NAME: JOYADDAR MD JOBAYER

MATRIC NO:1731833

EXERCISE CLUSTERING

**Answer of number 1**

Clustering is a data mining technique for grouping data components into similar categories. That is the process of grouping data (or objects) into the same class. Those in one cluster are more similar to each other than data in another. Clustering is the process of dividing data items into subclasses. A cluster is made up of data objects that have a high level of inter similarity but a low level of intra similarity. Clustering has the advantage of being adaptive to changes and assisting in the identification of relevant qualities that separate various groupings.

**Answer of number 2**

Clustering is a technique for evaluating data that does not include pre-labeled groups. The notion of maximizing inters - class similarity while decreasing similarity across classes is used to group data instances together. This means that the clustering method will locate and group examples that are quite similar to one another, as opposed to ungrouped instances that are very dissimilar. Clustering is a type of unsupervised learning since it does not involve the pre-labeling of classes.

**Answer of number 3**

**1.** Cluster analysis is frequently utilized in market research, pattern identification, data analysis, and image processing, among other applications. Clustering can assist marketers identify unique groups in their consumer bases and describe them based on purchase behaviors in the business world.

**2.** Clustering can also aid in the identification of similar land use areas in an earth observation database, as well as the identification of groups of houses in a city based on house type, value, and geographic location, as well as the identification of groups of auto insurance policyholders with a high average claim **cost.**

**3.** In other applications, clustering is referred to as data segmentation since it divides big data sets into categories based on their similarities. Outlier identification may also be done via clustering, with outliers (values that are "far away" from any cluster) potentially being more interesting than usual examples. Outlier detection is used in the identification of credit card fraud and the monitoring of illegal activity in the electronic commerce industry.

**Answer of number 4:**

After download the dataset we have transformed the file to CSV, and we have created a new attribute from the quality attributes which is type of quality. The new attribute contains three types of data which are low (1-3), medium (4-7) and good (8-10). We have attached a screenshot as a reference bellow:

Table

Description automatically generated

***Figure 1: After pre-process of white wine dataset***

We have used Weka for providing a simple explanation about how good or bad the dataset is. In Weka, we have used three types of clustering algorithm to justify the quality of white wines. Clustering algorithms are as follows:

* Simple k-Means Cluster
* Density-Based Cluster
* Hierarchical Cluster
* This is the interface of Weka after opening the white wine dataset.

Graphical user interface

Description automatically generated

* **Simple k-Means Cluster**

Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

***Figure 1: Weka Explorer- classification using K-means cluster***

For k-means has three cluster which are cluster 0 stand for medium, cluster 1 stand for good and cluster 2 stand for low. As we can see in figure 1, incorrectly clustered instances 2790 or 56.962% From the mean distance of the centroid, for fixed acidity of cluster 0(medium) value is 7.0453, cluster 1 (good) value is 6.8442 and cluster 2(low) value is 6. 5626.So, we can conclude that using k-means cluster algorithm incorrectly instances is 56.962%

* **Density-Based Cluster**

Graphical user interface, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

***Figure 2: Weka Explorer- classification using Density-based Cluster***

For density has three cluster which are cluster 0 stand for medium, cluster 1 stand for good and cluster 2 stand for low. As we can see in figure 2, incorrectly clustered instances 2916 or 59.5345% From the mean distance of the centroid, for volatile acidity of cluster 0(medium) value is 0.2897, cluster 1 (good) value is 0.2914 and cluster 2(low) value is 0. 2429.So, we can conclude that using density-based cluster algorithm incorrectly instances is 59.5345%

* **Hierarchical Cluster**

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, application

Description automatically generated

***Figure 3: Weka Explorer- classification using hierarchical Cluster***

For hierarchical cluster we can see, incorrectly clustered instances only are 200 or 4.0833%

**DECISION**

After using three types of cluster algorithm (**Simple k-Means Cluster, Density-Based Cluster &** Hierarchical Cluster) we have found incorrect instances as follows:

**Simple k-Means Cluster =** 2790 or 56.962%

**Density-Based Cluster =** 2916 or 59.5345%

**Hierarchical Cluster** = 200 or 4.0833%

Therefore, hierarchical cluster accuracy is good for the white wine dataset. Because this algorithm shows only 200 incorrected cluster instances whereas, density-based cluster is the worst accuracy for the dataset, so it shows 2916 incorrectly cluster instances.

Graphical user interface, scatter chart

Description automatically generatedGraphical user interface, application, Word

Description automatically generated

***Figure 4: Distribution of Simple k-Means Cluster***

Graphical user interface

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

***Figure 5: Distribution of Density-Based Cluster***

Graphical user interface, application, scatter chart

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

***Figure 6: Distribution of Hierarchical Cluster***