**Name: Aftab Khan ID: 12985**
**Subject: Data warehouse**

## Answer (1):

## Classification:

It is the process of learning a model that elucidate different predetermined classes of data. It is a two-step process, comprised of a **learning** step and a **classification** step. In learning step, a classification model is constructed and classification step the constructed model is used to prefigure the class labels for given data.

**Clustering:** It is a technique of organizing a group of data into classes and clusters where the objects reside inside a cluster will have high similarity and the objects of two clusters would be dissimilar to each other. Here the two clusters can be considered as disjoint. The main target of clustering is to divide the whole data into multiple clusters. Unlike classification process, here the class labels of objects are not known before, and clustering pertains to unsupervised learning.

## Key Differences Between Classification and Clustering:

1) Classification is the process of classifying the data with the help of class labels. On the other hand, Clustering is similar to classification but there are no predefined class labels.

2) Classification is geared with supervised learning. As against, clustering is also known as unsupervised learning.

3) Training sample is provided in classification method while in case of clustering training data is not provided.

## Answer (2):

### Clustering:

Clustering is defined as the grouping of objects such that the objects in a group (cluster) are similar or related to one another and different from the objects in other groups.

## Difference between Density and Center-based clustering:

| Density-based Clustering | Center-based Clustering |
|---|---|
| Here, a cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. | Here, a cluster is a set of objects such that an object in a cluster is closer to the “center” of a cluster than to the center of any other cluster. |
| It is used when noise and outliers are present and when the clusters are irregular or intertwined. | The center of a cluster is the centroid, the minimizer of distances from all the points in the cluster, or a medoid, the most “representative” point of a cluster. |
| One of the most popular density-based clustering methods is DBSCAN. | One of the approximate methods that provides a formal definition as an optimization problem is Lloyd’s algorithm, which is commonly known as the k-means algorithm |

## Answer (3):

## Objective function in clustering:

The role of the objective function in clustering is to determine the quality of the cluster. Quality of cluster can be computed e.g as the compactness of the cluster. Cluster compactness can be computed as the total distance of each cluster member to cluster centroid.

**For example:** The more closer each cluster member is to the cluster center, the more compact is the cluster, so it mean better group and better cluster solution. There are a lot of objective function for clustering such as Dunn index. SSE (Sum Square of Error).

## Answer (4):

By its very essence, a data warehouse is a centralized storage of physical data. The inputs that enter the data warehouse, usually your run-the-business application databases, but also external sources, such as credit score data or product line data, are data source streams.

### Characteristics of input data:

- It offers a centralized utility of resources of organizational data and information.
- It is located within a well-managed environment.
- It has clear and replicable procedures that are specified for functional large data.
- It's also built on even flexible, open standards that can manage potential data extensions.
- It offers tools to enable its clients, without even a high level of intellectual help, to turn information efficiently into data.

- For simplistic analysis and to enhance performance, some information is denormalized.
- There are vast volumes of historical data used.
- Sometimes, queries retrieve vast quantities of information.
- It is normal to have both expected and ad hoc queries.
- The load on the data is monitored.
- **Subject Oriented:** The data warehouse becomes subject-oriented since it presents knowledge about a topic instead of the current activities of the company. Such topics can include products, consumers, suppliers, revenues, income, etc. The data warehouse doesn't really concentrate on ongoing tasks, but rather on representing data and interpretation for decision-making.
- **Integrated:** By integrating data from heterogeneous sources such as database systems, flat files, etc, a private cloud is created. The successful analysis of the data is enhanced by this incorporation.
- **Time-Variant:** A fixed time frame is used to classify the data stored in a data warehouse. From a historical standpoint, the information in a data warehouse gives data.
- **Non-volatile:** Non-volatile indicates that, as new content is loaded to it, the past work is not deleted. A data warehouse is completely separated from the database system, which ensures that the data warehouse does not represent regular changes in the database system.

## Answer (5):

## Kmeans algorithm:

Kmean algorithm is an iterative algorithm that tries to partition the data set into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

## The way kmeans algorithm works is as follows:

1) Specify number of clusters K.

2) Initialize centroids by first shuffling the data set and then randomly selecting K data points for the centroids without replacement.

3) Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

* Compute the sum of the squared distance between data points and all centroids.

* Assign each data point to the closest cluster (centroid).

* Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

The approach kmeans follows to solve the problem is called **Expectation-Maximization**.
The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster.

## Limitation:

The most important limitations of K-means are:

- The user has to specify k (the number of clusters) in the beginning.
- K-means can only handle numerical data.
- K-means assumes that we deal with spherical clusters and that each cluster has roughly equal numbers of observations.

## Answer (6):

**Scaling/normalization:** It is a process of labeling the data of an attribute so that it belongs to a similar range. In a data warehouse, it is mainly performed as the pre-processing step before manipulating or initiating actual processing.

## Formula:

x'=(x-xmin)/(xmax-xmin) where x is the data to be modified,

xmin and xmax represent the lowest and highest values in the range.

**Standardization:** It is also a scaling technique that includes the mean along with a unit standard deviation of the featured range.

## Formula:

x' = (x-mean)/std. Deviation

## Purpose:

Scaling through Normalization and Standardization is mainly utilized when the data warehouse contains attributes on the various scales. There is a possibility of losing the effectiveness of some attributes(low scale) when larger scale attributes are present in the warehouse.

Hence, all the data should belong to the same scale to avoid creating poor data models. Thus, the data should be normalized or standardized to achieve similar scale attributes. This leads to enhancing all the database operations including data storage and manipulation. Again, normalization can be achieved through three techniques like Decimal Scaling, Min-Max Normalization, and z-Score Normalization.

## Example:

**Showing Decimal Scaling method:**

## Formula:

**nor_data = data/10^j where j is the number of digits in the highest absolute data.**

Consider a data store with several data {-20, 201, -40, 50, 300}.

The highest absolute value is 300. Therefore, all the values in the data frame will be divided by 1000 to scale them.

Now, using the Decimal Scaling method, the normalized data set would be {-.020, .201, -.040, .050, .300}. Any comparison would be easier with this set of normalized data.

**Answer (7):**

## Six Steps to Data Cleaning:

### 1) Monitor Errors:

Keep a record and look at trends of where most errors are coming from, as this will make it a lot easier to identify fix the incorrect or corrupt data. This is especially important if you are integrating other solutions with your fleet management software, so that errors don't clog up the work of other departments.

### 2) Standardize Your Processes:

It's important that you standardize the point of entry and check the importance of it. By standardizing your data process you will ensure a good point of entry and reduce the risk of duplication.

### 3) Validate Accuracy:

Validate the accuracy of your data once you have cleaned your existing database. Research and invest in data tools that allow you to clean your data in real-time. Some tools now even use AI or machine learning to better test for accuracy.

### 4) Scrub for Duplicate Data:

Identify duplicates, since this will help you save time when analyzing data. This can be avoided by researching and investing in different data cleaning tools, as mentioned above, that can analyze raw data in bulk and automate the process for you.

### 5) Analyze:

After your data has been standardized, validated, and scrubbed for duplicates, use third-party sources to append it. Reliable third-party sources can capture information directly from first-party sites, then clean and compile the data to provide more complete information for business intelligence and analytics.

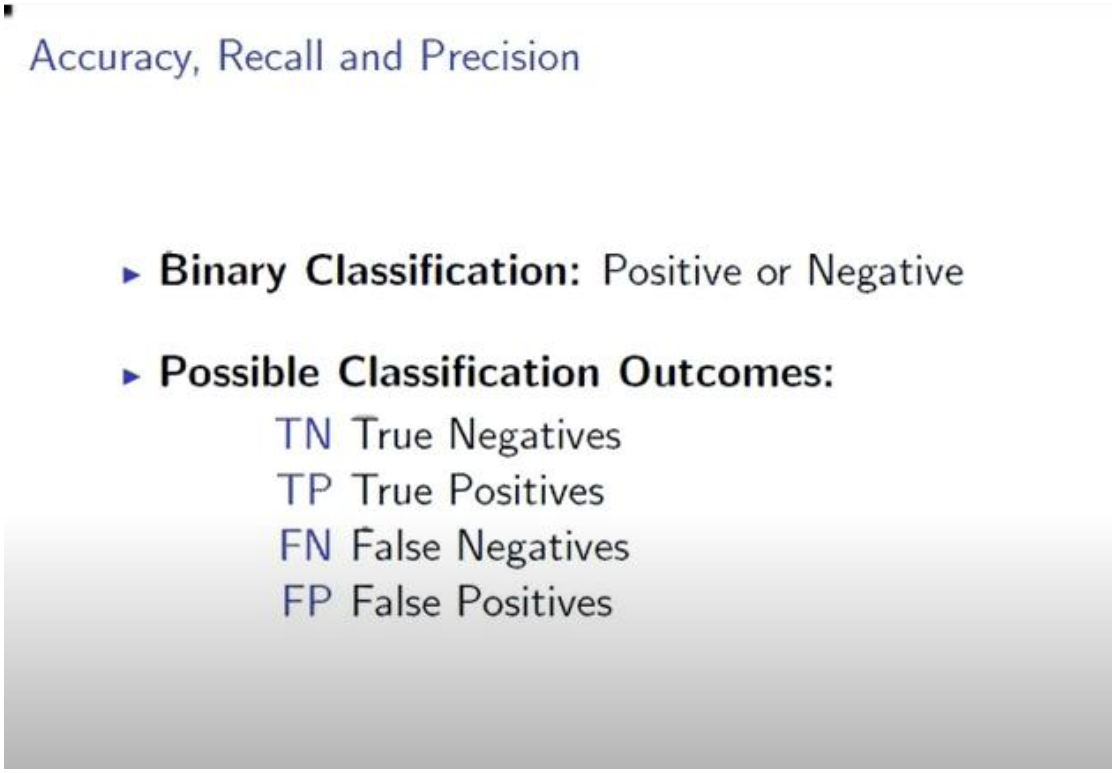
## 6) Communicate with the Team

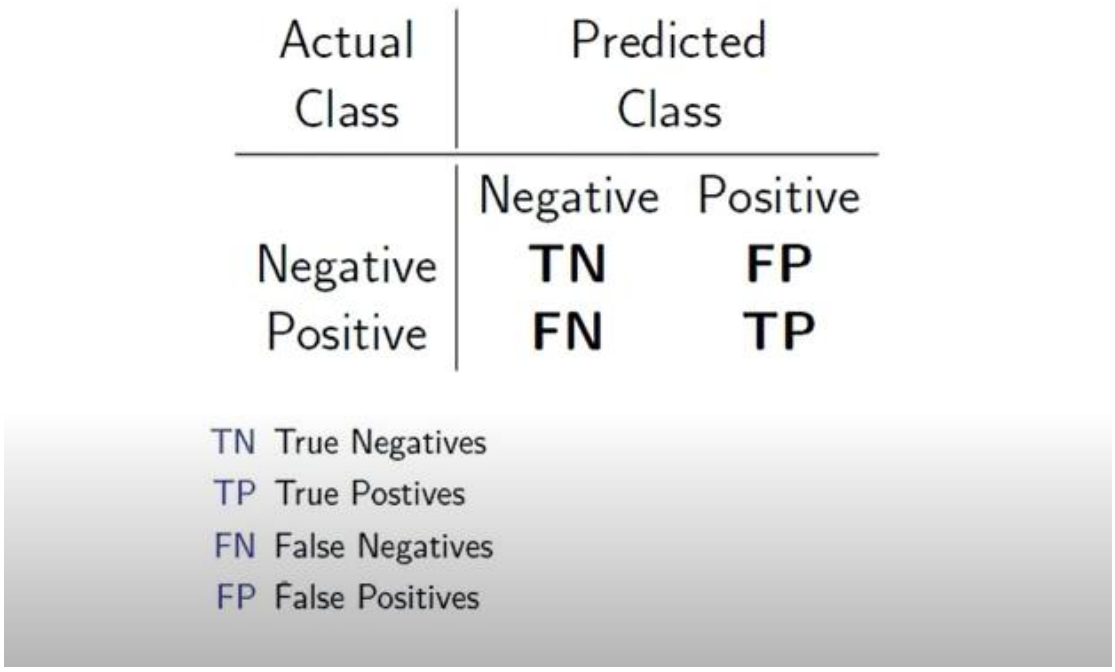
Communicate the new standardized cleaning process to your team. Now that you've scrubbed down your data, it's important to keep it clean. This will help you develop and strengthen your customer segmentation and send more targeted information to customers and prospects, so you want to make sure you get your team in line with it.

## 7) Get Your ROI from Data:

When you have the task of managing data, keeping on top of consistency and accuracy are two underlying jobs you have to deal with everyday. These steps should help make it easier to create a daily protocol. Once you have completed your data cleaning process, you can confidently move forward using the data for deep, operational insights now that your data is accurate and reliable.

**Answer(8):** Precision /recall and accuracy are explain with formulas and example in the pictures given
below.



## Key Metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Number of cases: **100,000**

| Actual State | Predicted Negative | Predicted Positive |
|---|---|---|
| Negative | TN **97750** | FP **150** |
| Positive | FN **330** | TP **1770** |

## Accuracy, Recall and Precision

**Accuracy:**
Proportion of correct classifications (**true positives and negatives**) from **overall** number of cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

## Accuracy, Recall and Precision

**Accuracy:**
Proportion of correct classifications (**true positives and negatives**) from **overall** number of cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Accuracy} = \frac{97,750 + 1,770}{100,000} = 0.9952$$

## Accuracy, Recall and Precision

**Recall:**
Proportion of correct positive classifications (**true positives**) from cases that are **actually positive**.

$$\text{Recall} = \frac{TP}{TP + FN}$$

## Accuracy, Recall and Precision

**Recall:**
Proportion of correct positive classifications (**true positives**) from cases that are **actually positive**.

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Recall} = \frac{1770}{1770 + 330} = 0.8428$$

## Accuracy, Recall and Precision

**Precision:**
Proportion of correct positive classifications (**true positives**) from cases that are **predicted as positive**.

$$\text{Precision} = \frac{TP}{TP + FP}$$

## Accuracy, Recall and Precision

**Precision:**
Proportion of correct positive classifications (**true positives**) from cases that are **predicted as positive**.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Precision} = \frac{1770}{1770 + 150} = 0.9219$$

# THE END