FRONTIER TECHNOLOGY INSTITUTE

DATA SCIENCE CERTIFICATION

MODULE -II: EXPLORATOTY DATA ANALYSIS

LAB- EXAM (80 Marks , 30 Points)

A) Using SKLEARN , load diabetes dataset (diabetes = datasets.load_diabetes()**) , and do the following:**

1. This dataset is in the form of a data dictionary with keys and values in the form of arrays. Extract data, target, feature names and DESCR and make a complete data frame which contains data, feature names and target in a single data frame. Extract DESCR so that it's readable. Bundle all these tasks in one heading " Diabetes Data Preparation" **(5 marks)**
2. **Add a heading "Data Description " and do the following in this section: (10 marks)**
    i. Display shape of the data
    ii. Display top 20 rows
    iii. Display data types
    iv. Display statistical properties like MCT and Dispersion
    v. Check for Null Values
3. **Add another heading and call it as Pre-Processing and do the following : (5 marks)**
    i. Round all the numeric values to 3 decimal places
    ii. Apply Z-Score Normalization to all the variables excluding target
4. **Add another heading as " Univariate / Bivariate Analysis " and do the following: (15 marks)**
    i. Plot the histograms for all the numeric variables
    ii. Plot the boxplots for all the numeric variables
    iii. Plot the scatter plots of all the variables with the target
    iv. Compute and plot the Pearson Correlation matrix for all numeric attributes
5. **Add another heading and name it as " Outlier Detection and Removal" , and do the following: (15 marks)**
    i. For age, sex , bmi and bp attributes , display the outliers using Z-Score
    ii. For age, sex , bmi and bp attributes , remove the outliers using Z-Score
    iii. For all other attributes excluding target use IQR to display the outliers
    iv. For all other attributes excluding target use IQR to remove outliers .

B) Use the crimes_fti dataset (attached) to perform the following tasks. (30 marks)

1. **Add a heading "Data Description " and do the following in this section: (5 marks)**
    i. Display shape of the data

ii. Display top 20 rows
iii. Display data types
iv. Display statistical properties like MCT and Dispersion
v. Check for Null Values

2. **Add a heading " Dealing with Data Quality Problems" and do the following: (10 marks)**
   i. Display values counts of all unique values in a column (All columns)
   ii. Detect values which are incorrect
   iii. Handle all such incorrect values (Only one or two columns may have incorrect values)
   iv. Replace all incorrect values with NaN
   v. Remove all the records having Null / NaN Values
   vi. Remove all the columns which have more that 50 % null values

3. **Add a heading "Feature Encoding and Discretization " and do the following: (15 marks)**
   i. Apply label encoding to OFFENSE_CODE_GROUP and DAY_OF_WEEK
   ii. Apply One-Hot Encoding to UCR_PART
   iii. Apply Discretization to HOUR (0-8 Early Morning , 8-12 Morning 12-16 Afternoon, 16-19 Evening, 19-23 Night) or you can create your own levels.

SUBMISSION INSTRUCTIONS:

Submit only the Python Notebook on the Google Classroom by Friday, 16th October, 2020

Make sure to name your notebook as YourName_M2_Final.