|  |  |  |
| --- | --- | --- |
| D:\NAAC (Anoop Tiwari)\IQAC\unnamed.png | | **SAGAR INSTITUTE OF SCIENCE & TECHNOLOGY**  **DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE**  **MST Paper Solation** |
| **BRANCH** | **AIDS** |
| **SESSION** | **2023-24** |
| **SEMESTER** | **IV** |
| **SUBJECT/CODE: AD -402 Data Science** | | |

|  |  |  |
| --- | --- | --- |
| **Q. No.** | **Question** | **Bloom’s**  **Taxonomy**  **Level** |
|  | (a)What is Data Wrangling in data science **(5 marks)** and why it is important? Explain. **(2 marks)**  Data wrangling, also known as data munging, is the process of cleaning, structuring, and enriching raw data into a desired format for better decision-making in data science. It involves several steps, including data collection, cleaning, transformation, and integration.  Data Collection: Gathering data from various sources such as databases, APIs, spreadsheets, or web scraping.  Data Cleaning: Identifying and rectifying errors, inconsistencies, missing values, and outliers in the data. This step ensures that the data is accurate and reliable for analysis.  Data Transformation: Restructuring and reshaping the data to suit the analysis requirements. This may involve converting data types, standardizing formats, and creating new features or variables.  Data Integration: Combining data from multiple sources into a single dataset for comprehensive analysis. This step ensures that all relevant information is considered and inconsistencies are resolved.  Data Enrichment: Enhancing the dataset by adding external data or deriving additional insights through feature engineering or data augmentation techniques.  Data wrangling is crucial in data science for several reasons:  Improves Data Quality: By cleaning and preprocessing the data, data wrangling ensures that the data is accurate, consistent, and reliable, which is essential for making informed decisions.  Facilitates Analysis: Well-structured and cleaned data is easier to analyze and interpret, leading to more accurate insights and predictions.  Reduces Bias and Errors: Data wrangling helps identify and mitigate biases and errors present in the raw data, ensuring the fairness and validity of analytical results.  Saves Time and Resources: Automating repetitive data cleaning and transformation tasks streamlines the data preparation process, saving time and resources for more advanced analysis.  Enables Effective Data Utilization: By organizing and enriching the data, data wrangling maximizes the value extracted from the available data, enabling organizations to make strategic decisions based on comprehensive insights.  Or  (b) Classify measures of central tendency of quantitative data. **(5 marks)** And outline the measures of variability of data. **(2 marks)**  **Measures of Central Tendency:**   1. **Mean:** **The arithmetic average of a set of values, calculated by summing all values and dividing by the total number of values.**      * + **Example: Using the same dataset of exam scores: {85, 90, 75, 92, 88}, first, sort the scores: {75, 85, 88, 90, 92}. Since there are 5 scores, the median is the middle value, which is 88.**  1. **Median:**   **Example: Using the same dataset of exam scores: {85, 90, 75, 92, 88}, first, sort the scores: {75, 85, 88, 90, 92}. Since there are 5 scores, the median is the middle value, which is 88.**   1. **Mode:**   **The most frequently occurring value(s) in a dataset.**   * + **Example: Continuing with the exam scores dataset: {85, 90, 75, 92, 88}. Here, the mode is 90, as it appears most frequently in the dataset.**   **Measures of Variability:**   1. **Range:**   **The difference between the maximum and minimum values in a dataset, providing a simple measure of spread.**     * + **Example: Using the same exam scores dataset: {85, 90, 75, 92, 88}. The range is calculated as the difference between the maximum and minimum values: Range = 92 - 75 = 17.**  1. **Variance:**   **The average of the squared differences from the mean, representing the dispersion of data points around the mean.**     * + **Example: Let's consider a dataset of daily temperatures in Celsius for a week: {20, 22, 19, 25, 21, 23, 20}. First, find the mean temperature (mean = 21.43). Then, calculate the squared differences from the mean for each temperature, sum them up, and divide by the total number of observations to get the variance.**  1. **Standard Deviation:**   **The square root of the variance, indicating the average distance of data points from the mean. It provides a measure of the spread of data, with higher values indicating greater variability.**   * + **Example: Using the same dataset of daily temperatures: {20, 22, 19, 25, 21, 23, 20}. After finding the variance, take the square root to get the standard deviation, which measures the average deviation from the mean temperature.** | Understanding (2) |
|  | (a)What do you mean by structured data? **(3 marks)** Give advantages, disadvantages of structured data**(2 marks)** and various tools used to process structured data. **(2 marks)**  Structured data refers to data that is organized and stored in a fixed format, typically within a database or a spreadsheet, where each data element is identified and can be accessed using predefined methods. This organization is based on a well-defined schema or model, which specifies the format, type, and relationships of the data elements within the dataset.  Characteristics of Structured Data:  Data conforms to a data model and has easily identifiable structure  Data is stored in the form of rows and columns Example : Database  Data is well organised so, Definition, Format and Meaning of data is explicitly known  Data resides in fixed fields within a record or file  Similar entities are grouped together to form relations or classes  Entities in the same group have same attributes  Easy to access and query, So data can be easily used by other programs  Data elements are addressable, so efficient to analyse and process  Advantages of structured data:   1. **Ease of Storage and Retrieval**: Structured data is organized in a predictable format, making it easier to store, query, and retrieve using standard database management systems. 2. **Efficient Analysis**: The organized nature of structured data allows for efficient data analysis, as it can be easily processed using various statistical and analytical tools. 3. **Interoperability**: Structured data formats such as CSV, XML, or relational databases facilitate interoperability between different systems and applications, enabling seamless data exchange.   Disadvantages of structured data:   1. **Limited Flexibility**: Structured data is constrained by its predefined schema, making it less adaptable to changes or variations in data structure or format. 2. **Complexity in Handling Unstructured Data**: Structured data formats may not be suitable for handling unstructured or semi-structured data types such as text, images, or multimedia content, which require more flexible data models.   Tools used to process structured data:   1. **Relational Database Management Systems (RDBMS)**: Examples include MySQL, PostgreSQL, Oracle, SQL Server, which are widely used for storing, querying, and managing structured data in relational databases. 2. **Spreadsheet Software**: Tools like Microsoft Excel, Google Sheets, and LibreOffice Calc are commonly used for managing and analyzing structured data in tabular format. 3. **Data Analysis and Visualization Tools**: Software such as Tableau, Power BI, and Google Data Studio provide capabilities for analyzing and visualizing structured data to extract insights and generate reports.   Or  (b) What is overfitting and under fitting in machine learning, **(5 marks)** and how does it affect the performance of a model? **(2 marks)**  Overfitting and underfitting are common phenomena in machine learning models:   1. **Overfitting**:    * Overfitting occurs when a model learns the training data too well, capturing noise and random fluctuations in the data rather than the underlying patterns.    * This results in a model that performs very well on the training data but fails to generalize to new, unseen data.    * Signs of overfitting include excessively high accuracy on the training data but poor performance on validation or test data. 2. **Underfitting**:    * Underfitting happens when a model is too simple to capture the underlying structure of the data.    * The model may fail to capture important patterns and relationships, resulting in poor performance on both training and validation/test data.    * Signs of underfitting include low accuracy on both training and validation/test data.   **Effects on Model Performance:**   * **Overfitting**:   + Overfitting leads to poor generalization performance, where the model performs well on the training data but poorly on new, unseen data.   + The model may fail to make accurate predictions on real-world data, as it has memorized noise and irrelevant details from the training data.   + Overfitting can result in misleading insights and erroneous conclusions if the model is deployed in practice. * **Underfitting**:   + Underfitting results in a model that fails to capture the underlying patterns and relationships in the data.   + The model may produce inaccurate predictions and fail to provide meaningful insights.   + Underfitting indicates that the model is too simple to represent the complexity of the data, and it may require a more sophisticated algorithm or additional features to improve performance | Understanding (2) |
| 3. | (a) Give an overview of Macros**(5 marks)** with suitable example. **(2 marks)**  **Macros in the context of computing generally refer to a way of automating repetitive tasks by recording a sequence of actions and then replaying them with a single command. Macros are commonly used in various software applications to streamline workflows, increase efficiency, and reduce manual labor.**  **Overview of Macros:**   1. **Recording Actions: The process typically begins by recording a series of user actions, such as mouse clicks, keystrokes, menu selections, or commands issued in a software application.** 2. **Scripting or Programming Language: The recorded actions are then translated into a script or programming language that represents the sequence of steps required to perform the task.** 3. **Replay and Automation: Once the macro is created, it can be replayed or executed with a single command, automating the task without requiring manual intervention.** 4. **Customization and Flexibility: Macros can often be customized and edited to suit specific requirements, allowing users to tailor them to their needs and preferences.** 5. **Productivity Enhancement: By automating repetitive tasks, macros can significantly increase productivity and efficiency, freeing up time for more valuable or creative activities.**   **Example**  A list of data with different names is available in the table below. Some names have “.” symbol. We want to replace the “.” symbol with “\_” by using macros in Excel.  You can download this Macro Excel Template here – [**Macro Excel Template**](https://www.wallstreetmojo.com/macros-in-excel/#popmake-95356)  Macro Example 1  The steps to add an excel macro are listed as follows:   * Click the “record macro” option in the Developer tab.   Macro Example 1-1   * The “record macro” window will pop out. Name the macro “ReplaceDot**”**in the “macro name” box**.**To assign a [**keyboard shortcut**](https://www.wallstreetmojo.com/keyboard-shortcuts-excel/), type “Ctrl+q” in the “shortcut key” box.   Select the option “This Workbook” in the “store macro in” box, which will ensure the macro is stored in the particular workbook.  It is optional to fill the “description” box explaining the task. Finally, click the “ok” button.  Macro Example 1-2   * The “ReplaceDot” macro will start recording the user actions in Excel. The user will observe the “stop recording” button appearing in the Developer tab.   Macro Example 1-3   * Let us now start replacing the “.(dot)” in the names with “\_(underscore)” by using the “find and replace” option. Enter “.” in the “find” and “ \_” in the “replace” option, respectively. Then click the “replace all” button.   **Note:** Use the shortcut key “Ctrl+H” to use “find and replace” option.  Shortcut key to FindMacro Example 1-4   * The “replace all” option replaces all the “.” (dots) with the “\_” (underscores). The number of replacements and the resulting output is shown in the succeeding image.   Macro Example 1-5   * The final output is displayed in the below image.   Macro Example 1-6   * In the end, click the “stop recording” button on the Developer tab to stop the macro recording.   Or  (b) What is pivoting? **(5 marks)** Explain the types of pivoting. **(2 marks)**  A pivot table is a statistics tool that summarizes and reorganizes selected columns and rows of data in a [spreadsheet](https://www.techtarget.com/whatis/definition/spreadsheet) or database table to obtain a desired report. The tool does not actually change the spreadsheet or database itself, it simply “pivots” or turns the data to view it from different perspectives.  Pivot tables are especially useful with large amounts of data that would be time-consuming to calculate by hand. A few data processing functions a pivot table can perform include identifying sums, averages, ranges or outliers. The table then arranges this information in a simple, meaningful layout that draws attention to key values.  Pivot table is a generic term, but is sometimes confused with the [Microsoft](https://www.techtarget.com/searchwindowsserver/definition/Microsoft) trademarked term, PivotTable. This refers to a tool specific to [Excel](https://www.techtarget.com/searchenterprisedesktop/definition/Excel) for creating pivot tables. How pivot tables work When users create a pivot table, there are four main components:   1. Columns- When a field is chosen for the column area, only the unique values of the field are listed across the top. 2. Rows- When a field is chosen for the row area, it populates as the first column. Similar to the columns, all row labels are the unique values and duplicates are removed. 3. Values- Each value is kept in a pivot table cell and display the summarized information. The most common values are sum, average, minimum and maximum. 4. Filters- Filters apply a calculation or restriction to the entire table.   For example, a store owner might list monthly sales totals for a large number of merchandise items in an Excel spreadsheet. If they wanted to know which items sold better in a particular financial quarter, they could use a pivot table. The sales quarters would be listed across the top as column labels and the products would be listed in the first column as rows. The values in the worksheet would show the sum of sales for each product in each quarter. A filter could then be applied to only show specific quarters, specific products or averages. Uses of a pivot table A pivot table helps users answer business questions with minimal effort. Common pivot table uses include:   * To calculate sums or averages in business situations. For example, counting sales by department or region. * To show totals as a percentage of a whole. For example, comparing sales for a specific product to total sales. * To generate a list of unique values. For example, showing which states or countries have ordered a product. * To create a 2x2 table summary of a complex report. * To identify the maximum and minimum values of a dataset. * To query information directly from an online analytical processing ([OLAP](https://www.techtarget.com/searchdatamanagement/definition/OLAP)) server.   For example, you may have hundreds of entries in your worksheet with sales figures of local resellers: The source data for a summary table  One possible way to sum this long list of numbers by one or several conditions is to use formulas as demonstrated in [SUMIF](https://www.ablebits.com/office-addins-blog/excel-sumif-function-formula-examples/) and [SUMIFS](https://www.ablebits.com/office-addins-blog/excel-sumifs-multiple-criteria/) tutorials. However, if you want to compare several facts about each figure, using a Pivot Table is a far more efficient way. In just a few mouse clicks, you can get a resilient and easily customizable summary table that totals the numbers by any field you want. Excel Pivot Table examples  The screenshots above demonstrate just a few of many possible layouts. And the steps below show how you can quickly create your own Pivot Table in all versions of Excel. | Apply(3) |
| 4. | (a) Discuss the usage of VLOOKUP and XLOOKUP**(5 marks)** operations used for data analysis. **(2 marks)**  **VLOOKUP (Vertical Lookup):**   1. **Usage**:    * VLOOKUP is used to search for a value in the first column of a table (known as the lookup table) and return a corresponding value from a specified column within the same row.    * It is commonly used to perform approximate or exact matches, making it useful for tasks like retrieving prices, names, or other information associated with specific identifiers or codes. 2. **Syntax**:   =VLOOKUP(lookup\_value, table\_array, col\_index\_num, [range\_lookup])   * + **lookup\_value**: The value to search for in the first column of the table.   + **table\_array**: The range of cells that contains the lookup table.   + **col\_index\_num**: The column number in the table from which to retrieve the value.   + **range\_lookup**: Optional. If TRUE or omitted, VLOOKUP searches for approximate matches. If FALSE, it searches for exact matches.   **XLOOKUP:**   1. **Usage**:    * XLOOKUP is a newer and more versatile function available in modern versions of Excel and Google Sheets.    * It allows users to perform horizontal and vertical lookup operations, as well as handle array operations and search for values in multiple columns or rows simultaneously.    * XLOOKUP is especially useful when dealing with non-contiguous ranges, dynamic arrays, and more complex lookup scenarios. 2. **Syntax**:   =XLOOKUP(lookup\_value, lookup\_array, return\_array, [if\_not\_found], [match\_mode], [search\_mode])   * + **lookup\_value**: The value to search for.   + **lookup\_array**: The range of cells to search within.   + **return\_array**: The range of cells containing the values to return.   + **[if\_not\_found]**: Optional. Specifies the value to return if no match is found.   + **[match\_mode]**: Optional. Specifies the match type (exact match, less than, greater than, etc.).   + **[search\_mode]**: Optional. Specifies the search direction (first to last, last to first).   **Usage in Data Analysis:**   * Both VLOOKUP and XLOOKUP are widely used in data analysis tasks such as:   + Merging datasets by matching key identifiers.   + Extracting and combining information from different tables.   + Performing data validation and error checking.   + Creating dynamic reports and dashboards.   + Handling data from databases or external sources.   Or  (b) How can I use the INDEX and MATCH formula in Excel to find **(5 marks)** and retrieve specific data from a table? **(2 marks)**   1. **MATCH Function**:    * MATCH is used to search for a specified value in a range of cells and returns the relative position of that item.    * Syntax:   =MATCH(lookup\_value, lookup\_array, [match\_type])   * + - **lookup\_value**: The value you want to find.     - **lookup\_array**: The range of cells where you want to search for the value.     - **[match\_type]**: Optional. Specifies the type of match (exact match, less than, greater than).  1. **INDEX Function**:    * INDEX is used to retrieve the value of a cell in a specified row and column within a table or range.    * Syntax:   =INDEX(array, row\_num, [column\_num])   * + - **array**: The range of cells containing the data.     - **row\_num**: The row number from which to retrieve the data.     - **[column\_num]**: Optional. The column number from which to retrieve the data. If omitted, INDEX returns the entire row.   **Steps to Retrieve Specific Data:**   1. **Identify the Lookup Value**:    * Determine the value you want to search for in the table. 2. **Use MATCH to Find the Row Number**:    * Apply the MATCH function to find the relative position of the lookup value in the row or column.    * For example:   =MATCH(lookup\_value, lookup\_array, 0)   * + This will return the row number where the lookup value is found.  1. **Use INDEX to Retrieve the Data**:    * Apply the INDEX function to retrieve the data from the table based on the row number obtained from MATCH.    * For example:   =INDEX(data\_range, match\_result, column\_number)   * + **data\_range**: The range of cells containing the data.   + **match\_result**: The result of the MATCH function (row number).   + **column\_number**: (Optional) If you want to retrieve data from a specific column, specify the column number. Otherwise, omit this argument to retrieve the entire row.   **Example:**  Suppose you have a table of sales data with product names in column A, and sales figures in column B. You want to retrieve the sales figure for a specific product, "Product X".   1. Use MATCH to find the row number of "Product X":   =MATCH("Product X", A:A, 0)   * + This will return the row number where "Product X" is found.  1. Use INDEX to retrieve the sales figure for "Product X":   =INDEX(B:B, match\_result)   * + **B:B** refers to the sales figures column.   + **match\_result** is the result obtained from the MATCH function. | Apply(3) |