DSBDA ORAL QUATIONS WITH ANSWERS

1) What is Pandas library in Python?
Pandas is a Python library used for data manipulation and analysis. It provides data structures like Series and DataFrame.

2) List some key features of Pandas.

- DataFrame and Series structures
- Handling missing data
- Data alignment
- Reading/writing CSV, Excel, SQL, etc.
- Grouping, filtering, merging, and reshaping data

3) What is Numpy Library in Python?
NumPy is a library for numerical computations in Python. It supports arrays, matrices, and high-level mathematical functions.

4) What is matplotlib library?
Matplotlib is a Python plotting library used to create static, interactive, and animated visualizations.

5) What is the difference between Seaborn and Matplotlib?
Seaborn is built on top of Matplotlib and offers simpler syntax and better-looking, statistical plots.

6) Is Sklearn and Scikit-Learn the same library? What is its use in data science?
Yes, sklearn is the import name for Scikit-Learn. It is used for machine learning tasks like classification, regression, clustering, etc.

7) What are functions available in Pandas and NumPy libraries?

- **Pandas**: read_csv(), head(), dropna(), groupby(), merge(), pivot_table()
- **NumPy**: array(), mean(), sum(), reshape(), arange(), linspace()

8) What is a DataFrame in Python?
A DataFrame is a 2-dimensional labeled data structure with columns of potentially different types (like a table).

9) How to find duplicates in Python?

python

CopyEdit

df.duplicated()

10) What is the use of the describe command?
It gives statistical summary like mean, std, min, max, and quartiles of numerical columns.

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

- GaussianNB
- MultinomialNB
- BernoulliNB (from sklearn.naive_bayes)
-

12) What is the significance of a Confusion Matrix?
It evaluates classification model performance by comparing predicted and actual labels.

13) What is TP, TN, FP, FN in a confusion matrix?

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative
-

14) What is Recall?
Recall = TP / (TP + FN) – Measures the ability to find all relevant cases.

15) What is Precision?
Precision = TP / (TP + FP) – Measures the accuracy of the positive predictions.

16) What is F1 Score?
F1 Score = 2 * (Precision * Recall) / (Precision + Recall) – Harmonic mean of precision and recall.

17) What is the need for data visualization in data science?
It helps to understand data trends, patterns, and outliers for better decision-making.

18) What is an outlier?
An outlier is a data point that is significantly different from other data points.

19) When to use a histogram and pie chart?

- Histogram: To show frequency distribution of numerical data
- Pie Chart: To show proportions or percentage of categories

20) What are the challenges in big data visualization?

- Handling massive volume and variety
- Performance issues
- Real-time visualization
- Scalability

21) What is a jointplot and distplot?

- jointplot(): Combines scatter plot and histograms
- distplot() (deprecated): Plots univariate distribution (use displot() now)

22) What are tools used for data visualization?
Matplotlib, Seaborn, Plotly, Power BI, Tableau, D3.js

23) What is data wrangling?
It is the process of cleaning, restructuring, and enriching raw data for analysis.

24) What is data transformation?
It involves converting data from one format/structure to another for analysis or storage.

25) What is the use of StandardScaler in Python?
It standardizes features by removing the mean and scaling to unit variance.

26) What is Hadoop?
Hadoop is an open-source framework for storing and processing large datasets using distributed computing.

27) What is HDFS and MapReduce?

- **HDFS**: Hadoop Distributed File System, stores data across multiple machines
- **MapReduce**: A programming model for distributed data processing

28) What are the components of the Hadoop Ecosystem?
HDFS, YARN, MapReduce, Hive, Pig, HBase, Oozie, Flume, Sqoop, Spark

29) What is Scala?
Scala is a high-level programming language combining object-oriented and functional programming.

30) What are features of Scala?

- Concise syntax
- Functional programming support
- Type inference
- Interoperable with Java

31) How is Scala different from Java?

- Scala is more concise and supports functional programming
- Java is more verbose and mainly object-oriented

32) List applications of Scala.

- Big data processing with Spark
- Web development
- Real-time systems
- Backend services

33) What is Data Science?
Data science is a field that uses scientific methods, statistics, and algorithms to extract insights from data.

34) What is Big Data?
Big data refers to extremely large datasets that are too complex for traditional data-processing tools.

35) What are the characteristics of Big Data?
Volume, Velocity, Variety, Veracity, Value

36) List phases in the data science life cycle.

- Data Collection
- Data Cleaning
- Data Exploration
- Modeling

- Evaluation
- Deployment

37) What is Standard Deviation?
Standard Deviation is a measure of the amount of variation or dispersion in a dataset. A low standard deviation indicates that values are close to the mean, while a high standard deviation indicates wide spread.

38) What is meant by Posterior Probability in Naive Bayes Theorem?
Posterior probability is the probability of a class given a feature — written as **P(Class|Feature)**. It is what Naive Bayes ultimately calculates using Bayes' Theorem.

39) What is meant by Likelihood Probability in Naive Bayes Theorem?
Likelihood is the probability of a feature given a class — written as **P(Feature|Class)**. It shows how likely the observed data is under each class.

40) How can we deal with missing values or null values?
Ways to handle missing values:

- **Remove** rows/columns using dropna()
- **Replace** using fillna() with mean, median, or a fixed value
- **Interpolation** using interpolate()
- **Model-based Imputation** like KNN or regression
-

41) What is NLTK?
NLTK (Natural Language Toolkit) is a Python library used for working with human language data (text). It supports tasks like tokenization, stemming, tagging, parsing, and classification.

42) What is Tokenization in NLP?
Tokenization is the process of splitting text into smaller units like words or sentences.
Example: "Hello world" → ['Hello', 'world']

43) What is Stemming?
Stemming is the process of reducing words to their root form.
Example: "running", "runs" → "run"

44) What is Lemmatization?
Lemmatization also reduces words to their root form but uses dictionary-based proper words (lemmas).
Example: "better" → "good"

45) What is Corpus in NLP?
A corpus is a large and structured set of texts used for training and testing NLP models.

46) What is Spark framework?
Apache Spark is an open-source, distributed computing system used for big data processing. It supports in-memory computation, making it faster than Hadoop MapReduce, and works with languages like Python (PySpark), Scala, Java, and R.

47) List phases in data science life cycle?

- **Data Collection**
- **Data Cleaning (Data Wrangling)**
- **Exploratory Data Analysis (EDA)**
- **Feature Engineering**
- **Model Building**
- **Model Evaluation**
- **Deployment**
- **Monitoring & Maintenance**

48) What is Central Tendency?
Central Tendency refers to measures that represent the center or average of a dataset. The main measures are **mean**, **median**, and **mode**.

49) What is Dispersion?
Dispersion refers to the extent to which data values vary around the central value. Common measures: **range**, **variance**, and **standard deviation**.

50) What is Mean, Mode, Mid-range, Median? Calculate for: 10, 22, 13, 10, 21, 43, 77, 21, 10

- **Mean** = (10 + 22 + 13 + 10 + 21 + 43 + 77 + 21 + 10) / 9 = **25.22**
- **Mode** = **10** (occurs 3 times)
- **Median** = Middle value in sorted list [10, 10, 10, 13, 21, 21, 22, 43, 77] → **21**
- **Mid-range** = (Min + Max) / 2 = (10 + 77) / 2 = **43.5**

51) What is Variance?
Variance measures the average squared deviation of each number from the mean. It shows how spread out the data is.

For the same data:
Mean = 25.22,
Variance formula:

Variance=1n∑(xi−mean)2\text{Variance} = \frac{1}{n} \sum (x_i - \text{mean})^2Variance=n1∑(xi −mean)2

Manual calculation gives approx. **Variance ≈ 493.4**