Assignment 1

Problem statement

Perform the following operations using R/Python on suitable data sets:
a) read data from different formats (like CSV, xls)
b) indexing and selecting data, sorting data,
c) describe attributes of data, checking data types of each column,
d) counting unique values of data, the format of each column, converting variable data type (e.g. from long to short, and vice versa),
e) identifying missing values and filling in the missing values.

S/W Packages and H/W apparatus used:

OS: Ubuntu/**Windows**
Tools: Google Colab
Packages: **Numpy** and **Pandas**

Theory :

1. Data Preparation:

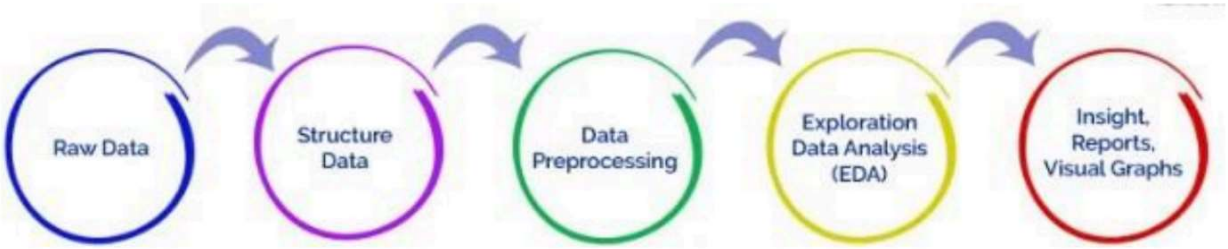
Data preparation also referred as “data preprocessing” is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and combining of data sets to enrich data.

2. Importance of Data preparation:

- Because most machine learning algorithms require data to be structured in a specific way, datasets must be prepared before they can offer useful insights. a number of databases having missing, invalid, or otherwise difficult-to-process values for an algorithm. If you're looking for information, The algorithm will be

unable to use it if it is missing. If the data is incorrect, the algorithm will produce inaccurate or even incorrect results. the outcomes are deceiving.

- Some datasets are relatively clean but need to be shaped (e.g., aggregated or pivoted) and many datasets just lack useful business context (e.g., poorly defined ID values), hence the need for feature enrichment. Good data preparation produces clean and well-curated data which leads to more practical, accurate model outcomes.
- Before entering the data into the machine learning model, this is the most important step. The reason for this is that the data set must be unique and specific to the model, thus we must identify the data's required characteristics. The data preparation process provides a mechanism for preparing data for project definition as well as project evaluation of machine learning algorithms.
- There are a variety of predicting machine learning models available, each with its own method. However, some processes are common to all models, and they allow us to identify the underlying business problem and its solutions. The following are some of the data preparation procedures:
    a. Determine the problems
    b. Data cleaning
    c. Feature selection
    d. Data transformation
    e. feature engineering
    f. Dimensionality reduction



Conclusion: Data preparation is recognized for helping businesses and analytics to get ready and prepare the data for operations.