**TEST**

**Machine Learning**

**1.Can you explain the difference between supervised and unsupervised learning?**

**Supervised learning** : sample labeled data are provided to the machine learning system for training, and the system then predicts the output based on the training data.The mapping of the input data to the output data is the objective of supervised learning.In supervised learning, sample labeled data are provided to the machine learning system for training, and the system then predicts the output based on the training data. Ex: Spam filtering

**Unsupervised learning** : a learning method in which a machine learns without any supervision.The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

**2.What is overfitting in machine learning, and how can it be prevented?**

Overfitting occurs when the model is very complex and fits the training data very closely. This will result in poor generalization of the model. This means the model performs well on training data but it will not be able to predict accurate outcomes for new, unseen data. Feature selection, Cross validation and Regularization is used to prevent overfitting.

**3.Explain the bias-variance tradeoff in model training.**

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time.To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.An optimal balance of bias and variance would never overfit or underfit the model.

**4.What is cross-validation, and why is it important in machine learning?**

Cross validation is a technique for evaluating a machine learning model and testing its performance. Cross validation is commonly used in applied ML tasks. It helps to compare and select an appropriate model for the specific predictive modelling problem.Cross validation is easy to understand, easy to implement, and it tends to have a lower bias than other methods used to count the model’s efficiency scores.

**5)Describe the concept of feature engineering and its significance in model building.**

Feature engineering is the process of creating new features or transforming existing ones in a dataset to improve the performance of machine learning models.It involves missing value handling, selecting necessary data, or manipulating raw data features to make them more informative and relevant for the model to learn from.It improves model's performance and [preventing over/under fitting.

**Spark**

**1)What is PySpark, and how does it relate to Apache Spark?**

PySpark is the Python API for Apache Spark, an open source, distributed computing framework and set of libraries for real-time, large-scale data processing. Apache Spark is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.

**2).Explain the advantages of using PySpark for big data processing.**

* Fast processing
* In-memory computing
* Fault tolerance
* Batch & real time processing
* Multilanguage support
* Less lines of code
* Distributive in nature

**3)Describe the main components of PySpark's architecture.**

PySpark's architecture comprises several key components like driver, spark context, cluster mmanager, work nodes

1. Driver-The driver program is the main control node that manages the overall execution of the PySpark application.It initializes the SparkContext where spark context/session is the application code
2. Cluster manager-The cluster manager is responsible for allocating resources (CPU, memory) and managing the nodes in the Spark cluster.spark cluster manager is based on YARN architecture.
3. Worker node-Each worker node runs executor processes that execute tasks as directed by the SparkContext.

**4)What is an RDD, and how does it differ from traditional data structures in Python?**

Resilient Distributed Dataset (RDD) is the fundamental data structure of Spark.It is the first API of spark. They are immutable Distributed collections of objects of any type. it has features of Fault-tolerance ,parallel processing ,lazy evaulator and type safety.

**5)Explain the concept of lineage in RDDs.**

Transformation are operation applied on RDD. But these operations don’t execute immediately instead they build a sequence of operations that will execute only ehen action is invoked.

**6)How can you create an RDD in PySpark?**

RDDs can created by applying transformations to existing RDDs. For transformation operations like map(), filter() creates a new RDD.

**7)Difference btw Transformation and Actions. Explain with example**

Transformation : These are operations performed on RDD that creates a new RDD but they don't trigger computation immediately; instead, they build a lineage of transformations. Ex: filter()

Action: These are operations that return a value after running a computation on an RDD. Ex: groupby()