Synapse – ML – Task 3.1

(Q) Let’s say you are given a large amount of textual data- messages, emails, books, etc. Before performing any operations on this data, it is necessary to clean and preprocess the data (removing unnecessary words or symbols, etc.)

Data Engineering (Introduction)

Data engineering is the process of designing and building systems to collect and analyse data to gain new insights that can transform a business. This is done with the help of raw data from multiple sources and formats. These systems empower people to find practical applications of the data which the business can utilize to thrive in dynamic conditions. It consists of three main parts, namely, data preprocessing , data analytics and data post processing.

Data Preprocessing

Data preprocessing is a **crucial** part of **data engineering and mining**. The goal of data preprocessing is to **improve the quality of the data** and make it more suitable for the task at hand. It begins with **identifying the business problem** which is to be overcome. Once **suitable data sources** have been identified, **selection of said data** for further processing begins.

Common steps involved in data preprocessing:

**Data Cleaning:**This involves **identifying and correcting errors or inconsistencies in the data such as missing values, outliers, and duplicates**.

**Data Integration:**This involves **combining data from multiple sources to create a unified dataset**. Data integration can be challenging to deal with as it involves handling data with different formats, structures, and semantics.

**Data Transformation:**This involves **converting the data into a suitable format for analysis**. Commonly used techniques in data transformation include normalization, standardization, and discretization.

**Data Reduction:**It involves **reducing the size of the dataset while preserving important information**.

**Data Discretization:**It involves **dividing continuous data into discrete categories or intervals**. Discretization is often used in data mining and machine learning algorithms that require categorical data.

**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**.

Need for Data Preprocessing

1. Establishing Data Quality Standards: Clear data quality standards including, criteria for accuracy, completeness, consistency and timeliness. The standards act as benchmarks.
2. Performing Regular Data Profiling: It involves analysing the structure, content and quality of data to identify anomalies and patterns. It helps in understanding the data characteristics, detecting issues and planning data cleaning strategies.
3. Implementing Automated Data Cleaning: Leveraging automation tools and techniques to efficiently complete the data cleaning process.
4. Document Data Cleaning Processes: Maintain comprehensive documentation of data cleaning processes and techniques employed. It enables transparency, facilititates reproducibility and supports data governance.