**Assignment #2: Clustering**

Student Name

Date

**Assignment #2: Clustering**

**Dataset 1: The Iris Dataset**

**Dataset Analysis**

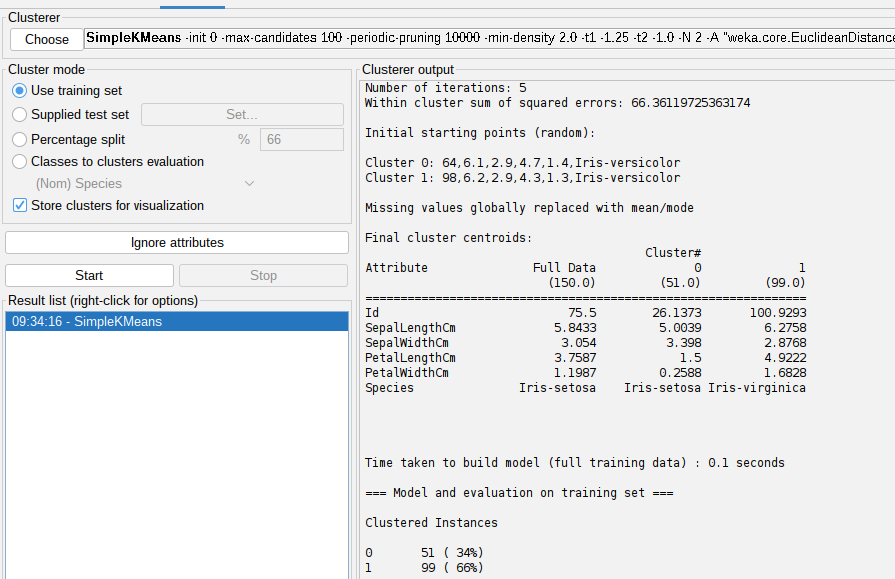
The Iris dataset is one of the widely used benchmarks in multi-class classification problems. It has 150 instances, and each instance contains measurements from a flower belonging to one of three classes: Setosa, Versicolor, or Virginica. Each record contains four numeric attributes, sepal length, sepal width, petal length, and petal width, in centimeters. This dataset provides a clear-cut classification problem with three classes evenly split (50 instances each) and is perfect for testing machine learning algorithms. It has been an extremely popular tool in pattern recognition and algorithmic testing since it was first devised by R.A. Fisher in 1936 because it is easy to visualize and clear-cut.

Data Preparation

No heavy processing was required because the Iris dataset is provided in a preprocessed .arff format that can be read by WEKA. Upon import into the WEKA Explorer, all the attributes were labeled as numeric, and the target attribute, flower type, was defined automatically as the class attribute. The data is clean with no missing values or errors. Since its classes were well-distributed and well-formatted, there was no need for any extra preprocessing tasks such as normalization, cleaning the data, or feature selection before model training.

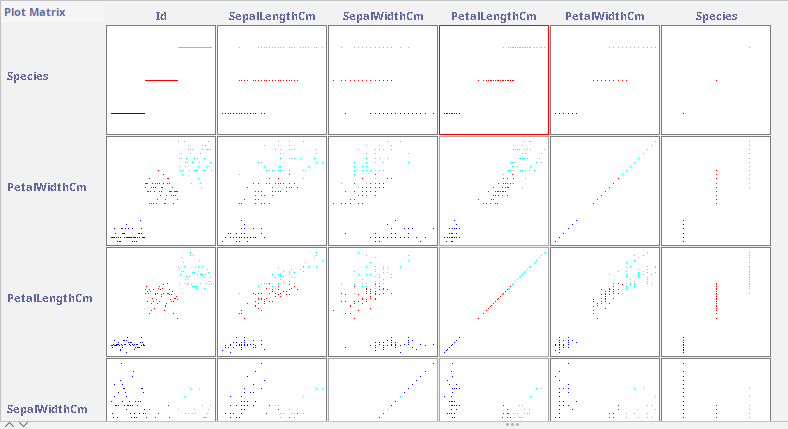
**Model Development**

The k-Means clustering algorithm on the Iris dataset correctly separated the instances into two clusters. The cluster result does not strictly correspond to the real three species in the dataset. Cluster 0 consists of 51 instances, which are all from the species Iris-setosa, and Cluster 1 consists of 99 instances, which are from the two species Iris-versicolor and Iris-virginica. The model recorded a within-cluster sum of squared errors (SSE) of 66.36, indicating spread within the clusters. Cluster centroids show different variations in characteristics, especially in petal length and petal width, which enabled the separation of the two clusters. However, the inability of the model to uniquely distinguish between Iris-versicolor and Iris-virginica suggests that the clustering process could have failed to capture the inherent pattern in the dataset.



*Figure 1: Simple Kmeans results*

When the clusters are plotted, the samples of Iris-setosa are well separated from the remaining two species and establish their own group. The samples of Iris-versicolor and Iris-virginica are mixed, meaning that more sophisticated clustering techniques or more clusters would be required to achieve greater precision. This plot reflects the complexity of distinguishing between these two species from the given features.



*Figure 2: K-means visualization*

**Dataset 2: Wine Quality Dataset**

**Dataset Analysis**

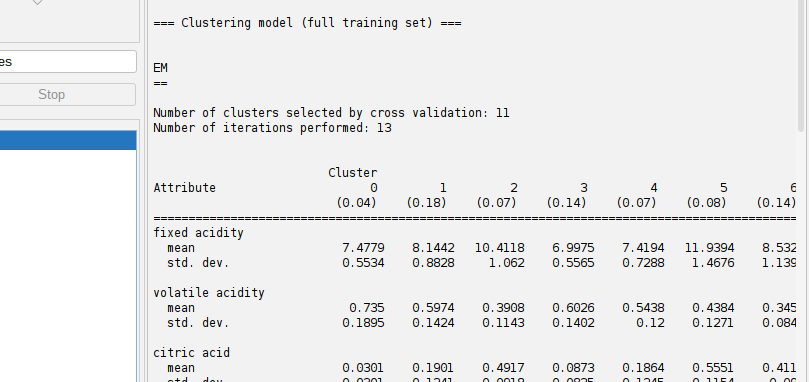
The Wine Quality dataset consists of red wines from the "Vinho Verde" region of Portugal and is primarily used for predicting wine quality based on several chemical characteristics. It has 11 features measured in laboratory tests, such as acidity, alcohol, residual sugar, sulfur dioxide content, and pH. The target feature, quality, is the wine quality score from 0 to 10 given by expert tasters. This is a regression task since quality scores are continuous but not uniformly distributed, with more wines in the middle quality and fewer in the extremes. Even though the task is straightforward, the nature of complexity is the presence of nonlinear relationships among the variables and the limited sample size, so it is an ideal candidate for testing whether machine learning models can predict product quality variation or not.

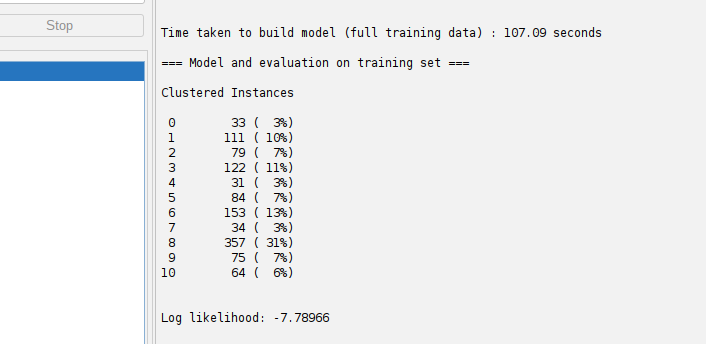
**Data Preparation**

The data was run through a basic cleaning procedure to identify and correct missing and invalid values. Initially provided in CSV format, it was converted to ARFF (Attribute-Relation File Format), which is WEKA compatible. Each attribute was verified to categorize it as numeric, and the quality variable was not converted to a numeric value to be able to perform regression analysis. The ARFF file was imported into WEKA to start building the model.

**Model Development**

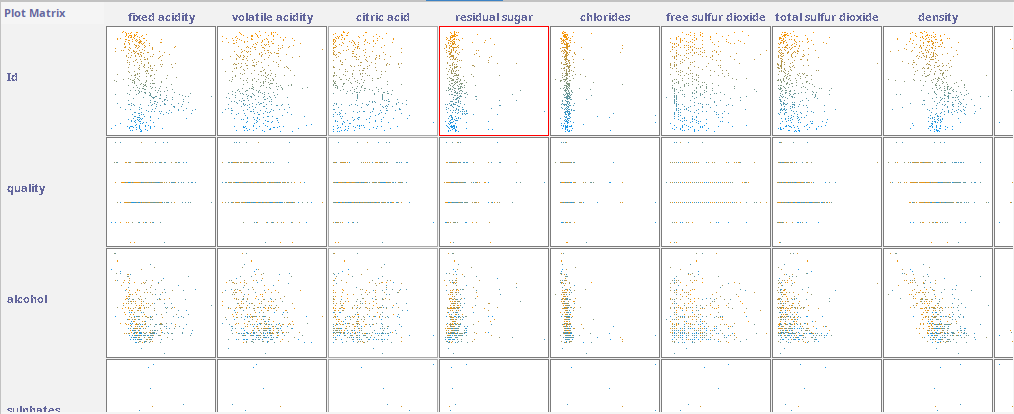
The algorithm yielded 11 clusters out of the 1,143 examples. The largest cluster is 8 at 31% of examples, signifying a preponderant subgroup in the data, and the clusters 0, 4, and 7 are relatively small. The clusters significantly differ along such dimensions as volatile acidity, citric acid, alcohol, and quality, evincing significant segmentation. For instance, Cluster 6 is high in mean alcohol (11.45) and quality (6.37), and Cluster 4 is low in alcohol (9.57) and quality (5.30), indicating probable clustering by quality levels of wines. The model log likelihood is -7.78966, and the model converged at 13 iterations, indicating good efficiency and convergence.





*Figure 3: Clustering results*

For visualization, a 2D PCA scatter plot by cluster color or a cluster distribution bar chart could quite possibly show the extent to which instances are split between clusters, showing visually where overlaps and dominant clusterings take place.



*Figure 4: Clustering visualization*