DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

Course Machine Learning

Semester: 8th, Batch

2020S

Assignment No.01

Assigned date: 03/11/2023 **Submission date: 10/11/2023**

Name: Sheikh Abdul Rehman
Roll No: 2020s-CS-236
Section: 8E

Question No.01 **Marks.03**

(i) Why we need Data Preprocessing for any research work. Discuss some major technique with suitable example for data preprocessing.

Answer:

Data Preprocessing:

Data preprocessing is an essential step in any research work because it helps to clean, transform, and organize raw data before it can be analyzed. This is important because real-world data is often noisy, incomplete, inconsistent, or contains irrelevant information. Data preprocessing enables researchers to improve the quality of data and enhance the accuracy of analytical models and algorithms.

Here are some major techniques used in data preprocessing along with suitable examples:

1. Data Cleaning: It involves removing noise and handling missing or erroneous data. For example, if a dataset contains missing values in certain attributes, various approaches can be used, such as substituting missing values with the attribute's mean or median, or using sophisticated techniques like multiple imputation or expectation-maximization.

2. Data Integration: This technique combines data from different sources into a coherent dataset. For instance, consider a research project that involves analyzing consumer behavior. The data might be collected from various sources, such as surveys, social media, and customer records. Data integration would involve merging these different sources into a unified dataset for analysis.

3. Data Transformation: This technique involves transforming the data into a suitable format for analysis. This may include assigning numerical values to categorical variables, log transformation of skewed data, or standardizing variables to have similar scales. For example, in a study that analyzes housing prices, a categorical variable like "neighborhood" can be transformed into numerical values using techniques like one-hot encoding or label encoding.

4. Data Reduction: This technique reduces the data's dimensionality while preserving its important characteristics. Principal Component Analysis (PCA) is a common technique used for data reduction. For instance, in a genetics research project, if a dataset contains thousands of genetic markers, PCA can be used to reduce the dimensionality and identify the most influential markers.

5. Data Discretization: This technique transforms continuous variables into discrete intervals. This is mainly used to handle numerical data that may be highly granular or to reduce noise. For instance, in a study analyzing customer age groups, continuous age variables can be discretized into categories like "young," "middle-aged," and "senior."

(ii) Please download suitable dataset from Kaggle, GitHub or UCI Machine Learning repository. Represent all major data preprocessing techniques in python like removing missing values, Column transformation, removing outliers, drop duplicates, data integration and data reduction.
Answer:
https://github.com/Abdul-Rehman745/DATA-PROCESSING-TECHNIQUE-ML-USING-PYTHON/blob/main/DATA%20PROCESSING%20TECHNIQUE%20ML%20ASSIGNMENT%201.ipynb