**Student Name: Dedeepya Yarra (11594115)**

**Project Title: K-Means Clustering using RapidMiner**

**Clustering:**

Clustering, also known as cluster analysis, is a method that groups entities according to the similarities they share. Defined as a type of unsupervised learning problem in which the goal is to generate training data using a given set of inputs but without any target values, the word problem comes from the word "unsupervised." The process of finding similar structures within a set of unlabeled data to make the data set more understandable and manipulable is known as data mining. It does this by separating the available heterogeneous datasets into subgroups in such a way that each individual cluster is more like each other than the overall set. To put it another way, these clusters are groups of objects that are like one another but are distinct from the objects found in other clusters. In the process of clustering, the machine learns the characteristics and patterns all by itself, without any input-output mapping being provided. The clustering algorithms first extract patterns and inferences from the different types of data objects, and then cluster those data objects into discrete categories according to how well they fit together.

We have various type of clustering techniques, they are:

1. (Hierarchical clustering) Connectivity-based-Clustering
2. (Partitioning methods) Centroids-based-Clustering
3. Distribution-based-Clustering
4. (Model-based methods), Density-based-Clustering
5. (Supervised Clustering). Fuzzy-Clustering, Constraint-based

**How Clustering Techniques were used:**

Clustering has a wide range of applications in a variety of industries. The following are some examples of common clustering applications::

* Market-segmentation
* Medical-imaging
* social-network-analysis
* image-segmentation
* search-result-grouping
* anomaly-detection

After the clustering process, a number that is referred to as a cluster ID is given to each cluster. You are now able to reduce the entire example set for an example into the cluster-ID for that example. The power of clustering lies in its ability to reduce a complicated example to a straightforward cluster ID. Taking this concept further, clustering data can help to reduce the complexity of large datasets.

You could, for instance, classify items according to their various characteristics, as illustrated by the following examples:

• Sort the stars according to their luminosity.

• Organize the documents based on the subject.

• Organize the documents based on the subject.

After that, machine learning systems can make use of cluster IDs to make the processing of large datasets more straightforward. As a result, the output of clustering is used as feature data for subsequent machine learning systems.

**A generalization:**when some examples within a cluster lack feature data, you can infer the data that is missing from other examples within the cluster.

To provide better video recommendations, for instance, less popular videos could be grouped together with more popular videos**.**

**Data Compression:** As was previously mentioned, the feature data for each and every example contained within a cluster may be substituted with the pertinent cluster ID. This replacement makes the feature data simpler while also saving storage space. When applied to large datasets, these benefits reveal themselves to be significant. Additionally, machine learning systems are able to use the cluster ID instead of the complete dataset as the input for the system. The complexity of the input data can be reduced to make the machine learning model easier and quicker for training**.**

**Example:**Data about a single video’s feature could include the following: information about viewers' locations, viewing times, and demographics; information about comments, including timestamps, text, and user IDs

**Privacy Protection:** If you cluster users and associate user-data with cluster-IDs rather than specific users,you can protect users' privacy without compromising their experience. It is necessary for the cluster to contain enough users for there to be no way to connect the user data to a particular user.

Take, for instance, the scenario in which we want to incorporate the history of YouTube users into model. You can group users into clusters and rely on the ID of the cluster instead of the user ID. Now, your model is only able to combine the history with a cluster ID rather than an individual user because the video history cannot be associated with an individual user**.**

**References:**

[1] Tang, N., & Engelbrecht, A. (2022). *Data Clustering*. IntechOpen.

[2] Olson, R. A. (2012). *Comparing Clustering Algorithms for Use with Genomic and Proteomic Data*. BiblioScholar.

**Select Attribute:** This Operator takes an Example Set's Attributes, selects a subset of those Attributes, and then removes the remaining Attributes. During the process, we used the select attribute to select the reviews column from the given dataset in order to identify words contained within it.

**Process Documents from Data:** Produces word vectors based on the attributes of string objects. It processes the data and generates word vectors based on the string attributes that are provided as input.

**K-Means Clustering (Mention the k value you used):** Clustering is accomplished with this operator by utilizing the kernel k-means algorithm. The k-means algorithm is a type of clustering known as an exclusive clustering algorithm. This means that each object is only ever placed in exactly one of a set of clusters. The items that are grouped together in the same cluster have a lot in common. A comparison of two objects' distances from one another is used to determine how closely they are related to one another.

**Cluster Distance Performance:** The Cluster Distance Performance operator uses the centroid cluster model and the clustered set as inputs to evaluate the model's performance based on the cluster centroids. The average distance traveled within a cluster and the Davies-Bouldin index are two performance measures that can be used.

**Final Model:**

Graphical user interface, diagram, application

Description automatically generated

Section2: Final Analysis

Centroid table Image:

A screenshot of a computer

Description automatically generated

Folder View Image:

Graphical user interface, text, application

Description automatically generated

From the cluster analysis we can say that there are

556 items in cluster 0, 316 items in cluster 1, 448 items in cluster 2, 180items in cluster 3

Total 1500 items divided into 4 clusters.

Cluster output statistics Image:

A screenshot of a computer

Description automatically generated

Pie chart cluster Image: for word abdominal describing cluster distances

Chart, pie chart

Description automatically generated

1. By sorting with cluster\_0 in decreasing order I have seen that the series words are more related to medication weakness strength

By sorting with cluster\_1 in decreasing order I have seen that the series words are more related to diseases like migraines headaches

By sorting with cluster\_2 in decreasing order I have seen that the series words are more related to symptoms of diseases like termors, anxiety shake .. et.c

By sorting with cluster\_3 in decreasing order I have seen that the series words are more related to intensity of diseases like abdominal, abnormal, accomplish, accustomed.. et.c

4.Cluster 0 – medication and persons strength for different diseases

Cluster 1 – different kinds of diseases

Cluster 2- symptoms of different diseases

Cluster 3- Intensity of different diseases

5.

Before when k=4 image:

Graphical user interface, text, application

Description automatically generated

After when k= 9 Image:

Graphical user interface, text, application

Description automatically generated