ON 7th Nov 2024

DEPARTEMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

OPTION: INFORMATION TECHNOLOGY

PROGRAM: B-TECH

MODULE: MACHINE LEARNING

ASSIGNMENT GROUP

GROUP MEMBERS:

1. MEREWENEZA John 24RP15451
3. NIYIGABA Claude 24RP14647

Data Analysis and Preparation Concepts

1. Importance of Data Analysis

Data analysis is crucial because it transforms raw data into valuable insights, helping businesses and researchers make data-driven decisions. It aids in identifying patterns, trends, and correlations, improving accuracy in predictions and strategies. Data analysis is also essential for problem-solving, risk assessment, and process optimization in various fields such as marketing, finance, healthcare, and engineering.

2. Approaches for Data Cleaning

Data cleaning is the process of preparing data by fixing or removing errors and inconsistencies. Common approaches include:

- Removing duplicates: Identifying and deleting repeated entries.
- Handling missing values: Using methods like imputation, removal, or substitution.
- Outlier detection and handling: Identifying unusual data points and deciding on treatment.
- Standardizing formats: Ensuring uniform data formats (e.g., dates, strings).
- Correcting errors: Fixing typos, misspellings, or incorrect values.
- Data type conversion: Converting data into the proper types for analysis.
- Removing irrelevant data: Dropping columns or rows not needed for analysis.

3. Univariate vs. Multivariate Analysis

- Univariate Analysis: Involves the analysis of a single variable, primarily to understand its distribution, central tendency, and spread. Example: Examining the age distribution of a population using histograms or box plots.
- Multivariate Analysis: Involves the analysis of multiple variables to understand relationships and interactions between them. Example: Analyzing the relationship between age, income, and spending habits using scatter plots or regression analysis.

4. Why Data Wrangling is Used & Its Steps

Data wrangling, also known as data preprocessing, is used to transform and prepare raw data into a clean and usable format for analysis. It ensures data quality and readiness for analysis or modeling.

- Steps in Data Wrangling:
- Data collection: Gathering data from different sources.
- Data cleaning : Removing inconsistencies, duplicates, and handling missing values.
- Data transformation: Normalizing, scaling, or encoding data as necessary.
- Data integration: Merging data from different sources or tables.
- Data validation: Ensuring data accuracy and consistency.
- Data structuring: Organizing data into appropriate formats for analysis (e.g., converting data into tables or datasets).

5. Removing Duplicate Entries from a Dataset

To remove duplicates, you can use functions like drop_duplicates() in Python's pandas library or similar commands in R, SQL, and Excel. These functions typically allow for specifying columns to check for duplicates and provide options for keeping the first, last, or removing all duplicates.

6. Fundamentals of Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an initial step in data analysis where analysts investigate datasets to discover patterns, anomalies, and relationships. Key fundamentals include:

- Data visualization: Using graphs and charts to spot trends, distributions, and relationships.
- Descriptive statistics: Calculating mean, median, mode, variance, and skewness to summarize data.
- Identifying outliers: Recognizing extreme values that may impact analysis.
- Checking correlations: Analyzing relationships between variables (e.g., using correlation matrices).
- Data cleaning: Detecting missing values, duplicates, and potential errors in the data.

7. Types of Exploratory Data Analysis

- Univariate Analysis: Examining the distribution and summary statistics of each variable independently.
- Bivariate Analysis: Analyzing the relationship between two variables (e.g., scatter plots, correlation).
- Multivariate Analysis: Examining interactions among multiple variables to understand complex relationships.