**Batch: H2-4 Roll No.:16010122257**

**Experiment 02**

|  |
| --- |
| **Title:**  Dataset pre-processing |

# Objective:

# 1. To learn how to prepare the dataset

# 2. To learn various steps in Data -Preprocessing

# Course Outcome:

# CO1: Learn how to locate and download datasets, extract insights from that data and present their findings in a variety of different formats.

# Books/ Journals/ Websites referred:

<https://www.kaggle.com/>

<file:///C:/Users/DELL/Downloads/visual-data-storytelling-with-tableau_LindyRyan%20(2).pdf>

# Resources used:

# <https://www.kaggle.com/datasets/kaggleprollc/suicides-in-india-data-collection>

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Theory (About Data Preprocessing): Data preprocessing is one of the important steps in the data mining process. It refers to the cleaning, transforming, and integrating of data for making it analysis-ready. The goal of data preprocessing is to improve the quality of the data and making it more suitable for specific data mining tasks.

# Following points should be written by students

# Different steps in Data Preprocessing:

# Finding missing, null values

# Replacing missing, null values with statistical parameters

# Encoding categorical data

# Normalization

# Finding Missing and Null Values:

# Missing values: These are the values that are absent from the dataset. They can occur due to various reasons such as data corruption, human error, or incomplete data collection.

# Null values: In programming, null represents a variable or data that does not exist in memory. In the context of data, null values are often used interchangeably with missing values.

# The first step is to identify where these missing or null values exist in the dataset. This can be done using various functions or methods provided by the programming language or libraries you are using, like isnull() in Python's pandas library.

# Replacing Missing and Null Values with Statistical Parameters:

# One way to handle missing or null values is to replace them with statistical parameters like mean, median, or mode. This is called imputation.

# For numerical data, you can replace missing values with the mean (average) or median (middle value). Mean is sensitive to outliers, while the median is more robust in the presence of outliers.

# For categorical data, you can replace missing values with the mode (most frequent value) of that column.

# Encoding Categorical Data:

# Machine learning algorithms require numerical input, so categorical variables (variables that can take on a limited, fixed number of values) need to be converted into numerical form.

# One common technique is one-hot encoding, where each unique category is converted into a binary vector. Each category becomes a separate column, and a 1 or 0 indicates the presence or absence of that category for a particular data point.

# Another method is label encoding, where each category is assigned a unique integer. However, this method should be used carefully, as the algorithms might interpret these integers as having some sort of ordinal relationship, which might not be the case for all categorical variables.

# Normalization:

# Normalization is the process of scaling and centering numerical variables. It's essential when features have different ranges, as some machine learning algorithms are sensitive to the scale of the input features.

# Common techniques include Min-Max scaling, where values are scaled to fall within a specific range (usually 0 to 1), and Z-score normalization (standardization), where values are scaled to have a mean of 0 and a standard deviation of 1.

# Normalization ensures that all features contribute equally to the analysis and prevents features with larger scales from dominating the learning algorithm.

# Note: Student can use any technology like Tableau, Tableau-Prep, PowerBI, Google spreadsheet, excel, R programming, Python, Java any other technology for preprocessing.

# Platform used by the student:Microsoft excel

# Working (Paste the code and Output for each Data Preprocessing task):

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# Conclusion (Students should write in their own words):

# In this experiment, we focused on a specific dataset and delved into various data processing techniques, including cleaning, integration, reduction, and transformation. Throughout the experiment, we explored methods to handle data such as identifying and dealing with missing values, combining data from different sources, reducing the dimensionality of the dataset, and applying transformations to make the data suitable for analysis and modeling purposes.

# Post Lab Question:

# Write the importance of Data Preprocessing

# Data preprocessing,as the name suggests,is a means to pre-process the data,that is, is a means to pre-process the data, i.e., clean, transform, and organize raw data into a format that is suitable for analysis or machine learning tasks.

# Reasons why data preprocessing is important:

# Improvement in quality of data:Raw data often contains errors,missing values,inconsistencies etc.Data preprocessing helps to identify and rectify those issues,to ensure the accuracy and reliability of the final dataset.

# - Data preprocessing provides methods to impute or remove missing values while minimizing their impact on the analysis.

# Enhanced Accuracy: Machine learning algorithms and statistical analyses rely heavily on the input data. Preprocessing ensures that the data is in the right format and distribution, leading to more accurate and reliable results.

# Normalization and Standardization: Different features in the dataset might be measured in different units or have varying scales. Normalization and standardization techniques bring all features to a common scale, preventing domination of certain features in the analysis due to their larger values.

# Noise Reduction: Noise in data can arise from various sources, such as measurement errors or inconsistencies. Preprocessing techniques like smoothing and filtering can help reduce the noise, revealing underlying patterns and trends.

# Time and Cost Efficiency: Proper preprocessing reduces the need to backtrack during analysis or modeling due to unforeseen data issues. This saves time and resources by avoiding rework and repeated analyses.