****

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**ODD SEMESTER (JULY 2024 - NOV 2024)**

**DATA WAREHOUSING AND DATA MINING (U20ITT511)**

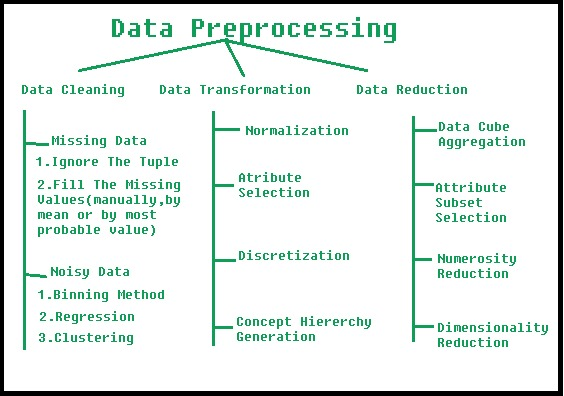
**UNIT II DATA MINING**

**Data Preprocessing :**

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

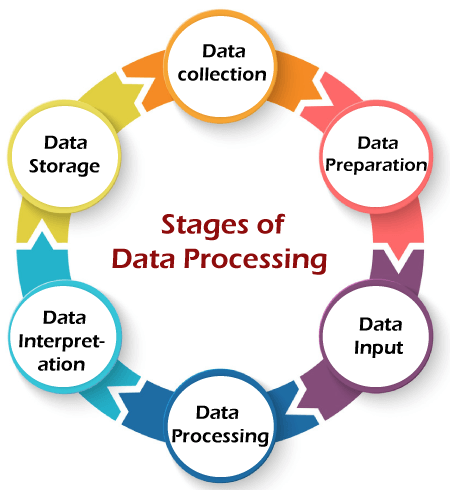
**Steps for Data Preprocessing:**

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.



Data preprocessing plays a crucial role in ensuring the quality of data and the accuracy of the analysis results. The specific steps involved in data preprocessing may vary depending on the nature of the data and the analysis goals.

By performing these steps, the data mining process becomes more efficient and the results become more accurate.



**1.Data Cleaning**

**2.Data Integration**

**3.Data Transformation**

**4.Data Reduction**

**5.Data Discretization**

**1.Data Cleaning:**

The collection of raw data is the first step of the data processing cycle. The raw data collected has a huge impact on the output produced. Hence, raw data should be gathered from defined and accurate sources so that the subsequent findings are valid and usable. Raw data can include monetary figures, website cookies, profit/loss statements of a company, user behavior, etc.

**2.Data Integration:**

Data preparation or data cleaning is the process of sorting and filtering the raw data to remove unnecessary and inaccurate data. Raw data is checked for errors, duplication, miscalculations, or missing data and transformed into a suitable form for further analysis and processing. This ensures that only the highest quality data is fed into the processing unit.

**3.Data Transformation:**

This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.

**4.Data Reduction:**

This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.

**5.Data Discretization:**

This involves dividing continuous data into discrete categories or intervals. Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.

**Data Normalization:**

This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.

**Data Cleaning**:

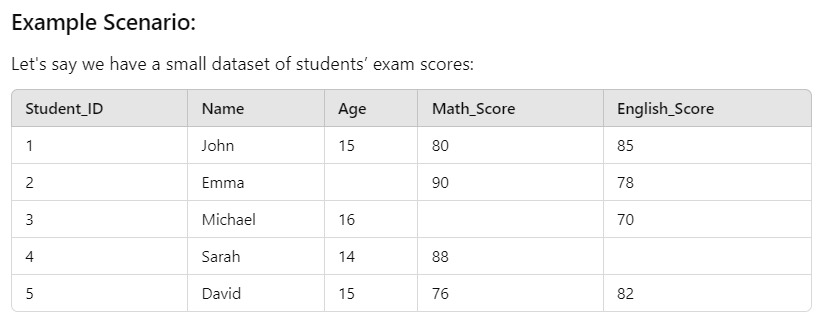
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

1. **Missing Data**
2. **Noisy Data**

**1.Missing Data:**

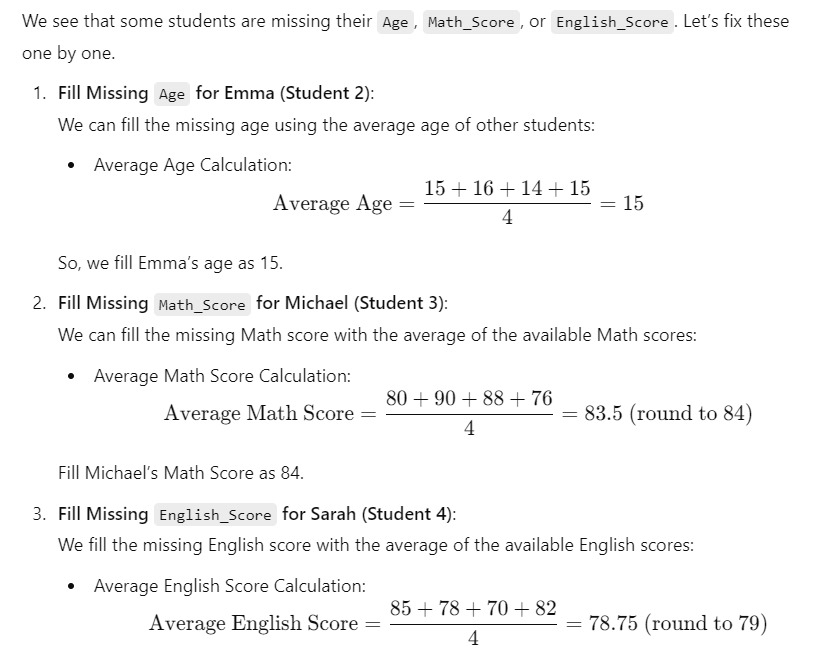
This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

* **Ignore the tuples:** This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
* **Fill the Missing values**: There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.



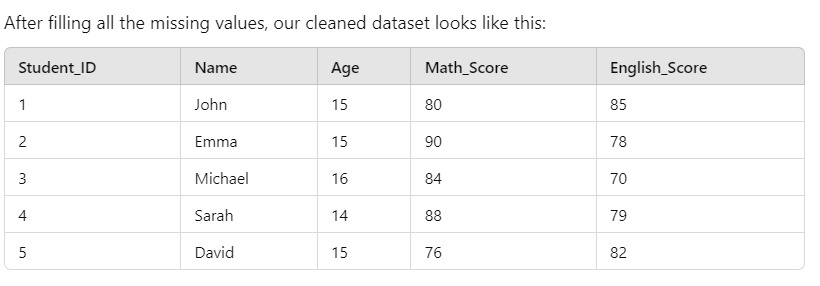
**Problem:** This data has some missing values, which means it needs to be cleaned before any meaningful analysis can be performed.

**Step 1:** Handling Missing Data



Fill Sarah’s English Score as 79.

**Step 2:** Final Cleaned Dataset:

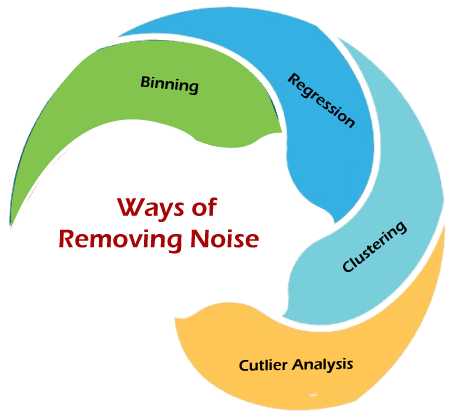


**Step 3: Why is Data Cleaning Important?**

Now that the missing values are handled, we have a complete dataset with no gaps. This is essential because:

* **Consistency:** Ensures that every student has the same type of data, making analysis easier.
* **Accuracy:** Missing values can lead to incorrect analysis results. Filling in gaps improves accuracy.
* **Readability:** Cleaned data is easier to understand and work with.

**2.Noisy Data:**



Noisy data is a meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

1. **Binning Method:** This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

* **Smoothing by bin mean method:** In this method, the values in the bin are replaced by the mean value of the bin.
* **Smoothing by bin median:** In this method, the values in the bin are replaced by the median value.
* **Smoothing by bin boundary:** In this method, the using minimum and maximum values of the bin values are taken, and the closest boundary value replaces the values.

1. **Regression** : This is used to smooth the data and help handle data when unnecessary data is present. For the analysis, purpose regression helps decide the suitable variable. ***Linear regression*** refers to finding the best line to fit between two variables so that one can be used to predict the other. ***Multiple linear regression*** involves more than two variables. Using regression to find a mathematical equation to fit into the data helps to smooth out the noise.
2. **Clustering**: This is used for finding the outliers and also in grouping the data. Clustering is generally used in unsupervised learning.

**2.Data integration:**

The process of combining data from multiple sources (databases, spreadsheets,text files) into a single dataset. Single and consistent view of data is created in this process. Major problems during data integration are Schema integration(Integrates set of data collected from various sources), Entity identification(identifying entities from different databases) and detecting and resolving data values concept.

**Data Sources:**

Data can come from various sources, including databases, spreadsheets, cloud storage, IoT devices, and APIs. In data warehousing, this may include operational databases, external datasets, and legacy systems.

**Integration Techniques:**

* **ETL (Extract, Transform, Load):** This is the most common approach used in data warehousing. It involves extracting data from source systems, transforming it to fit operational needs (e.g., cleaning, aggregating), and loading it into a data warehouse.
* **Data Federation:** This technique provides a virtual view of the data, allowing users to access data from multiple sources without physically moving it.
* **Data Warehousing:** Involves consolidating data from multiple sources into a central repository (data warehouse) for analysis and reporting.
* **Data Lakes:** For unstructured or semi-structured data, data lakes allow for storing raw data that can be processed and analyzed later.

**EXAMPLE:** Problems Faced in Data Integration:

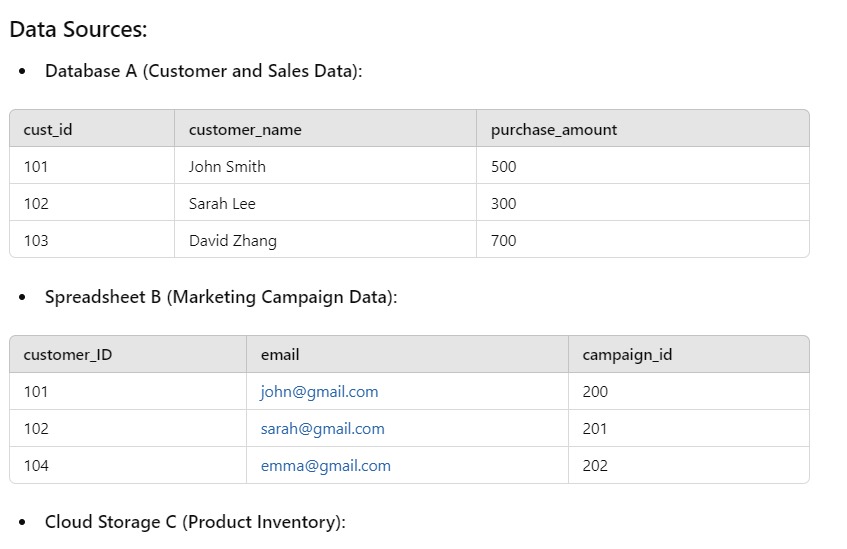
1. Schema Integration:

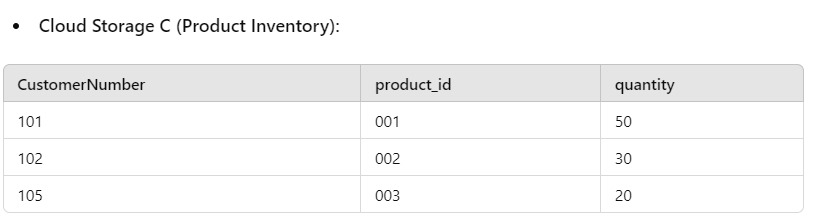
Each data source may have different structures (schemas). For example:

Database A has a column cust\_id (Customer ID) to identify customers.

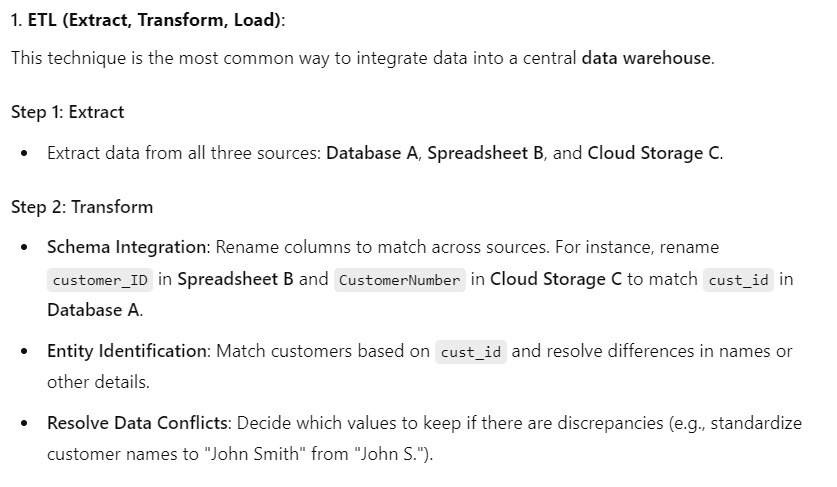
Spreadsheet B has a column customer\_ID.

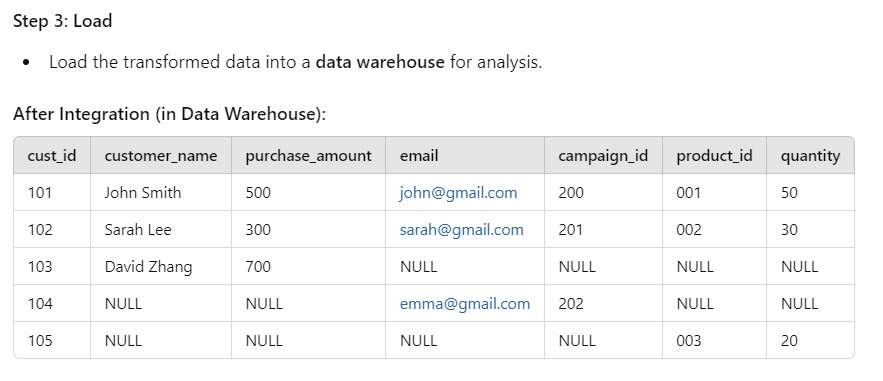
Cloud Storage C has a column CustomerNumber.





**Integration Techniques:**

****

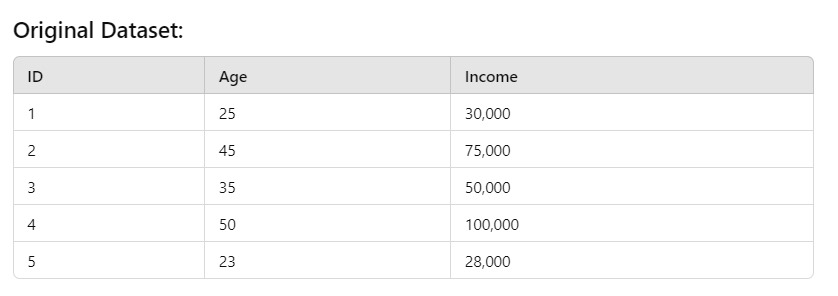
****

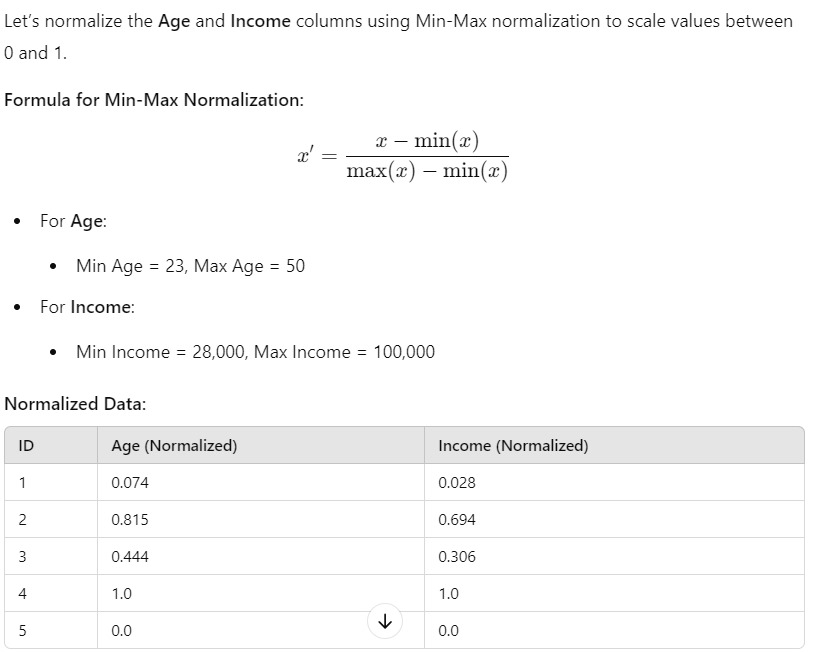
**3.Data Transformation:**

Data Transformation is the process of converting raw data into a suitable format that is easier to analyze and use in data mining or warehousing. It involves applying specific operations to the data, such as scaling, aggregating, or encoding, in order to make it compatible with the algorithms or analytical models**.**

* **Smoothing:** With the help of algorithms, we can remove noise from the dataset and helps in knowing the important features of the dataset. By smoothing we can find even a simple change that helps in prediction.
* **Aggregation**: In this method, the data is stored and presented in the form of a summary. The data set which is from multiple sources is integrated into with data analysis description. This is an important step since the accuracy of the data depends on the quantity and quality of the data. When the quality and the quantity of the data are good the results are more relevant.
* **Normalization:** It is a technique used to scale data so that it fits within a smaller, specific range, often between -1.0 to 1.0 or 0 to 1. This helps in eliminating issues where features with larger values dominate those with smaller values.

**Normalization (Min-Max Normalization):**

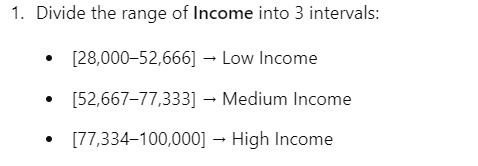


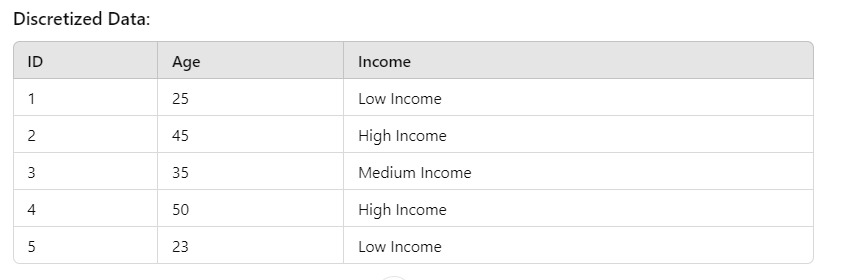


* **Discretization:** It is the process of converting continuous data into discrete intervals or categories, which reduces the data size and simplifies it for analysis. This technique is commonly used when dealing with continuous numerical data that needs to be transformed into categories for classification or association rule mining.

**Discretization (Equal-Width Discretization)**

Let’s discretize the Income column into 3 categories using equal-width discretization.





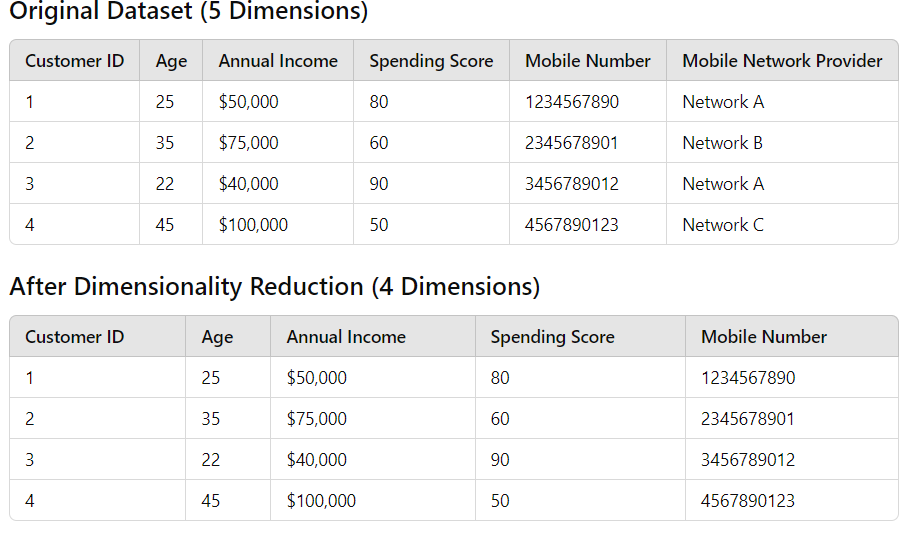
**Importance of Data Transformation:**

* **Improves Model Accuracy:** Properly transformed data ensures that models and algorithms can perform more efficiently and accurately.
* **Increases Compatibility**: Some algorithms require data in a particular format (e.g., numerical, scaled), and transformation ensures that data meets these requirements.
* **Reduces Complexity:** Simplifying data by aggregating or discretizing helps reduce the dimensionality and complexity, improving computational performance.

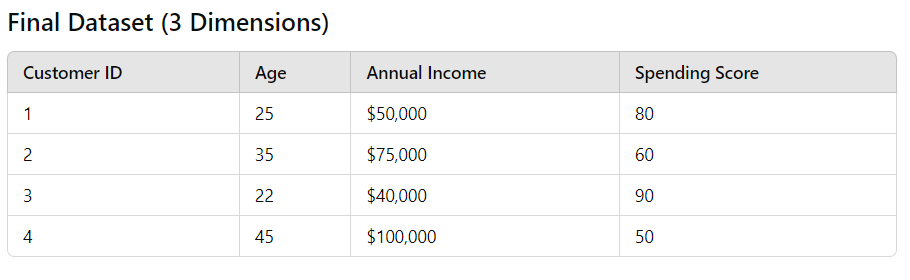
**3. Data Reduction:**

If the data is very large, data reduction is performed. Sometimes, it is also performed to find the most suitable subset of attributes from a large number of attributes. This is known as dimensionality reduction. Data reduction also involves reducing the number of attribute values and/or the number of tuples. Various data reduction techniques are:

* **Data cube aggregation:** In this technique the data is reduced by applying OLAP operations like slice, dice or rollup. It uses the smallest level necessary to solve the problem.
* **Dimensionality reduction:** The data attributes or dimensions are reduced. Not all attributes are required for data mining. The most suitable subset of attributes are selected by using techniques like forward selection, backward elimination, decision tree induction or a combination of forward selection and backward elimination.

****

**Explanation:** Since the mobile number implies the mobile network provider, we can eliminate the "Mobile Network Provider" column without losing significant information. This reduces the dimensionality of the dataset from 5 dimensions to 4.

****

**Explanation**: By removing both the Mobile Network Provider and Mobile Number, we focus only on the key variables that are more relevant for customer segmentation or prediction tasks, reducing dimensions to 3**.**

**5.Data Discretization:**

This involves dividing continuous data into discrete categories or intervals. Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.

**1. Equal-Width Discretization:**

Equal-width discretization divides the range of continuous values into equal-sized intervals. The width of each interval is determined by the overall range of the data divided by the number of intervals.

**Process:**

* **Determine Range:** Find the minimum and maximum values in the dataset.
* **Calculate Interval Width**: Divide the range by the number of desired intervals.
* **Create Intervals:** Generate intervals using the calculated width.

**2. Equal-Frequency Discretization:**

Equal-frequency discretization divides the dataset into intervals that contain an equal number of data points. This method ensures that each interval has the same number of observations, which helps in balancing the representation of different ranges.

**Process:**

* **Sort the Data**: Arrange the continuous data in ascending order.
* **Determine the Number of Intervals:** Decide how many intervals are required.
* **Divide Data Points:** Allocate data points evenly across the intervals.

**3. Clustering-Based Discretization:**

This method utilizes clustering algorithms to group similar continuous values into intervals based on their distribution. The goal is to capture natural clusters in the data, allowing for more meaningful discretization.

**Process:**

* **Choose a Clustering Algorithm:** Common algorithms include K-means, DBSCAN, or hierarchical clustering.
* **Cluster the Data:** Apply the chosen algorithm to group the continuous values.
* **Define Intervals Based on Clusters:** Create intervals that encompass the clusters.

**4. Decision Tree-Based Discretization:**

This method employs decision tree algorithms to identify optimal cut points for creating intervals based on the relationship between the continuous variable and the target variable. It aims to maximize information gain or minimize entropy.

**Process:**

* **Build a Decision Tree:** Train a decision tree on the dataset where the continuous variable is the feature.
* **Identify Cut Points**: Use the tree structure to find the optimal cut points that create the intervals.
* **Define Intervals Based on Cut Points:** Create intervals that represent the splits made by the decision tree.

**Benefits of Discretization:**

* **Improved Interpretability:** Discretized data is often easier to interpret, as it provides clear categories instead of continuous values.
* **Reduction of Noise:** Discretization can help filter out noise from the data, leading to better model performance.
* **Compatibility with Algorithms:** Some algorithms (e.g., decision trees, Naïve Bayes) work better with categorical data than with continuous data.