**Synapse Task 3**

**Q. 3.1) 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.). Explain how you would go about preprocessing. What different steps would be followed? Why are they necessary?**

Ans) Data Cleaning is an essential step that we cannot miss before creating good Machine Learning models. This is because machine learning uses data and program to create an algorithm which in turn can be used on other datasets to yield desired output. If the data is not up to the mark, we cannot expect good algorithms. Let’s take an example. Consider the data collected from an online form for selections in an interview and our goal is to predict the chances he/she will get placed. The primary data needed from a student would be: name, resume, mobile number, email-id, address, overall CGPA, and in case of too many applicants, father’s name, mother’s name, github id, linkedIn id and D.O.B. But what if the applicants don’t fill D.O.B section, or the applicants apply twice using two different accounts, or instead of writing name in the format “first-name last-name” he/she ends up writing “first-name” or “first-name father-name last-name”. What if the student applied for the form and never sat for the interview? These can create change in the prediction of the model and reduce the accuracy of the same. It’s like finding the factorial of 10 but we miss out 5, or add 1 twice or thrice, or add 9 instead of 6, leading to output with less accuracy. That’s why data cleaning comes into picture. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Dataset can contain useless or missing data not only because of user’s mistake, but also due to different sources of data. Collecting a Data from an interviewee can be through forms, company registries any many more.

Now, how can we go about cleaning the data? There is no specific way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset, but considering the above example, we can do it the following way:

1. Remove duplicate or irrelevant observations: Irrelevant data are those which has no relation with our aimed prediction. For Example: We don’t need parent’s name
2. Fix structural errors:
3. Filter unwanted outliers:
4. Handle missing data:

**Q. 3.2) Imagine using a random prompt like "a cat riding a bicycle on Mars" and seeing an AI generate an image that matches your description perfectly. This is made possible by using advanced models like DALL-E, which use various machine learning techniques, including the diffusion processes. For this task, explain what basic diffusion is, how it works, and why it is used in generating such impressive output.**

**OR**

**Have you ever wondered how streaming platforms like Netflix work and how they recommend movies or shows based on your current watch? How does a bank decide which customers get loans and which do not? This all is done using Unsupervised learning. Machine Learning is internally subdivided into different parts- one of them is Unsupervised learning. The technique used for these kinds of problems is known as Clustering. So, for this task, explain what clustering is and describe any two types of clustering.**

**OR**

**You must have heard of or used ChatGPT at some point in the last year, maybe even before. It's like Janet from the Good Place, it always has the answers to your question. Unlike Janet, this may not be 100% correct but does get the job done. But it wasn’t always like this. Earlier stages of ChatGPT used to give bogus answers all the time, but over time it learnt how to give appropriate answers, relevant to the context, being as accurate as it possibly can. This is done through a method called Reinforcement Learning. For this task, you must understand and explain the working of reinforcement learning. Additionally, list some other examples and explain how they work.**

Ans)