DSA Laboratory (21AD62)-Viva Answers

1) What is Data Science?

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It involves a combination of mathematics, statistics, computer science, and domain-specific knowledge to analyze and interpret complex data.

2) Types of Data Science

Data science encompasses various techniques and methods, primarily categorized into:

- **Descriptive Analytics**: Summarizing historical data to understand what has happened.
- **Diagnostic Analytics**: Examining data to understand why something happened.
- **Predictive Analytics**: Using models to predict future outcomes based on historical data.
- **Prescriptive Analytics**: Recommending actions based on predictions to influence future outcomes.

3) Explain Supervised, Unsupervised, and Reinforcement Learning

- **Supervised Learning**: A type of machine learning where the model is trained on labelled data. The goal is to learn a mapping from input to output (e.g., regression, classification).
- **Unsupervised Learning**: The model is trained on unlabelled data and aims to find hidden patterns or intrinsic structures in the input data (e.g., clustering, association).
- **Reinforcement Learning**: An agent learns to make decisions by performing actions and receiving rewards or penalties. The goal is to maximize the cumulative reward.

4) Applications of Types of Data Science

- **Supervised Learning**: Email spam detection, image recognition, medical diagnosis.
- **Unsupervised Learning**: Customer segmentation, anomaly detection, recommendation systems.
- **Reinforcement Learning**: Robotics, game playing, autonomous vehicles.

5) What is the NumPy Library?

NumPy is a fundamental package for scientific computing with Python. It provides support for arrays, matrices, and a large collection of mathematical functions to operate on these data structures.

6) What is the pandas Library?

Pandas is an open-source data analysis and manipulation library for Python. It provides data structures like DataFrames and Series, which make it easier to handle and analyze large datasets.

7) What is the Matplotlib Library?

Matplotlib is a plotting library for Python that provides a MATLAB-like interface. It is used for creating static, interactive, and animated visualizations in Python.

8) Name 3 Types of Visualization

- **Scatterplot**: Used to show the relationship between two variables.
- **Histogram**: Used to represent the distribution of a dataset.
- **Bar Chart**: Used to compare different groups or categories.

9) What is Logistic Regression?

Logistic regression is a statistical method for modeling the relationship between a dependent binary variable and one or more independent variables. It is used for binary classification problems and provides probabilities for the classes.

10) Explain SVM Classifier

Support Vector Machine (SVM) is a supervised learning algorithm that finds the hyperplane that best separates data points of different classes in a high-dimensional space. It is effective in high-dimensional spaces and used for classification and regression tasks.

11) Explain Decision Tree

A decision tree is a supervised learning algorithm used for classification and regression. It models decisions and their possible consequences, representing them in a tree-like graph of decisions and their possible outcomes.

12) Explain Clustering

Clustering is an unsupervised learning technique used to group similar data points into clusters. The goal is to partition data into homogeneous subsets where items in each subset are more similar to each other than to those in other subsets (e.g., k-means clustering).

13) Explain scikit-learn (sklearn) Library

Scikit-learn is an open-source machine learning library for Python. It provides simple and efficient tools for data mining and data analysis, including classification, regression, clustering, and dimensionality reduction algorithms.

14) How Do You Split Train and Test Dataset?

You can split a dataset into training and testing sets using the train_test_split function from scikit-learn:

python

```
{from sklearn.model_selection import train_test_split


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42) }
```

This function splits the data into training and testing sets with a specified test size.

15) Types of Supervised Algorithms

- **Linear Regression**: Predicts a continuous dependent variable based on independent variables.
- **Logistic Regression**: Used for binary classification problems.
- **Support Vector Machines (SVM)**: Finds the optimal hyperplane to classify data points.
- **Decision Trees**: Uses a tree structure to make decisions based on feature values.
- **Random Forest**: An ensemble of decision trees to improve prediction accuracy.
- **k-Nearest Neighbors (k-NN)**: Classifies data points based on the majority class of their k-nearest neighbors.

16) Difference Between Regression and Classification

- **Regression**: Predicts a continuous outcome. Examples include predicting house prices or temperature.
- **Classification**: Predicts a discrete label or category. Examples include email spam detection or image classification.