School of Computer Science Engineering and Technology

Course-B. Tech	**Type**- Specialization Core-II
Course Code- CSET228	**Course Name**- Data Mining and Predictive Modeling (Lab)
Year- 2025	**Semester**- EVEN
Date- 10/01/2025	**Batch**- 2024-25

CO-Mapping

	CO1	CO2	CO3
Q1	√		
Q2	√		
Q3	√		

Objectives

1. Students will be able to learn data scaling and distribution.
2. Students will be able to learn data preparation and imputation.
3. Students will be able to learn missing value handling in data preprocessing.

Questions:

1. This problem emphasizes on Data Scaling, Binarization and Standardization. Use the Pima Indians diabetes dataset which is often used for binary classification problem. Do the following using suitable python library using this dataset. Note: use only the feature information of this dataset (exclude the label/class information) (30 Mins)
   (a) Since the attributes present in this dataset is of different scales, first rescale the attributes of this dataset so that all the attributes have the same scale. Use the scale 0 to 1 for the feature information of this dataset.
   (b) The next is to binarize the data. Transform the whole data set using a binary threshold. Allvalues above the threshold are marked 1 and all equal to or below are marked as 0. Use thethreshold value 0.0 for all the features of this dataset. Print the head of the transformed data.
   (c) The last step is to standardize the data, which is essential for analysing data distribution. Standardize the data associated with this dataset to a standard Gaussian distribution with amean of 0 and standard deviation of 1. (Hint. Use the scikit-learn library)

   Expected Output:

```
[[ 0.64   0.848  0.15   0.907 -0.693  0.204  0.468  1.426]
 [-0.845 -1.123 -0.161  0.531 -0.693 -0.684 -0.365 -0.191]
 [ 1.234  1.944 -0.264 -1.288 -0.693 -1.103  0.604 -0.106]
 [-0.845 -0.998 -0.161  0.155  0.123 -0.494 -0.921 -1.042]
 [-1.142  0.504 -1.505  0.907  0.766  1.41   5.485 -0.02 ]]
```

2. Use the 'DataPreprocessing.csv' dataset to perform the following pre-processing activities. Separate the features and the labels' part from the dataset and store them in variables X and Y.It is also a binary class dataset. (30 Mins)
   (a) Use the imputation technique to fill out the missing values present in the dataframe. (makeuse of

SimpleImputer function). Transform the data using imputing strategy defined through fit function.

(b) Make use of the label encoder to handle the categorical data present in the dataframe. the Region variable should consist of a 3-bit binary variable. The left most bit represents India,2nd bit represents Brazil and the last bit represents USA. If the bit is 1 then it represents data for that country otherwise not. For Online Shopper variable, 1 represents Yes and 0 represents No.

(c) Use the StandardScaler() to scale the dataset, where all the attributes are of same sacale.

Expected Output:

```
array([[0., 1., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0.,
        0., 0., 0., 0., 1., 0., 0.],
       [1., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0.,
        0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 1., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1.,
        0., 0., 0., 0., 0., 0., 0.],
       [1., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0.,
        0., 1., 0., 0., 0., 0., 0.],
       [0., 0., 1., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0.,
        0., 0., 1., 0., 0., 0., 0.],
       [0., 1., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
        1., 0., 0., 0., 0., 0., 0.],
       [1., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 1., 0.,
        0., 0., 0., 0., 0., 0., 0.],
       [0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0., 0.,
        0., 0., 0., 0., 0., 1., 0.],
       [0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 0., 0., 0.,
        0., 0., 0., 0., 0., 0., 1.],
       [0., 1., 0., 0., 0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
        0., 0., 0., 1., 0., 0., 0.]])
```

3. Read the 'Employee_list' file using pandas and implement following questions. The assistant of the company has missed to enter the salary of some employees. Now help him to predict the missing salary so that he can use the data for analysis. Consider the mean of salary to complete the data.

Expected Output:

	Sno.	Name	Age	Profession	Salary	Empid
0	1	Rahul [dr]	38	Engineer	86567.000000	15
1	2	Vipul	29	Doctor	77298.000000	9
2	3	Saurav	33	Doctor	81302.000000	11
3	4	Niyaz	39	Teacher	30456.000000	6
4	5	Franklin	28	Engineer	67553.142857	21
5	6	Niroja	34	Engineer	79000.000000	12
6	7	Meetesh	29	Engineer	67553.142857	23
7	8	Shashank	28	Teacher	45000.000000	31
8	9	Chauhan	41	Doctor	73249.000000	44

Owner is not approving your interpreted salary. He decided to fire those employees whose salary was not entered in the dataset. Write a code to help owner. (30 mins)