BSIT 41043 - Data Mining Data Warehousing

Assignment 03

BSIT 211007 – Sameera Pramodi Jayakody

BCI Campus Negombo

Abstract

This report brings the analysis and modeling of the quality of wines based on physicochemical properties using the WEKA software. The dataset provided contains samples of red and white Vinho Verde wines from northern Portugal. The main goal is the prediction of wine quality from the given 11 real-valued features. The tasks to be completed for this assignment include handling missing data, comparison of clustering methods—K-means and DBSCAN—and experimentation with data discretization techniques in order to improve clustering accuracy. Analysis of the different parameters to be used in clustering and application of various discretization techniques will be done in order to compare the effect of those approaches in the quality of clustering and accuracy. Final report elaborates discussions on results, methods, and comparison among diverse techniques.

Dataset Overview

The dataset contains 4898 instances of two types of Vinho Verde wines: red and white. It has 11 features, all real-valued, describing various physicochemical properties of the wines.

These features are, Fixed acidity, Volatile acidity, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Density, pH, Sulphates, Alcohol

The target variable is wine quality, which is represented by an integer value ranging from 0 to 10. The dataset is used to explore how these physicochemical attributes influence the quality of wine and to build models to predict wine quality based on these features.

Table of Contents

1. **Handling Missing Values in Two Datasets**

Apply the *ReplaceMissingValues* filter to the dataset. This will replace Numerical missing values with the mean of the attribute. Categorical missing values with the mode.

ReplaceMissingValues appropriateness for numerical data. The features in the wine datasets (e.g., pH, alcohol, density) are numerical. Replacing missing values with the mean, avoids bias and maintains consistency in the dataset.

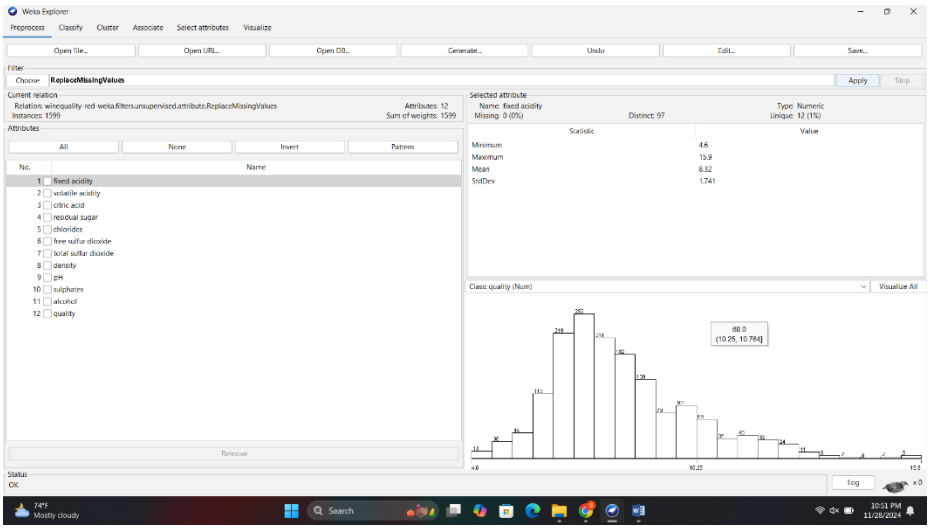

Figure 1: Apply ReplaceMissingValues filter to Winequality – Red dataset

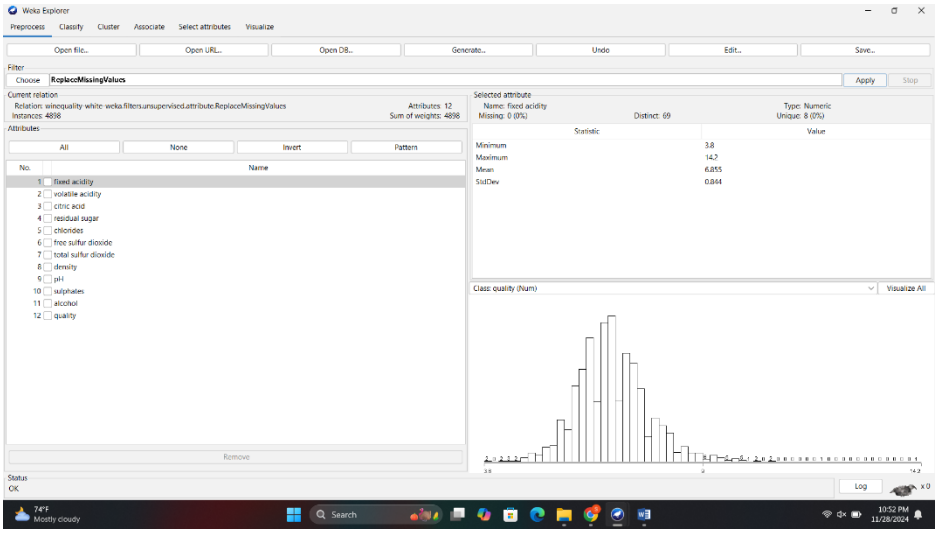

Figure 2: Apply ReplaceMissingValues filter to Winequality - White dataset

2. Experiment with different parameter changes affect the clustering outcomes

2.1 Configure parameters in SimpleKMeans Algorithm

- Number of clusters

Applying k=2, the clusters broadly separate wines into two categories (high vs. low quality).

Applying k=5, the clusters become more specific in wine characteristics. Increasing k provides more detail but can risk overfitting, making the model complex without improving accuracy.

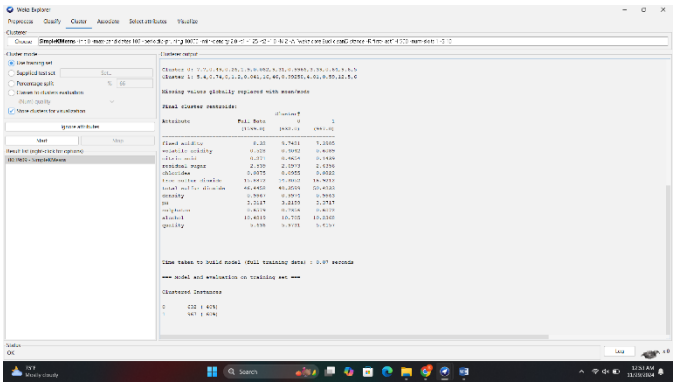

Figure 4: Set k=2 in winequality-red dataset

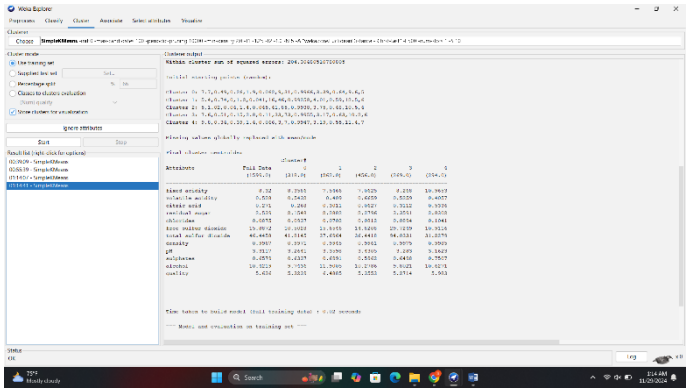

Figure 3: Set k=4 winequality-red dataset

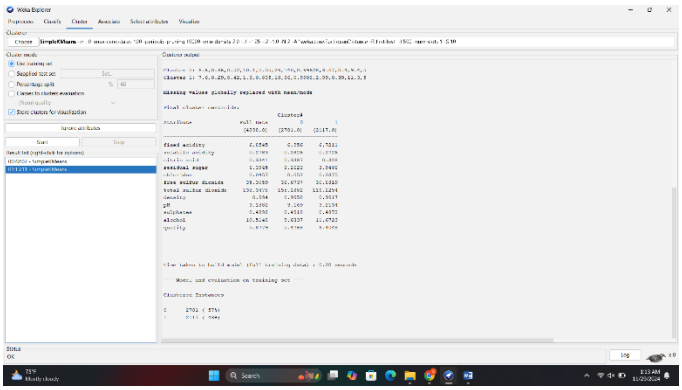

Figure 5: Set k=2 winequality-white dataset

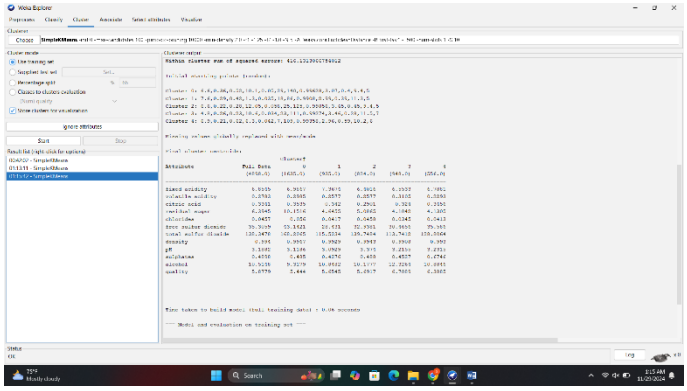

Figure 6: Set k=4 winequality-white dataset

- Distance Function

Euclidean Distance - The dataset is split into 2 clusters as Cluster 0 containing 632 instances 40% and Cluster 1 containing 967 instances 60%.

Manhattan Distance - There are 2 clusters with 628 instances 39% in Cluster 0 and 971 instances 61% in Cluster 1.

Impact is the cluster sizes remain identical between both distance functions and the overall partitioning remains consistent though the centroids slightly change.

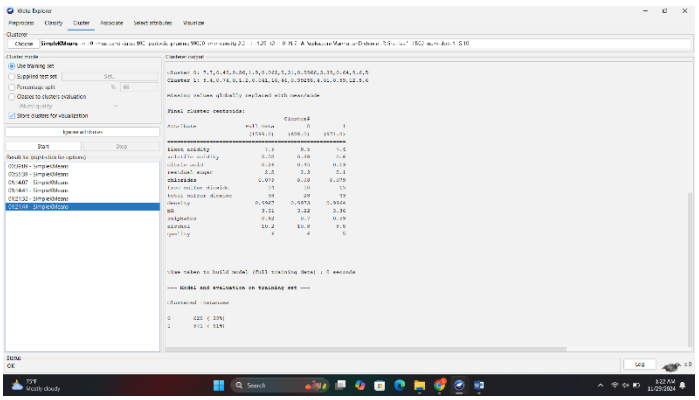

Figure 7: Apply Manhattan Distance as Distance – winequality-red dataset

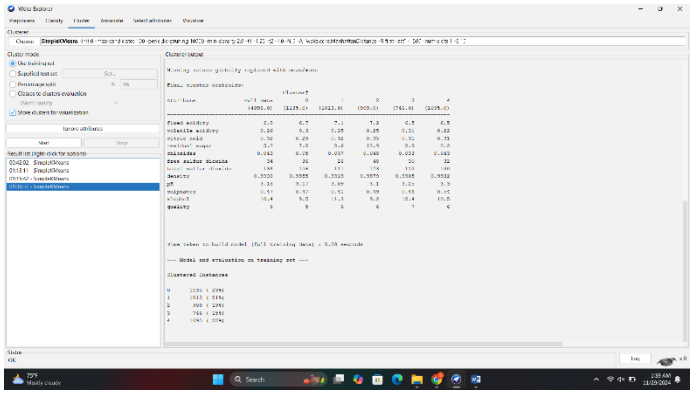

Figure 8: Apply Manhattan Distance as Distance Function - winequality-white dataset

Euclidean Distance performs better in evenly distributed clusters and the data is not sparse.

Manhattan Distance is effective when there are sparse data, less sensitive to outliers.

2.2 Configure parameters in DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

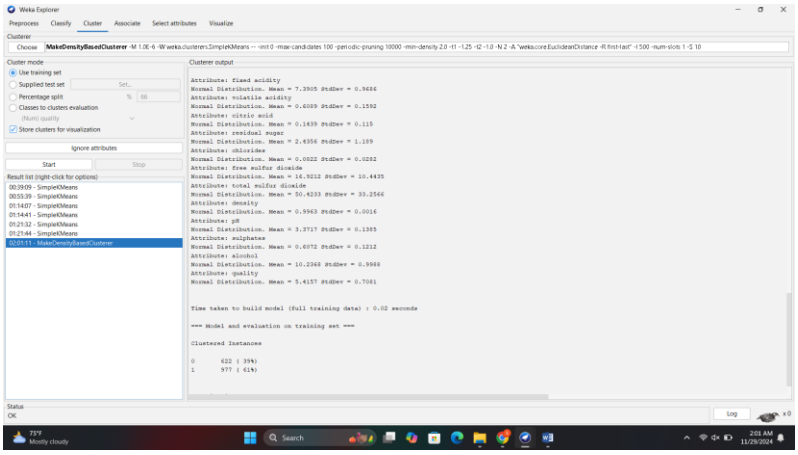

Figure 9: Apply MakeDensityBasedCluster

- Epsilon (eps)

- MinPts

- Distance Function

Conclusion

The analysis was realized with the help of software WEKA, running on a number of data pre-processing and clustering techniques in predicting wine quality. First, handling missing values: it enabled imputation methods like mean substitution, enabling the completion of the dataset for further analysis. The comparison of the clustering methods, K-means versus DBSCAN, showed large effects of the algorithm's choice and its parameters on the accuracy of clustering. K-means was sensitive to the number of clusters, while DBSCAN was more flexible with varying densities and was capable of finding outliers. The application of data discretization techniques like Equal Width and Equal Frequency methods improved clustering performance since the continuous features were transformed into discrete bins, making the clusters more distinguishable. Generally, the results of clustering have proven that the selection of proper parameters and preprocessing methods is important in increasing the effectiveness of the clustering process.