TASK 3.1

Data preprocessing is the critical first step in analyzing data. It lets you transform raw data into an understandable and usable format for analysis. It’s a comprehensive process that ensures the data is primed and ready for the subsequent exploration, modeling, and interpretation stages.

Common problems in data preprocessing

1. **Missing Data:** Raw data may contain missing values, which can be due to various reasons such as data entry errors or system issues.
2. **Inaccurate Data:**Raw data may contain errors, inconsistencies, or inaccuracies, which can occur due to several reasons such as human error, measurement error, or technical issues.
3. **Outliers:**Outliers are data points that deviate significantly from the majority of the data points. Outliers can occur due to measurement errors, data entry errors, or unusual circumstances.
4. **Duplicates:**Raw data may contain duplicates, which can occur due to data entry errors or technical issues.
5. **Inconsistent Data:**Raw data may contain inconsistent data, which can be due to differences in data formats or standards used in different sources.
6. **Non-standard Data:**Raw data may be in non-standard formats or data types that are not suitable for analysis.

Importance Of Data Preprocessing

* **Improves reliability and accuracy**

Preprocessing the data helps in removing unwanted and inconsistent data that results from any errors. Removing those data improves the accuracy, reliability and quality of the data set.

* **Data Preprocessing in Data Science**

provides the system with processed and reliable data. This would help to improvise the efficiency of the system and provide more accurate results.

* **Enhances Data algorithm’s readability**

Professionals with **Data Science training** use **Data Preprocessing**to improve the data to such an extent that it makes it easier for the Machine Learning algorithms to understand, analyse and process it.

* **Provides consistent data**

When the unwanted data is removed from the pile of data, you are left with the quality data that would help in providing accurate results. When the data is cleaned through **Data Preprocessing**, it helps to streamline the process. A ‘Data-driven’ decision can only prove to be beneficial if the data is properly processed before being analysed.

* **Removes duplicity**

**Data Preprocessing** helps to remove duplicate data from a dataset. If the system analyses the repeated data, the results might vary. This would affect the entire project. So, it is important to remove duplicity, which can be easily done through **Data Preprocessing**.

* **Identify and sort missing data**

When the raw and uncompiled data is preprocessed, there would be instances of missing data. **Data Preprocessing in Data Science** helps to identify missing data. It also helps to sort the missing data by providing the required data. This would help in proper data analysis.

Steps involved in Data Preprocessing

* Data profiling

Data profiling is the first step in Data Preprocessing. This involves examining, analysing, and reviewing the collected data for its quality. This step starts with identifying different data sets that are relevant to the project and then preparing the inventories for the significant attributes. This helps to form a hypothesis that would help in data analysis or making the Machine Learning task easier.

* Data cleansing

The second step of Data Preprocessing in Data Science is Data cleansing. The objective of this process is to identify an easier way to correct quality issues. This is done by removing bad data and providing the missing data. This makes the raw data suitable for Machine Learning projects.

* Data reduction

There are many bits of raw data that are not required for your computation in Machine Learning, Artificial Intelligence or any type of analytical task. This process removes the data that are not required for the particular project.

* Data transformation

Data transformation helps the Data Scientists analyse how the data can be organised in different aspects to make it more relatable to the goal. This includes structuring the unstructured data, merging the required variables and identifying the important ranges. This would help the Data Scientists to focus on the data in a proper way.

* Data Enrichment

One important step of Data Preprocessing in Data Science is data enrichment. Data Scientists use different features of engineering libraries to the raw data to get their desired transformations.

Data enrichment arranges the data to achieve the optimised balance between the training time for the new model and the time required for computation. Data Scientists spend a considerable amount of time on it so that the data received after Data enrichment would help them in their project.

* Data validation

Data validation is the last step of Data Preprocessing. In this process, the data is split into two different sets. The first set trains the Machine Learning or the Deep Learning model. The second set is used to test the data and used to analyse the accuracy and the robustness of the particular model.

Task 3.2

**Diffusion models** are a class of generative models used in machine learning to create new data samples that resemble a given dataset. Unlike traditional models that generate data directly, diffusion models operate by gradually transforming a simple noise distribution into complex data through a series of steps.

These steps can be broadly divided into two processes:

1. **Forward Process**: Start with real data and progressively add noise to it. This process gradually transforms the data into pure noise.
2. **Reverse Process**: Learn to reverse this process by training a neural network to convert noise back into data. The model learns to gradually remove noise step-by-step, reconstructing the original data from noise.
3. **Generating New Data:** Once trained, the model can start with pure noise and iteratively remove noise to generate new, realistic data that resembles the training data.

**Why is it Impressive?**

* **High-quality Output:** Diffusion models often produce highly detailed and realistic outputs, surpassing previous methods like GANs in terms of image quality.
* **Versatility:** They can be applied to various data types, including images, text, audio, and even video.
* **Controllability:** While still under development, researchers are exploring ways to control the generation process more precisely, allowing for more specific outputs.