**Hands-On Labs**

Data Mining: WEKA

Assignment #2

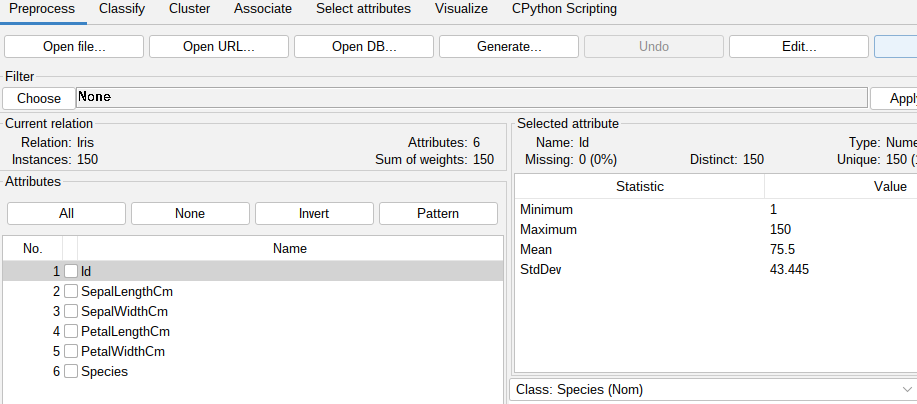
**Assignment #2**

The selected data sets are Iris, a classic dataset, and Wine, which contains chemical analysis of wines. The first step is to analyze each dataset in detail.

**Dataset Analysis**

Iris Dataset

The Iris dataset is one of the most used machine learning datasets and was first introduced in R.A. Fisher's 1936 publication, The Use of Multiple Measurements in Taxonomic Problems. It is available in the UCI Machine Learning Repository and is widely utilized for classification and clustering tasks. The data consists of three iris species: Setosa, Versicolor, and Virginica. There are a total of 150 observations, with 50 observations for each of the species. There are five features in the dataset: SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm, and Species.

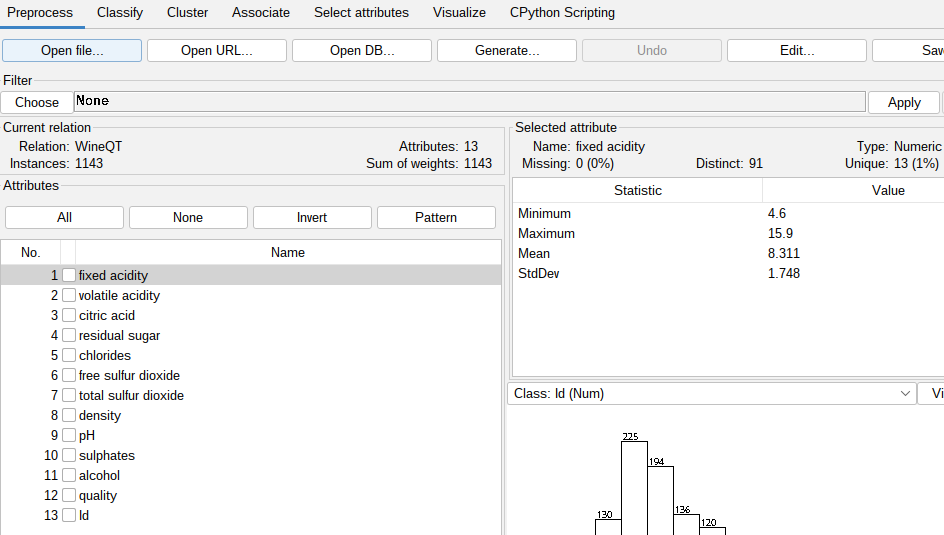


*Figure 1: Iris Dataset*

In clustering, this data is of particular interest because one of the species, Setosa, can be linearly separated from the other two species, but the other two species (Versicolor and Virginica) cannot be separated linearly from each other. This renders the dataset an excellent candidate for unsupervised learning methods such as clustering. Clustering algorithms will reveal underlying patterns in the data based on the sepal and petal measurements of the flowers, even without the knowledge of the species label. Domain-specific knowledge derived from this data set can be like plant species classification or plant research, where clustering can be used to identify similarities and dissimilarities in the flower features to be explored in detail.

Wine Dataset

The Wine Quality dataset is a standard dataset located in the UCI Machine Learning Repository. It contains information regarding the quality of red wine with 11 input variables from physicochemical tests and one output variable for the sensory quality of the wine. The input variables contain measurements of fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The output variable, quality, is a human subjective score between 0 and 10, which reflects the quality of the wine as established by human sensory assessment.

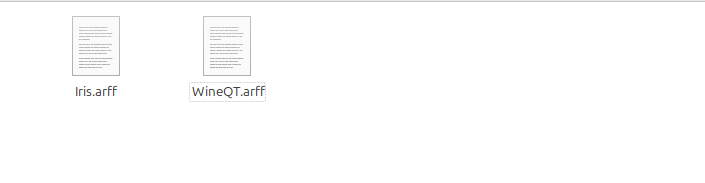


*Figure 2: Wine quality dataset*

This data set is particularly suitable for clustering since it can identify groupings of wines with similar physicochemical properties and quality scores that might be of interest to identify various types of wine or potential product categories. Clustering techniques can assist in identifying patterns or correlations between the chemical makeup of the wine and its perceived quality, enabling manufacturers to cater to some consumer preferences or improve wine-making processes. The dataset is commonly used in the field of food science and the wine industry, where clustering can be employed to create products, inspect quality, and segment wine types based on chemical characteristics.

**Data Preparation**

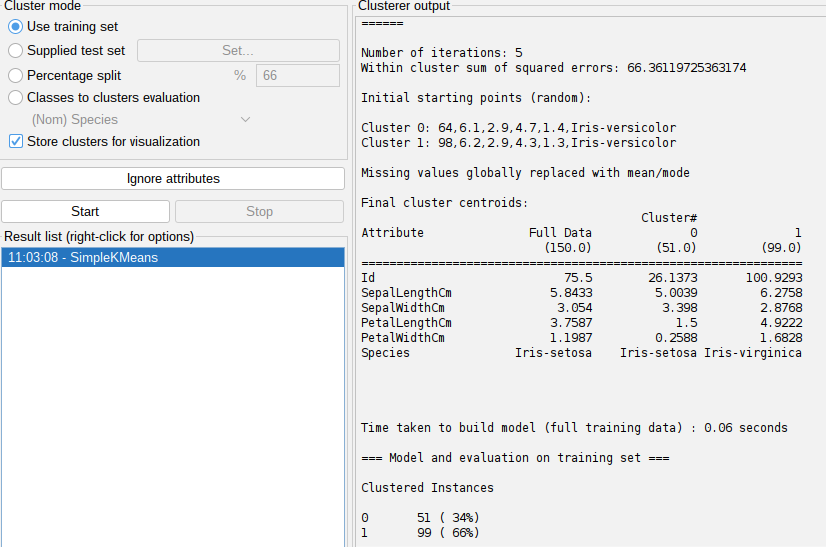
The most important step in this section is to convert CSV to ARFF format, WEKA’s native format. For this, the WEKA application was used. This was achieved by loading the CSV files in WEKA and saving them as ARFF files. The two files are as shown below.



*Figure 3: Both datasets in ARFF format*

**Model Development**

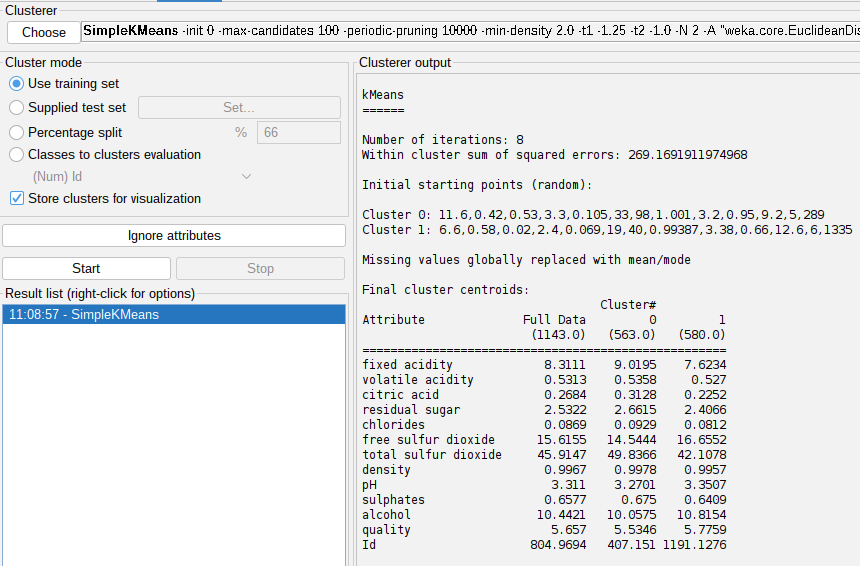
In this section, we implement three different models for each dataset and compare the results. The first model to be implemented is K-means clustering. For the Iris dataset, the following results were obtained.

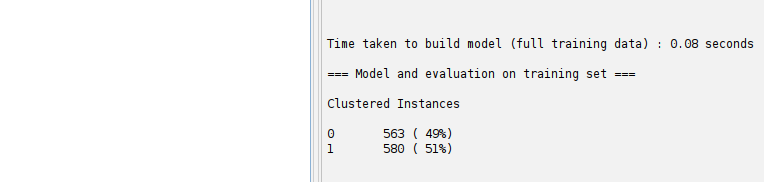


*Figure 4: K-Means clustering on the Iris dataset*

The K-Means clustering algorithm was performed on the Iris dataset with k=2 and resulted in a model that converged in 5 iterations with within-cluster sum of squared errors (SSE) of 66.36, indicating the overall cluster compactness. Centroids at the end show a clear differentiation between Cluster 0, which is very similar to Iris-setosa, and Cluster 1, which is a mix of Iris-versicolor and Iris-virginica, both being less separable linearly. Cluster 0 contains 51 samples with very small petal values (e.g., PetalLength ~1.5 cm), closely marking it as Setosa. Cluster 1, which contains 99 observations, contains larger petal sizes, which align with the higher overlap complexity of Versicolor and Virginica. This reinforces that although K-Means works well in distinguishing the easily distinguishable Setosa, it struggles to distinguish the other two species because they are close in feature space. The model ran with ease, taking only 0.06 seconds, and demonstrates K-Means' suitability in working with data having clean separations like this.

For the Wine Quality dataset, the following results were obtained after running the K-means algorithm.

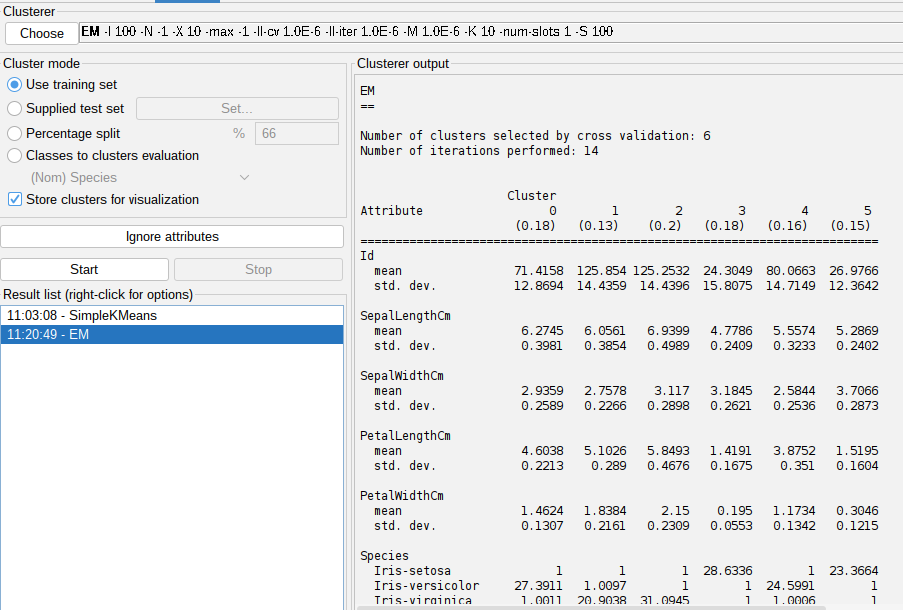


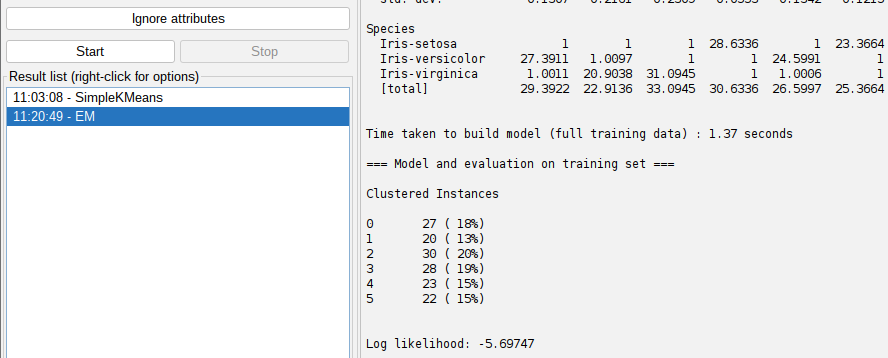


*Figure 5: Wine Quality K-means results*

The K-Means algorithm was executed on the Wine Quality dataset with 2 clusters (k=2) and yielded a pretty balanced split: Cluster 0 with 563 instances and Cluster 1 with 580. It converged to 8 iterations with a within-cluster sum of squares of errors (SSE) of 269.17, which is the measure of the internal homogeneity of the clusters. Cluster centroids reveal that Cluster 0 corresponds to wines of greater fixed acidity, citric acid, and density but lower alcohol and slightly lower quality ratings. Cluster 1, on the other hand, is wines of typically lower acidity, greater alcohol, and slightly greater quality ratings, which means alcohol content and acidity are potentially major factors in consumers' perceptions of wine quality. The almost equal distribution suggests natural clustering with small imbalance, and the model efficiency (0.08 seconds of training) accentuates the relevance of K-Means to large numeric datasets like this one.

The next model that will be implemented is EM (Expectation Maximization). After running the algorithm, the following results were obtained for the Iris dataset.

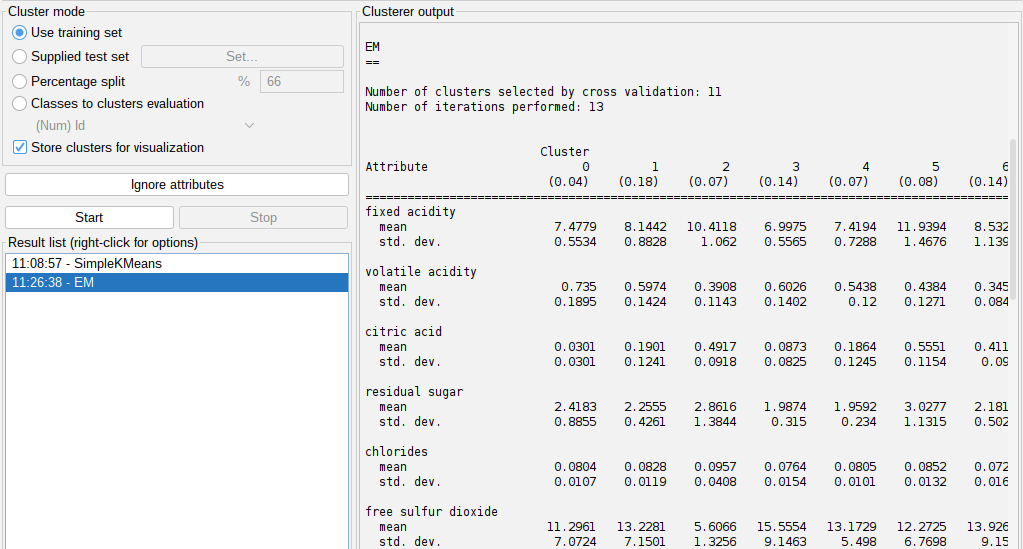


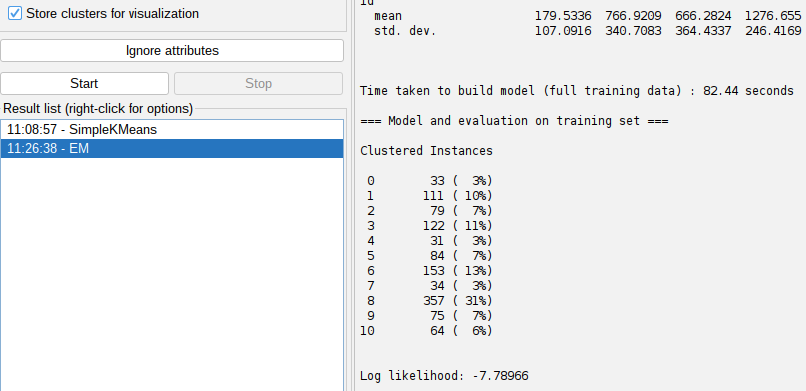


*Figure 6: EM on Iris dataset*

The EM clustering algorithm on the Iris dataset automatically chose six clusters by using cross-validation and completed training in 1.37 seconds within 14 iterations. The resulting clusters represent a more detailed grouping of the 150 flower instances, with a fairly even split across the clusters, between 13% and 20% of the data in each cluster. The centroids of petal length and petal width played a big role in the separation, corresponding to known biological disparities among species. For instance, Cluster 3 and Cluster 5 have the lowest petal values and are dominated by Iris-setosa, which is linearly separable. Conversely, clusters with large petal values like Cluster 2 and Cluster 1 have more Iris-virginica instances, which are harder to separate from Iris-versicolor. The model did reveal the underlying structure of the data beyond the three-class labels that were known, with a log likelihood of -5.69747 indicating the overall fit.

The EM clustering algorithm was also applied to the Wine Quality dataset and the following results were obtained.





*Figure 7: EM clustering on Wine Quality dataset*

Expectation Maximization (EM) algorithm for clustering the WineQT data automatically detected eleven clusters within 82.44 seconds in 13 iterations by cross-validation. The resultant clusters indicate different groupings of the wine samples of 1143 between 3% and 31% of the total data. Cluster 8, with nearly one-third of the samples, has medium acidity, higher total sulfur dioxide, and relatively low quality scores, while Cluster 10, with the highest average alcohol (12.32) and mean quality of 6.01, likely represents the most refined samples. In contrast, clusters like 2 and 6 combine high citric acid and alcohol with greater than average quality, implying higher-grade wines. The application of persistent chemical attributes such as fixed acidity, sulphates, and alcohol unearthed by the model well separated the latent structure in the data set to exhibit rich wine profiles from basic quality labels. On a log likelihood of -7.78966, the cluster results capture EM's capacity to model real-world, multidimensional data with complex, overlapping distributions.

**Evaluation and Findings**

The contrast between EM and K-Means on the Iris and Wine Quality datasets shows complementary advantages of both methods. K-Means performed satisfactorily on the Iris data, distinguishing the linearly separable Setosa species and merging Versicolor and Virginica due to feature overlap. In contrast, the EM algorithm, being computationally more intensive, uncovered a more profound structure by discovering six subtle clusters, yielding better insight into latent class distributions with a decent log likelihood value. In the case of the Wine Quality dataset, both models proved effective at clustering similar chemical wines, but EM surpassed by easily identifying eleven clusters and projecting them to different wine profiles based on variables like alcohol and acidity. The more subtle segmentation and probabilistic character of EM suited it better to the richness of the Wine dataset, whereas K-Means provided a fast, interpretable baseline for initial exploration.