GOVERNMENT COLLEGE
KASARAGOD

Micro Course in Data Analysis

Week - 3
Session 6: Introduction to Preprocessing

By
Reshma Soosan George

Open Data Lab
GCK

Introduction to Pre-Processing

Advantages of Pre-Processing.

- Technique for converting raw data into clean data.
- Pre-processing makes the data feasible for analysis.
- To make data into better format in order to get better result.
- Data mining, Data cleaning, Data exploration and feature engineering are data pre-processing.

Steps Involved in Data Pre-processing:

1. **Data Cleaning**
   - Handling of missing values
   - Handling of duplicates
   - Handling of noisy data
2. **Data Transformation**
   - Transform data into suitable format for mining process
   - Normalization
   - Attribute selection
3. **Data Reduction**
   - Dimensionality Reduction

Handling of Missing Values

- Missing Data can occur when no information is provided.
- Failed to record data values.

Functions with which we find out whether we have null values in data:

- ❑ is.na().sum()
- ❑ isnull().sum()

Methods of Dealing With Missing Data

- Deleting the entire column/Row

  - ❑axis=1 is used to drop the column with NaN values.
  - ❑axis=0 is used to drop the row with NaN values.

- Imputation method

  - ❑Imputation using mean
  - ❑Imputation using median
  - ❑Imputation using mode

Handling of Outliers

- Outlier is a data point that differs significantly from other observations. i.e, It is extremely low or data point.

An outlier has to satisfy either of the following two conditions:

```
outlier < Q1 - 1.5(IQR)
```

```
outlier > Q3 + 1.5(IQR)
```

An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.

Encoding

- Encoding is converting categorical data to numbers before we use it for fitting and evaluating a model.
- All input output variables have to be converted into numerical variable.

E.g.:

	ENCODED COLUMN
Work_type	**Work_type**
Private Sector	0
Government Sector	1
Children	2
Government Sector	1

Types of Encoding

- One hot encoding
  - Dataset contains column that has no specific order
  - Data in column denote a category

- Label Encoding
  - Label encoding converts the categorical data into numerical ones, but it assigns a unique number to each class of data.

Scaling

- Min-max Scaling
  - ❖MinMaxScaler scales the data to a fixed range, typically between 0 and 1.

- Standard Scaling
  - ❖Rescale feature value such that it has mean 0 and variance 1
  - ❖It follows Std normal distribution

Thank you