Sec-A

Assignment-4

1. What do you mean by clustering in machine learning?

ans Clustering is an unsupervised learning technique in machine learning used to group similar data points together based on their features. It helps in pattern discovery and segmentation without predefined labels.

2. What is the role of of feature selection in machine learning?

ans feature selection involves choosing the most relevant features from the dataset to improve model accuracy, reduce overfitting and shorten training time.

3. What is the decision Tree algorithm?

ans A Decision Tree is a supervised learning algorithm used for classification and regression. It splits data into branches based on feature values, forming a tree structure that leads to decision outcomes.

4. What is Apache Spark

ans Apache Spark MLlib is a scalable machine learning library built on Spark. It provides

distributed algorithms and tools for classification, regression, clustering, and more, making it suitable for handling Big Data.

Sec - B

1. How does forest improve classification tasks? Explain with an example.

ans. Random forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and reduce overdrifting. For example, in a customer churn prediction, multiple trees consider different subsets of features and samples. The final decision is based on the majority vote, leading to more reliable predictions.

2. What are the key steps in evaluating a machine learning model?

ans key steps include:

- Splitting data into training and testing sets.
- Choosing appropriate metrics (e.g. accuracy, precision, recall).
- Performing cross-validation.

- Analyzing confusion matrix.
- Fine-tuning hyperparameters.
- Validating with unseen data to check generalization.

Sec-C

1. Discuss the roll of Distributed computing in machine learning.

ans Distributed Computing allows machine learning tasks to be Performed across multiple nodes, enabling Parallel data Processing faster training, and handling of large datasets. It is essential for Big Data environments where traditional single-machine systems are insufficient.

2. How does Apache Spark support Scalable machine learning?

ans Apache Spark Supports Scalable machine learning through:
- In-memory computation for speed.
- MLlib for distributed algorithms.
- Resilient Distributed Datasets (RDDs) for fault tolerance.

- Integration with Hadoop and various data sources.
- Scalability across clusters for large datasets.