Academic Year: 2022-23 **Semester: VI**

Class / Branch: TE IT A/B **Subject: DS using Python Lab**

Experiment No.1

Aim: To understand the process of data preparation using NumPy and Pandas

CO Mapped:
CO1: To apply the process of data preparation for the given dataset to solve real-world problems

Prerequisites: Python3, basic syntax of NumPy and Pandas

Theory:

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labelling raw data into a form suitable for machine learning (ML) algorithms and then exploring and visualizing the data.

Derive an index field and add it to the data set

Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of data-centric python packages. Pandas is one of those packages and makes importing and analysing data much easier.

Pandas set_index() is a method to set a List, Series or Data frame as index of a Data Frame. Index column can be set while making a data frame too. But sometimes a data frame is made out of two or more data frames and hence later index can be changed using this method.
Syntax:

DataFrame.set_index(keys, drop=True, append=False, inplace=False, verify_integrity=False)

Parameters:
keys: Column name or list of column name.
drop: Boolean value which drops the column used for index if True.
append: Appends the column to existing index column if True.
inplace: Makes the changes in the dataframe if True.
verify_integrity: Checks the new index column for duplicates if True.

Find out the missing values

Missing Data can occur when no information is provided for one or more items or for a whole unit. Missing Data is a very big problem in a real-life scenarios. Missing Data can also refer to as NA(Not Available) values in pandas. In DataFrame sometimes many datasets simply arrive with missing data, either because it exists and was not collected or it never existed. For Example,

Academic Year: 2022-23 **Semester: VI**

Class / Branch: TE IT A/B **Subject: DS using Python Lab**

Suppose different users being surveyed may choose not to share their income, some users may choose not to share the address in this way many datasets went missing. In Pandas missing data is represented by two value:

- None: None is a Python singleton object that is often used for missing data in Python code.
- NaN : NaN (an acronym for Not a Number), is a special floating-point value recognized by all systems that use the standard IEEE floating-point representation

Pandas treat None and NaN as essentially interchangeable for indicating missing or null values. To facilitate this convention, there are several useful functions for detecting, removing, and replacing null values in Pandas DataFrame :

- isnull()
- notnull()
- dropna()
- fillna()
- replace()
- interpolate()

Finding outliers using statistical methods

Since the data doesn't follow a normal distribution, we will calculate the outlier data points using the statistical method called interquartile range (IQR) instead of using Z-score. Using the IQR, the outlier data points are the ones falling below Q1–1.5 IQR or above Q3 + 1.5 IQR. The Q1 is the 25th percentile and Q3 is the 75th percentile of the dataset, and IQR represents the interquartile range calculated by Q3 minus Q1 (Q3–Q1).

Using the convenient pandas .quantile() function, we can create a simple Python function that takes in our column from the dataframe and outputs the outliers:

Conclusion: -

In this experiment, we studied how using pandas and NumPy library we can pre-process the data.