Machine Learning Assignment

General Instructions:

1. There is no programming language/ tool/ technology barrier for this assignment. Although, it is preferable to use a single programming language/platform throughout the assignment to perform all the tasks
2. The links to the datasets are provided in this document itself. You can access the dataset by clicking on the name of the dataset.
3. You may use external sources (blogs/videos/books) to get familiar with the concepts of this assignment.
4. Please avoid plagiarism. When you take facts, thoughts, ideas, viewpoints and short or long code snippets from others and use them in your own work, the sources you have used must be clearly stated - the links to which must be saved in a text file named references.txt. In other words you must not give the impression that others' thoughts, ideas, viewpoints and results are your own if they are not.
5. Please follow the following hierarchy and legible naming conventions for your code files to submit the assignment.

```
Parent Folder (Named after the assignment)
        -Child Folder One (Name: Code)
                -Your code/script files here
        -Child Folder Two (Name: Output)
                -Your output files here
        -references.txt
```

You may then zip the parent folder and share it with the concerned authority for evaluation.

Datasets:

You have been provided with the following datasets:

1. Data Cleaning: Melbourne Housing Dataset *(click to access)*
2. NLP: The Blog Authorship Corpus *(click to access)*

Listed on the following pages are the tasks that you must perform as a part of this assignment.

1. Data Preprocessing

1.1 Data Cleaning

You are to perform data cleaning on the Melbourne Housing dataset. After every operation, save your output in a .csv file, the names for which have been described below. You must specifically perform the following three operations to deal with the missing values:

1. Drop the columns with missing values. (Output file(s): 1 code file + drop.csv)
2. Impute the missing values
    a. Using Mean/Median/Mode to populate the missing value fields (Output file(s): 1 code file + mmClean.csv)
    b. Using Linear Interpolation to populate the missing value fields (Output file(s): LiClean.csv)
    c. Finally, extract only the imputed **tuples** from the dataset and save it in an external .csv file. (Output file(s): 1 code file + extracted.csv)

1.2 Categorical Data Encoding

You are to perform categorical data encoding for the features '*CouncilArea*' and '*Regionname*' of the Melbourne Housing dataset. (Output file(s): 1 code file + encoded.csv)

1.3 NLP Text-Corpora Preprocessing

You are to perform the following operations on each paragraph for each of the xml files(data from one blog) of the The Blog Authorship Corpus dataset. After every operation, save your output in a .csv file.

1. Tokenization
    a. Sentence tokenization (Output file(s): 1 code file + sentence.csv)
    b. Word tokenization (Output file(s): 1 code file + word.csv)
2. Frequency Distribution - capture the frequency of every word per paragraph (Output file(s): frequency.csv)
3. Stopwords and Non-stop words
    a. Extract only the stopwords from each paragraph of the dataset (Output file(s): 1 code file + stopwords.csv)
    b. Extract only the non-stop words from each paragraph of the dataset (Output file(s): 1 code file + nonstopwords.csv)
4. Lexicon Normalization
    a. Perform stemming for each word of the paragraph and save the stems (Output file(s): 1 code file + stems.csv)
    b. Perform lemmatization for each word of the paragraph and save the lemmas (Output file(s): 1 code file + lemmas.csv)

2. Building Model

2.1 Selecting Prediction Target

Refer to the Melbourne Housing dataset. You must determine the prediction target from amongst all the attributes provided in the dataset. Also provide a plausible explanation supporting the grounds for your prediction target (in not more than 100 words).
(Output file: predictionTarget.txt)

2.2 Choose Features

Refer to the Melbourne Housing dataset and your prediction target. You must determine a set of features that you will use to train your model on, from all the attributes provided in the dataset. Also provide a plausible explanation supporting the grounds for each attribute being a part of the feature set (in not more than 250 words).
(Output file: features.txt)

2.3 Build Model

Refer to your prediction target, feature set and the Melbourne housing dataset. You must build a model to predict your prediction target based on your features. In accordance with the same, you may create a train dataset and a test dataset. Finally, you must evaluate your model at least 3 times, every time on a random chunk from the dataset and record the accuracies for each run. Save your predictions in the predict.csv file and suffix the filename by the iteration number. (eg: predict_1.csv, predict_2.csv)
(Output files: 1(or more) code file(s) + train.csv, test.csv, predict.csv)
NOTE: sample file sample_predict.csv *(click to access)*