A blue and grey circle with a black background

Description automatically generated

CSE422

Project Report

On

“Wine Quality Prediction and Analysis Using Machine Learning”

By: Mahinur Rashid & Kazi Iftekharul Adnan

Table of contents

1. Introduction ---------------------------------------- page 3
2. Dataset description -------------------------------- page 3
3. Dataset pre-processing --------------------------- page 6
4. Feature scaling ------------------------------------ page 8
5. Dataset splitting ---------------------------------- page 9
6. Model training & testing ------------------------ page 11
7. Model selection/comparison analysis --------- page 14
8. Conclusion ---------------------------------------- page 16
9. **Introduction**

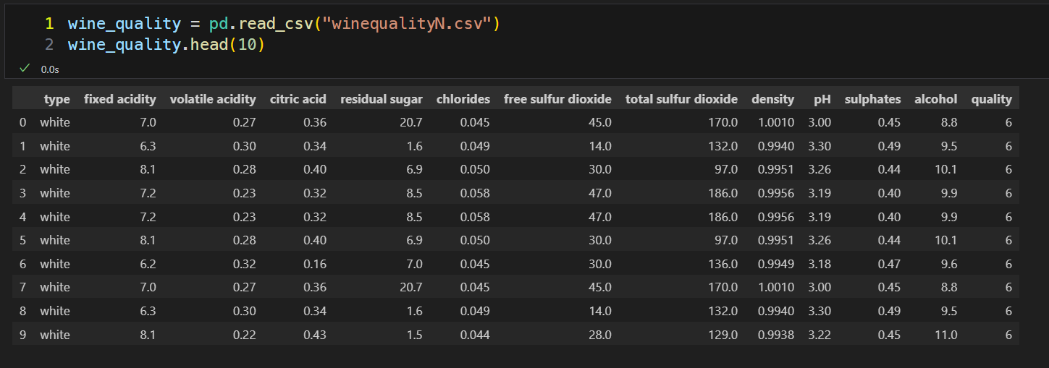
This project aims to analyze and predict the quality of wine using the "winequalityN.csv" dataset through machine learning techniques. The primary objective is to develop a predictive model that can accurately classify the quality of wine based on these attributes. The motivation behind this project stems from the significant impact that wine quality has on consumer satisfaction and the wine industry. By leveraging machine learning algorithms, this project seeks to provide a reliable and automated method for wine quality assessment, which can aid wine producers in maintaining high standards and improving their products. The project involves data preprocessing, feature engineering, model training, and evaluation to achieve optimal prediction accuracy.

1. **Dataset description**

Dataset source link: <https://www.kaggle.com/datasets/rajyellow46/wine-quality>

Reference: Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Wine Quality [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C56S3T>.

About dataset:

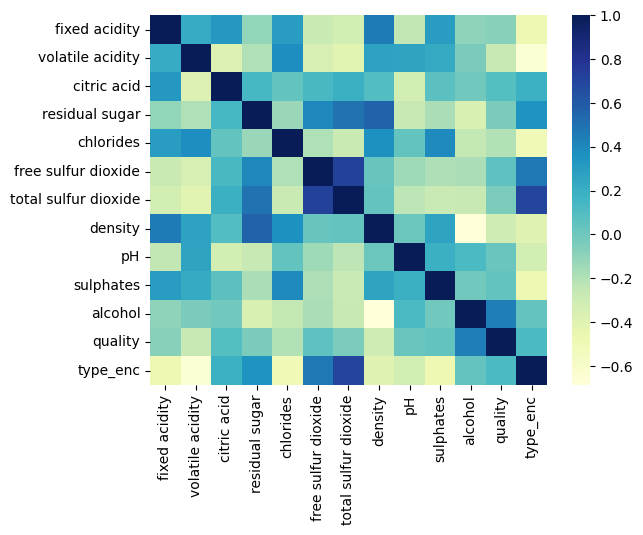


There are 12 features in the dataset namely - type, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol.

This is a classification problem. Because, the target variable “quality” contains 7 different values (3, 4, 5, 6, 7, 8, 9) and we want to classify to see how many wine samples belong to these seven classes. Then for the prediction also, we want to check for the test samples to be classified into these seven classes.

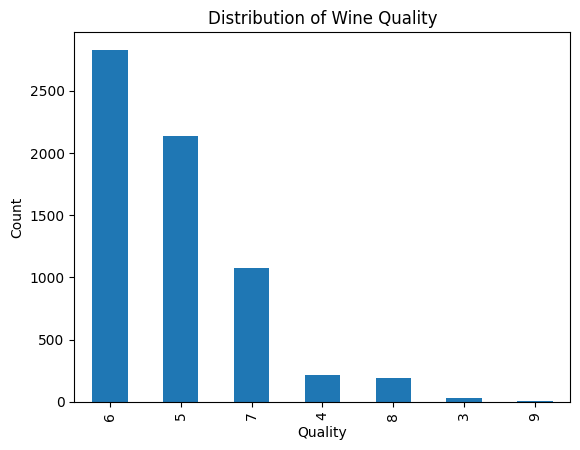
There are 6497 datapoints in the dataset. Except the values of the feature variable “type” (contains categorical value), all the other values in the dataset are quantitative value.

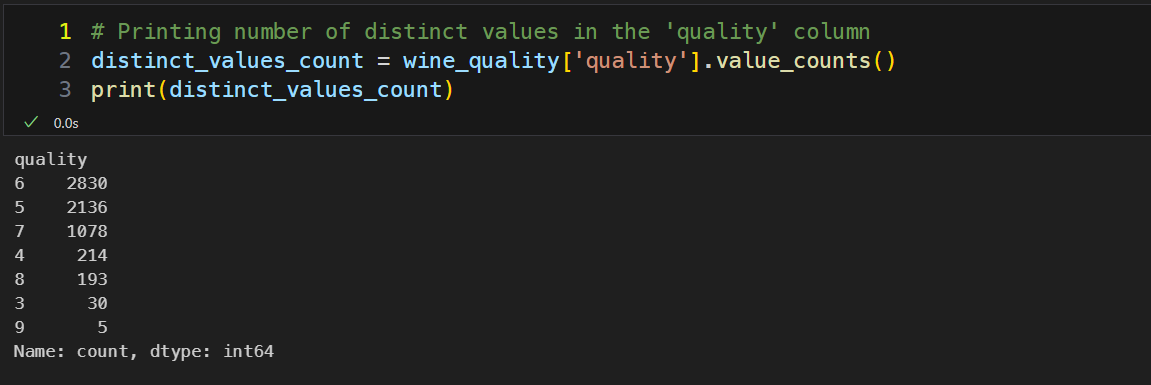
After finding the correlation of all the features by applying heatmap using the seaborn library, we can see that two input features have high correlation and they are “free sulfur dioxide” and “total sulfur dioxide”.



Imbalanced Dataset:

For the output feature “quality”, the unique classes do not have equal number of instances.

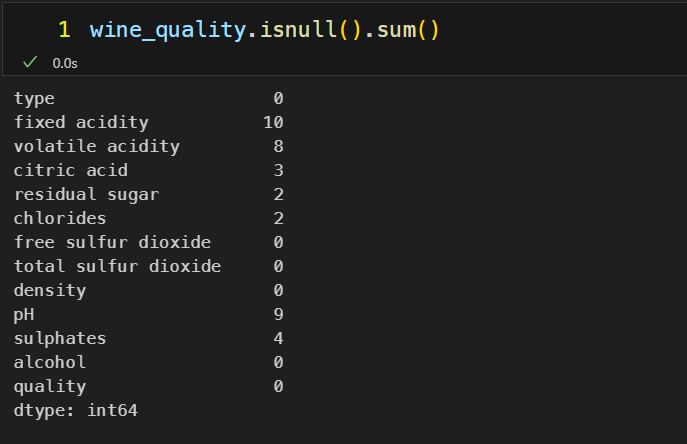




1. **Dataset pre-processing**

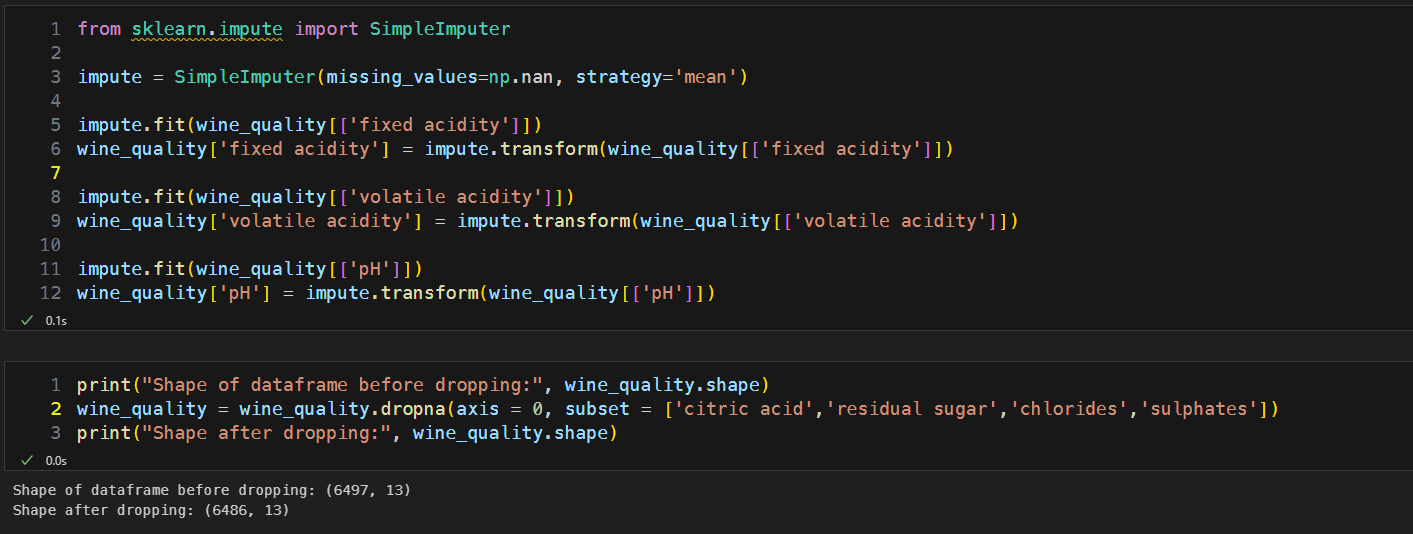
Faults:

NULL values:

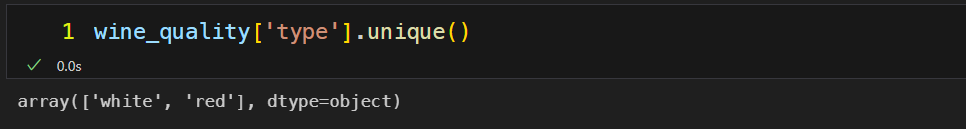


Solution for NULL values:

For NULL values, we replaced missing values with the mean to prevent errors during model training. And dropped rows with missing values in critical columns to maintain data quality.



Categorical values:

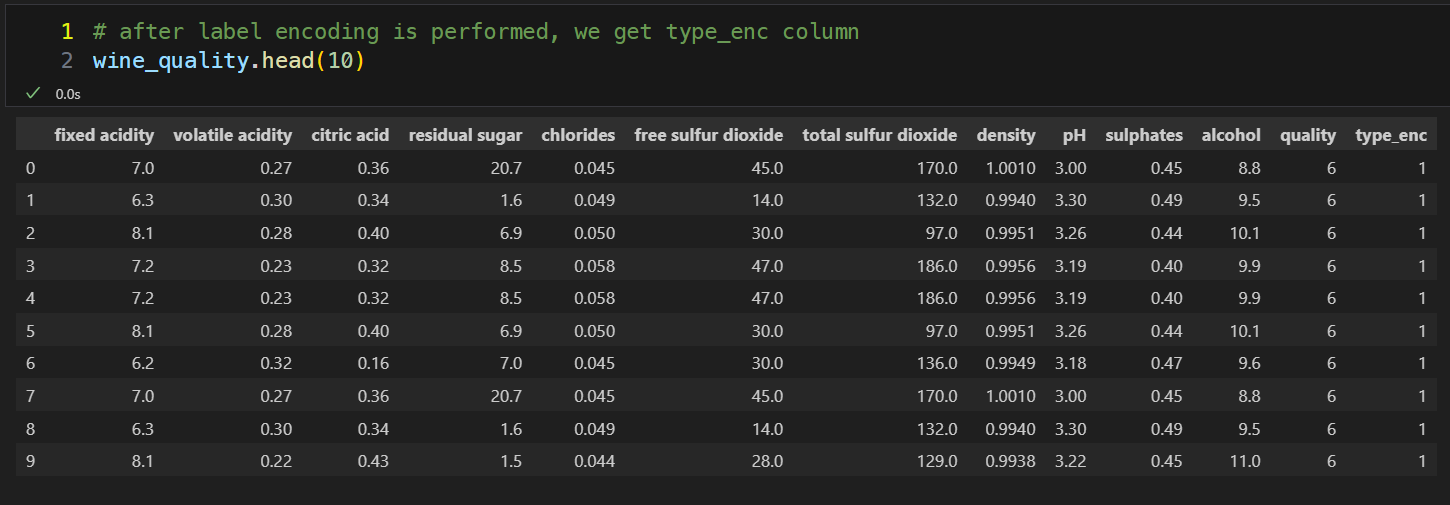


Solution for Categorical values:

For categorical values, we converted the type column to numeric using Label Encoding, which is essential for machine learning models to process non-numeric data. Moreover, we did these for ensuring the dataset is suitable for machine learning models. This improved model performance and accuracy by ensuring the data was properly prepared.



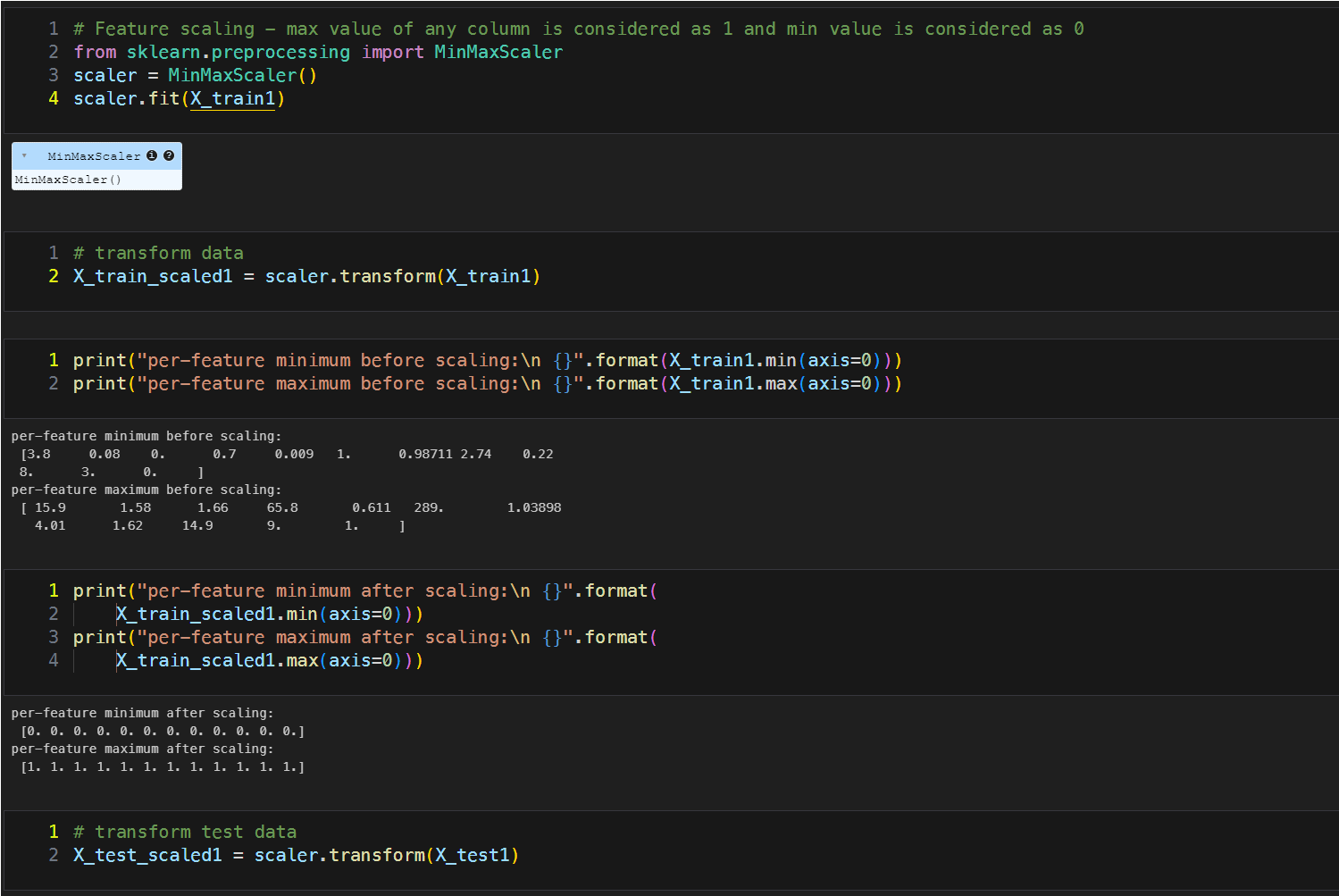




1. **Feature scaling**

Except the feature variables “free sulfur dioxide” and “total sulfur dioxide”, all the other quantitative feature variables contain floating point values. So, there is an absence of cohesion between all the feature variables. For this reason, we require feature scaling.

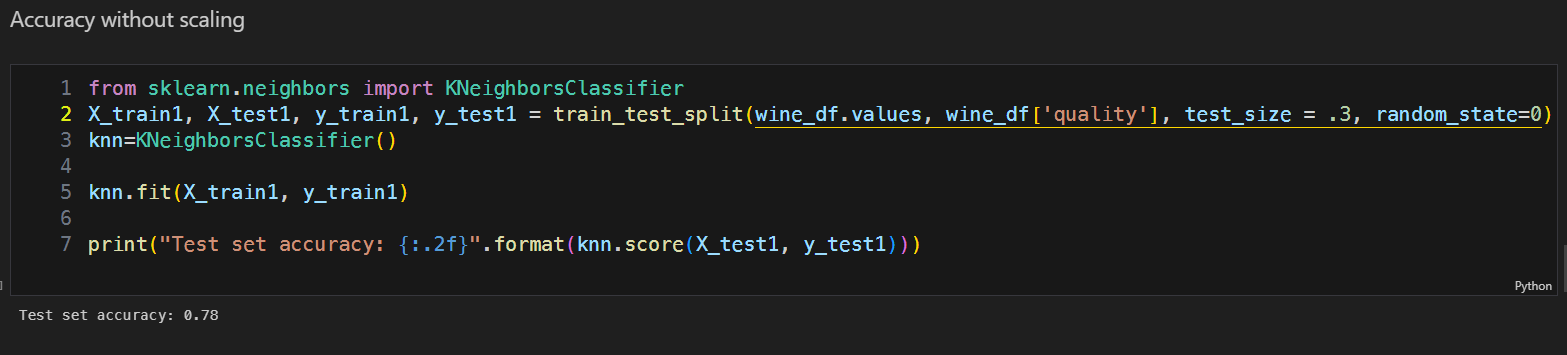
For our dataset, we have applied scaling methods such as Min Max Scalar and Standard Scalar.



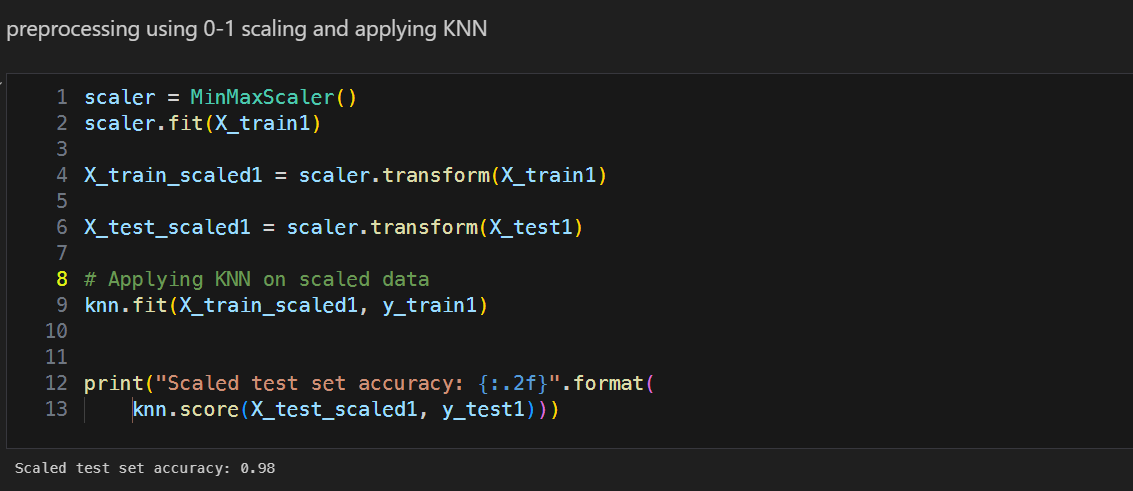
1. **Dataset splitting**

Stratified dataset splitting ensures that the training and testing sets have the same proportion of each target class as the original dataset. For our dataset, we saw some classes under the target variable “quality” are very less in number, which may remain underrepresented if we split our dataset randomly. As a result, accuracy result may drop significantly as well. But in such cases, Stratified splitting helps maintain the class distribution, leading to more reliable model evaluation. That’s why, we have used this Stratified splitting method to split our entire dataset into 70% for Train set and rest of the 30% for the Test set.

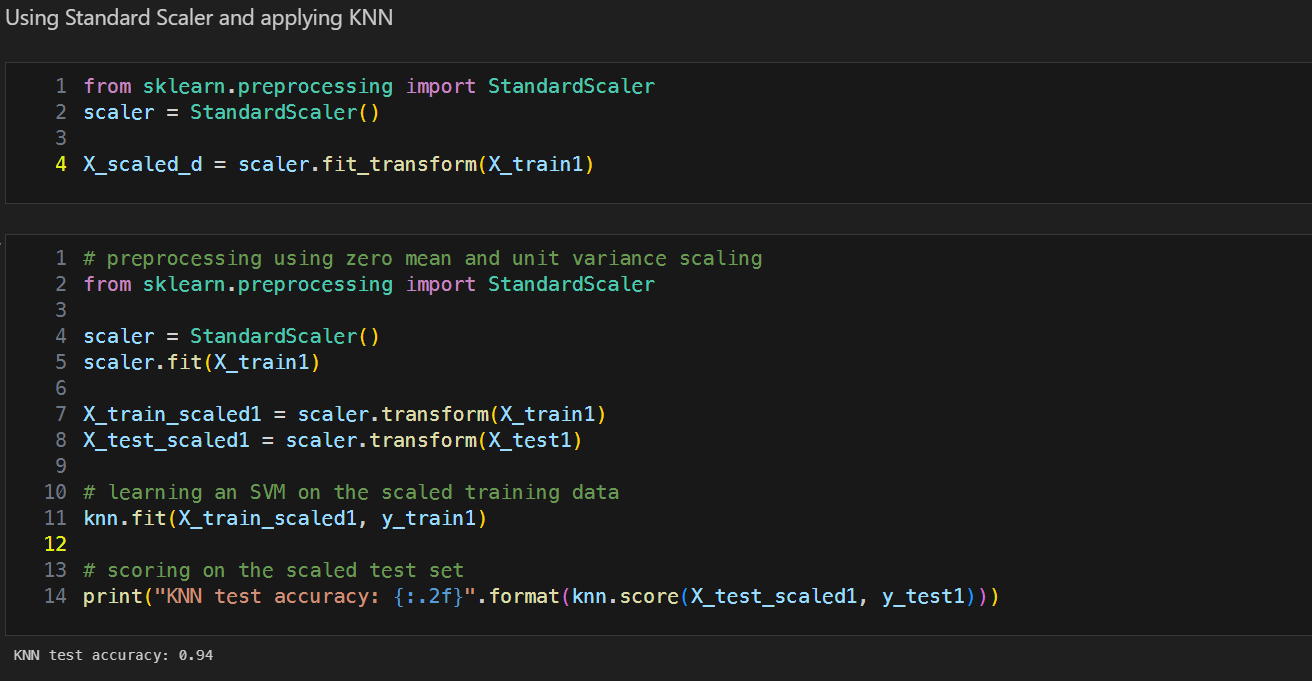
So, before scaling, accuracy score for the Test set is = 0.78.



But, after using Min Max Scaler and applying through KNN, the accuracy of the Scaled Test set is = 0.98

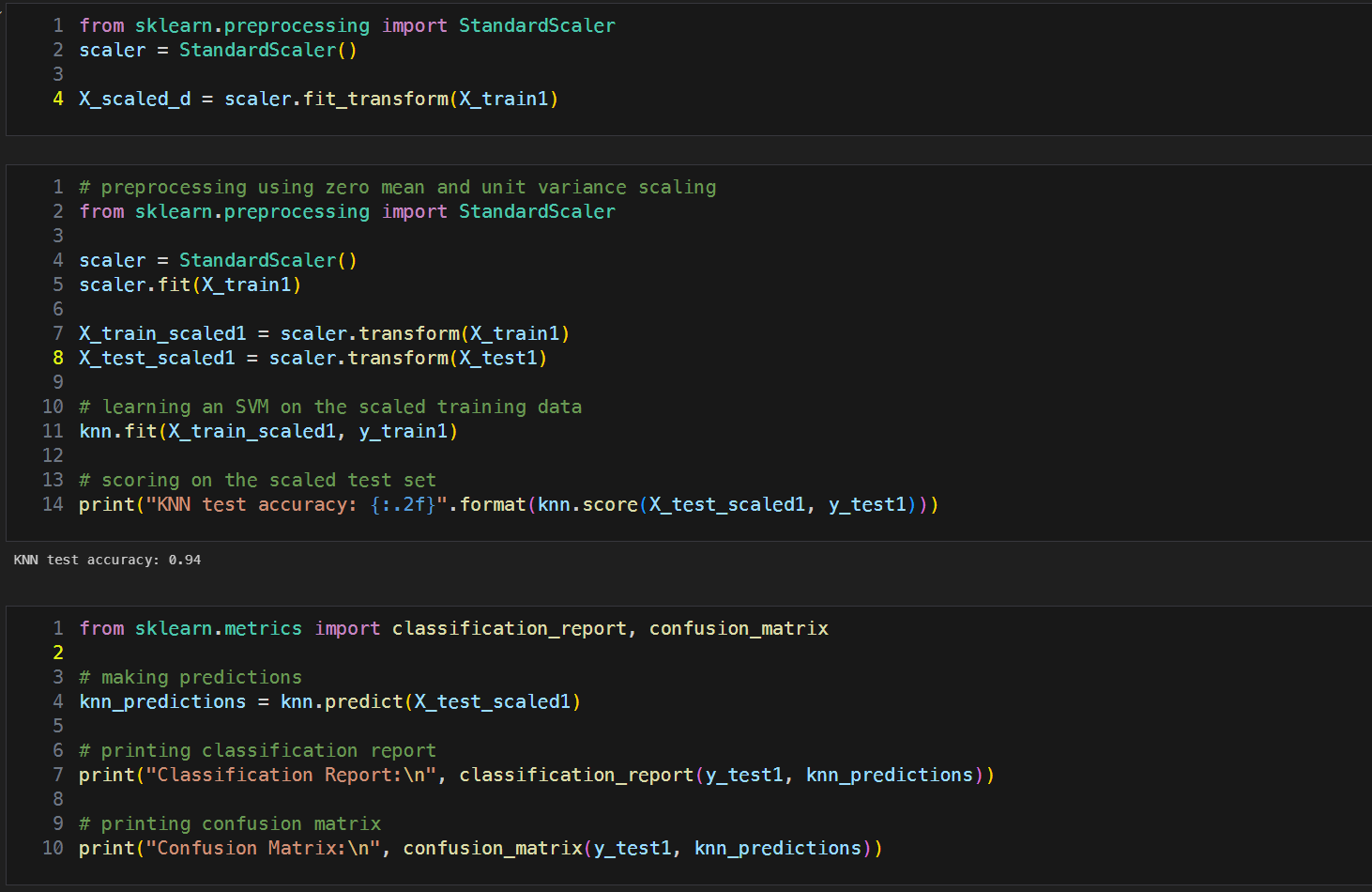


Also, after using Standard Scaler and applying through KNN, the accuracy of the Scaled Test set is = 0.94

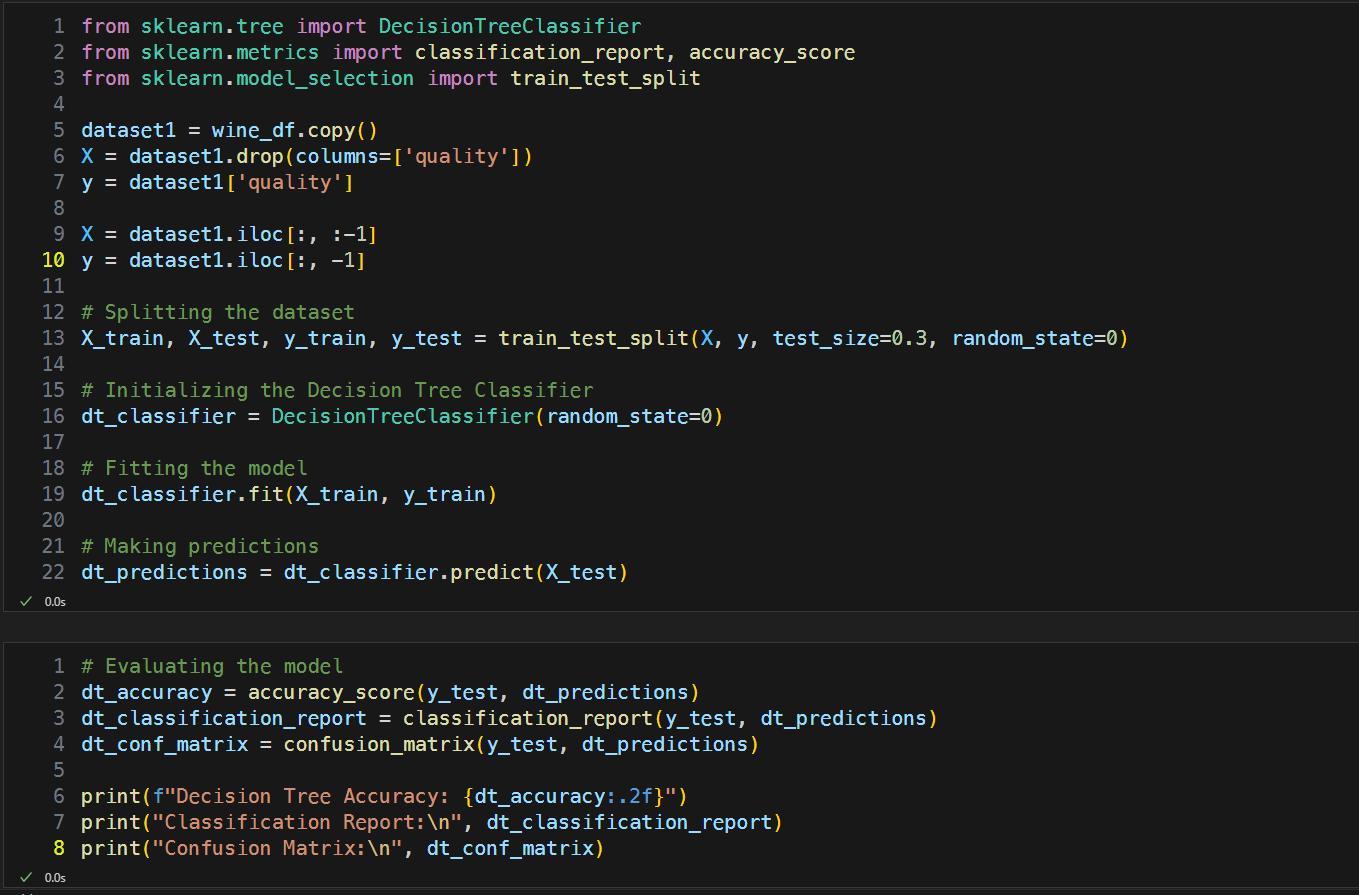


1. **Model training and testing**

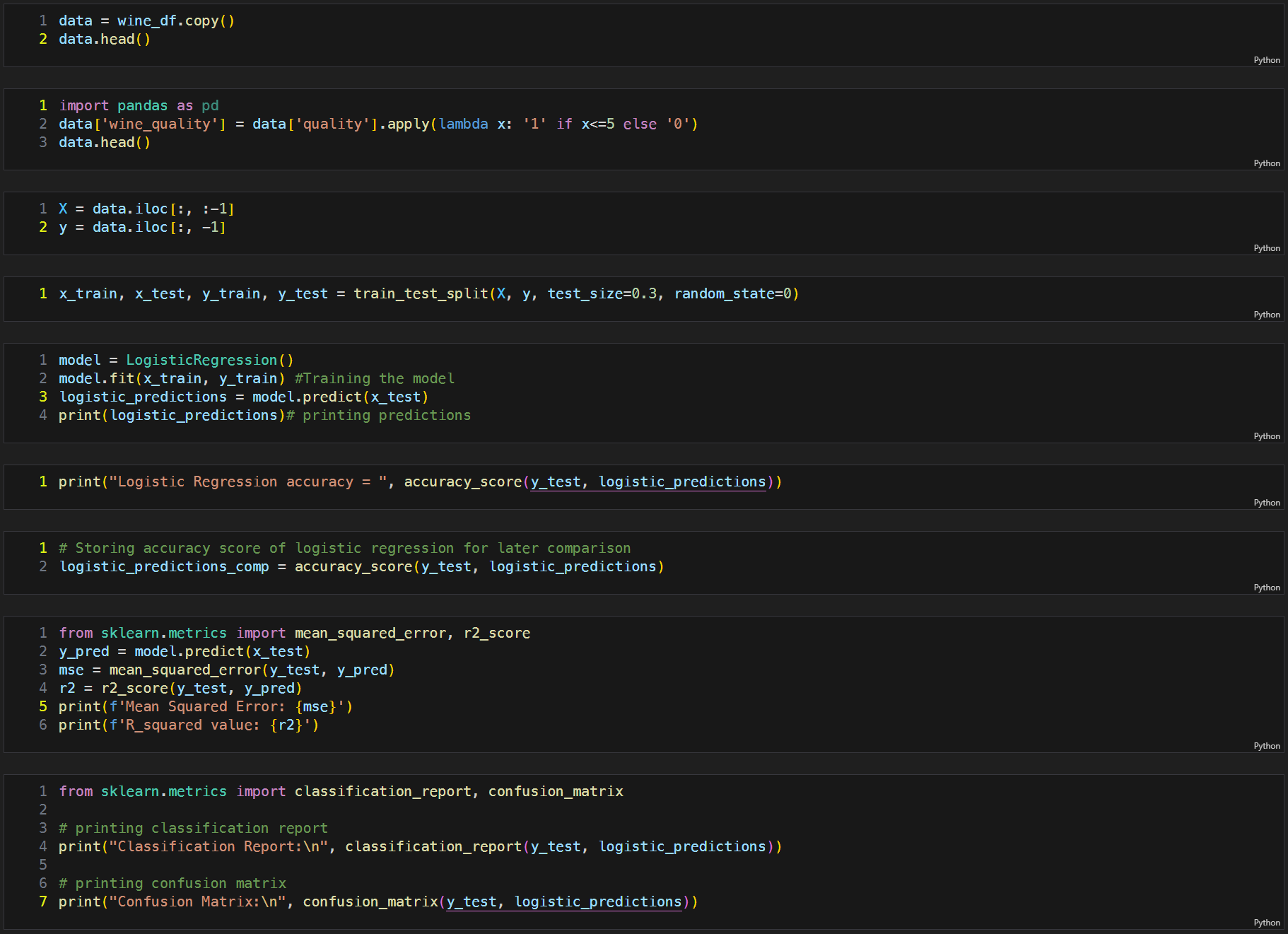
KNN



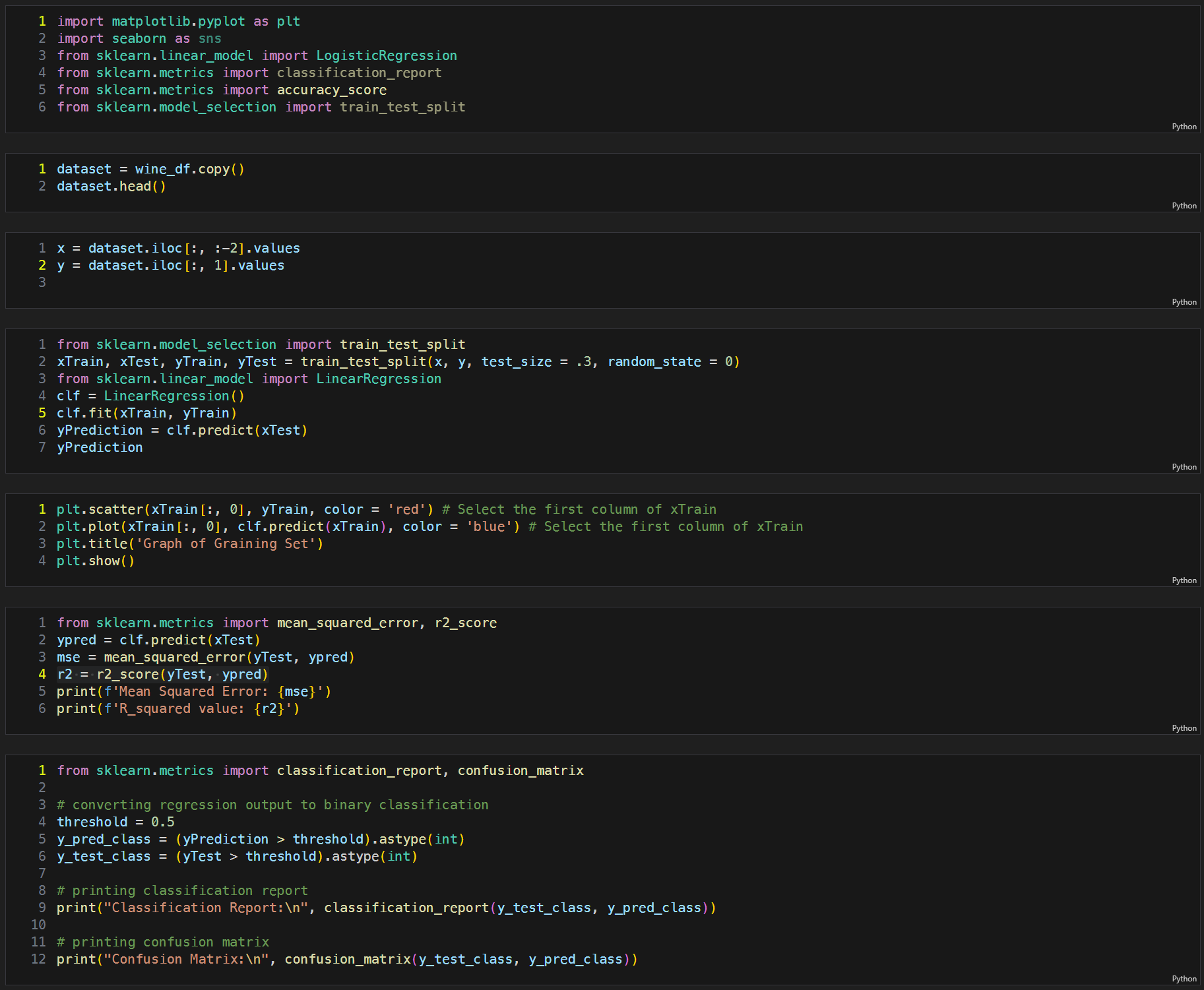
Decision Tree



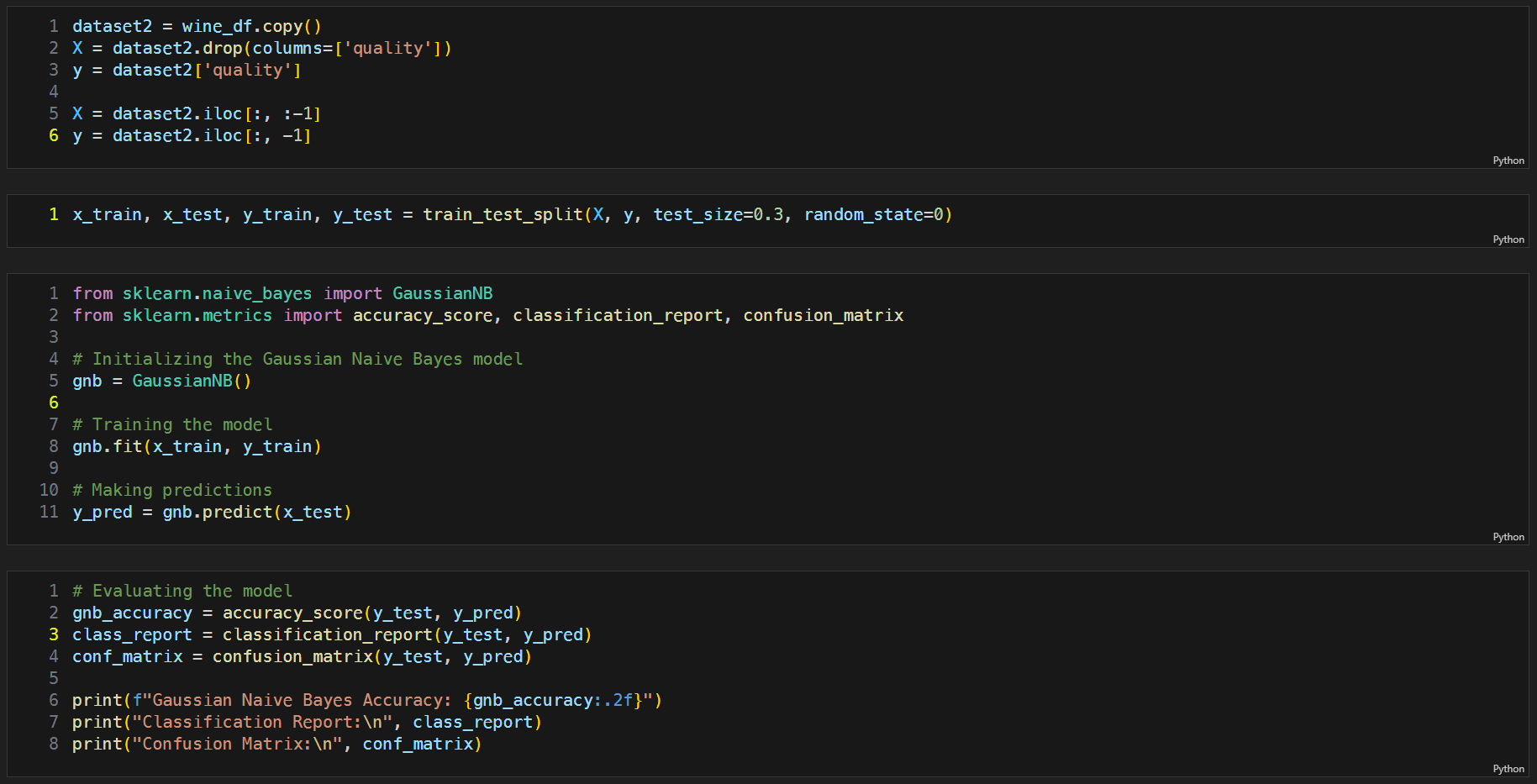
Logistic Regression



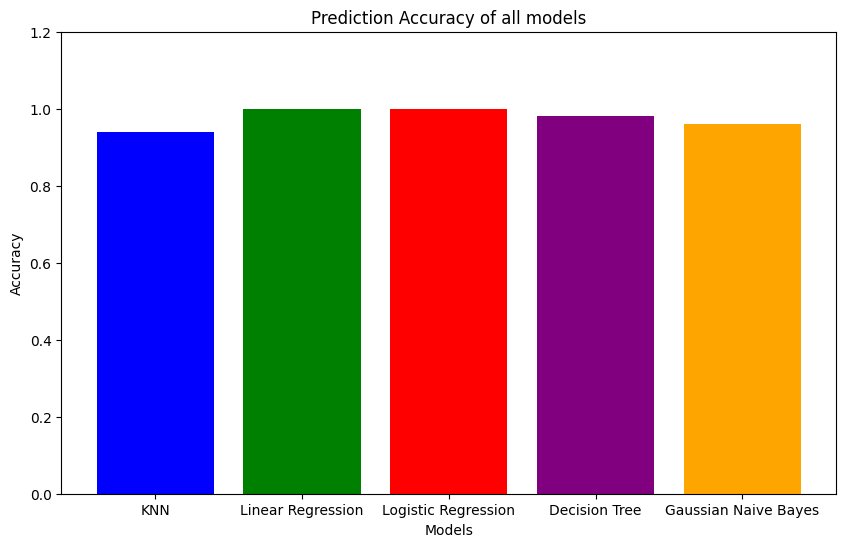
Linear Regression



Naive Bayes

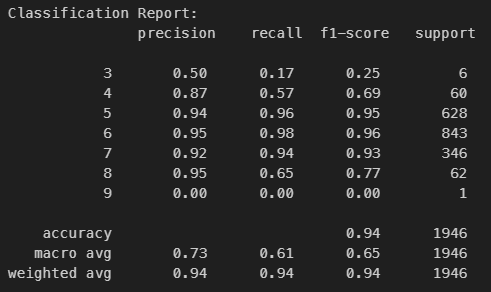


1. **Model selection/Comparison analysis**

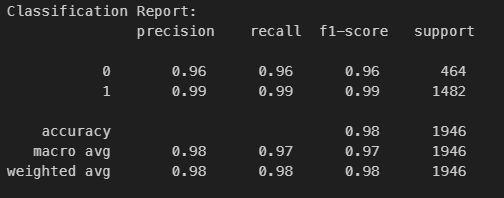


Precision, recall comparison

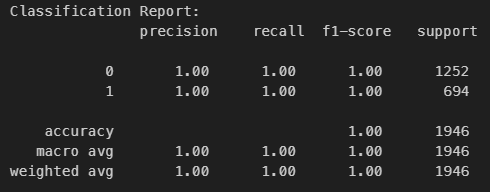
KNN



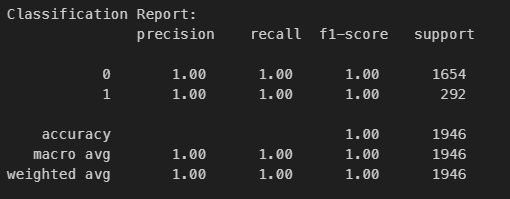
Decision Tree



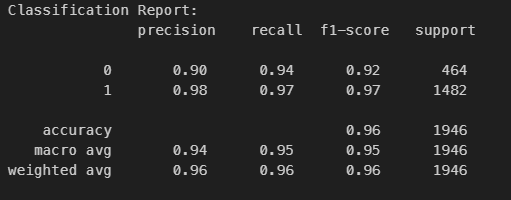
Logistic Regression



Linear Regression



Naive Bayes



1. **Conclusion**

After building and evaluating multiple models (KNN, Decision Tree, Logistic Regression, Linear Regression and Naive Bayes) for wine quality prediction, both **Linear Regression** and **Logistic Regression model** have showed the highest accuracy (100%). The classification reports highlight the models' precision, recall, and F1-scores, indicating their performance in handling both positive and negative predictions. Overall, the project successfully handled missing values and categorical data, and the models demonstrated reliable predictive performance for wine quality classification.