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

Pandas is a Python library used mainly for data analysis and data manipulation. It helps in handling structured data like tables, and provides tools to read, clean, filter, and analyze data easily.

2) List some key features of Pandas.

Some main features of Pandas are:It has powerful data structures like Series and DataFrame.Easy handling of missing data.It supports filtering, grouping, merging, reshaping of data.You can easily read/write data from Excel, CSV, SQL, etc.

3) What is Numpy Library in Python?

NumPy is used for numerical and mathematical operations in Python. It works mainly with arrays and is very fast and efficient for performing operations like matrix multiplication, statistics, etc.

4) What is matplotlib library?

Matplotlib is a library for creating static, animated, and interactive visualizations in Python. It’s mostly used for plotting graphs like line charts, bar graphs, scatter plots, etc.

5) What is the difference between Seaborn and Matplotlib?

Matplotlib is the basic plotting library, while Seaborn is built on top of it. Seaborn gives better-looking and more advanced charts easily, like heatmaps and violin plots, with less code.

6) Are Sklearn and Scikit-learn the same? What is its use in Data Science?

Yes, Sklearn and Scikit-learn are the same. Scikit-learn is used in Data Science for machine learning — it provides algorithms for classification, regression, clustering, and also tools for model selection and evaluation.

7) What functions come in Pandas and Numpy libraries?

In Pandas, functions like `read_csv()`, `groupby()`, `merge()`, `dropna()`, and `describe()` are commonly used.

In NumPy, we use functions like `array()`, `mean()`, `std()`, `reshape()`, and `linspace()`.

8) What is a DataFrame in Python?

A DataFrame is a 2-dimensional labeled data structure — kind of like a table with rows and columns. It’s part of the Pandas library and is used to store and work with data in a structured form.

9) How to find duplicates in Python?

You can use this command in Pandas:`pythonCopyEditdf.duplicated()`And to drop duplicates:`pythonCopyEditdf.drop_duplicates()`

10) What is the use of the `describe()` command?

The `describe()` function gives summary statistics of numeric columns — like count, mean, min, max, and standard deviation. It helps understand the distribution of data.

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

In Scikit-learn, we have:**GaussianNB** for continuous data,**MultinomialNB** for text or count data,**BernoulliNB** for binary or yes/no features.**11) What is the significance of Confusion Matrix?**

A confusion matrix is used to evaluate how well a classification model is performing. It shows the actual vs predicted values in a table format, and helps in calculating accuracy, precision, recall, and F1 score.

12) What is TP, TN, FP, FN in confusion matrix?TP (True Positive): Model predicted positive, and it was actually positive.**TN (True Negative)**: Model predicted negative, and it was actually negative.**FP (False Positive)**: Model predicted positive, but it was actually negative.**FN (False Negative)**: Model predicted negative, but it was actually positive.

13) What is Recall?

Recall tells us how many actual positive cases the model was able to identify.

Formula:

Recall = TP / (TP + FN)

14) What is Precision?

Precision tells us how many predicted positive cases were actually correct.

Formula:

Precision = TP / (TP + FP)

15) What is F1 Score?

F1 score is the harmonic mean of precision and recall. It balances both metrics, especially when the data is imbalanced.

Formula:

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

16) What is the need for data visualization in Data Science?

Data visualization helps to understand patterns, trends, and outliers in data easily. It turns complex data into visuals like graphs and charts, which makes analysis and communication simpler.

17) What is an outlier?

An outlier is a data point that is very different from the rest of the data. It can affect the average and other statistics, and sometimes may indicate an error or special case.

18) When to use Histogram and Pie Chart?Use a **histogram** to show the distribution of numerical data — like marks or ages.Use a **pie chart** to show parts of a whole — like percentage of sales by region.

19) What are the challenges in Big Data visualization?

Some challenges are:Handling huge volumes of data.Slow performance while loading or rendering graphs.Making the visualization interactive and readable.Choosing the right type of chart for complex data.

20) What is a joint plot and dist plot?Distplot is used to show the distribution of a single variable (like a histogram with a curve).**Jointplot** combines two plots: scatter plot and histograms. It shows the relationship between two variables and their distributions.**22) What is Data Wrangling?**

Data wrangling means cleaning, organizing, and transforming raw data into a usable format for analysis or machine learning.

23) What is Data Transformation?

Data transformation means changing the format, structure, or values of data — like converting text to numbers, normalizing, or aggregating data — to prepare it for analysis.

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

StandardScaler is used to standardize features by removing the mean and scaling to unit variance. It's useful when we want all features to be on the same scale in machine learning models.

25) What is Hadoop?

Hadoop is an open-source framework for storing and processing large amounts of data in a distributed manner across multiple machines.

26) What is HDFS and MapReduce?HDFS (Hadoop Distributed File System) is the storage part of Hadoop — it stores large data across multiple nodes.**MapReduce** is the processing part — it breaks the data into chunks and processes it in parallel.

27) What are the components of Hadoop Ecosystem?

Main components include:**HDFS** (Storage)**MapReduce** (Processing)**YARN** (Resource Management)**Hive, Pig, Sqoop, Flume, HBase, Zookeeper, Oozie** — for querying, importing, scheduling, etc.

28) What is Scala?

Scala is a high-level programming language that combines object-oriented and functional programming. It's used with Apache Spark and other big data tools.

29) What are features of Scala?Supports both object-oriented and functional style.Statically typed language.Concise syntax, compared to Java.Interoperable with Java (can use Java libraries).Used in Big Data and concurrent applications.

30) How is Scala different from Java?Scala code is shorter and more expressive.Scala supports functional programming fully, Java only partially.Scala has features like pattern matching and immutability by default.Java is more verbose and older, but still widely used.

31) List applications of Scala. Used in Big Data tools like Apache Spark.Web development (Play Framework).Real-time analytics.Data science and Machine Learning.Backend services and APIs.

32) What is Data Science?

Data Science is a field that uses statistics, programming, and machine learning to extract insights and knowledge from data. It helps in making better decisions using data.**31) What is Big Data?**

Big Data refers to extremely large and complex datasets that can't be handled using traditional tools. It includes structured, semi-structured, and unstructured data.

32) What are the characteristics of Big Data?

They are called the **5 V's**:**Volume** – large amount of data**Velocity** – speed of data generation**Variety** – different types of data (text, video, etc.)**Veracity** – data quality and accuracy**Value** – useful insights from data

---

33) List phases in Data Science life cycle. Business UnderstandingData CollectionData Cleaning / WranglingData Exploration / VisualizationFeature EngineeringModel BuildingEvaluationDeploymentMonitoring & Maintenance

---

34) What is Central Tendency?

Central tendency tells us the central or average value in a dataset. Common measures are **mean**, **median**, and **mode**.

---

35) What is Dispersion?

Dispersion tells us how spread out the data is — how much values differ from the average. It includes **range**, **variance**, and **standard deviation**.

---

36) What is Mean, Mode, Median, Mid-Range? Calculate for:

Data: 10, 22, 13, 10, 21, 43, 77, 21, 10**Mean (Average):** (10+22+13+10+21+43+77+21+10) / 9 = **25.22Mode (Most frequent): 10** (appears 3 times)**Median (Middle value):** Sorted → 10,10,10,13,21,21,22,43,77 → Middle = **21Mid-Range:** (Min + Max) / 2 = (10 + 77) / 2 = **43.5**

---

37) What is Variance?

Variance measures how far each number is from the mean. It shows the **spread** of the data.

---

38) What is Standard Deviation?

Standard deviation is the **square root of variance**. It tells how much data varies from the average — in the same units as the data.

---

39) What is Posterior Probability in Naive Bayes?

Posterior is the **final probability** we want — the chance of a class **given** the data.

Formula:

P(Class | Data)

---

40) What is Likelihood in Naive Bayes?

Likelihood is the **chance of seeing the data** for a given class.

Formula:

P(Data | Class)

---

41) How can we deal with missing values?

We can handle missing or null values by:**Removing** rows or columns with too many nulls**Filling** with mean, median, mode**Forward-fill or back-fillUsing prediction models** to guess missing values

41) What is NLTK?
NLTK stands for **Natural Language Toolkit**. It's a Python library used for working with human language data — useful in tasks like tokenization, stemming, parsing, tagging, etc., in NLP (Natural Language Processing).

---

42) What is Tokenization in NLP?
Tokenization means **breaking text into smaller parts**, usually words or sentences.
For example, breaking a sentence into individual words.

---

43) What is Stemming?
Stemming is the process of **removing suffixes** from words to get the root form.
Example: "playing", "played" → stemmed to **"play"** (even if the root isn't always a valid word).

---

44) What is Lemmatization?
Lemmatization also gives the root form of a word, but it returns **a real word**, and considers the context or part of speech.
Example: "running" → **"run"**

---

45) What is Corpus in NLP?
A corpus is a **collection of text data** used for NLP tasks. It's like a dataset of language — for training or testing models.

---

46) What is Spark framework?
Apache Spark is an open-source **Big Data processing framework**. It’s fast and supports batch and real-time data processing, mainly used with large-scale data and machine learning tasks.