* **Task 3.1-**

Different steps I would follow are –

1. Acquiring the dataset-

Data collection is the first step in any machine learning project. The collected data for a particular problem in a proper format is known as Dataset. Dataset may be of different format for different purposes. To use the dataset in our code we usually put it into a CSV(comma-seperated values) file format.

1. Importing the libraries-

Library is a collection of functions that an algorithm can call and utilize. In order to perform data preprocessing using python we need to import some predefined python libraries like Numpy(used for including any type of mathematical operation in the code), Matplotlib(used to plot any type of charts for the code) and Pandas(used for importing and managing datasets).

1. Importing the datasets-

Before importing a dataset, we need to set the current directory as the working directory. To import the dataset, we use the read\_csv() function of panda’s library.

1. Check for missing values-

By employing techniques such as removal, the gap in information is effectively mitigated. Either the row containing the missing data is removed or it is estimated to the value of mean.

1. Encoding the data-

Machine learning modules cannot understand non-numeric data. Hence, we have to arrange the data in numerical form to prevent any problems later. For this, the LabelEncoder() function is used.

1. Scaling-

Scaling is a technique that can convert data values into shorter ranges. Rescaling and standardization can be used for scaling.

1. Splitting the dataset into training, evaluation and validation sets-

The final step is to distribute data in three different sets namely – training set(to train the data), validation set(to validate data) and evaluation set(to evaluate the data).

* **Task 3.2-**

Clustering-

Clustering is a popular type of analysis under unsupervised learning. The task of grouping data points based on their similarity with each other is called Clustering or Cluster Analysis. It aims at gaining insights from unlabelled data points and don’t have a target variable. It is commonly used for statistical data analysis. After applying the clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets.

Types of Clustering algorithms-

There are many types of clustering algorithms like-

1.Hierarchical Method-

It is an alternative approach to partitioning. It does not require prespecifying the number of clusters to be generated. The result of hierarchical clustering is a tree-based representation of objects known as dendogram. There are two approaches as bottom-up approach and top-down approach. In bottom-up approach, we start with each object forming a separate group. Then it keeps merging objects or groups that are close to each other. In top-bottom approach, we start with all of the objects in the same cluster. In continuous iteration, a cluster split up into smaller clusters. It is done until the termination condition holds.

2.Centroid-based clustering –

These are also called Partitioning methods. They group data points on the basis of their closeness. It will classify the data into k groups, each group containing at least one object. The partitioning method will create an initial partitioning. Then uses the iterative relocation technique to improve partitioning by moving objects. The general criterion of a good partitioning is that objects in same cluster are close to each other whereas in different clusters are far from each other.