# 📚 DSBDA Oral Que Ans

1) What is Pandas library in Python?

Pandas is a powerful Python library used for data manipulation and analysis. It provides data structures like Series and DataFrame.

2) List some key features of Pandas.

Fast and efficient DataFrame object, handling of missing data, data alignment, powerful group-by functionality, and easy file I/O.

3) What is Numpy Library in Python?

NumPy is a Python library used for numerical computations, offering support for large multi-dimensional arrays and matrices.

4) What is matplotlib library?

Matplotlib is a Python plotting library used for 2D graphics, charts, and plots.

5) What is the difference between seaborn and matplotlib?

Seaborn is built on top of Matplotlib and provides better visualizations with less code, especially for statistical plots.

6) Is Sklearn and Scikit-learn the same library? What is its use in data science?

Yes, they are the same. Scikit-learn is used for machine learning, including classification, regression, clustering, etc.

7) What are functions available in Pandas and Numpy?

Pandas: read_csv(), head(), describe(), dropna(), merge().

NumPy: array(), mean(), std(), reshape(), linspace().

8) What is DataFrame in Python?

A DataFrame is a 2D labeled data structure similar to a table in SQL or Excel.

9) How to find duplicates in Python?

Use df.duplicated() or df[df.duplicated()].

10) What is the use of describe command?

It gives a statistical summary of numerical columns including count, mean, std, min, max, etc.

11) Which are Naive Bayes classification algorithms used in Python?

GaussianNB, MultinomialNB, BernoulliNB from sklearn.naive_bayes.

12) What is the significance of Confusion Matrix?

It evaluates classification models by showing true vs predicted classifications.

13) What is TP, TN, FP, FN?

TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

14) What is recall?

Recall = TP / (TP + FN)

15) What is precision?

Precision = TP / (TP + FP)

16) What is F1 score?

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

17) Why is data visualization needed in data science?

To understand patterns, trends, and insights in data visually.

18) What is an outlier?

An observation point that is distant from other observations.

19) When to use histogram and pie chart?

Histogram: to show frequency distribution. Pie chart: to show percentage composition.

20) What are challenges in Big Data Visualization?

Large volume, high velocity, variety of data, real-time visualization, and scalability.

21) What are joint plot and dist plot?

Joint plot: shows scatter + distribution. Dist plot: shows the distribution of a variable.

22) What are tools used for data visualization?

Matplotlib, Seaborn, Plotly, Power BI, Tableau.

23) What is data wrangling?

Process of cleaning and transforming raw data into a usable format.

24) What is data transformation?

Changing data format, structure, or values for analysis.

25) What is the use of StandardScaler function in Python?

It standardizes features by removing the mean and scaling to unit variance.

26) What is Hadoop?

An open-source framework for storing and processing Big Data using distributed computing.

27) What is HDFS and MapReduce?

HDFS: Hadoop Distributed File System. MapReduce: processing model for large data sets.

28) What are the components of Hadoop Ecosystem?

HDFS, MapReduce, YARN, Hive, Pig, HBase, Spark, Oozie, etc.

29) What is Scala?

A high-level programming language combining object-oriented and functional programming.

30) What are features of Scala?

Type inference, immutability, concise syntax, interoperability with Java, support for functional programming.

31) How is Scala different from Java?

Scala supports functional programming, has concise syntax, and better concurrency features.

32) Applications of Scala:

Used in Apache Spark, backend development, Big Data processing, and concurrent systems.

33) What is data science?

A field that uses scientific methods, algorithms, and systems to extract insights from data.

34) What is Big Data?

Extremely large data sets that may be analyzed computationally to reveal patterns and trends.

35) Characteristics of Big Data:

Volume, Velocity, Variety, Veracity, Value.

36) Phases in data science life cycle:

Data Collection, Data Cleaning, Data Exploration, Modeling, Evaluation, Deployment.

37) What is Central Tendency?

It refers to mean, median, and mode—measures to identify the center of a data set.

38) What is Dispersion?

The spread of data—includes range, variance, and standard deviation.

39) Calculate Mean, Mode, Mid Range, Median for: 10, 22, 13, 10, 21, 43, 77, 21, 10

Mean: 25.2, Mode: 10, Median: 21, Mid Range: (10+77)/2 = 43.5

40) What is Variance?

A measure of how far values spread from the mean.

Variance ≈ 470.96

41) What is Standard Deviation?

Square root of variance.

Std Dev ≈ 21.7

42) What is posterior probability in Naive Bayes?

Probability of a class given a feature set (using Bayes theorem).

43) What is likelihood probability in Naive Bayes?

Probability of feature set given the class.

44) How to deal with missing values?

Use dropna(), fillna(), or imputation techniques.

45) What is NLTK?

Natural Language Toolkit: a Python library for working with human language data.

46) What is Tokenization in NLP?

Breaking text into words or sentences.

47) What is Stemming?

Reducing words to their root form (e.g., "playing" → "play").

48) What is Lemmatization?

Reduces words to their dictionary form (e.g., "better" → "good").

49) What is Corpus in NLP?

A large collection of text used for training NLP models.

50) What is Spark framework?

Apache Spark is a fast, in-memory data processing engine for large-scale data analytics.

How to make CSV File ?

CSV= Comma separated values

1) Using text editor ( notepad, vs code )

2) Using Excel sheet or Google sheet

3) Using python library CSV

4) using command prompt Cmd= echo