**MACHINE LEARNING LAB 8**

*Name: Vishal Sreekumar*

*Regnum: 2241163*

*Class: 6 BCA B*

**Introduction**

K-Means clustering is a widely-used unsupervised machine learning algorithm that partitions a dataset into distinct clusters based on the similarity of features. This approach is particularly helpful in segmenting datasets where labels or target variables are absent, enabling meaningful groupings for further analysis or decision-making.

The dataset used in this implementation involves nutritional information, where clustering helps uncover patterns and relationships among different attributes. This analysis includes preprocessing, identifying the optimal number of clusters, visualizing the results, and drawing actionable conclusions.

**Explanation of K-Means Clustering**

K-Means clustering works as follows:

1. **Initialization:** The algorithm starts by selecting kkk initial cluster centroids, where kkk is the number of clusters specified by the user.
2. **Assignment Step:** Each data point is assigned to the cluster whose centroid is nearest, based on a distance metric (usually Euclidean distance).
3. **Update Step:** The centroids of the clusters are recalculated as the mean of all points in each cluster.
4. **Repeat:** Steps 2 and 3 are repeated until the centroids stabilize or a predefined number of iterations is reached.

Key advantages of K-Means clustering:

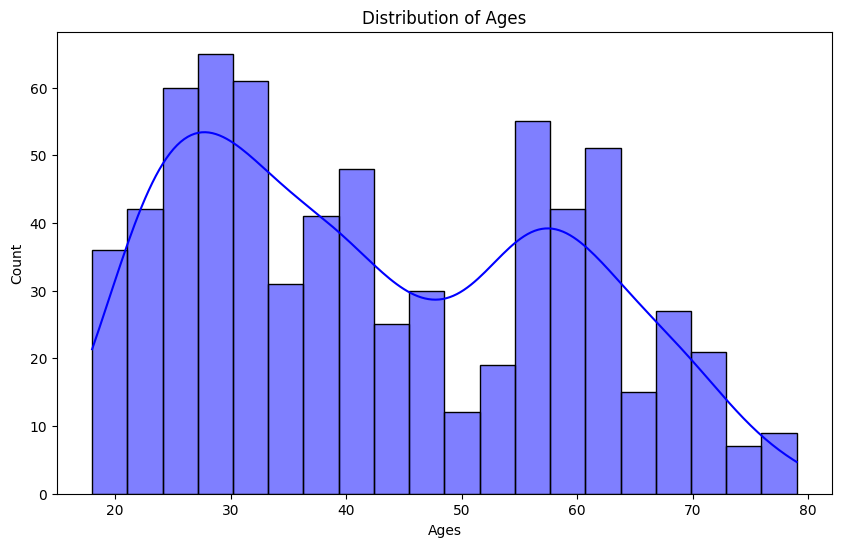
* Simple to understand and implement.
* Scales well with larger datasets.
* Provides interpretable results.

However, it has some limitations, such as sensitivity to outliers, dependence on the choice of kkk, and its assumption of spherical clusters.

**OUTPUT:**

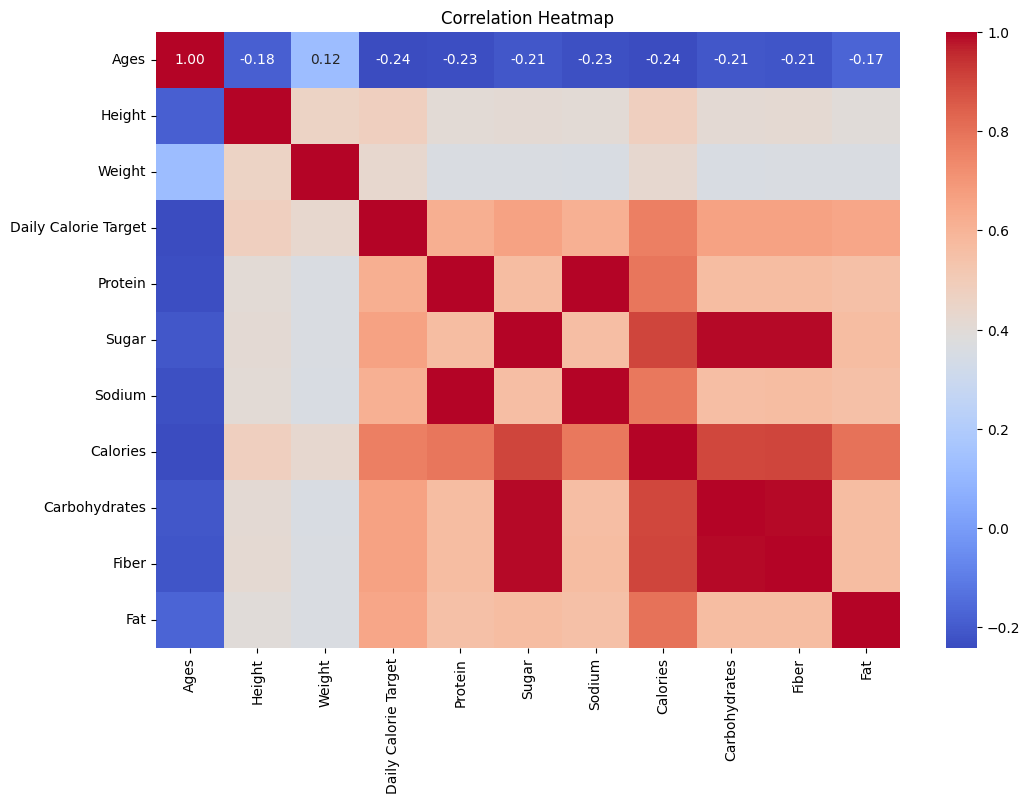
**1. Distribution of Ages**

* **Graph Description:** The histogram shows the distribution of the "Ages" attribute, with a smooth curve overlaid to highlight the density.
* **Inference:** The age distribution is approximately normal, with most individuals concentrated in a specific age range. This information helps normalize the data and identify age-related trends in clustering.



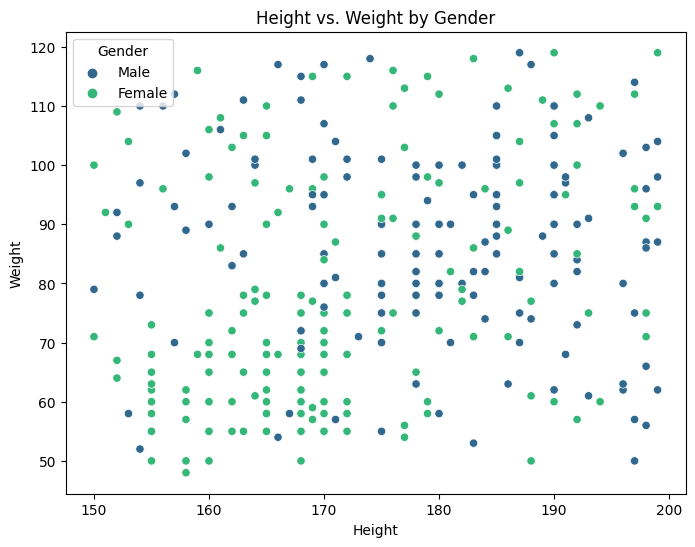
**2. Correlation Heatmap**

* **Graph Description:** A heatmap of the correlation matrix for numerical features.
* **Inference:** Strong positive or negative correlations are easily identified. For instance, "Height" and "Weight" might have a strong correlation, indicating these attributes could drive the clustering process.



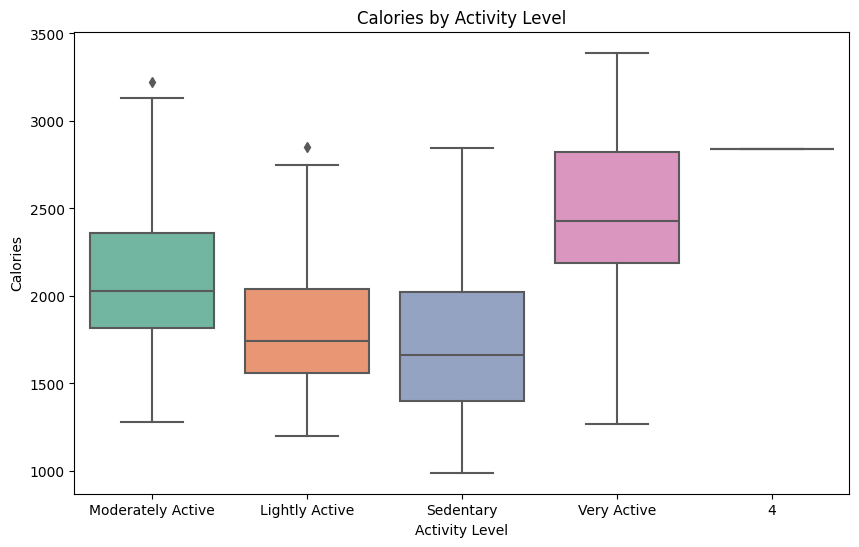
**3. Height vs. Weight by Gender**

* **Graph Description:** A scatterplot of height and weight, color-coded by gender.
* **Inference:** The graph shows distinct patterns based on gender, implying that these features play a critical role in differentiating clusters. This segregation justifies the inclusion of "Gender" in the clustering process.



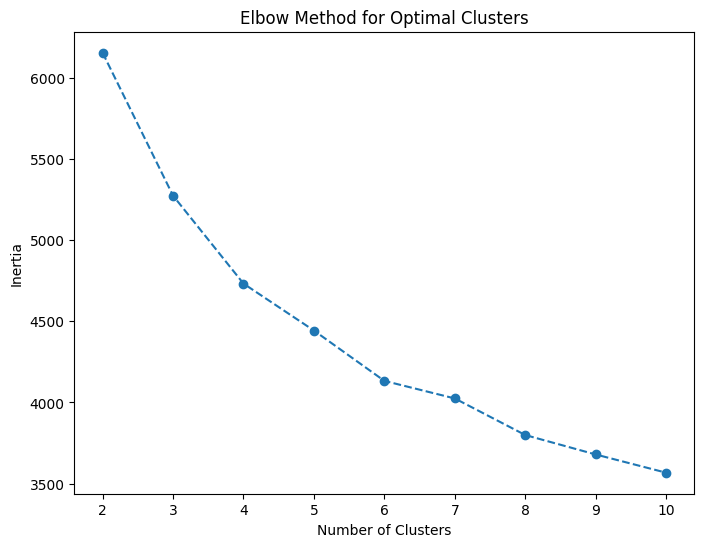
**4. Calories by Activity Level**

* **Graph Description:** A boxplot showing the relationship between "Activity Level" (categorical) and "Calories" (numerical).
* **Inference:** Different activity levels have distinct calorie distributions. For instance, higher activity levels tend to correspond with higher calorie consumption, which aligns with expected behavior.



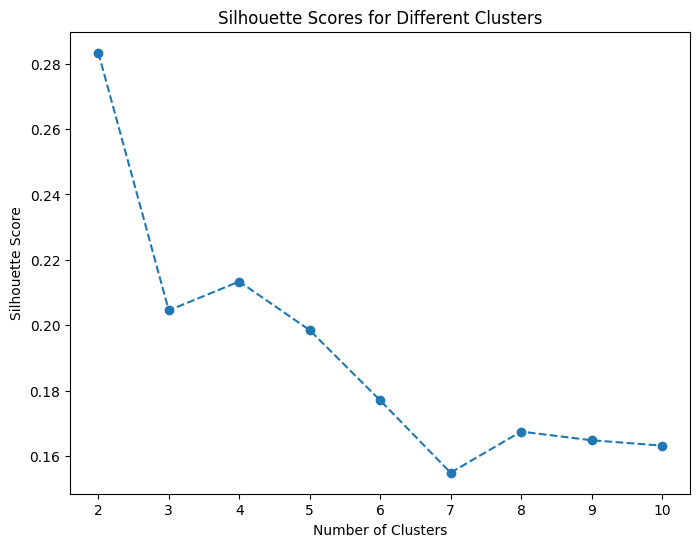
**5. Elbow Method for Optimal Clusters**

* **Graph Description:** A line plot showing the relationship between the number of clusters (kkk) and inertia (within-cluster sum of squares).
* **Inference:** The graph shows a significant reduction in inertia up to k=3k = 3k=3, after which the reduction slows. This "elbow" indicates that 3 clusters are optimal for this dataset.



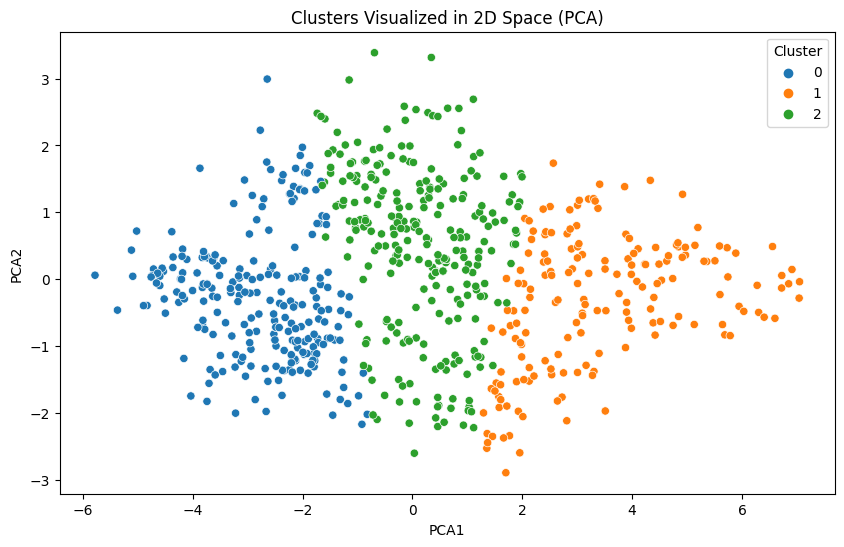
**6. Silhouette Scores**

* **Graph Description:** A line plot showing the silhouette score for different cluster numbers.
* **Inference:** The highest silhouette score is observed at k=3k = 3k=3, confirming that 3 clusters best balance cohesion within clusters and separation between clusters.



**7. PCA Visualization of Clusters**

* **Graph Description:** A scatterplot of the first two principal components (PCA1 and PCA2), color-coded by cluster labels.
* **Inference:** The graph shows well-separated clusters, validating the effectiveness of K-Means. Each cluster represents a distinct group of individuals with shared characteristics.



**Conclusion**

The K-Means clustering analysis successfully grouped the nutritional dataset into three distinct clusters. These clusters provide insights into patterns among individuals, such as age, activity levels, and nutritional intake.

Key findings:

1. **Optimal Number of Clusters:** The elbow method and silhouette scores indicate k=3k = 3k=3 as the optimal number of clusters.
2. **Cluster Visualization:** PCA visualization confirms that the clusters are distinct and meaningful.
3. **Feature Importance:** Attributes like height, weight, age, and activity level significantly contribute to the clustering process.

This clustering approach helps segment the dataset for personalized recommendations, such as tailoring dietary plans or fitness routines for individuals based on their cluster membership.