ITM740 Assignment 1

Winter 2002

Objective:

Learn the basics of exploratory data analysis (EDA) in Python with Pandas, NumPy, Matplotlib, and seaborn (optional)

Dataset:

LoanTrain.csv

Tasks:

1. Import necessary libraries
2. Load the data into dataframe
    - Read data from train.csv file
    - print the names of the headers
    - print the shape of the data
    - get each attribute's data type
    - print the first three line of data
3. Check for missing data and drop rows and columns with missing data

    Hint: use isnull() and dropna()
4. Obtain summary (descriptive statistics) of the data
5. Univariate Analysis:
    - Review skew of attribute distribution
    - Plot histogram for continuous variables
    - Plot bar graph for discrete variables
    - Plot density plot for continuous variables
    - Plot bar plot (hint: bar()) for continuous variables
    - Plot boxplot for continuous variables
6. Bivariate Analysis:
    - Review correlation between attributes
    - Plot correlation matrix
7. Data Rescaling:
    - rescale the data use MinMaxScaler class
    - use L1 Normalize technique to normalize the data
    - use L1 Normalize technique to normalize the data
    - binarize the data
    - standardize the data
8. Feature Selection

- select 4 of the attributes having best features with the help of chi-square statistical test and print them
- implement RFE feature selection technique to select the best 3 attributes and print them
- implement PCA feature selection technique
- implement this feature selection technique with the help of ExtraTreeClassifier class

Submission:

You must use Jupyter Notebook and provide an ipynb file with both markdown and codes.

The name of the files shall be ITM740Assignment1byAliceBrown.ipynb when “AliceBrown” shall be replaced by your first name and last name.