**DEPARTMENT OF INFORMATION TECHNOLOGY**

**COURSE CODE:** DJ19TEL7014  **DATE:**

**COURSE NAME:** Machine Learning  **CLASS:** Final Year B.Tech

**EXPERIMENT NO. 1**

**CO1** Solve real-world problems using suitable machine learning techniques.

**TITLE:** Data Preprocessing

**AIM / OBJECTIVE:**

To perform data preprocessing in terms of handling, missing data, removing outliers, eliminating duplicate rows and modifying the datatype, etc.

**DESCRIPTION OF EXPERIMENT:**

Python is an easy-to-learn programming language, which makes it the most preferred choice for beginners in Data Science, Data Analytics, and Machine Learning. It also has a great community of online learners and excellent data-centric libraries. With so much data being generated, it becomes important that the data we use for Data Science applications like Machine Learning and Predictive Modeling is clean. But what do we mean by clean data? And what makes data dirty in the first place? Dirty data simply means data that is erroneous. Duplicacy of records, incomplete or outdated data, and improper parsing can make data dirty. This data needs to be cleaned. Data cleaning (or data cleansing) refers to the process of “cleaning” this dirty data, by identifying errors in the data and then rectifying them. Data cleaning is an important step in and Machine Learning project, and we will cover some basic data cleaning techniques (in Python) in this article.

**Cleaning Data in Python**

We will now separate the numeric columns from the categorical columns.

**Missing values**

We will start by calculating the percentage of values missing in each column, and then storing this information in a DataFrame.

**Drop observations**

One way could be to drop those observations that contain any null value in them for any of the columns. This will work when the percentage of missing values in each column is very less.

**Remove columns (features)**

Another way to tackle missing values in a dataset would be to drop those columns or features that have a significant percentage of values missing.

**Impute missing values**

There is still missing data left in our dataset. We will now impute the missing values in each numerical column with the median value of that column.

**Outliers**

An outlier is an unusual observation that lies away from the majority of the data. Outliers can affect the performance of a Machine Learning model significantly.

**Duplicate records**

Data can sometimes contain duplicate values. It is important to remove duplicate records from your dataset before you proceed with any Machine Learning project. In our data, since the ID column is a unique identifier, we will drop duplicate records by considering all but the ID column.

**Fixing data type**

Often in the dataset, values are not stored in the correct data type. This can create a problem in later stages, and we may not get the desired output or may get errors while execution.

**OBSERVATIONS:**

**CONCLUSION:**

Data cleaning and preprocessing are fundamental steps in the data science and machine learning pipeline. By using pandas and the techniques covered in this experiment, we can ensure that our data is clean, consistent, and ready for further analysis and modeling. These skills are essential for anyone working with real-world datasets, as they enable us to uncover meaningful insights and build robust data-driven solutions.

In conclusion, our experiment on data cleaning and preprocessing using Python's pandas library has provided us with valuable insights and techniques for preparing raw data for analysis and modeling. We employed several important data cleaning and preprocessing methods, including df.dropna(), df.drop\_duplicates(), handling outliers, and managing data types.