Tech Mahindra Data Science Interview Questions

1. How to deploy a model on the cloud?

Ans. Deploying a model on the cloud typically involves:

- **Selecting a cloud provider (e.g., AWS, Azure, GCP).**
- **Packaging the model with its dependencies.**
- **Creating a Docker container if needed.**
- **Using cloud services (e.g., AWS SageMaker, Azure ML) to deploy and manage the model.**

2. Azure vs. AWS:

Ans.

- **Azure: Strong integration with Microsoft products, better for enterprises already using Microsoft services, offers a wide range of AI and machine learning tools.**
- **AWS: Market leader with a vast array of services, strong ecosystem for machine learning (e.g., SageMaker), extensive global infrastructure.**

3. Dockerfile vs. Docker Compose YAML:

Ans.

- **Dockerfile: A script containing a series of instructions on how to build a Docker image.**
- **Docker Compose YAML: A configuration file for defining and running multi-container Docker applications, managing multiple Docker containers.**

4. Pipeline from code development to model deployment:

Ans.

- Code Development: Write and test your code locally.
- Version Control: Use Git to manage code versions.
- Continuous Integration: Use CI tools (e.g., Jenkins) to run tests.
- Containerization: Package the application using Docker.
- Deployment: Deploy the container to the cloud using services like Kubernetes or cloud-specific offerings.

5. What is Jenkins pipeline?

Ans. Jenkins Pipeline: An automated process for building, testing, and deploying code, defined as code using the Groovy-based DSL.

6. How do you use Jenkins to automate CI/CD?

Ans. By setting up Jenkins pipelines to automate the build, test, and deployment processes, integrating with version control, and configuring various stages of the pipeline.

7. When will a model be redeployed?

Ans. A model will be redeployed when:

- The model is retrained with new data.
- Improvements or updates are made to the model.
- There are changes in the production environment or dependencies.
- Performance issues or bugs are detected in the current deployment.

8. How does AWS handle multiple models deployed in production simultaneously?

Ans. AWS handles multiple models in production using services like:

- AWS SageMaker Endpoints: Create multiple endpoints for different models.

- Load Balancing: Use Elastic Load Balancing to distribute traffic among multiple models.
- Container Orchestration: Use ECS or EKS to manage multiple model containers efficiently.
- Monitoring and Scaling: Use CloudWatch for monitoring and auto-scaling groups for automatic scaling based on demand.

SQL Basic Queries

9. How do you write a query to select all columns from a table named employees?

Ans. SELECT * FROM employees;

10. How do you filter records in the employees table where the salary is greater than 50000?

Ans. SELECT * FROM employees WHERE salary > 50000;

11. How do you find the average salary of employees grouped by department?

Ans. SQL Code

```
SELECT department, AVG(salary) AS average_salary
FROM employees
GROUP BY department;
```

12. How do you join two tables, employees and departments, on the department_id column?

Ans. SQL code

```
SELECT e.*, d.department_name
FROM employees e
JOIN departments d ON e.department_id = d.department_id;
```

Unnecessary Definition from SQL

13. What is an unnecessary definition in SQL?

Ans. An unnecessary definition in SQL refers to defining redundant or non-essential elements in a query, such as selecting columns that are not used, using subqueries when a JOIN is sufficient, or including complex expressions that do not enhance the query's functionality or performance.

Linear and Logistic Regression

14. What is Linear Regression?

Ans. Linear Regression is a supervised learning algorithm used for predicting a continuous dependent variable based on one or more independent variables by fitting a linear relationship between them.

15. What is Logistic Regression?

Ans. Logistic Regression is a supervised learning algorithm used for binary classification tasks. It predicts the probability of a binary outcome using a logistic function to model the relationship between the dependent variable and one or more independent variables.

Decision Tree and Random Forest

16. What is a Decision Tree?

Ans. A Decision Tree is a supervised learning algorithm used for both classification and regression tasks. It splits the data into subsets based on the value of input features, creating a tree-like model of decisions.

17. What is a Random Forest?

Ans. A Random Forest is an ensemble learning method that combines multiple decision trees to improve the model's accuracy and prevent overfitting. It creates a 'forest' of random decision trees and aggregates their predictions.

Naive Bayes Theory

18. What is Naive Bayes Theory?

Ans. Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem. It assumes independence between the features and calculates the probability of each class based on the input features, selecting the class with the highest probability.

How to Handle Null Values in a Dataset

19. How do you handle null values in a dataset?

Ans. Handling null values can be done in several ways:

- **Removal: Remove rows or columns with null values.**
- **Imputation: Fill null values with a specific value like the mean, median, mode, or a fixed value.**
- **Prediction: Use predictive models to estimate and replace null values.**
- **Flagging: Create a separate binary feature indicating the presence of null values.**

20. What are your favorite libraries in Python?

Ans. One of my favorite libraries in Python is Pandas. It's great for data manipulation and analysis. With Pandas, you can easily read data from different file formats, clean it, and perform various operations to get insights quickly. Another favorite is NumPy, which is essential for numerical computations and handling arrays efficiently. Lastly, I really like Matplotlib for creating visualizations; it makes it easy to generate plots and charts to understand data better. For example, I often use Pandas to clean my datasets, NumPy to perform calculations, and Matplotlib to visualize the results.

22. What is the difference between Logistic and Linear Regression?

Ans.

Linear Regression:

- **Purpose: Predicts a continuous outcome based on input variables.**

- **Output: Gives a straight-line prediction (like predicting house prices).**
- **Equation: Uses a simple linear equation to find a relationship between variables.**
- **Use: Best for tasks where the result is a number.**

Logistic Regression:

- **Purpose: Predicts the probability of a categorical outcome.**
- **Output: Provides probabilities that map to binary outcomes (like yes/no or spam/not spam).**
- **Equation: Uses a logistic function to model the probability.**
- **Use: Ideal for tasks where the result is a category.**

Key Differences:

- **Output Type: Linear regression predicts numbers; logistic regression predicts probabilities.**
- **Application: Linear regression for numbers, logistic regression for categories.**

In essence, linear regression predicts numbers (like house prices), while logistic regression predicts probabilities and maps them to categories (like yes/no).

23. What is the difference between data science and big data?

Ans.

Data Science:

- **Focus: Using data to solve problems and make decisions.**
- **Tasks: Analyzing, visualizing, and modeling data to find insights.**
- **Tools: Statistics, machine learning, Python/R programming.**

Big Data:

- **Focus: Dealing with large and complex datasets.**
- **Characteristics: High volume, speed, and variety of data.**

- **Challenges: Storing, managing, and analyzing massive amounts of data efficiently.**

Key Differences:

- **Focus: Data science analyzes data; big data handles large datasets.**
- **Tools: Data science uses stats and ML; big data uses special tech for storage and processing.**

In essence, data science applies tools to analyze data for insights, while big data deals with storing and managing large, varied datasets efficiently.

24. What is Word2Vec?

Ans. Word2Vec is a technique in natural language processing (NLP) used to convert words into vectors of numerical values. It captures semantic relationships between words based on their contexts in large datasets.

25. What is TF-IDF?

Ans. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents. It reflects how frequently a term appears in a document adjusted by how often it appears across all documents.

26. How do you use regex to remove special characters from a text?

Ans. Python code:

```
import re

text = "Hello! This is a sample text with @special characters #included."
clean_text = re.sub(r'[^a-zA-Z0-9\s]', '', text)
print(clean_text)
```

Explanation:

- **r'[^a-zA-Z0-9\s]': This regex pattern matches any character that is not alphanumeric (a-z, A-Z, 0-9) or whitespace (\s).**

- **re.sub(r'[^a-zA-Z0-9\s]', '', text): This function call replaces all characters matching the pattern with an empty string, effectively removing them from the text.**