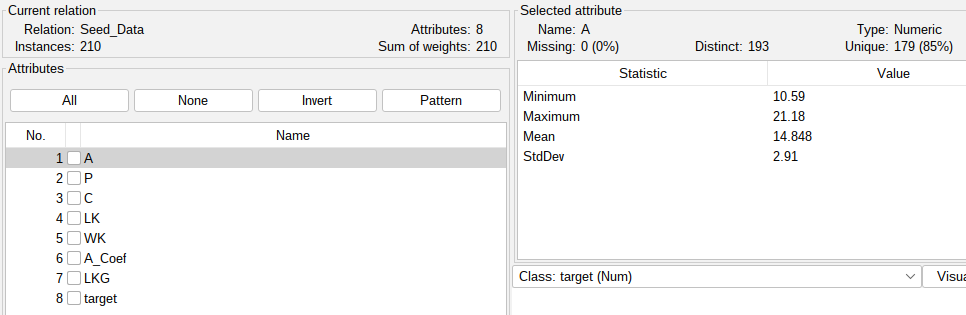
**Assignment #2: UCI Dataset Analysis and Model Development**

Student Name

Date

**Seeds Dataset**

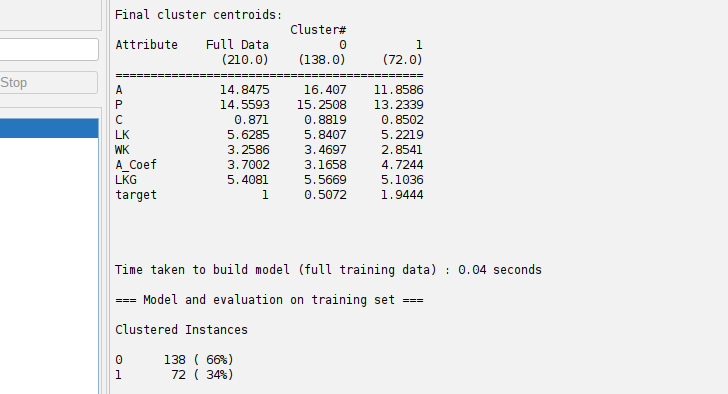
The Seeds dataset includes measurements of geometrical properties of kernels from three types of wheat: Kama, Rosa, and Canadian. There are 70 randomly selected kernels for each type. Data was captured through a non-destructive soft X-ray technique that provides high-resolution internal kernel images at a lower cost compared to other imaging methods. The seven numeric attributes determining kernel geometry in the dataset make it suitable for clustering and classification applications. The well-preprocessed dataset gives a realistic scenario to analyze clustering techniques in agricultural data analysis.



*Figure 1: Seeds dataset*

Preprocessing of the data was done by converting all attribute values to numeric and consistent, and removing any formatting abnormalities. Class attribute was nominalized to allow clustering, and the data was checked for missing values to ensure quality input into the model.

The K-Means model was trained on the Seed\_Data dataset, having 210 instances and 8 features, and split the data into 2 clusters. After 5 iterations, the model produced a within-cluster sum of squares value of almost 59.43, which indicated the clusters' moderate compactness. The clusters are well-defined with centroid values, such that Cluster 0 contains larger means in features like A, P, and WK, and Cluster 1 contains larger values in A\_Coef and target. Cluster 0 contains 66% of the instances, and Cluster 1 contains 34%, showing an even distribution of instances among clusters. Global means are used to replace missing values to ensure data integrity or modes.

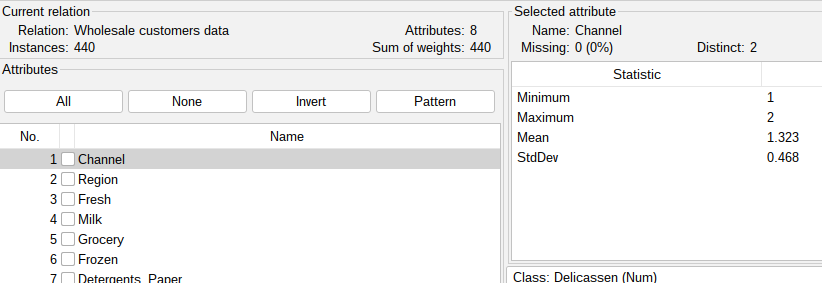


*Figure 2: Clustering results*

Since clustering is an unsupervised process, usual measures of accuracy do not apply; however, the minimum sum of squared errors and distinct cluster centroids indicate the model is effective in separating similar seed kernel measurements, which could be useful for further classification or pattern recognition.

**Wholesale Customers Dataset**

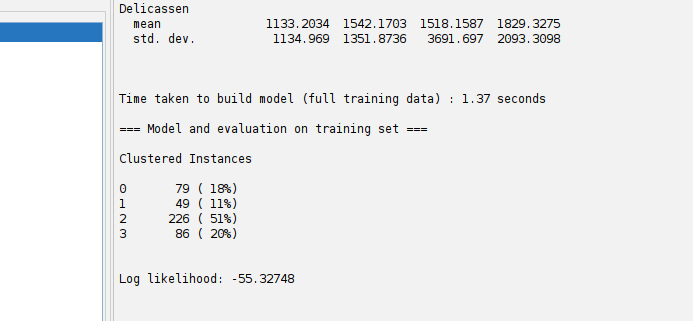
The Wholesale Customers dataset is a set of data on the annual spending of customers of a wholesale distributor on various product categories like fresh products, milk, grocery, frozen foods, detergents and paper, and delicatessen. The dataset contains 440 instances with 7 continuous numerical features and 2 nominal features for customer channel and region. It is a multivariate dataset suitable for classification and clustering problems to analyze customer purchasing behavior. There are no missing values, and the data has a considerable amount of variation in spending across categories, providing a rich ground for the identification of distinct customer segments or patterns.



*Figure 3: Wholesale customers dataset*

The data preparation involved cleaning the Wholesale Customers dataset by verifying no missing values and that all features were well formatted for analysis. The categorical variables were well encoded and converted into a compatible format for WEKA. This preprocessing guaranteed that the data was in an appropriate format for effective clustering and classification modeling.

EM clustering algorithm on Wholesale Customers dataset converged with four clusters after three iterations, where cross-validation was employed to identify the ideal number of clusters. The clusters vary substantially concerning customer spending patterns as well as channel/region distribution. For example, Cluster 1 and Cluster 3 are representative of customers with higher spending in milk, grocery, and detergents, while Cluster 0 and Cluster 2 are representative of lower spending on those product categories but with differentiation on fresh and frozen product spending. The instance distribution shows that the majority of customers (51%) are in Cluster 2, representing a dominant customer segment of moderate spending.



*Figure 4: Clustering results using EM*

The model's log likelihood value, -55.33, reflects the fit quality, where convergence was reached in a swift 1.37 seconds. The EM algorithm partitioned the customers into meaningful groups successfully, with significant implications for targeted marketing or business.