Department of Computer Science and Engineering (Data Science)

Lab Manual

Subject: Foundations of Data Analysis Laboratory (DJ19DSL303)

Semester: III

Experiment 7

(Data Preprocessing)

Aim: Perform Data cleaning on a given dataset.

Theory: Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset.

After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

Missing Values may involve removal of those entries (Usually if number of missing values is low and/or the field is important for analysis), estimated (If high correlation exists, low number of missing values), or that field/column may be dropped (large number of missing values and/or)

1. Remove duplicate or irrelevant observations
2. Fix structural errors
3. Filter unwanted outliers
4. Handle missing data
5. Validate

Dataset: Reservations.csv

Perform the following if required

1. Remove Duplicate Values

2. Imputation of missing values

3. Remove outliers

4. Correlation analysis

5. Data Transformation.