S. B. JAIN INSTITUTE OF TECHNOLOGY, MANAGEMENT & RESEARCH, NAGPUR.

Practical No. 7

Aim: Comparative Analysis of Clustering Techniques Using K-Means and Hierarchical Clustering.

Name of Student: Shrutika Pradeep Bagdi

Roll No.: CS22130

Semester/Year: 6th/3rd

Academic Session: 2024-2025

Date of Performance:

Date of Submission:

AIM: Comparative Analysis of Clustering Techniques Using K-Means and Hierarchical Clustering

OBJECTIVE/EXPECTED LEARNING OUTCOME:

The objectives and expected learning outcome of this practical are:

- For simple k-means include understanding clustering, implementing the k-means algorithm, evaluating clustering results, handling algorithm limitations, applying k-means to real-world datasets, and gaining practical experience in data analysis and machine learning.
- Understanding the concept of hierarchical clustering, implementing the algorithm, interpreting dendrograms, evaluating clustering results, handling different linkage methods, and applying hierarchical clustering to real-world datasets.

HARDWARE AND SOFTWARE REQUIRMENTS:

Hardware Requirement: Computer System with high configurations

Software Requirement: Weka Tool-3.6.9

THEORY:

Simple k-means Clustering Algorithm:

The simple k-means clustering algorithm is an iterative algorithm used to partition a dataset into k distinct clusters. Here is a high-level overview of the algorithm:

Initialization: Randomly select k initial cluster centroids from the dataset.

Assignment: For each data point, calculate the distance to each centroid and assign it to the nearest centroid's cluster.

Update: Recalculate the centroids by taking the mean of all data points within each cluster.

Repeat: Repeat steps 2 and 3 until convergence, which occurs when the centroids no longer change significantly or a maximum number of iterations is reached.

Output: The final cluster assignments and centroids represent the clustering result.

The steps involved in implementing the simple k-means algorithm include choosing the number of clusters (k), initializing centroids, assigning data points to clusters, updating the centroids, and iterating until convergence. Evaluation metrics such as the within-cluster sum of squares (WCSS) can be used to assess the quality of the clustering.

It's important to note that simple k-means has some limitations, such as sensitivity to initialization and the assumption of spherical clusters. However, it remains a popular and straightforward algorithm for clustering tasks.

Here are some important points to note about the simple k-means clustering algorithm:

Initialization: The choice of initial centroids can affect the clustering result. Different initialization methods, such as random selection or k-means++, can be used to improve convergence and avoid local optima.

Distance Metric: The choice of distance metric, such as Euclidean distance or Manhattan distance, impacts how the data points are assigned to clusters. The appropriate distance metric depends on the nature of the data and the problem domain.

Convergence: The algorithm iteratively updates cluster assignments and centroids until convergence. Convergence occurs when the centroids no longer change significantly or a maximum number of iterations is reached. It's important to monitor convergence and set appropriate stopping criteria.

Trade-off between Complexity and Performance: The algorithm's time complexity is influenced by the number of data points and the number of clusters. As the dataset grows, the computational cost increases. Additionally, increasing the number of clusters leads to finer granularity but also increases the risk of overfitting.

Handling Outliers: Simple k-means is sensitive to outliers as they can significantly impact the centroid calculation. Outliers can distort cluster boundaries and affect the overall clustering result. Preprocessing steps, such as outlier detection or outlier removal, can be employed to mitigate this issue.

Determining the Optimal Number of Clusters: The choice of the optimal number of clusters (k) is subjective and problem-specific. Various techniques, such as the elbow method, silhouette score, or hierarchical clustering, can be used to find an appropriate value of k.

Limitations: Simple k-means assumes that clusters are spherical and have equal variance. It may struggle with non-linear or irregularly shaped clusters. To handle such scenarios, more advanced clustering algorithms like DBSCAN or Gaussian Mixture Models can be considered.

The hierarchical clustering algorithm:

Agglomerative and Divisive: Hierarchical clustering can be performed using two main approaches: agglomerative and divisive.

Agglomerative clustering starts with each data point as an individual cluster and progressively merges the closest clusters until a desired number of clusters is reached.

Divisive clustering begins with all data points in a single cluster and iteratively splits the clusters until each data point forms its own cluster.

Dendrograms: Hierarchical clustering produces dendrograms, which are tree-like structures that illustrate the merging or splitting of clusters. The dendrogram visually represents the similarity or dissimilarity between clusters or data points.

Distance Metric: The choice of distance metric plays a crucial role in hierarchical clustering. Common distance metrics include Euclidean distance, Manhattan distance, and correlation distance. The appropriate distance metric depends on the nature of the data and the problem domain.

Linkage Methods: Linkage methods determine how the distance between clusters is calculated during the merging or splitting process. Common linkage methods include:

Single linkage: Measures the distance between the closest points in two clusters.

Complete linkage: Measures the distance between the furthest points in two clusters.

Average linkage: Measures the average distance between all pairs of points in two clusters.

Ward's linkage: Minimizes the increase in variance when merging clusters.

Determining the Number of Clusters: Hierarchical clustering does not require specifying the number of clusters beforehand. Instead, the number of clusters is determined based on the dendrogram or by setting a threshold on the distance or dissimilarity measure.

Interpretation of Dendrograms: Dendrograms help interpret the clustering results. The height of the vertical lines in a dendrogram represents the distance or dissimilarity between merged or split clusters. The horizontal axis represents the data points or clusters.

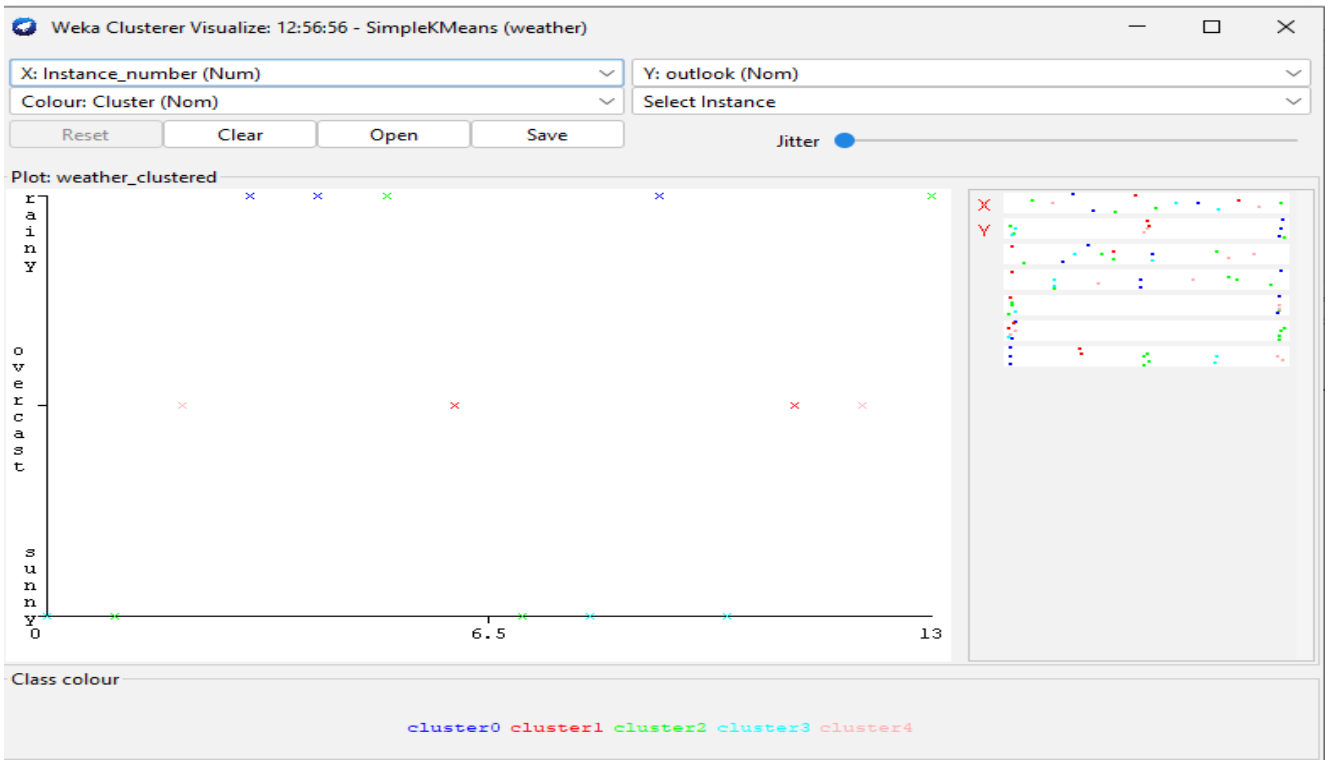
Handling Large Datasets: Hierarchical clustering can become computationally expensive for large datasets due to its quadratic time complexity. In such cases, techniques like sampling or dimensionality reduction can be employed to make the algorithm more feasible.

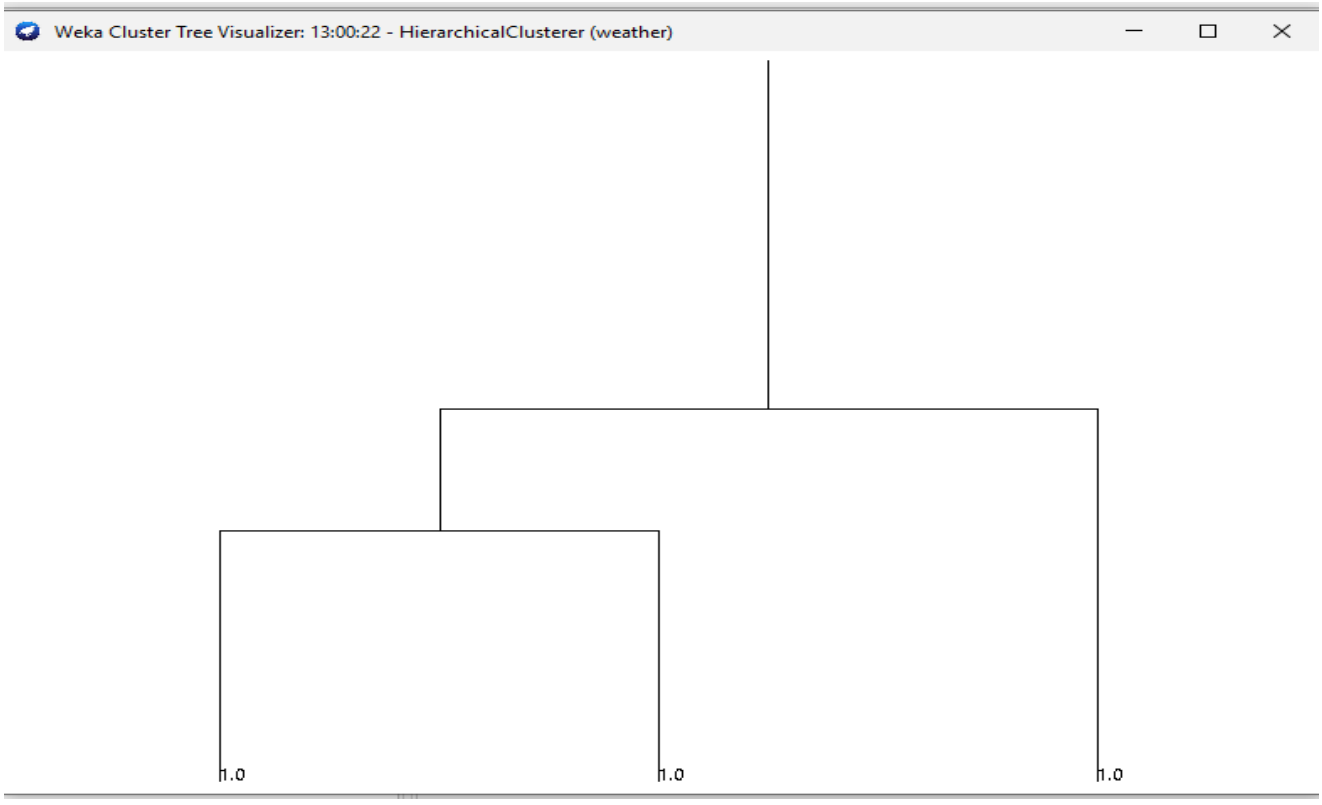
Evaluation: Hierarchical clustering does not have a clear objective function like k-means. Therefore, evaluating the quality of the clustering can be subjective. Visual inspection of dendrograms, silhouette scores, or other domain-specific criteria can be used to assess the clustering quality.

Procedure:

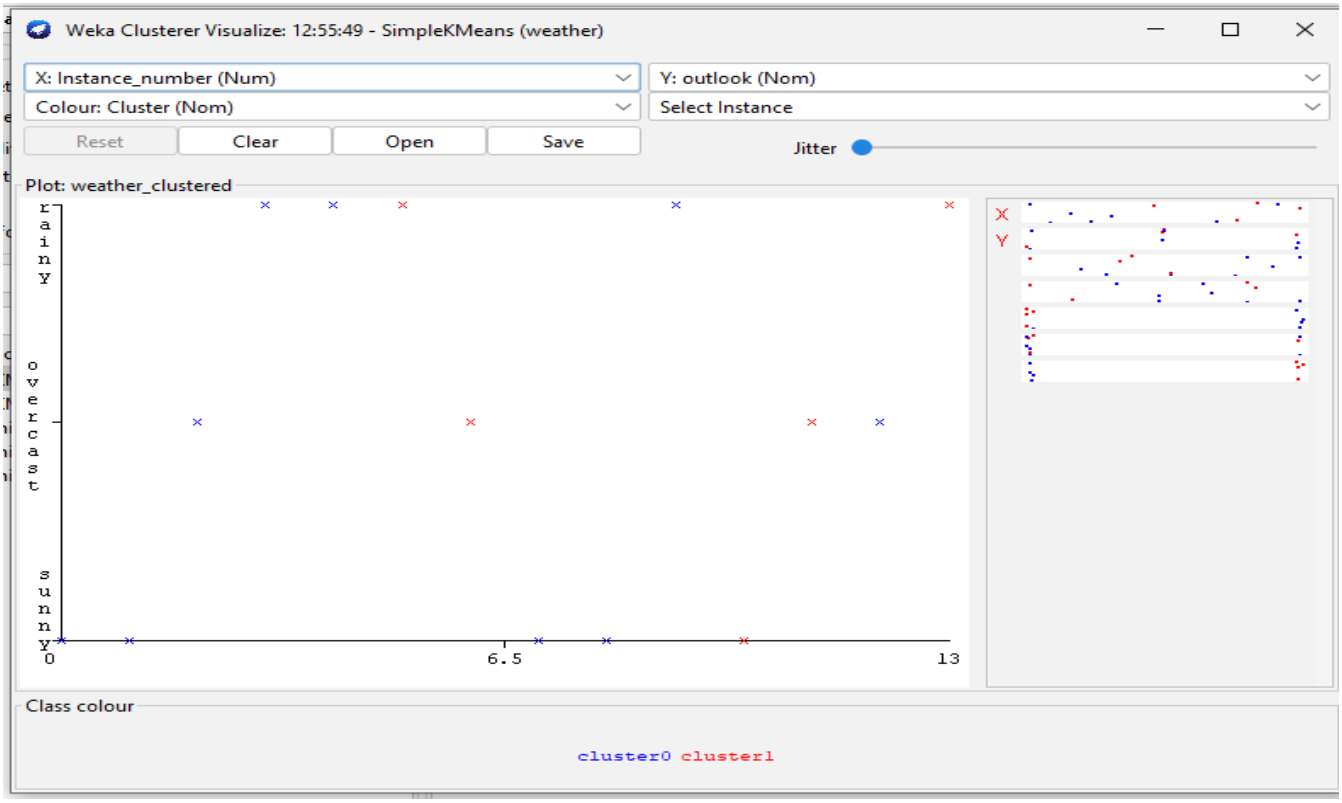
OUTPUT (SCREENSHOTS):

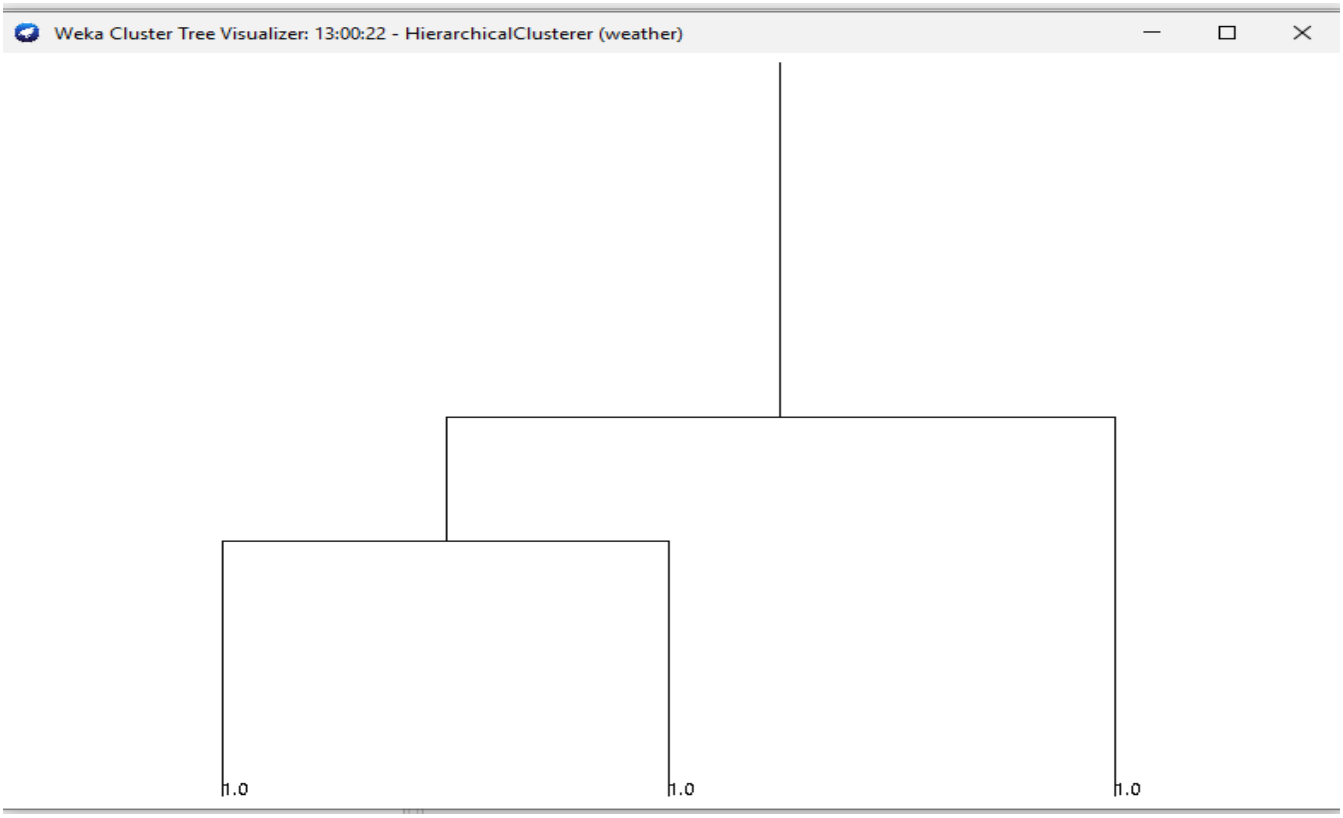
Value=5





Value=2


Weka Clusterer Visualize: 12:55:49 - SimpleKMeans (weather)
X: Instance_number (Num)
Y: outlook (Nom)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: weather_clustered
rainy
overcast
sunny
0
6.5
13
X
Y
Class colour
cluster0 cluster1


Weka Cluster Tree Visualizer: 13:00:22 - HierarchicalClusterer (weather)
1.0
1.0
1.0

CONCLUSION:

DISCUSSION AND VIVA VOCE:

- What is the objective of the k-means clustering algorithm?
- Describe the process of updating cluster centroids in the k-means algorithm.
- What are some limitations of the k-means algorithm?
- What are the main differences between agglomerative and divisive hierarchical clustering approaches?
- How does the choice of linkage method affect the clustering result?
- What are the limitations of hierarchical clustering?

REFERENCE:

- https://www.datacamp.com/
- https://towardsdatascience.com/
- Data Mining – Concepts and Techniques, Jiawei Han & Micheline Kamber, Morgan Kauf-mann Publishers, Elsevier, 2nd Edition, 2006.

Observation book: (3)	Viva-Voce (3)	`Quality of Submission and timely Evaluation (4)
Total:		**Sign with date:**