Data Preprocessing-Data Aggregation

Data Aggregation

Data aggregation is a process of combining multiple data values into a single value, such as sum, average, count, or minimum/maximum. It's an essential step in data preprocessing that helps to simplify and transform the data into a format suitable for analysis.

Types of Data Aggregation:

1. **Summary Functions**: Calculate a summary statistic from a group of values, e.g., `SUM`, `AVG`, `MAX`, `MIN`.
2. **Grouping Operations**: Combine multiple rows with similar characteristics, e.g., grouping by category or date.
3. **Roll-up Operations**: Collapse data to higher levels of aggregation, e.g., summing sales by region.

Example:

Suppose we have a dataset of sales transactions:

Order ID	Customer Name	Product	Quantity	Price
1	John Doe	Book	2	$10.99
2	Jane Smith	Book	3	$15.99
3	John Doe	Magazine	1	$5.99

We want to calculate the total sales for each customer and product.

Data Aggregation Steps:

1. **Group by**: Group the data by Customer Name and Product.
2. **Apply aggregation function**: Use SUM to calculate the total Quantity and Price for each group.

Aggregated Data:

Customer Name	Product	Total Quantity	Total Price
John Doe	Book	2	$21.98
Jane Smith	Book	3	$47.97
John Doe	Magazine	1	$5.99

Benefits of Data Aggregation:

1. Simplifies data analysis by reducing the number of rows.
2. Provides insights into trends and patterns in the data.
3. Helps to identify relationships between variables.

In this example, data aggregation helped us transform the original dataset into a more meaningful format for analysis, allowing us to see which customers are buying which products.

Curated by Brajesh Kumar