**Assignment No 1**

**Code:**

# Assignment number 1:

# Load the dataset

# Display basic information

# Display statistical information

# Display null values

# Fill the null values

# Change datatype of variable

# Quantization (Encoding): Convert categorical to numerical variable

# Normalization

#---------------------------------------------------------------------------------------

# Importing libraries

import pandas as pd

import numpy as np

#---------------------------------------------------------------------------------------

# Reading dataset

df = pd.read\_csv('Placement.csv')

#---------------------------------------------------------------------------------------

# Display basic information

print('Information of Dataset:\n', df.info)

print('Shape of Dataset (row x column): ', df.shape)

print('Columns Name: ', df.columns)

print('Total elements in dataset:', df.size)

print('Datatype of attributes (columns):', df.dtypes)

print('First 5 rows:\n', df.head().T)

print('Last 5 rows:\n',df.tail().T)

print('Any 5 rows:\n',df.sample(5).T)

#---------------------------------------------------------------------------------------

# Display Statistical information

print('Statistical information of Numerical Columns: \n',df.describe())

#---------------------------------------------------------------------------------------

# Display Null values

print('Total Number of Null Values in Dataset:', df.isna().sum())

#---------------------------------------------------------------------------------------

# Fill the missing values

df['gender'].fillna(df['gender'].mode()[0])

df['ssc\_p'].fillna(df['ssc\_p'].mean())

print('Mode of ssc\_b: ', df['ssc\_b'].mode())

df['ssc\_b'].fillna(df['ssc\_b'].mode()[0])

print('Total Number of Null Values in Dataset:', df.isna().sum())

#---------------------------------------------------------------------------------------

# changing data type of columns

# see the datatype using df.dtypes

# change the datatype using astype

df['sl\_no']=df['sl\_no'].astype('int8')

print('Change in datatype: ', df['sl\_no'].dtypes)

#---------------------------------------------------------------------------------------

# Converting categorical (qualitative) variable to numeric (quantitative) variable

# 1. Find and replace method

# 2. Label encoding method

# 3. OrdinalEncoder using scikit-learn

# Find and replace method

df['gender'].replace(['M','F'],[0,1])

# Label encoding method

df['ssc\_b']=df['ssc\_b'].astype('category') #change data type to category

df['ssc\_b']=df['ssc\_b'].cat.codes

# Ordinal encoder using Scikit-learn

from sklearn.preprocessing import OrdinalEncoder

enc = OrdinalEncoder()

df[['hsc\_b']]=enc.fit\_transform(df[['hsc\_b']])

print('After converting categorical variable to numeric variable: ')

print(df.head().T)

#---------------------------------------------------------------------------------------

# Normalization of data

# converting the range of data into uniform range

# marks [0-100] [0-1]

# salary [200000 - 200000 per month] [0-1]

# Min-max feature scaling

# minimum value = 0

# maximum value = 1

# when we design model the higher value over powers in the model

df['salary']=(df['salary']-df['salary'].min())/(df['salary'].max()-df['salary'].min())

# (x - min value into that column)/(max value - min value)

# Maximum absolute scaler using scikit-learn

from sklearn.preprocessing import MaxAbsScaler

abs\_scaler=MaxAbsScaler()

df[['mba\_p']]=abs\_scaler.fit\_transform(df[['mba\_p']])

#---------------------------------------------------------------------------------------

print(df.head().T)

**Output:**

Information of Dataset:

 <bound method DataFrame.info of      sl\_no gender  ssc\_p    ssc\_b  ...  specialisation  mba\_p      status    salary

0        1      M  67.00   Others  ...          Mkt&HR  58.80      Placed  270000.0

1        2      M  79.33  Central  ...         Mkt&Fin  66.28      Placed  200000.0

2        3      M  65.00  Central  ...         Mkt&Fin  57.80      Placed  250000.0

3        4      M  56.00  Central  ...          Mkt&HR  59.43  Not Placed       NaN

4        5    NaN  85.80  Central  ...         Mkt&Fin  55.50      Placed  425000.0

..     ...    ...    ...      ...  ...             ...    ...         ...       ...

210    211      M  80.60   Others  ...         Mkt&Fin  74.49      Placed  400000.0

211    212      M  58.00   Others  ...         Mkt&Fin  53.62      Placed  275000.0

212    213      M  67.00   Others  ...         Mkt&Fin  69.72      Placed  295000.0

213    214      F  74.00   Others  ...          Mkt&HR  60.23      Placed  204000.0

214    215      M  62.00  Central  ...          Mkt&HR  60.22  Not Placed       NaN

[215 rows x 15 columns]>

Shape of Dataset (row x column):  (215, 15)

Columns Name:  Index(['sl\_no', 'gender', 'ssc\_p', 'ssc\_b', 'hsc\_p', 'hsc\_b', 'hsc\_s',

       'degree\_p', 'degree\_t', 'workex', 'etest\_p', 'specialisation', 'mba\_p',

       'status', 'salary'],

      dtype='object')

Total elements in dataset: 3225

Datatype of attributes (columns): sl\_no               int64

gender             object

ssc\_p             float64

ssc\_b              object

hsc\_p             float64

hsc\_b              object

hsc\_s              object

degree\_p          float64

degree\_t           object

workex             object

etest\_p           float64

specialisation     object

mba\_p             float64

status             object

salary            float64

dtype: object

First 5 rows:

                        0         1          2           3          4

sl\_no                  1         2          3           4          5

gender                 M         M          M           M        NaN

ssc\_p               67.0     79.33       65.0        56.0       85.8

ssc\_b             Others   Central    Central     Central    Central

hsc\_p               91.0     78.33       68.0        52.0       73.6

hsc\_b             Others    Others    Central     Central    Central

hsc\_s           Commerce   Science       Arts     Science   Commerce

degree\_p            58.0     77.48       64.0        52.0       73.3

degree\_t        Sci&Tech  Sci&Tech  Comm&Mgmt    Sci&Tech  Comm&Mgmt

workex                No       Yes         No          No         No

etest\_p             55.0      86.5       75.0        66.0       96.8

specialisation    Mkt&HR   Mkt&Fin    Mkt&Fin      Mkt&HR    Mkt&Fin

mba\_p               58.8     66.28       57.8       59.43       55.5

status            Placed    Placed     Placed  Not Placed     Placed

salary          270000.0  200000.0   250000.0         NaN   425000.0

Last 5 rows:

                       210       211        212        213         214

sl\_no                 211       212        213        214         215

gender                  M         M          M          F           M

ssc\_p                80.6      58.0       67.0       74.0        62.0

ssc\_b              Others    Others     Others     Others     Central

hsc\_p                82.0      60.0       67.0       66.0        58.0

hsc\_b              Others    Others     Others     Others      Others

hsc\_s            Commerce   Science   Commerce   Commerce     Science

degree\_p             77.6      72.0       73.0       58.0        53.0

degree\_t        Comm&Mgmt  Sci&Tech  Comm&Mgmt  Comm&Mgmt   Comm&Mgmt

workex                 No        No        Yes         No          No

etest\_p              91.0      74.0       59.0       70.0        89.0

specialisation    Mkt&Fin   Mkt&Fin    Mkt&Fin     Mkt&HR      Mkt&HR

mba\_p               74.49     53.62      69.72      60.23       60.22

status             Placed    Placed     Placed     Placed  Not Placed

salary           400000.0  275000.0   295000.0   204000.0         NaN

Any 5 rows:

                        201         91         135        140       121

sl\_no                  202          92        136        141       122

gender                   M           M          F          M         F

ssc\_p                 54.2        52.0       72.0       65.0      64.0

ssc\_b              Central     Central    Central    Central   Central

hsc\_p                 63.0        57.0       56.0       64.8      67.0

hsc\_b               Others     Central     Others     Others    Others

hsc\_s              Science    Commerce    Science   Commerce   Science

degree\_p              58.0        50.8       69.0       69.5      69.6

degree\_t         Comm&Mgmt   Comm&Mgmt  Comm&Mgmt  Comm&Mgmt  Sci&Tech

workex                  No          No         No        Yes       Yes

etest\_p               79.0        67.0       55.6       56.0     55.67

specialisation      Mkt&HR      Mkt&HR     Mkt&HR    Mkt&Fin    Mkt&HR

mba\_p                58.44       62.79      65.63      56.94     71.49

status          Not Placed  Not Placed     Placed     Placed    Placed

salary                 NaN         NaN   200000.0   265000.0  250000.0

Statistical information of Numerical Columns:

             sl\_no       ssc\_p  ...       mba\_p         salary

count  215.000000  213.000000  ...  213.000000     148.000000

mean   108.000000   67.372911  ...   62.266103  288655.405405

std     62.209324   10.811464  ...    5.837303   93457.452420

min      1.000000   40.890000  ...   51.210000  200000.000000

25%     54.500000   60.800000  ...   57.900000  240000.000000

50%    108.000000   67.000000  ...   62.000000  265000.000000

75%    161.500000   76.000000  ...   66.230000  300000.000000

max    215.000000   89.400000  ...   77.890000  940000.000000

[8 rows x 7 columns]

Total Number of Null Values in Dataset: sl\_no              0

gender             2

ssc\_p              2

ssc\_b              1

hsc\_p              1

hsc\_b              2

hsc\_s              2

degree\_p           2

degree\_t           1

workex             2

etest\_p            2

specialisation     2

mba\_p              2

status             0

salary            67

dtype: int64

Mode of ssc\_b:  0    Central

Name: ssc\_b, dtype: object

Total Number of Null Values in Dataset: sl\_no              0

gender             2

ssc\_p              2

ssc\_b              1

hsc\_p              1

hsc\_b              2

hsc\_s              2

degree\_p           2

degree\_t           1

workex             2

etest\_p            2

specialisation     2

mba\_p              2

status             0

salary            67

dtype: int64

Change in datatype:  int8

c:\Users\Himanshu Lonkar\OneDrive\Documents\Academics\TE\DSBDA\A1\A1.py:56: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infer\_objects(copy=False)`. To opt-in to the future behavior, set `pd.set\_option('future.no\_silent\_downcasting', True)`

  df['gender'].replace(['M','F'],[0,1])

After converting categorical variable to numeric variable:

                       0         1          2           3          4

sl\_no                  1         2          3           4          5

gender                 M         M          M           M        NaN

ssc\_p               67.0     79.33       65.0        56.0       85.8

ssc\_b                  1         0          0           0          0

hsc\_p               91.0     78.33       68.0        52.0       73.6

hsc\_b                1.0       1.0        0.0         0.0        0.0

hsc\_s           Commerce   Science       Arts     Science   Commerce

degree\_p            58.0     77.48       64.0        52.0       73.3

degree\_t        Sci&Tech  Sci&Tech  Comm&Mgmt    Sci&Tech  Comm&Mgmt

workex                No       Yes         No          No         No

etest\_p             55.0      86.5       75.0        66.0       96.8

specialisation    Mkt&HR   Mkt&Fin    Mkt&Fin      Mkt&HR    Mkt&Fin

mba\_p               58.8     66.28       57.8       59.43       55.5

status            Placed    Placed     Placed  Not Placed     Placed

salary          270000.0  200000.0   250000.0         NaN   425000.0

                       0         1          2           3          4

sl\_no                  1         2          3           4          5

gender                 M         M          M           M        NaN

ssc\_p               67.0     79.33       65.0        56.0       85.8

ssc\_b                  1         0          0           0          0

hsc\_p               91.0     78.33       68.0        52.0       73.6

hsc\_b                1.0       1.0        0.0         0.0        0.0

hsc\_s           Commerce   Science       Arts     Science   Commerce

degree\_p            58.0     77.48       64.0        52.0       73.3

degree\_t        Sci&Tech  Sci&Tech  Comm&Mgmt    Sci&Tech  Comm&Mgmt

workex                No       Yes         No          No         No

etest\_p             55.0      86.5       75.0        66.0       96.8

specialisation    Mkt&HR   Mkt&Fin    Mkt&Fin      Mkt&HR    Mkt&Fin

mba\_p           0.754911  0.850944   0.742072    0.762999   0.712543

status            Placed    Placed     Placed  Not Placed     Placed

salary          0.094595       0.0   0.067568         NaN   0.304054