Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

T.Y.B.Tech (CSE)

Data Warehousing And Data Mining

Lab Assignment No – 1

Name: Aniruddha Shende

Roll number: PE04

Batch: E1

Panel: E

Name:- Aniruddha Arun Shende
Roll no:- PE04
Batch:- E1
Panel:- E

Subject:- Data Warehousing & Data Mining (DWDM)

ASSIGNMENT - 1

Aim: Preprocess Data Using Python

Objective:
- To clean data & make it noise free.
- Prepare data for analysis.

Problem Statement:
Perform Preprocessing on a Dataset of your choice. I have taken Melbourne dataset.

Case Study:
Download different dataset from kaggle, UCI like diamond.csv, adult.csv, & WorldCupMatches.csv Use Python libraries like pandas, numpy, sklearn, etc.
Perform following tasks on suitable dataset.
Perform different data preprocessing techniques for following categories like
- Data cleaning: Missing Data, Noisy Data (Binning), Regression Analysis.
- Data Integration: Redundancy (Correlation Analysis Pearson, Chi Square).
- Tuple Duplication

- Data Reduction: Histogram on larger Datasets, Sampling
- Data Transformation: Min-Max normalization, Z-score Normalization.
- Data Encoding Methods: Label Encoding, One hot encoding.

Theory:

Explain following points with standard function used in python:

1) Data Preprocessing

→ (a) Data preprocessing is a data mining technique that contains transforming raw data into an understandable format.

(b) Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data Preprocessing is a proven method of resolving such issues.

2) Data cleaning

→ In data cleaning, data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

3) Data Integration

→ In the process of Data Integration, data with different representations are put together and conflicts within the data are resolved.

s

4) Data Redundancy

→ During data integration in data mining, various data stores are used. This can lead to the problem of redundancy in data. An attribute is called redundant if it can be derived from any other attribute or set of attributes. Inconsistencies in attribute or dimension naming can also lead to the redundancies in data set.
Data Redundancy also occurs when same piece of code is stored in 2 or more separate places.

5) Data Transformation :

→ ⓐ Data transformation is a data pre-processing technique used to re-organize or restructure the raw data in such a way that the data mining retrieves strategic information efficiently & easily.
ⓑ So basically, in data transformation, data is normalized, aggregated & generalized.

Input : Dataset

Output: Cleaned, Integrated, transformed dataset.

Platform: Windows

Conclusion :

Thus, we have learned different preprocessing techniques using python on .csv file.

FAQ's (Continued on next page.....)

FAQ's

1) What are different data types in data mining?

Ans 1) The data types in data mining are spatial data, graph data, data streams, engineering design data, multi-media data, etc.

2) What are the applications of correlational analysis? Illustrate Pearson correlation with an example?

Ans 2) Correlation analysis helps to determine how strongly one attribute implies other attribute. Correlation coefficient always lies between -1 & $+1$.

For eg: ① Time spent on E-commerce website vs Money spent by a customer.
② Salary / income of person vs Area of Home
③ No. of years of study vs income.

Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes. For eg:- Up till certain age For eg:- a child's height will keep increasing as his / her age increases.

$$r_{A,B} = \frac{\Sigma (A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A \sigma_B}$$

3) What is noise in data?

Ans 3) Noise is a random error or variance in measured variable. This term has often been used as a synonym for corrupt data.