Better ways to clean the data (task 3.1)

The large amount of data seems like a messy cupboard, which feels like a burden. But once it is arranged properly it feels good and things can be accessed easily.

Similarly, when the data is properly managed , dealing with data becomes easy.

Data is something through which we can obtain information relevant to our use.

So one should try to arrange the data in the proper format so that one can easily use the data.

Step-by-Step Data Cleaning and Preprocessing:

- Importing Data:
  Load the dataset from the secondary memory(disk) into primary memory(RAM) for faster processing. This can be done using libraries like Pandas in Python.
- Data Structuring:
  Use appropriate data structure to organize the data efficiently. Data Frames in Pandas are well-suited for this purpose as they allow easy manipulation and visualization.
- Data Formatting:
  Ensure that the data is arranged in a tabular format with consistent data types for each column. Convert columns to the appropriate data types if necessary.
- Handling Missing Values:
  Identify and handle missing values. You can either remove rows with missing values or impute them using appropriate methods (mean, median, mode, etc.).
- Removing Unnecessary Columns:
  Remove columns that are not needed for analysis to reduce complexity and improve performance.
- Filtering Data:
  Use filtering to focus on specific subsets of the data. The filter() function can be used to work on particular datasets based on conditions.
- Sorting Data:
  Sort the data to identify the highest or lowest values in a particular column.
- .Data Visualization:
  Use visualization libraries like Matplotlib etc to visualize the data and understand patterns and distributions.

Task 3.2: Machine Learning solution to our daily problems

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.

Machine Learning:

The key idea is that the model improves its performance over time as it is exposed to more data. It is divided into two sub categories supervised and unsupervised learning.

Supervised Learning:

Supervised Learning has no target variable. It is a type of machine learning where the model is trained using a dataset so that the model can predict the output for new, unseen inputs.

Unsupervised Learning:

Unsupervised learning comes in picture when there is no target variable , it's basic goal is to discover the group of similar examples within the data.

Example: Clustering

Now what is clustering?

Clustering data mining techniques group data elements into clusters that share common characteristics. We can cluster data pieces into categories by simply identifying one or more attributes. Some of the well-known clustering techniques are k-means clustering, hierarchial clustering and Gaussian mixture models.

Explaining the clustering techniques:

1. K-Means Clustering:
   K-Means clustering is a popular and widely used unsupervised learning algorithm that partitions a dataset into K distinct, non-overlapping subsets or clusters. The goal is to group data points into clusters such that those within the same cluster are more similar to each other than to those in other clusters.

2. Hierarchical Clustering:

   Hierarchical clustering is a method of clustering data points in a way that builds a hierarchy or tree of clusters. It is particularly useful when you want to see the nested structure of the data, where smaller clusters are part of larger clusters..

   Yes I have definitely wondered how streaming platforms like Netflix work and how they recommend movies or shows based on our current watch.
   At that verry instant it feels magical but now I have basic idea what’s happening at backend.Basically users are grouped into K clusters based on their preferences, and recommendations are made by suggesting popular items within the user's cluster.