Task 3.1

1.The first step would be to remove unnecessary punctuations from the text to simplify the text and to lowercase the text. (str.strip(punctuation), str.lower())

2.The next step would be to remove a few numbers wherever unnecessary, and also remove some special characters. (str.replace())

3.Then we can split the text into individual words or phrases for further processing.

4.We could eliminate words such as a, an, the, etc. wherever irrelevant.

5.Remove the unnecessary blank and white spaces. (str.strip())

6.Handle the contractions like don’t becomes do not(for efficient Noiranalysis).

7.Take care of the emojis in the text, use them only if relevant or necessary.

8.Normalize the text.

These things are necessary to ensure consistency, reduce noise, simplify analysis and improve accuracy.

Task 3.2

Machine Learning can be divided into supervised and unsupervised learning. Supervised learning requires labelled data whereas unsupervised learning involves training without labeled responses. The goal of unsupervised learning is to discover hidden patterns, groupings or features in the hidden data. Common algorithms include k-means clustering, hierarchical clustering,etc.

Clustering in simple terms is grouping similar data points together. The similiarity is often measured using distance metrices. It does not require labelled data. This algorithm identifies patterns and structures with the data by itself. Clustering is a powerful technique for data analysis and exploration.

K-Means Clustering- K-means clustering is a popular and simple clustering algorithm used to partition a dataset into a specified number of clusters. It partitions the data into ‘k’ clusters where each data point belongs to the cluster with the nearest mean. The number of clusters ‘k’ is specified by the user before running the algorithm. Each cluster is defined by a centre point known as the centroid(mean). Assign each data point to the nearest centroid based on a distance metric. Use techniques like K-means++ to select better starting points.

K-means is easy to implement and computationally efiicient. It works with large datasets.

Applications- Used in grouping customers according to their purchases.

Grouping similar documents or texts.

Image compression- Reducing the no of colors in an image.

Hierarchical Clustering- Hierarchical clustering is a method of cluster analysis that builds a hierarchy of clusters. Unlike K-means clustering, which requires you to specify the number of clusters in advance, hierarchical clustering creates a nested structure of clusters that can be visualized as a tree-like diagram called a dendrogram.

Agglomerative Hierarchical Clustering- It follows a Bottom-Up Approach. It begins with each data point as a separate cluster and then merges the closest clusters based on a distance metric until all points are in a single cluster.

Divisive Hierarchical Clustering- It follows a Top-Down Approach. It begins with all data points in a single cluster and then divides the cluster into smaller clusters until each point is in its own cluster. It is less commonly used than agglomerative cluster due to its higher computational complexity.

In hierarchical clustering, you don’t need to specify the number of clusters in advance.

Applications- Used in gene expression analysis.

Used in customer segmentation.

Used to organize documents into topics or categories.

Used in Image Segmentation and object recognition.