EXPT. NO. | NAME | Page No.: 10 | Date: 4/2/25 | YOUVA

M T W T F S S

EXPERIMENT NO. 4.

AIM: To treat missing values with different techniques in python.

THEORY:

Missing data is a common issue in real-world datasets and can arise due to various reasons such as human error, data corruption or system failures. Handling missing values effectively is essential to ensure accurate data analysis and reliable machine learning models.

Types of missing Data:

1. Missing Completely at Random (MCAR): The missing values occur randomly without any pattern.
2. Missing at Random (MAR): The missing values are related to some observed data but not the missing data itself.
3. Missing Not at Random (MNAR): The missing values have a specific pattern and are dependent on unobserved factors.

Teacher's Signature: _____

EXPT. NO. | NAME | Page No: 11 | Date: 4/2/25 | YOUVA

M T W T F S S

Techniques to Handle Missing values

(1) Mean Imputation : Replaces missing values with the mean of the column.

(2) Median Imputation : Replaces missing values with the median of the column, useful when the data contains outliers.

(3) Most Frequent (Mode) Imputation : Replaces missing values with the most frequent value in the column, effective for categorical data.

First, we will extract the required columns from the dataset that contain missing values

The median strategy is useful for handling numerical data that contains outliers. Missing values in Age and Experience columns are replaced with their respective median values.

The most_frequent strategy replaces missing values with the most common value in the column. This method is effective for categorical data but can also be applied to numerical data.

EXPT. NO. | NAME

M T W T F S S

Page No: 12

Date: 4/2/25

YOUVA

CONCLUSION:

Handling missing values is an important step in data preprocessing. The choice of imputation strategy depends on the nature of the data:

- Median imputation is useful when data contains outliers while most frequent imputation is effective for categorical data.

By applying these techniques, we ensure that missing data does not impact the overall analysis or machine learning model performance.