***DATA PREPROCESSING***

Data preprocessing is the procedure for making raw data into a suitable form, so it is ready for machine learning. it must be in an organised format to apply machine learning algorithms to get correct predictions.

Why Do We Need Data Preprocessing?

For machine learning to give correct outputs, it must go through a series of steps so the data is organised and is in a standard format. Cleaned data increases the accuracy and efficiency of the learning model.

**SUBTOPICS IN DATA PREPROCESSING**

**1.Data Cleaning:**

Data cleaning, also known as data cleansing or data wrangling, is the process of identifying, correcting, or removing inaccurate, incomplete, or irrelevant data from a dataset.

2. **Data Integration:**

Data integration is the process of combining data from different sources to provide a unified view.

3. **Data Transformation**

Data transformation is the process of converting data from one format or structure into another. This is often necessary to ensure that data from different sources can be integrated, analyzed, and used effectively. Here are some key aspects of data transformation:

1. **Normalization**: Adjusting data to eliminate redundancy and ensure consistency.
2. **Aggregation**: Summarizing detailed data into a more usable form, such as totals or averages.
3. **Data type conversion**: Converting data from one type to another, such as from text to numbers.
4.  **Filtering**: Removing irrelevant or unnecessary data.
5.  **Splitting and merging**: Dividing a dataset into smaller parts or combining multiple datasets into one.

**DATA REDUCTION:**

Data reduction is the process of reducing the amount of data that needs to be stored, managed, and analyzed without losing significant information. This can help improve the efficiency of data processing and analysis. Here are some key methods of data reduction:

1. **Dimensionality reduction**: Reducing the number of variables or features in a dataset. Techniques include:
   * Principal Component Analysis (PCA)
   * Linear Discriminant Analysis (LDA)
   * t-Distributed Stochastic Neighbor Embedding (t-SNE)
2. **Data compression**: Encoding data in a way that uses fewer bits. Common methods are:
   * Lossless compression (e.g., ZIP, PNG)
   * Lossy compression (e.g., JPEG, MP3)
3. **Aggregation**: Summarizing detailed data into a more compact form, such as calculating the average, sum, or count.
4. **Sampling**: Selecting a representative subset of the data to analyze, which can be less time-consuming than working with the entire dataset.
5. **Data filtering**: Removing irrelevant or redundant data, keeping only what's necessary for the analysis.

**DATA DISCRETIZATION:**

It involves converting continuous data into discrete bins or intervals. This can help simplify data models, reduce noise, and improve the performance of certain algorithms that work better with discrete values