Oral Questions & Answers of DSBDA

By Aniii…

- **Data Wrangling 1:**

1. What are the different Python libraries?

Ans:

NumPy: Numerical Python. Provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.

pandas: Data manipulation and analysis library. Offers data structures like DataFrames and Series, along with functions to clean, manipulate, and analyze data.

matplotlib: Plotting library for creating static, interactive, and animated visualizations in Python. Often used for creating charts, histograms, scatter plots, etc.

Seaborn: Built on top of matplotlib, Seaborn provides a high-level interface for drawing attractive and informative statistical graphics.

scikit-learn: Machine learning library featuring various algorithms for classification, regression, clustering, dimensionality reduction, and more

SciPy: Scientific Python library that builds on NumPy and provides a large number of mathematical algorithms and functions for optimization, integration, interpolation, and more.

TensorFlow: An open-source machine learning framework developed by Google for building and training neural network model

2. What is data wrangling. Define data wrangling?

Ans:

Data Wrangling is a technique that is executed at the time of making an interactive model. In other words, it is used to convert the raw data into the format that is convenient for the consumption of data.

3. What are the different steps involved in the data wrangling process?

Ans:

Data wrangling involves acquiring, cleaning, transforming, integrating, reducing, validating, enriching, and documenting data to prepare it for analysis.

4. How you are going to deal with the missing values?

Ans:

Data Inspection: Start by identifying missing values in the dataset. Use functions like .info() or .isna() in Python's pandas library to inspect the presence of missing values in each column.

Imputation: One common approach is to fill missing values with a substitute. This can be done by replacing missing values with a measure of central tendency such as the mean, median, or mode of the column

5. How your going to deal with the outliers or noisy data?

Ans:

Detection: Identify outliers using statistical methods or visualization.

Removal: Consider removing outliers, but cautiously to avoid loss of valuable information.

Transformation: Apply mathematical functions to mitigate the effect of outliers.

6. What are the different data prepossessing techniques?

Ans:

Data cleaning involves identifying and correcting errors or inconsistencies in data.

Data integration combines data from different sources into a unified view.

Data transformation converts data from one format or structure to another.

Data reduction reduces the volume of data while retaining its meaningful information.

7. What is the difference between data normalization and the data standardization.?

Ans:

Data Normalisation:

The primary goal of data normalization is to rescale the features of the data to have similar ranges and distributions.

Normalization typically involves scaling numeric features to a specific range, often between 0 and 1 or -1 and 1

Data Standarization:

Data standardization, often referred to simply as standardization, is the process of transforming data in such a way that it has a mean of 0 and a standard deviation of 1.

Standardization involves subtracting the mean of each feature from the data and then dividing by the standard deviation of the feature.

8. How category variables are converted into quantitative variables?

Ans:

Label Encoding: Assigns each unique category a numerical label.

One-Hot Encoding: Converts categories into binary vectors where each category has its own column.

Dummy Coding: Similar to one-hot encoding but uses one fewer binary column per categorical variable to avoid multicollinearity.

Feature Hashing (Hash Encoding): Converts categories into fixed-length vectors using a hash function.

Target Encoding (Mean Encoding): Replaces each category with the mean of the target variable for that category.

9. What is data science and the applications of the data science?

Ans:

Data science is an interdisciplinary field that uses scientific methods, algorithms, processes, and systems to extract knowledge and insights from structured and unstructured data.

business Analytics: Utilizing data to analyze business performance, customer behavior, market trends, and optimize operations and strategies.

Healthcare Analytics: Analyzing medical records, patient data, and clinical trials to improve patient care, diagnosis, treatment plans, and healthcare operations.

10. What is the life cycle of data science?

Ans:

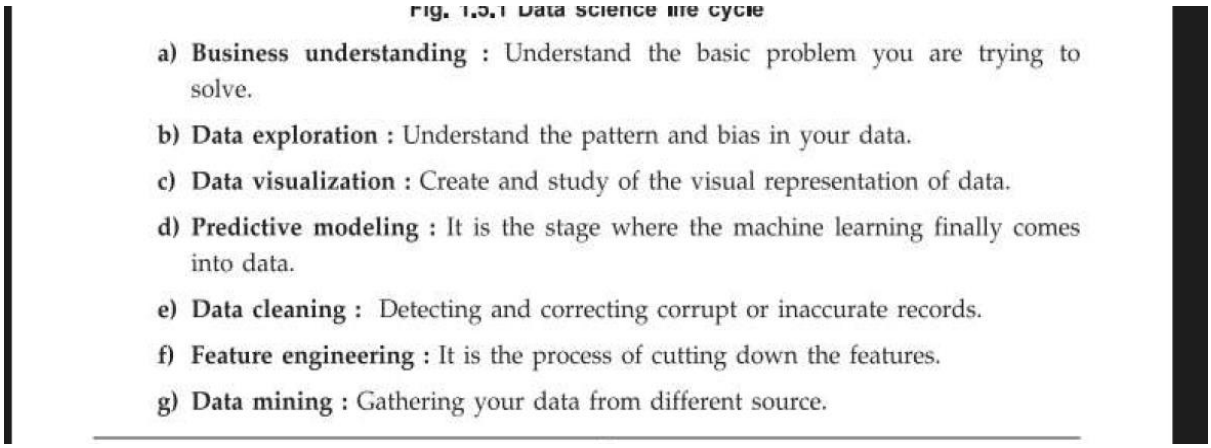

Fig. 1.5.1 Data science life cycle

a) **Business understanding :** Understand the basic problem you are trying to solve.

b) **Data exploration :** Understand the pattern and bias in your data.

c) **Data visualization :** Create and study of the visual representation of data.

d) **Predictive modeling :** It is the stage where the machine learning finally comes into data.

e) **Data cleaning :** Detecting and correcting corrupt or inaccurate records.

f) **Feature engineering :** It is the process of cutting down the features.

g) **Data mining :** Gathering your data from different source.

- **Data Wrangling II**

1. How to create csv file in Excel?

Ans:

i. Open Excel: Launch Microsoft Excel on your computer.
ii. Enter Data: Enter the data you want to save as a CSV file into the Excel worksheet. You can organize your data into rows and columns as needed.
iii. Save As: Once you have entered the data, click on the "File" menu in the top-left corner of the Excel window.
iv. Choose Save As: From the dropdown menu, select "Save As". This will open the "Save As" dialog box.
v. Select File Format: In the "Save As" dialog box, navigate to the folder where you want to save the CSV file. Then, in the "Save as type" dropdown menu, choose "CSV (Comma delimited) (*.csv)".
vi. Provide a File Name: Enter a name for your CSV file in the "File name" field. Make sure to include the ".csv" file extension at the end of the file name.
vii. Save: Click the "Save" button to save the Excel file as a CSV file.
viii. Confirmation: Excel may display a warning message about the features that aren't supported in the CSV format. Click "Yes" or "OK" to confirm and save the file as CSV.

2. How to fill the place of missing values?

Ans:

Fill missing values by replacing them with the mean/median/mode, forward/backward filling, linear interpolation, KNN imputation, regression imputation, multiple imputation, or domain-specific methods**.**

3. What are the different data transformation techniques are there?

Ans:

Normalization

Standardization

Feature Scaling

One-Hot Encoding

Label Encoding

4. What is skewness?

Ans:

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.

- **Descriptive Statistics**

1. What are the different measures of Central tendencies

Ans:

Mean

Median

Mode

2. How you are going to calculate mean, median, minimum, maximum, standard deviationvalues.

Ans:

Mean: Add up all the values in the dataset and divide by the total number of values.

Median: Arrange the values in ascending order and find the middle value. If there is an even number of values, take the average of the two middle values.

Minimum: Identify the smallest value in the dataset.

Maximum: Identify the largest value in the dataset.

Standard Deviation:

Calculate the mean of the dataset.

Subtract the mean from each value, square the result, and sum all squared differences.

Divide the sum by the total number of values.

Take the square root of the result.

3. What is hypothesis testing?

Ans:

Hypothesis testing is a statistical method to assess whether sample data provides enough evidence to reject or not reject a null hypothesis about a population parameter.

4. What is chi square test and T test?

Ans:

The chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables in a contingency table.

The t-test is a statistical test used to determine whether there is a significant difference between the means of two groups in a sample, typically when the sample size is small and the population standard deviation is unknown.

4. What is difference between inferential statistics and descriptive statistics?

Ans:

Descriptive statistics involves summarizing and describing the characteristics of a dataset, such as measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), and visualizations (histograms, box plots).

Inferential statistics involves making inferences or predictions about a population based on sample data, typically through hypothesis testing, confidence intervals, and regression analysis, among other methods. It allows us to draw conclusions and make generalizations about a population from which the sample was drawn.

- **Data Analytics**

1What is regression analysis what are the types of regression analysis.

Ans:

Regression analysis is a statistical method used to model the relationship between one or more independent variables and a dependent variable. Types include linear regression, logistic regression, polynomial regression, ridge regression, lasso regression, etc.

2. Compare linear regression and the Logistic regression give example of each.

Ans:

Linear regression is used to model the relationship between a dependent variable and one or more independent variables with a linear equation. Example: Predicting house prices based on features like size and number of bedrooms.

3. What are the data Analytics types.

Ans:

Data analytics types include descriptive analytics (summarizing data), diagnostic analytics (identifying patterns and relationships), predictive analytics (making predictions), and prescriptive analytics (providing recommendations).

4. What is the significance and importance of each Python library

Ans:

Python libraries like NumPy (for numerical computing), pandas (for data manipulation), Matplotlib (for data visualization), scikit-learn (for machine learning), and TensorFlow (for deep learning) are significant for their roles in data analysis, manipulation, visualization, machine learning, and deep learning tasks, respectively.

5. What is ridge and the lasso regression.

Ans:

Ridge and lasso regression are regularization techniques used to prevent overfitting in linear regression models by penalizing large coefficients.

Ridge regression adds a penalty term proportional to the squared magnitude of coefficients.

Lasso regression adds a penalty term proportional to the absolute magnitude of coefficients.

- **Data Analytics 2:**

1. What is classification in data science?

Ans :

Classification in data science is a supervised learning technique where the goal is to categorize data into predefined classes or labels based on its features.

2. What are the different algorithms used as classification algorithms?

Ans:

Different algorithms used for classification include Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Naive Bayes, and Neural Networks.

3. What is the use of the naive Bayes theorem?

Ans:

The Naive Bayes theorem is used for classification tasks, particularly in text classification and spam filtering. It calculates the probability of a hypothesis given the evidence using Bayes' theorem, assuming independence between features.

4. What is decision tree.

Ans:

A decision tree is a flowchart-like structure where each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. It's used for classification and regression tasks.

5. What is a confusion matrix? Draw the confusion matrix.

Ans:

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It's a 2x2 matrix for binary classification problems.

	Predicted Negative	Predicted Positive
Actual Negative	True Negative	False Positive
Actual Positive	False Negative	True Positive

6. What is the formula of accuracy error rate precision recall by using the configuration confusion method?

Ans:

The formulas for accuracy, error rate, precision, and recall using the confusion matrix are:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Error Rate = (FP + FN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall (Sensitivity) = TP / (TP + FN)

- Data Analytics 3:

1. Explain Naive Bayes classification algorithm.

Ans:

Naive Bayes classification algorithm: Naive Bayes is a probabilistic classifier based on Bayes' theorem with an assumption of independence between features. It calculates the probability of each class given a set of features and selects the class with the highest probability as the prediction.

2. What is confusion Matrix draw the confusion matrix.

Ans:

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It's a 2x2 matrix for binary classification problems

	Predicted Negative	Predicted Positive
Actual Negative	True Negative	False Positive
Actual Positive	False Negative	True Positive

3. What is the formula of accuracy error rate precision recall by using configuration confusion method.

Ans:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Error Rate = (FP + FN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall (Sensitivity) = TP / (TP + FN)

- Text Analytics:

1. What is tokenisation, stopword removal, stemming, limitization.

Ans:

Tokenization: Breaking text into individual words or tokens.

Stopword removal: Eliminating common words like "the", "and", "is" which carry little meaning.

Stemming: Reducing words to their root form, like "running" to "run".

Lemmatization: Similar to stemming but ensures resulting words are valid lemmas, like reducing "running" to "run".

2. How term frequency and Inverse document frequency is calculated.

Ans:

Term frequency (TF) is the count of a term in a document divided by the total number of terms in the document. Inverse Document Frequency (IDF) is the logarithm of the total number of documents divided by the number of documents containing the term.

3. What is text prepossessing techniques.

Ans:

Text preprocessing techniques include tokenization, stopword removal, stemming, lemmatization, lowercasing, removing punctuation, and handling special characters.

4. What is bag of word.

Ans:

A bag of words (BoW) is a representation of text data where each document is represented as a collection of words, ignoring grammar and word order but keeping track of word frequency.

- **Data Visualization I,II,III**

1. What are the challenges of big data visualization.

Ans:

The challenges of big data visualization include handling large volumes of data efficiently, ensuring scalability, dealing with high dimensionality, maintaining interactivity, and addressing data quality and variability issues.

2. What are the types

Ans:

Types of big data visualization include exploratory visualization for data exploration and analysis, explanatory visualization for communicating insights to others, and predictive visualization for presenting predictive models and future trends.

3. what are the different data visualization techniques.

Ans:

Different data visualization techniques include scatter plots, bar charts, line charts, pie charts, histograms, heatmaps, box plots, bubble charts, treemaps, network diagrams, and geographic maps.

4. what are the different plots.

Ans:

Different plots commonly used in data visualization include scatter plots, bar plots, line plots, pie charts, histograms, box plots, violin plots, and heatmap plots.

5. what is difference between seaborn library and matplotlib Library

Ans:

Seaborn is built on top of Matplotlib and provides a higher-level interface for creating attractive statistical graphics. It offers more aesthetically pleasing default settings and simplifies the process of creating complex visualizations compared to Matplotlib.