the key takeaways from each analysis:

**Regression Analysis**

The regression analysis aimed to predict wine quality using linear and polynomial regression models. The polynomial regression model outperformed the linear regression model, suggesting a nonlinear relationship between features and wine quality.

**Decision Tree Analysis**

The Decision Tree analysis successfully classified wines into different quality classes with high accuracy (94%). The visualizations of decision boundaries and the decision tree provided insights into feature importance and the decision-making process.

**kNN Analysis**

The kNN analysis focused on classifying wines using the kNN algorithm and a Random Forest classifier. Both models demonstrated reasonable accuracy, with the Random Forest adding an ensemble approach to improve classification. The visualizations of decision boundaries enhanced our understanding of how the models separate different wine classes.

This analysis provides valuable insights for winemakers and quality control teams. Polynomial regression can be a powerful tool for predicting wine quality, while Decision Trees and kNN algorithms are effective in classifying wines into quality classes. The ensemble approach, specifically the Random Forest classifier, enhances the classification performance.

# References

1. White Wine Quality Dataset. Retrieved from [UCI Machine Learning Repository](<https://archive.ics.uci.edu/dataset/186/wine+quality.zip>).(1)

2. Wine Classification Dataset. Retrieved from [UCI Machine Learning Repository]( https://archive.ics.uci.edu/dataset/109/wine).(2)