**1) What is Pandas library in Python?**

Pandas is an open-source Python library used for data manipulation and analysis, built on top of NumPy.

**2) Key features of Pandas:**

* DataFrame and Series objects
* Handling of missing data
* Label-based indexing
* Built-in data alignment
* GroupBy for data aggregation
* Reading/writing from various file formats (CSV, Excel, SQL, etc.)

**3) What is NumPy Library in Python?**

NumPy is a core scientific computing library that provides support for arrays and matrices, along with a collection of mathematical functions.

**4) What is matplotlib library?**

Matplotlib is a data visualization library used to create static, animated, and interactive plots in Python.

**5) Difference between Seaborn and Matplotlib:**

* Seaborn is built on top of Matplotlib and offers higher-level interface and attractive default styles.
* Matplotlib provides more control and is lower-level.

**6) Is Sklearn and Scikit-learn same?**

Yes, sklearn is the import name for the Scikit-learn library used in machine learning.

**7) Functions in Pandas and NumPy:**

* Pandas: read\_csv(), DataFrame(), groupby(), merge(), dropna(), fillna()
* NumPy: array(), mean(), std(), reshape(), linspace()

**8) What is DataFrame in Python?**

A DataFrame is a 2D labeled data structure with columns of potentially different types in Pandas.

**9) Find duplicates in Python:**

df.duplicated()

**10) Use of describe() command:**

Provides summary statistics (mean, count, std, min, max, etc.) for numeric columns.

**11) Naive Bayes classification algorithms in Python:**

* GaussianNB
* MultinomialNB
* BernoulliNB  
  (from sklearn.naive\_bayes)

**12) Significance of Confusion Matrix:**

It helps evaluate the performance of a classification model by showing true/false positives/negatives.

**13) TP, TN, FP, FN:**

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

**14) Recall:**

Recall=TPTP+FNRecall = \frac{TP}{TP + FN}

**15) Precision:**

Precision=TPTP+FPPrecision = \frac{TP}{TP + FP}

**16) F1 Score:**

F1=2×Precision⋅RecallPrecision+RecallF1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}

**17) Need of Data Visualization:**

To simplify complex data, identify patterns, and support better decision-making.

**18) What is an outlier?**

An outlier is a data point that differs significantly from other observations.

**19) Histogram vs Pie chart:**

* Histogram: To show frequency distribution.
* Pie Chart: To show percentage/part-to-whole relationships.

**20) Challenges in Big Data Visualization:**

* Data volume and variety
* Real-time processing
* Performance/scalability issues

**21) Joint Plot and Dist Plot:**

* jointplot(): Shows bivariate relationships with histograms.
* distplot(): Shows distribution of a single variable.

**22) Tools for Data Visualization:**

* Matplotlib
* Seaborn
* Plotly
* Tableau
* Power BI

**23) Data Wrangling:**

Cleaning and transforming raw data into a usable format.

**24) Data Transformation:**

Converting data into a proper format/structure for analysis.

**25) Use of StandardScaler:**

Standardizes features by removing the mean and scaling to unit variance.

**26) What is Hadoop?**

An open-source framework for distributed storage and processing of big data.

**27) HDFS and MapReduce:**

* HDFS: Hadoop Distributed File System for storage.
* MapReduce: Programming model for processing large datasets.

**28) Components of Hadoop Ecosystem:**

* HDFS
* MapReduce
* YARN
* Hive
* Pig
* HBase
* Sqoop
* Flume
* Oozie

**29) What is Scala?**

Scala is a high-level programming language combining functional and object-oriented programming.

**30) Features of Scala:**

* Static typing
* JVM interoperability
* Concise syntax
* Functional programming support

**31) Scala vs Java:**

* Scala is more concise
* Better functional programming support
* Java has broader community and tools

**32) Applications of Scala:**

* Data analytics (with Spark)
* Web development
* Backend services

**33) What is Data Science?**

Field that uses statistics, ML, and computing to extract insights from data.

**34) What is Big Data?**

Large and complex datasets that require new processing techniques.

**35) Characteristics of Big Data (5Vs):**

* Volume
* Velocity
* Variety
* Veracity
* Value

**36) Phases in Data Science Lifecycle:**

* Data Collection
* Data Cleaning
* Data Exploration
* Modeling
* Evaluation
* Deployment

**37) What is Central Tendency?**

A measure to identify the center of data (mean, median, mode).

**38) What is Dispersion?**

Describes the spread of data (range, variance, standard deviation).

**39) Calculate Mean, Mode, Mid-range, Median for:**

**Data:** 10, 22, 13, 10, 21, 43, 77, 21, 10

* Mean: 25.22
* Mode: 10
* Median: 21
* Mid-range: (10 + 77)/2 = 43.5

**40) What is Variance?**

Average squared deviation from the mean.

**41) Standard Deviation:**

Square root of the variance, shows data spread.

**42) Posterior Probability (Naive Bayes):**

The probability of a class given the features:

P(Class∣Data)P(Class|Data)

**43) Likelihood Probability (Naive Bayes):**

The probability of the data given a class:

P(Data∣Class)P(Data|Class)

**44) Handling Missing Values:**

* Remove rows
* Fill with mean/median/mode
* Use interpolation or predictive models

**45) What is NLTK?**

Natural Language Toolkit – a Python library for NLP tasks like tokenization, parsing, etc.

**46) What is Tokenization in NLP?**

Splitting text into words, phrases, symbols, or other meaningful elements.

**47) What is Stemming?**

Reducing words to their root form (e.g., running → run).

**48) What is Lemmatization?**

Like stemming but returns meaningful root words (e.g., better → good).

**49) What is Corpus in NLP?**

A large collection of text data used for training NLP models.

**50) What is Spark Framework?**

Apache Spark is a fast, distributed data processing engine used for big data analytics.