Last Name: _____ First Name: _____ Student ID: _____

AIDI 1002: Machine Learning Programming — Assignment - 1

Due Date : October 07, 2022, 11:59 PM

Note : Submit two files in the submission folder. First is your colab notebook including your code and outputs and second is the pdf of colab notebook with the following naming convention for both the files.
(File name : FirstName_LastName_Assignment_1.pdf or ipynb)

1. Consider the dataset 'noisy_data.csv' and apply the following pre-processing techniques and obtain the clean dataset.
   - Handling missing values by imputation (10 points)
   - Apply Normality tests to numerical columns and state the hypothesis clearly and comment on the normality of the data (10 points)
   - Apply encodings for categorical variable and scale the features (10 points)
2. Consider the text present in the file 'wiki.txt' and Answer the following questions :
   - Write a program to convert following text into tokens with two tokenization methods such as 'RegexpTokenizer()' and 'word_tokenize()' from NLTK library. (Note :The tokens should not have stop words and punctuation symbols. Feel free to decide about the correct list of stop words; e.g., negative words (don't) could be important for you. Execute both methods of tokenization along with your code of removing stop words and punctuation.) (10 points)
   - Write a regular expression to extract all the year mentions in the 'wiki.txt' file. (10 points)
   - State the differences observed in the output of tokenization methods. (10 points)
3. Consider this dataset from kaggle. (Download the dataset from following link : `https://www.kaggle.com/dansbecker/melbourne-housing-snapshot/home`) and answer the following questions :
   - Apply the feature selection techniques over the melbourne-housing -dataset namely (20 points):
     * Correlation
     * Chi-Square
     * Mutual-Information
     * Random Forest feature importance
   - Compare the importance of selected features using bar chart (10 points).
   - Comment on the results obtained from various feature selection techniques and which is the best and worst feature selection selection technique on the given dataset (10 points).