Name: Madhusudan D. Pangarkar

Class: TEIT Roll No.: 41

Sub: Laboratory Practice-I (Machine Learning) (314448)

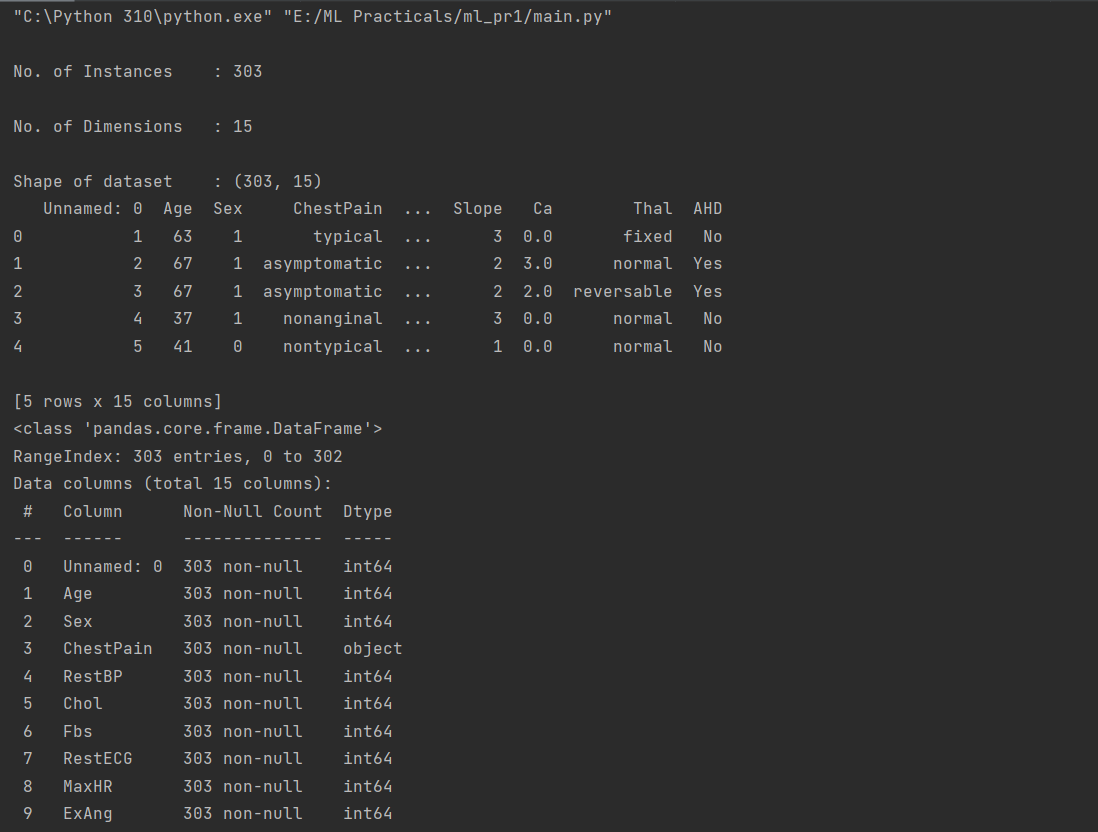
------------------------------------------------------------------------------------------------

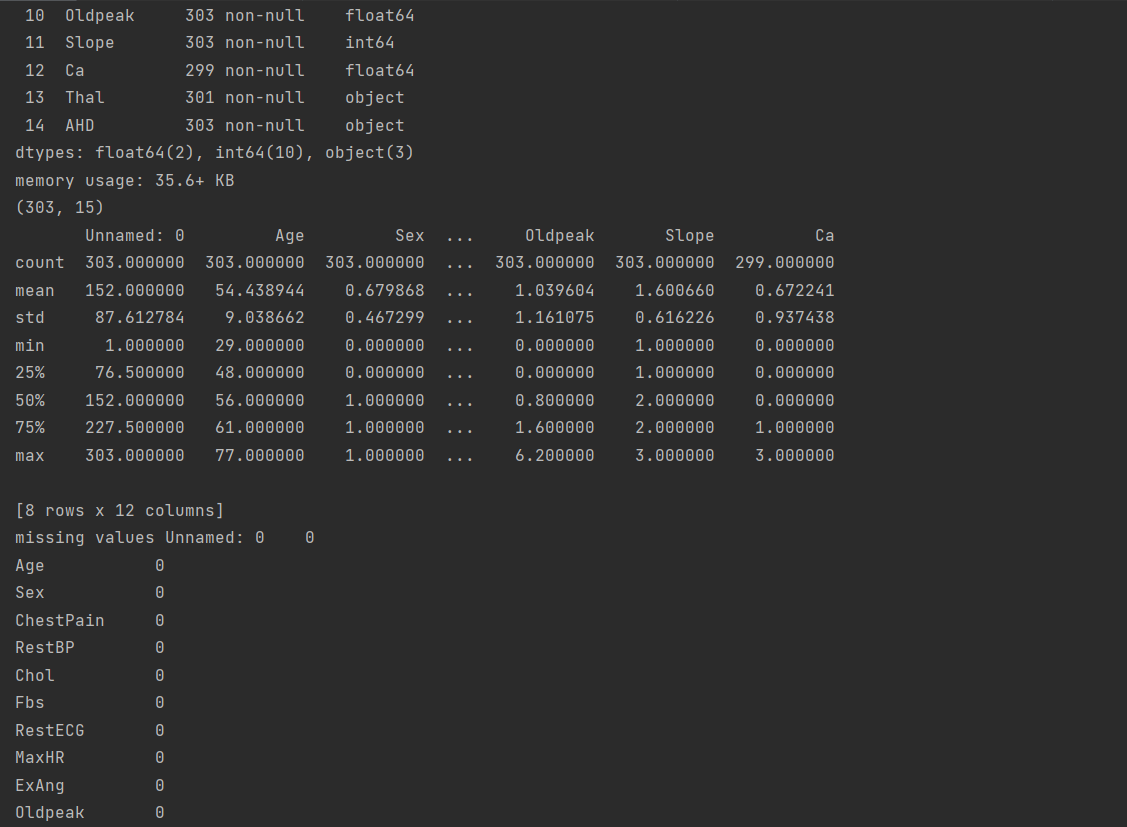
**Assignment No- 01) Data Preparation**

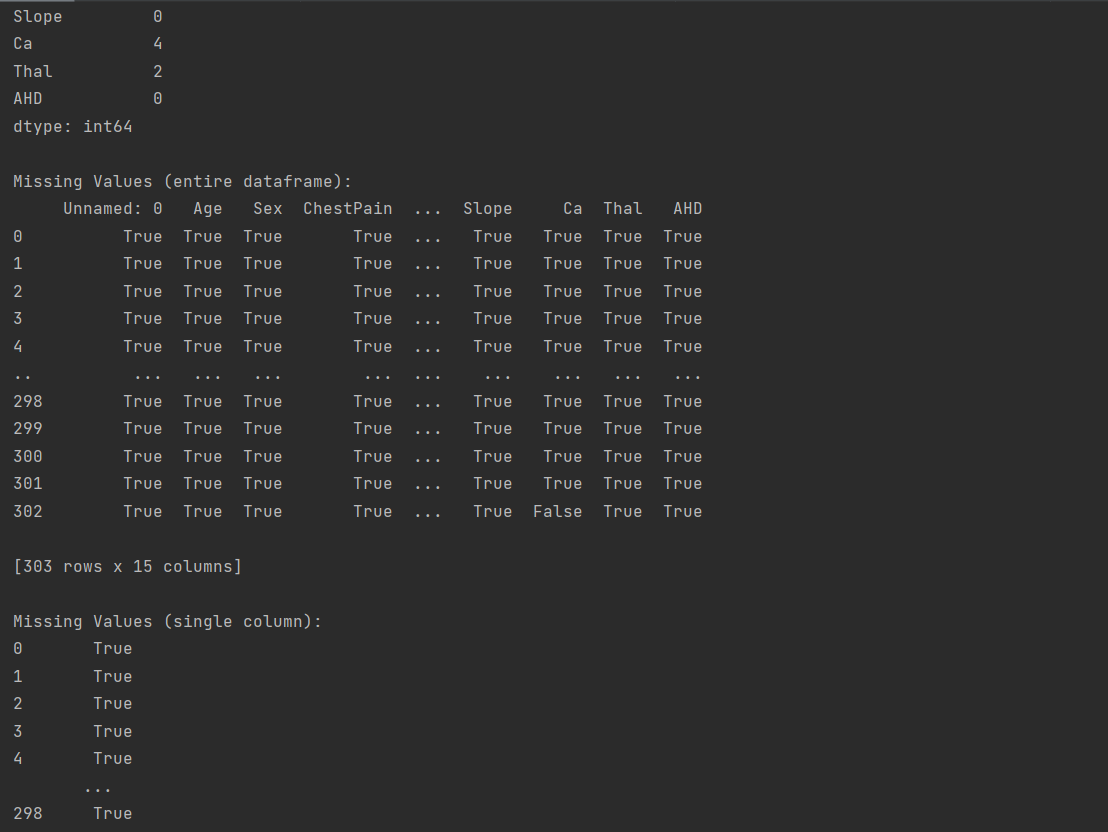
**Code:**

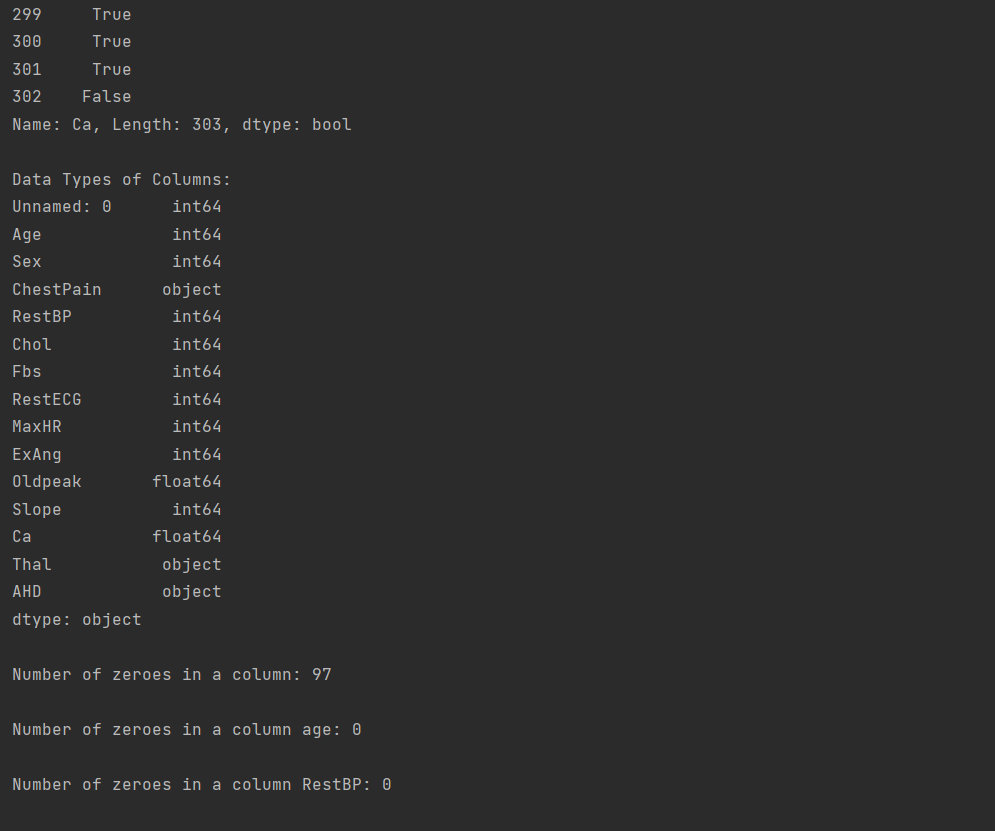
import pandas as pd  
from sklearn.model\_selection import train\_test\_split  
  
# load CSV file to use as a dataframe  
filename = 'E:\ML Practicals\Heart.csv'  
data = pd.read\_csv(filename)  
  
# Print number of instances (rows)  
print('\nNo. of Instances\t:', len(data))  
  
# Print number of dimensions (features/columns)  
print('\nNo. of Dimensions\t:', len(data.columns))  
  
# Print dimensions of the dataset,i.e., rows \* columns  
print('\nShape of dataset\t:', data.shape)  
  
# Print first 5 rows of the dataset  
  
print(data.head())  
  
# Print concise summary of the dataset  
data.info()  
  
# Print dimensions of the dataset,i.e., rows \* columns  
print(data.shape)  
  
# Generate descriptive statistics of dataset Descriptive statistics include those that summarize #the central tendency, dispersion and shape of a dataset’s distribution, #excluding NaN values.  
  
summary=data.describe()  
  
print(summary)  
  
# Print the missing values, i.e., NaN values  
  
print("missing values",data.isna().sum())  
miss1=data[data.isnull().any(axis=1)]  
  
  
# Print the missing values, i.e., NaN values  
# notnull() returns a dataframe with Boolean values stating True if the value  
# is not null and False if the value is null.  
print('\nMissing Values (entire dataframe):')  
print(data.notnull())  
print('\nMissing Values (single column):')  
print(pd.notnull(data['Ca']))  
  
# Print data type of values contained in each column of dataframe  
print('\nData Types of Columns:')  
print(data.dtypes)  
  
# Print number of zeroes in the columns  
# Similarly for all columns  
print('\nNumber of zeroes in a column:', (data['Sex'] == 0).sum())  
print('\nNumber of zeroes in a column age:', (data['Age'] == 0).sum())  
print('\nNumber of zeroes in a column RestBP:', (data['RestBP'] == 0).sum())  
print('\nNumber of zeroes in a column Fbs:', (data['Fbs'] == 0).sum())  
print('\nNumber of zeroes in a column RestECG:', (data['RestECG'] == 0).sum())  
print('\nNumber of zeroes in a column MaxHR:', (data['MaxHR'] == 0).sum())  
print('\nNumber of zeroes in a column ExAng:', (data['ExAng'] == 0).sum())  
print('\nNumber of zeroes in a column Ca:', (data['Ca'] == 0).sum())  
print('\nNumber of zeroes in a column Slope:', (data['Slope'] == 0).sum())  
print('\nNumber of zeroes in a column Oldpeak:', (data['Oldpeak'] == 0).sum())  
  
  
# Print mean of all values in age column  
print('\nMean Age:', data['Age'].mean(axis=0, skipna=True))  
print('Mean Age rounded to 2 decimal places:', round(data['Age'].mean(axis=0, skipna=True)))  
  
# Extract only 'Age', 'Sex', 'ChestPain', 'RestBP', 'Chol' columns from dataset  
y = data[['Age', 'Sex', 'ChestPain', 'RestBP', 'Chol']]  
  
# Split extracted data into training data and testing data  
splits\_train, splits\_test = train\_test\_split(y, test\_size = 0.25, random\_state=20)  
  
# Train Dataset  
print('\nTraining Dataset (75%): ')  
print('Length of Training Dataset:', len(splits\_train))  
print(splits\_train)  
  
# Test Dataset  
print('\nTesting Dataset (75%): ')  
print('Length of Testing Dataset:', len(splits\_test))  
print(splits\_test)

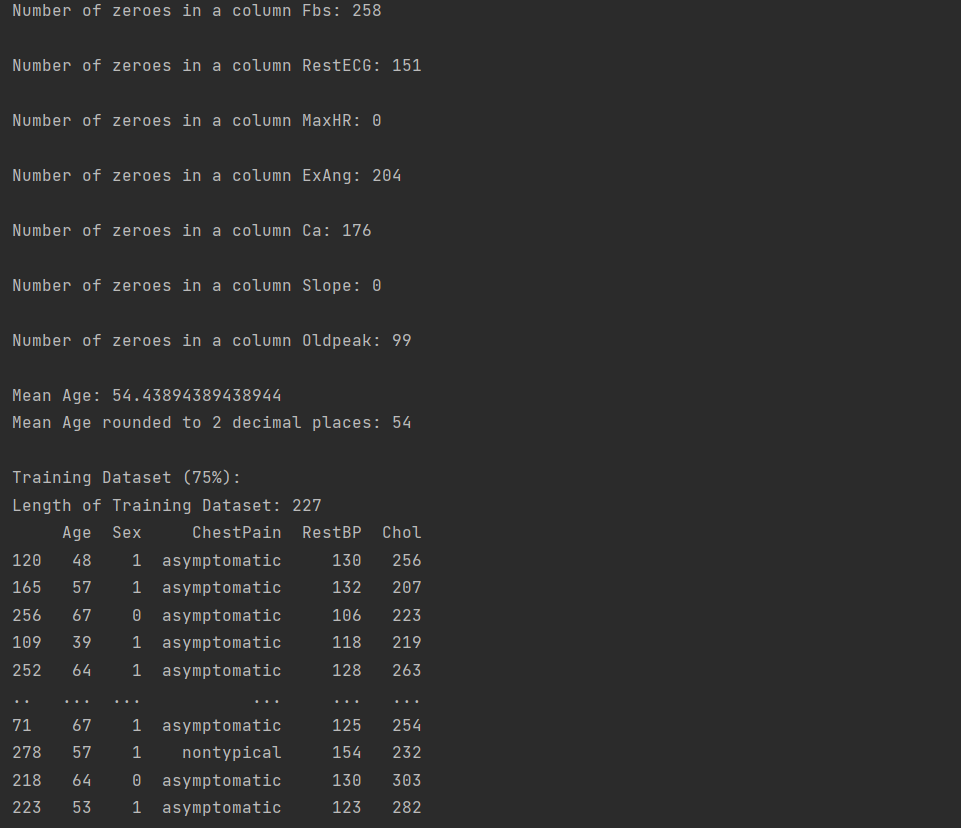
**Output:**

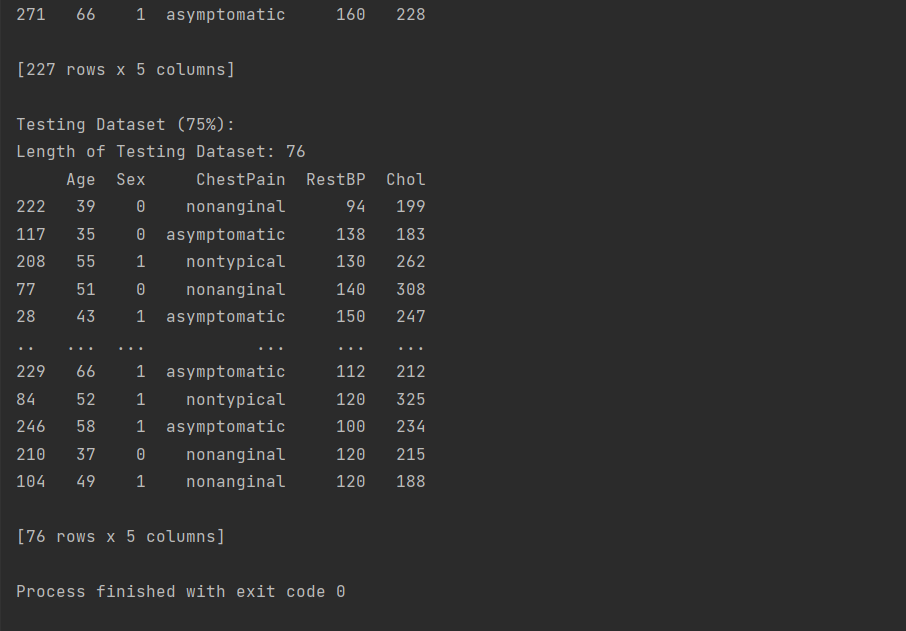
****

****

****

****

****

****