07 May 2025 01:22

Great! Here are some simple and important foundational DSBDA concepts that you may be asked during viva or should know while doing practicals:

◇ Basic Terms & Concepts

1. **What is the full form of CSV?**
   - **CSV**: Comma-Separated Values
     It is a plain text file format used to store tabular data, where each line is a row and columns are separated by commas.
2. **What is a Dataset?**
   - A dataset is a collection of data, often in table form with rows and columns. Rows are records; columns are features/attributes.
3. **What is Data Preprocessing?**
   - It is the process of cleaning and preparing raw data for analysis (e.g., removing missing values, encoding text, scaling numbers).
4. **What is a Feature?**
   - A feature is a measurable property or column in the dataset. Example: "age", "fare", or "species".
5. **What is a Label/Target?**
   - In supervised learning, a label is the output value we want to predict. Example: “survived” in Titanic dataset.
6. **What is Null / Missing Value?**
   - Data that is not available or empty in the dataset.
7. **What is EDA (Exploratory Data Analysis)?**
   - It is the process of summarizing the main characteristics of the data using statistics and visualizations (like plots, graphs).
8. **What is Visualization?**
   - It means creating visual representations like graphs and charts to understand patterns or relationships in data.
9. **What is Pandas?**
   - A Python library used for data manipulation and analysis. It provides data structures like DataFrames.
10. **What is NumPy?**
    - A Python library used for numerical operations. It supports arrays, matrices, and mathematical functions.
11. **What is Seaborn?**
    - A Python library built on top of matplotlib used for making statistical graphics and visualizations.
12. **What is Matplotlib?**
    - A Python plotting library used to create static, animated, and interactive visualizations.
13. **What is the difference between .head() and .info() in pandas?**
    - .head() shows first 5 rows of the dataset.
    - .info() shows structure, data types, and non-null counts.
14. **What is KDE in a histogram?**
    - KDE (Kernel Density Estimation) adds a smooth curve over a histogram to estimate the probability density of a variable.
15. **What is Boxplot?**
    - A plot that shows the distribution of data and detects outliers using quartiles.
16. **What is a DataFrame?**

- A 2D table in pandas with labeled rows and columns (like Excel).

17. **What is a Series in pandas?**

- A single column (1D array) of data in pandas.

18. **What is Label Encoding?**
    - Converting categorical text data into numbers.
      Example: "male" → 0, "female" → 1.
19. **What is One-Hot Encoding?**
    - Converting categories into multiple binary columns (0 or 1).
      Example: Color: red, blue → becomes: [1, 0], [0, 1]
20. **What is Scaling?**
    - Resizing numerical values to a fixed range (like 0 to 1).
      Example: MinMaxScaler.
21. **What is Normalization?**
    - Making all data values fall into a small range (like 0–1), usually row-wise.
22. **What is Standardization?**
    - Rescaling data to have a mean of 0 and a standard deviation of 1.
23. **What is a Correlation Matrix?**
    - A table showing how strongly two variables are related (from -1 to 1).
24. **What is Linear Regression?**
    - A method to model the relationship between a dependent and one/more independent variables using a straight line.
25. **What is Logistic Regression?**
    - A classification method used to predict binary outcomes (like yes/no, 0/1).
26. **What is Clustering?**
    - Grouping similar data points together. Example: K-Means clustering.
27. **What is TF-IDF?**
    - TF-IDF (Term Frequency-Inverse Document Frequency): a method to find important words in a document by considering frequency and uniqueness.
28. **What is Tokenization?**
    - Breaking text into words or sentences.
29. **What is Stop Words?**
    - Common words (like "is", "the", "and") that are removed during text processing.
30. **What is Stemming?**
    - Reducing words to their root form. Example: "running" → "run"
31. **What is Lemmatization?**
    - Converting words to their dictionary form. Example: "was" → "be"
32. **What is the difference between Stemming and Lemmatization?**
    - Stemming is faster but may cut words roughly.
    - Lemmatization is slower but gives correct root words.
33. **What is Overfitting?**
    - A model that performs well on training data but badly on new (test) data.
34. **What is Underfitting?**
    - A model that performs poorly even on training data.
35. **What is a Heatmap?**
    - A visual chart showing values in color (often used to show correlation).
36. **What is a Scatter Plot?**
    - A graph showing the relationship between two numerical variables.
37. **What is Outlier?**
    - A value that is very different from others in the dataset.
38. **What is the purpose of .dropna()?**
    - To remove missing/null values from the dataset.
39. **What is the use of groupby() in pandas?**
    - Used to group data by categories and apply functions like sum, mean, count.
40. **What is Data Preprocessing?**
    - Preparing raw data for analysis (cleaning, encoding, scaling, etc.).
41. **What is EDA (Exploratory Data Analysis)?**
    - A process to explore data patterns, trends, and relationships using statistics and plots.
42. **What is Data Cleaning?**
    - Fixing or removing incorrect, missing, or irrelevant data.
43. **What is Missing Data?**

- Data that is not recorded or unavailable in a dataset.

44. **How to handle missing data?**
    - Remove rows (dropna()) or fill missing values (fillna()).
45. **What is .describe() in pandas?**
    - Gives basic statistics (mean, min, max, std) for numeric columns.
46. **What is .value_counts()?**
    - Shows the frequency count of unique values in a column.
47. **What is .corr()?**
    - Finds correlation (relationship) between numerical columns.
48. **What is a Bar Plot?**
    - A chart showing comparison using rectangular bars.
49. **What is a Box Plot?**
    - A plot showing data distribution, median, and outliers.
50. **What is a Histogram?**
    - A plot to show frequency distribution of a numeric column.
51. **What is a KDE Plot?**
    - KDE (Kernel Density Estimation): A smooth curve showing data distribution.
52. **What is a Pair Plot?**
    - A grid of scatter plots showing relationships between all pairs of features.
53. **What is matplotlib used for?**
    - Creating static, interactive, and animated visualizations in Python.
54. **What is seaborn used for?**
    - High-level data visualization library based on matplotlib.
55. **Difference: seaborn vs matplotlib?**
    - Seaborn is built on top of matplotlib and provides prettier, easier plots.
56. **What is the Titanic dataset?**
    - A dataset of passengers from the Titanic ship, used for classification problems.
57. **What is the Iris dataset?**
    - A flower dataset with 3 species, used in classification and visualization.
58. **What is a classification problem?**
    - Predicting categories or classes (like spam or not spam).
59. **What is a regression problem?**
    - Predicting continuous values (like house price, temperature).
60. **What is Multicollinearity?**
    - When two or more input features are highly correlated.
61. **What is np.array()?**
    - A NumPy method to create arrays (like lists but faster and more powerful).
62. **What is pd.DataFrame()?**
    - A pandas method to create a table from data.
63. **What is the difference between NumPy and pandas?**
    - NumPy handles numeric arrays.
    - pandas is for tabular data with labeled rows and columns.
64. **What is a categorical variable?**
    - A variable with fixed categories (like male/female, Yes/No).
65. **What is a numerical variable?**
    - A variable with numbers (like age, salary, height).
66. **What is .mean(), .median(), .mode()?**
    - Mean: average
    - Median: middle value
    - Mode: most frequent value
67. **What is .groupby() + .agg() in pandas?**
    - Used together to apply multiple functions on grouped data.
68. **What is .map() in pandas?**
    - Used to replace or convert values in a column.
69. **What is .apply() in pandas?**
    - Apply a function to rows or columns of a DataFrame.
70. **What is train_test_split()?**

- Used to divide data into training and testing sets.

☑ CSV & Data Basics

1. **Q: What is the full form of CSV?**
   A: Comma-Separated Values.
2. **Q: What is a dataset?**
   A: A collection of data, usually in tabular form.
3. **Q: What is the difference between structured and unstructured data?**
   A: Structured data is in tables (like Excel); unstructured data includes text, images, videos.
4. **Q: What is metadata?**
   A: Data about data (e.g., column names, data types).

☑ Data Types

1. **Q: What are numeric data types?**
   A: Integers, floats — used for calculations.
2. **Q: What are categorical variables?**
   A: Variables with limited, fixed values (like 'male', 'female').
3. **Q: What is the difference between nominal and ordinal data?**
   A: Nominal has no order (e.g., colors), ordinal has order (e.g., grades A > B > C).

☑ Data Cleaning & Preprocessing

1. **Q: What is data cleaning?**
   A: Removing or correcting incorrect or missing data.
2. **Q: What are missing values?**
   A: Empty or null entries in a dataset.
3. **Q: How to handle missing values?**
   A: Drop them or fill them (e.g., using mean or median).
4. **Q: What is normalization?**
   A: Scaling data between 0 and 1.
5. **Q: What is standardization?**
   A: Scaling data so that it has mean 0 and standard deviation 1.

☑ Features and Target

1. **Q: What is a feature?**
   A: An input variable (like age, income) used to predict something.
2. **Q: What is a target?**
   A: The output we want to predict (like survived or not).
3. **Q: What is feature selection?**
   A: Choosing the most important features for a model.

☑ Machine Learning Concepts

1. **Q: What is overfitting?**
   A: When a model learns too much from training data and performs poorly on new data.
2. **Q: What is underfitting?**
   A: When a model is too simple and cannot capture patterns in data.
3. **Q: What is a model?**
   A: A mathematical function used to make predictions.
4. **Q: What is accuracy?**
   A: The percentage of correct predictions.
5. **Q: What is a confusion matrix?**
   A: A table showing true vs. predicted values.

☑ Libraries & Tools

1. **Q: What is pandas used for?**

A: Data manipulation and analysis.

2. **Q: What is NumPy used for?**
   A: Handling arrays and numerical calculations.
3. **Q: What is seaborn used for?**
   A: Creating attractive and informative data visualizations.
4. **Q: What is matplotlib?**
   A: A basic library for plotting graphs and charts.
5. **Q: What is scikit-learn?**
   A: A machine learning library for Python.

☑ More Common Terms

1. **Q: What is EDA?**
   A: Exploratory Data Analysis — understanding the data before modeling.
2. **Q: What is a correlation?**
   A: A measure of how two variables move together.
3. **Q: What is an outlier?**
   A: A data point that is very different from others.
4. **Q: What is a cluster?**
   A: A group of similar data points.
5. **Q: What is classification?**
   A: Predicting categories or labels.

Mean

- **Definition:** The **average** of a set of numbers.
- **Formula:** Mean=∑Xn\text{Mean} = \frac{\sum X}{n}Mean=n∑X
  - Where XXX is the data points, and nnn is the number of data points.
- **Example:** For the data [2, 4, 6, 8], the mean is 2+4+6+84=5\frac{2+4+6+8}{4} = 542+4+6+8=5.

◇ Median

- **Definition:** The **middle value** of a dataset when it's ordered.
- **Process:**
  1. Sort the data.
  2. If there is an odd number of data points, the median is the middle one.
  3. If there is an even number of data points, the median is the average of the two middle values.
- **Example:** For [1, 3, 7, 8, 9], the median is 7 (middle number).
  - For [1, 2, 3, 4], the median is 2+32=2.5\frac{2 + 3}{2} = 2.522+3=2.5.

◇ Mode

- **Definition:** The **value** that appears most frequently in a dataset.
- **Example:** For the data [1, 2, 2, 3, 4], the mode is 2 (since it appears twice).

◇ Standard Deviation (SD)

- **Definition:** A measure of how much the values in a dataset **deviate** from the mean.
- **Formula:**
  SD=∑(Xi−μ)2n\text{SD} = \sqrt{\frac{\sum (X_i - \mu)^2}{n}}SD=n∑(Xi−μ)2
  - Where XiX_iXi is each data point, μ\muμ is the mean, and nnn is the number of data points.
- **Interpretation:** A **low SD** means the data points are close to the mean, while a **high SD** indicates the data points are spread out.
- **Example:** If you have [2, 4, 6, 8], the SD tells you how spread out these numbers are from the average.

◇ Variance

- **Definition:** The average of the squared differences from the mean.
- **Formula:**
  Variance=∑(Xi−μ)2n\text{Variance} = \frac{\sum (X_i - \mu)^2}{n}Variance=n∑(Xi−μ)2
- **Relation with Standard Deviation:**
  SD=Variance\text{SD} = \sqrt{\text{Variance}}SD=Variance
- **Interpretation:** Variance is the square of the standard deviation. It shows how much the values deviate, but in squared units.

◇ Range

- **Definition:** The difference between the maximum and minimum values in a dataset.
- **Formula:**
  Range=Maximum Value−Minimum Value\text{Range} = \text{Maximum Value} - \text{Minimum Value}Range=Maximum Value−Minimum Value
- **Example:** For the data [1, 3, 5, 7], the range is 7−1=67 - 1 = 67−1=6.

◇ Interquartile Range (IQR)

- **Definition:** The difference between the 75th percentile (Q3) and the 25th percentile (Q1) of a dataset.
- **Formula:**
  IQR=Q3−Q1\text{IQR} = Q3 - Q1IQR=Q3−Q1
- **Purpose:** It helps to understand the spread of the middle 50% of the data and is less affected by outliers than the range.
- **Example:** If Q3 = 7 and Q1 = 3, then IQR = 7−3=47 - 3 = 47−3=4.

◇ Percentiles

- **Definition:** Percentiles divide a dataset into 100 equal parts. For example, the 25th percentile is the value below which 25% of the data lies.
- **Key percentiles:**
  - **25th percentile (Q1):** First quartile.
  - **50th percentile (Q2):** Median.
  - **75th percentile (Q3):** Third quartile.

◇ Z-Score

- **Definition:** A measure of how many **standard deviations** a data point is from the mean.
- **Formula:**
  Z=X−μσZ = \frac{X - \mu}{\sigma}Z=σX−μ
  - Where XXX is the data point, μ\muμ is the mean, and σ\sigmaσ is the standard deviation.
- **Interpretation:** A Z-score tells us whether a value is **above** or **below** the mean in terms of standard deviations.
  - Example: A Z-score of 2 means the value is 2 standard deviations above the mean.

◇ Skewness

- **Definition:** A measure of the asymmetry of the distribution of data.
  - **Positive skew**: Long tail on the right.
  - **Negative skew**: Long tail on the left.
  - **Zero skew**: Symmetric distribution.

◇ Kurtosis

- **Definition:** A measure of the **peakedness** or flatness of the data distribution.
  - **Leptokurtic:** High peak (positive kurtosis).
  - **Platykurtic:** Flat (negative kurtosis).
  - **Mesokurtic:** Normal peak (zero kurtosis).

◇ Correlation

- **Definition:** A statistical measure that describes the **relationship** between two variables.
  - **Positive correlation:** As one variable increases, the other increases.
  - **Negative correlation:** As one variable increases, the other decreases.
  - **Zero correlation:** No relationship between the variables.
- **Coefficient range:** -1 to 1.
  - **1** means perfect positive correlation.
  - **-1** means perfect negative correlation.
  - **0** means no correlation.

◇ Covariance

- **Definition:** Measures the joint variability of two random variables. It tells us the direction of the relationship between variables, but not the strength.
- **Formula:**
  $Cov(X,Y)=\sum(Xi-\mu X)(Yi-\mu Y)n\text{Cov}(X, Y) = \frac{\sum (X_i - \mu_X)(Y_i - \mu_Y)}{n}$
  $Cov(X,Y)=n\sum(Xi-\mu X)(Yi-\mu Y)$

◇ Probability Concepts

- **Definition of Probability:** The likelihood of an event occurring, ranging from 0 (impossible) to 1 (certain).
- **Random Variable:** A variable whose value is determined by the outcome of a random event.
- **Probability Distribution:** A function that describes the likelihood of obtaining the possible values that a random variable can take.

◇ Hypothesis Testing

- **Null Hypothesis (H0):** Assumes no effect or no difference.
- **Alternative Hypothesis (H1):** Suggests there is an effect or a difference.
- **P-Value:** The probability of observing the results, or something more extreme, under the null hypothesis. A p-value less than 0.05 typically means rejecting the null hypothesis.

1. Pandas

- **Purpose:** Used for data manipulation and analysis.
- **Key Functions:**
  - pd.read_csv(): Load data from CSV.
  - df.dropna(): Remove missing data.
  - df.groupby(): Group data for aggregation.

2. NumPy

- **Purpose:** Provides support for large, multi-dimensional arrays and mathematical functions.
- **Key Functions:**
  - np.array(): Create arrays.
  - np.mean(), np.std(): Calculate mean and standard deviation.

3. Matplotlib

- **Purpose:** For creating static visualizations like graphs and charts.
- **Key Functions:**
  - plt.plot(), plt.show(): Create and display plots.
  - plt.xlabel(), plt.ylabel(): Label axes.

4. Seaborn

- **Purpose:** Built on top of Matplotlib, for statistical data visualization.
- **Key Functions:**
  - sns.histplot(): Create histograms.
  - sns.boxplot(): Create box plots.

    - sns.heatmap(): Visualize correlations with heatmaps.

5. Scikit-learn

- **Purpose:** For implementing machine learning models.
- **Key Functions:**
    - LinearRegression(): Build regression models.
    - train_test_split(): Split data for training/testing.
    - fit(), predict(): Train and make predictions.

Key Concepts:

- **Data Preprocessing:** Cleaning and transforming data (using **Pandas** and **NumPy**).
- **Data Visualization:** Plotting graphs and analyzing data visually (using **Matplotlib** and **Seaborn**).
- **Machine Learning:** Building and evaluating models (using **Scikit-learn**).

Thoery concept

07 May 2025 01:32

Absolutely! Below, I'll provide answers to the theoretical questions mentioned earlier in **simple language** for your **viva preparation**:

What is a Confusion Matrix?

- **Definition**: A **Confusion Matrix** is a table used to evaluate the performance of a classification algorithm. It compares the predicted and actual values.
- **Components**:
  - **True Positives (TP)**: Correctly predicted positive cases.
  - **False Positives (FP)**: Incorrectly predicted as positive.
  - **True Negatives (TN)**: Correctly predicted negative cases.
  - **False Negatives (FN)**: Incorrectly predicted as negative.
    Example:

	Predicted Positive	**Predicted Negative**
Actual Pos	TP	FN
Actual Neg	FP	TN

- **From this, you can calculate**:
  - **Accuracy** = (TP + TN) / Total
  - **Precision** = TP / (TP + FP)
  - **Recall** = TP / (TP + FN)

7. What is Naive Bayes?

- **Definition**: **Naive Bayes** is a **probabilistic classification algorithm** based on **Bayes' Theorem**. It assumes that the features (predictor variables) are independent (hence the "naive" part).
- **Purpose**: It's used for classification problems, especially in text classification (e.g., spam detection).
  Bayes' Theorem formula:
  P(A|B) = (P(B|A) * P(A)) / P(B)

8. What is Tokenization?

- **Definition**: **Tokenization** is the process of splitting text into smaller units called **tokens** (e.g., words, sentences).
- **Purpose**: It's the first step in **text preprocessing** for NLP (Natural Language Processing).
  Example:
  "I love data science!" → ["I", "love", "data", "science"]

9. What is Stopwords Removal?

- **Definition**: **Stopwords** are common words (e.g., "the", "is", "and", "in") that don't contribute much to the meaning of a sentence and are typically removed during text preprocessing.
- **Purpose**: Helps in focusing on important words that carry the meaning in text analysis.

10. What is Stemming?

- **Definition**: **Stemming** is the process of reducing a word to its base/root form (e.g., "running" → "run").
- **Purpose**: To standardize words and reduce variations to a single form.

11. What is Lemmatization?

- **Definition**: **Lemmatization** is the process of reducing a word to its **lemma** (dictionary form), considering its meaning (e.g., "better" → "good").
- **Purpose**: More accurate than stemming because it looks at the word's context and meaning.

12. What is TF-IDF?

- **Definition**: **TF-IDF (Term Frequency-Inverse Document Frequency)** is a statistical measure used to evaluate how important a word is to a document in a collection of documents.
    - **TF (Term Frequency)**: How often a word appears in a document.
    - **IDF (Inverse Document Frequency)**: How rare the word is across the documents.
      Formula: TF-IDF = TF * IDF

13. What is a Heatmap?

- **Definition**: A **Heatmap** is a graphical representation of data where values are depicted by color.
- **Purpose**: It's commonly used for visualizing **correlations** between variables in a dataset.

14. What is a Boxplot?

- **Definition**: A **Boxplot** (also known as a **box-and-whisker plot**) is a statistical graph that shows the distribution of a dataset based on its **quartiles**.
- **Purpose**: Helps in identifying outliers, central tendency, and spread of the data.

Summary:

- **DataFrame**: 2D table-like structure in pandas.
- **Series**: A one-dimensional array (column of DataFrame).
- **NumPy Array**: Homogeneous array used for numerical computations.
- **Linear Regression**: Predicting continuous values using a straight line.
- **Logistic Regression**: Classification problem for binary outcomes.
- **Confusion Matrix**: Used for evaluating classification models.
- **Naive Bayes**: A probabilistic classifier based on Bayes' Theorem.
- **Tokenization**: Splitting text into smaller units (tokens).
- **Stopwords Removal**: Removing common, unimportant words.
- **Stemming & Lemmatization**: Reducing words to their root/base form.
- **TF-IDF**: Measure for word importance in a document.
- **Heatmap**: A colored representation of data.
- **Boxplot**: Displays distribution and outliers

1. General Questions about Python Libraries

1. **What is pandas, and how is it used in data science?**
    - **Answer:** Pandas is a Python library used for data manipulation and analysis. It helps us work with large datasets, clean data, and perform operations like filtering, merging, grouping, etc., using structures like DataFrames (tables) and Series (columns).
2. **What is the difference between a pandas DataFrame and a NumPy array?**
    - **Answer:** A pandas **DataFrame** is a table-like structure with rows and columns, where each column can have different data types (e.g., integers, strings). A **NumPy array** is more like a grid that can only hold elements of the same data type and is mostly used for numerical calculations.
3. **Explain the role of NumPy in scientific computing and how it is used in data analysis.**
    - **Answer:** NumPy is a library used for numerical computing. It provides efficient operations on arrays and matrices. It's faster and more flexible than regular Python lists and is used for mathematical operations like matrix multiplication, linear algebra, and statistical analysis.
4. **What is the difference between iloc[] and loc[] in pandas?**
    - **Answer:** iloc[] is used to access data based on **integer position** (index) while loc[] is used

to access data based on the **index label** (row or column names).

5. **What are the advantages of using pandas over traditional Python data structures (like lists and dictionaries)?**
   - **Answer:** Pandas makes it easier to handle large datasets, clean data, and perform complex operations. It provides built-in methods for operations like merging datasets, dealing with missing values, and data manipulation, which is hard to do with regular Python lists and dictionaries.
6. **What is the purpose of the apply() function in pandas, and how is it used?**
   - **Answer:** The apply() function is used to apply a function along an axis (rows or columns) of a DataFrame. It allows us to perform complex operations on each row/column in a simple way.
7. **What is a Series in pandas? How is it different from a DataFrame?**
   - **Answer:** A **Series** is a one-dimensional array-like object, whereas a **DataFrame** is two-dimensional with rows and columns. A DataFrame can contain multiple Series as its columns.
8. **Explain the concept of "vectorization" in NumPy and why it is important.**
   - **Answer:** Vectorization refers to performing operations on entire arrays or matrices without using loops. NumPy makes this possible, making it faster and more efficient than using traditional Python loops.
9. **What is the significance of the matplotlib library in data visualization?**
   - **Answer:** matplotlib is a basic plotting library in Python used for creating simple static, animated, or interactive plots. It's often used to visualize data in graphs like bar charts, histograms, and line plots.
10. **How does seaborn improve on matplotlib?**
    - **Answer:** seaborn is built on top of matplotlib and provides a higher-level, more user-friendly interface for creating attractive and informative visualizations. It has better default themes and easier handling of complex datasets.

2. Data Preprocessing and Wrangling

1. **What is data wrangling, and why is it an important step in data analysis?**
   - **Answer:** Data wrangling (or cleaning) is the process of cleaning, transforming, and organizing raw data into a format that is easy to analyze. It's crucial because raw data is often messy, incomplete, and inconsistent.
2. **Explain the difference between 'Data Cleaning' and 'Data Wrangling'.**
   - **Answer:** Data cleaning refers specifically to fixing errors or inconsistencies in data (e.g., handling missing values, correcting typos). Data wrangling is a broader process that includes cleaning and transforming data into a useful format.
3. **How do you handle missing values in a dataset using pandas?**
   - **Answer:** You can handle missing values in pandas by using methods like isnull() to detect missing values and fillna() to replace them with a specific value or method like mean or median.
4. **Explain how to normalize a dataset and why normalization is important.**
   - **Answer:** Normalization scales the data so that it falls within a specific range (e.g., 0 to 1). It's important when features have different units or scales to prevent certain features from dominating in machine learning models.
5. **What is the purpose of the isnull() function in pandas, and how is it used?**
   - **Answer:** The isnull() function detects missing values in a DataFrame. It returns a DataFrame with True for missing values and False for non-missing values.
6. **What are outliers, and how can they be detected and handled in data analysis?**
   - **Answer:** Outliers are data points that are significantly different from the rest of the data. They can be detected using statistical methods (like Z-scores or IQR). Handling outliers involves removing or transforming them.
7. **How do you convert categorical variables into numeric values in pandas?**
   - **Answer:** You can convert categorical variables using methods like get_dummies() to create dummy variables or LabelEncoder() to assign a unique number to each category.
8. **What is feature scaling, and why is it important in machine learning?**

- **Answer:** Feature scaling is the process of scaling data to a standard range (e.g., 0 to 1). It's important in machine learning because many algorithms (like KNN and gradient descent) are sensitive to the scale of data.

9. **Describe one method to deal with skewed data in a dataset.**
   - **Answer:** One method is to apply a **log transformation** to the skewed data, which helps make the distribution more symmetric.
10. **Explain the difference between get_dummies() and LabelEncoder() for encoding categorical data.**
    - **Answer:** get_dummies() creates binary columns for each category, while LabelEncoder() assigns a unique integer to each category.

3. Supervised Learning (Linear and Logistic Regression)

1. **What is Linear Regression, and how is it used for predicting continuous variables?**
   - **Answer:** Linear regression is a statistical method used to predict the value of a continuous variable based on one or more input features by fitting a linear relationship.
2. **What assumptions does Linear Regression make about the data?**
   - **Answer:** Linear regression assumes that the relationship between the dependent and independent variables is linear, and it assumes homoscedasticity (constant variance of errors), independence of errors, and normality of errors.
3. **Explain the concept of 'Overfitting' and how it affects the performance of a model.**
   - **Answer:** Overfitting happens when a model learns the noise in the training data rather than the underlying patterns. It leads to poor performance on new data because the model is too complex and specific to the training set.
4. **What is the significance of the coef_ and intercept_ attributes in a Linear Regression model?**
   - **Answer:** coef_ represents the coefficients (weights) of the features in the linear regression equation, and intercept_ represents the y-intercept of the regression line.
5. **What is Logistic Regression, and how is it different from Linear Regression?**
   - **Answer:** Logistic regression is used for classification problems, where the output is categorical (e.g., yes/no). It uses the sigmoid function to predict probabilities. Linear regression is used for predicting continuous values.
6. **Explain the sigmoid function and how it is used in Logistic Regression.**
   - **Answer:** The sigmoid function squashes the output of a linear equation into a value between 0 and 1, which is interpreted as a probability in Logistic Regression.
7. **What is the Confusion Matrix, and what does it tell us about a classification model?**
   - **Answer:** The confusion matrix is a table that compares the actual and predicted classifications, showing the counts of true positives, false positives, true negatives, and false negatives. It helps assess model performance.
8. **What are Precision, Recall, and F1-score? How do they differ from accuracy?**
   - **Answer:**
     - **Precision** measures how many predicted positives are actually positive.
     - **Recall** measures how many actual positives were correctly predicted.
     - **F1-score** is the harmonic mean of precision and recall.
     - **Accuracy** measures how many predictions (both positive and negative) were correct.
9. **What is the role of scikit-learn in supervised learning models?**
   - **Answer:** scikit-learn is a Python library that provides simple and efficient tools for data analysis and machine learning, including algorithms for classification, regression, clustering, and more.
10. **What is the purpose of the fit() function in machine learning models?**
    - **Answer:** The fit() function is used to train a machine learning model on the provided data. It learns the relationships in the data and adjusts the model's parameters.