DMDW LAB

ASSIGNMENT 3

1

TOPICS

- Data Preprocessing:
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

2

Data Preprocessing

- **Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.**

- **Data Preprocessing is that step in which the data gets transformed or Encoded to bring it to such a state that now the machine can easily parse it.**

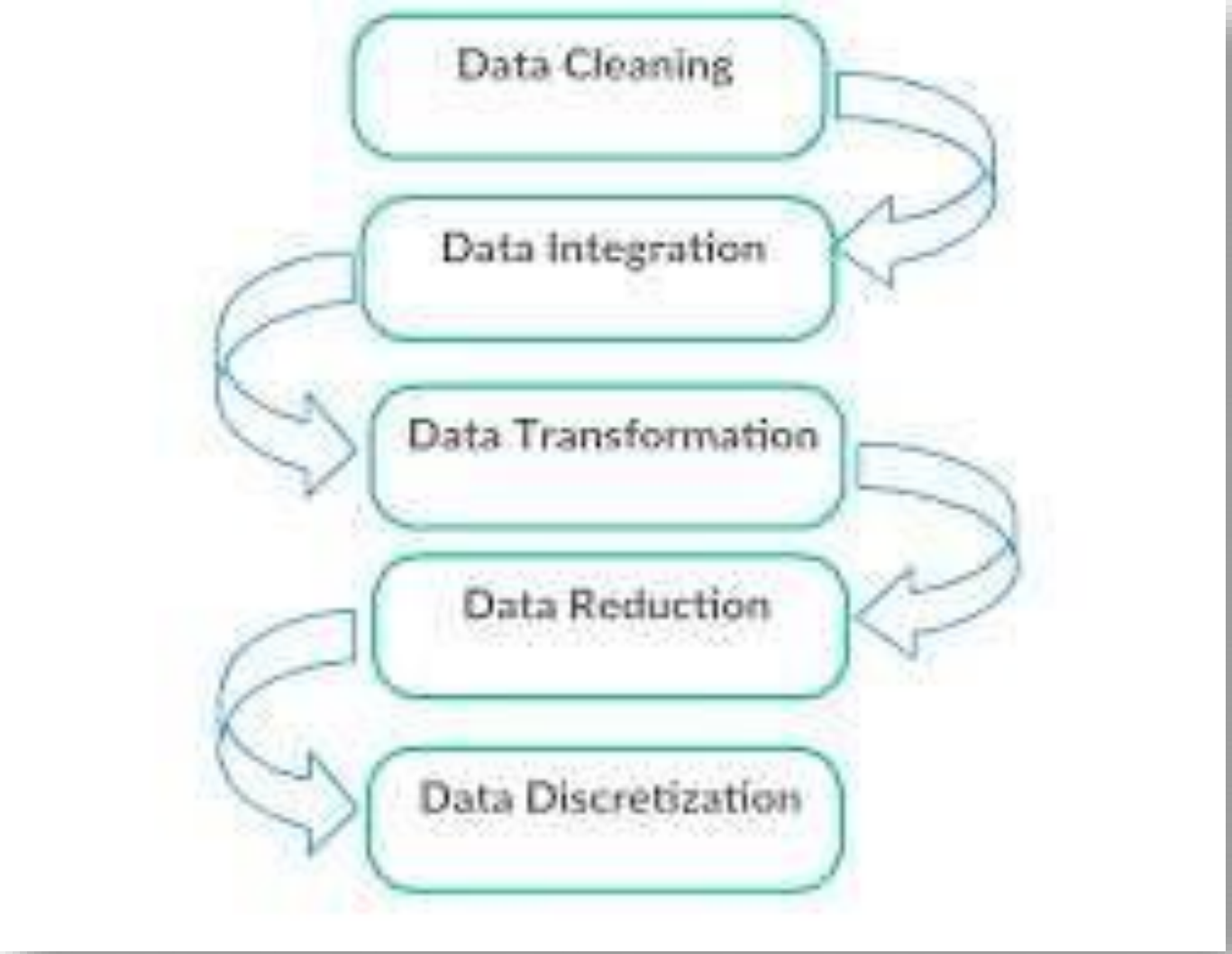
- **Data preprocessing includes cleaning , instance selection, normalization, transformation, feature extraction and selection.**

3

Why Preprocess the Data?

- **Accuracy:** correct or wrong, accurate ornot
- **Completeness:** not recorded, unavailable, …
- **Consistency:** some modified but some not,dangling, …
- **Timeliness:** timely update?
- **Believability:** how trustable the data are correct?
- **Interpretability:** how easily the data can be understood?

Major Tasks in Data Preprocessing



Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data reduction
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- Data transformation and data discretization
  - Normalization
  - Concept hierarchy generation

What is Data Cleaning?

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=“ ” (missing data)
  - Noisy: containing noise, errors, or outliers
    - e.g., *Salary*=“–10” (an error)
  - Inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*=“42”, *Birthday*=“03/07/2010”
    - Was rating “1, 2, 3”, now rating “A, B, C”
  - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone’s birthday?

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing which is not effective when the % of missing values per attribute varies considerably.
- Fill in the missing value manually
- Fill in it automatically with
  - A global constant : e.g., "unknown", a new class
  - The attribute mean

What is Noisy Data?

- **Noise:** Random error or variance in a measured variable
- **Incorrect attribute values** may be due to
  - Faulty data collection
  - Data entry problems
  - Data transmission problems
  - Technology limitation
  - Inconsistency in naming convention
- **Other data problems** which require data cleaning
  - Duplicate records
  - Incomplete data
  - Inconsistent data

Assignment 3

- **Upload the Toyota Dataset :**

path=https://raw.githubusercontent.com/archana1822/DMDW-Lab/main/Toyota.csv

```
import pandas as pd
data =pd.read_csv(path)
```

- **Use the following command on Toyota dataset**

1. type(data)
2. data.shape
3. data.info()
4. data.index
5. data.columns

Assignment 3

6. data.head()
7. data.tail()
8. data.head(5)
9. data[['Price',"Age"]].head(10)
10. data.isnull().sum()
11. data.dropna(inplace=True)
    data.isnull().sum()
12. data.shape
13. data.head(10)
14. data['MetColor'].mean()
15. data['MetColor'].head()

Assignment 3

16. import numpy as np

data['MetColor'].replace(np.NaN,data['MetColor'].mean()).head()

17. data.head(10)
18. data[‘CC'].mean()
19. data[‘CC'].head()
20. data[[‘Age',“KM"]].head(20)