UNIT 1

Data Pre-processing and Data Cleaning

Presented By :
Aditya Joshi
Assistant Professor
Graphic Era Deemed to be University, Dehradun

Data Pre-processing (Data in the real world is dirty and raw)

- Transform data into understandable form.
- It is done to improve the quality of data in dataset.
- Increases efficiency.
- Ease of Mining Process.
- Removes Noisy Data, Inconsistent Data and Incomplete Data.

Noisy Data : Containing errors and outliers

Salary="-1"

Inconsistent Data: Containing discrepancies in codes and names

Age ="42" dob="03-08-1970"

Was rating="1,2,3" now rating ="A, B, C"

Incomplete Data: Missing Attribute Values, lack of certain attribute of interest.

Occupation=""

Why data pre-processing in important

- No quality Data, no quality mining result

✓Quality result must be based on quality data

✓Duplicate or missing data may cause incorrect or even misleading statistics.

- Data preparation, cleaning, and transformation will do 90% of work

What is Data Pre-processing: The overall process of making data more suitable for statistical analysis.

It includes several tasks employed in the process to make data more relevant.

Steps involved in Data Pre-processing in data set(Data Pre-processing Task)

1. Data Cleaning
2. Data Integration
3. Data Transformation (normalization, Aggregation)
4. Data Reduction(to reduce data and does not effect useful information)
5. Data Discretization (it is also a part of data reduction) (categorized on the basis of Numeric/category)

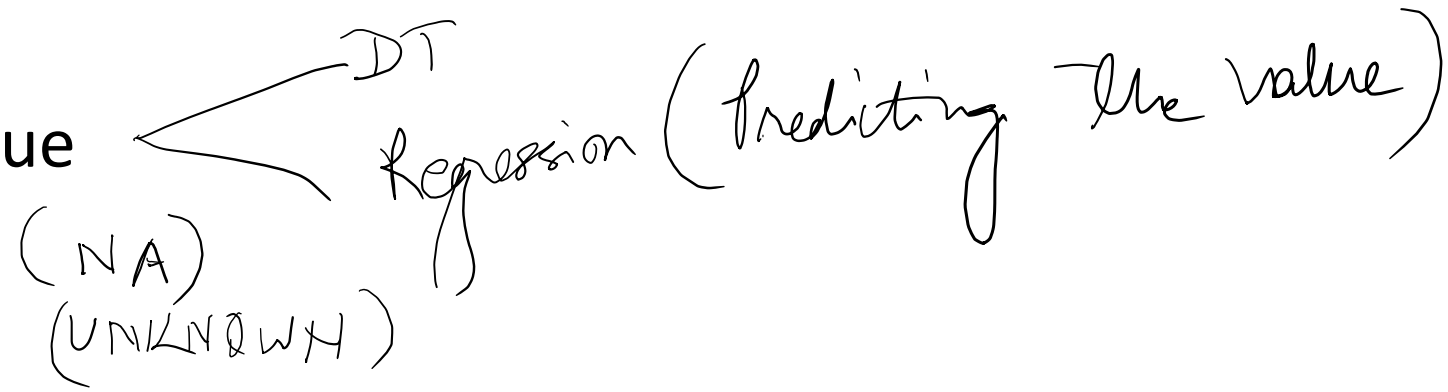
Data Cleaning

- It Cleans the Data by filling the missing values
- Smoothing Noisy Data (random Error or Variance in measures variable/meaning less data)
- Resolving the inconsistency and removing the outliers

↳ Naming Conventions

How to Handle Missing Data During Cleaning

- Manual Entry of missing Data
- Using Attribute Mean
- Using Most Probable Value
- Using Global Constant
- Ignore the Tuple



Steps in ML (example whether the drink in wine or beer)

- Gathering Data
- Preparing Data
- Choosing Data Model
- Training
- Evaluation
- Parameter Tuning (selecting best parameters for an algorithm to optimize its performance)
- Prediction