|  |  |
| --- | --- |
| Semester | T.E. Semester V – Computer Engineering |
| Subject | Data Warehousing and Mining |
| Subject Professor In-charge | Dr. Kavita P Shirsat |
| Laboratory | M-312A |

|  |  |
| --- | --- |
| Student Name | Musab Khan |
| Roll Number | 22102A0066 |

|  |  |  |
| --- | --- | --- |
| Experiment Number | 01 | |
| Experiment Title | Pre-processing of Dataset to remove missing values using central tendency (Mean, median mode) | |
| Resources / Apparatus Required | Hardware: Computer system | Software: Python |
| Description | * The terms mean, median and mode are used to describe the central tendency of a large data set. Range provides context for the mean, median and mode. * When working with a large data set, it can be useful to represent the entire data set with a single value that describes the "middle" or "average" value of the entire set. In statistics, that single value is called the central tendency and mean, median and mode are all ways to describe it. * To find the mean, add up the values in the data set and then divide by the number of values that you added. * To find the median, list the values of the data set in numerical order and identify which value appears in the middle of the list. * To find the mode, identify which value in the data set occurs most often. * Missing values are identified and replaced by the mean median or mode based on the requirement. | |
| Link of the | import pandas as pd  import numpy as np  data = pd.read\_csv('adult.csv')  data = data.applymap(lambda x: np.nan if str(x).strip() == '?' else x)  print(data.head(10))  missing\_values = data.isnull().sum()  print(missing\_values)  numerical\_cols = [  'age', 'fnlwgt', 'education-num', 'capital-gain', 'capital-loss',  'hours-per-week'  ]  for col in numerical\_cols:  data[col].fillna(data[col].mean(), inplace=True)  categorical\_cols = [  'workclass', 'education', 'marital-status', 'occupation', 'relationship',  'race', 'sex', 'native-country', 'income'  ]  for col in categorical\_cols:  data[col].fillna(data[col].mode()[0], inplace=True)  missing\_values = data.isnull().sum()  print(missing\_values)  new\_filename = 'cleaned\_dataset.csv'  data.to\_csv(new\_filename, index=False)  print(f"Cleaned data saved to {new\_filename}") | |
| Output |  | |
|  |  | |
|  |  | |
|  |  | |
|  |  | |