Introduction to data science - Assignment #3

The 'crime.csv', 'kidney_disease.csv', and 'email.csv' data files attached to the assignment file (available at Sakai) is taken from the UCI repository [1] and Kaggle website [2, 3]. The Crime Data file reports the number of violent crimes per 100,000 population for the communities within the United States. It also includes some socio-economic factors.
The variables are as follows:

- 'PctPopUnderPov': Percentage of people under the poverty level (numeric: from 0 to 1)
- 'PctUnemployed': Percentage of unemployed people (numeric: from 0 to 1)
- 'PolicPerPop': Ratio of police officers to the population (numeric: from 0 to 1)
- 'Pcthomeless': Percentage of homeless people (numeric: from 0 to 1)
- 'PctBSorMore': Percentage of people with a bachelor's degree or higher education (numeric: from 0 to 1)
- 'ViolentCrimesPerPop': Ratio of violent crimes to the population (numeric: from 0 to 1)

The Kidney Disease Data reports the age, blood pressure, and the results of the blood test factors for healthy people and kidney patients.
This data includes the following 9 variables:

- 'age': Age of the individual (numeric: from 2 to 90)
- 'bp': Blood pressure (numeric: from 50 to 180)
- 'sod': Blood sodium level test result (numeric: from 104 to 163)
- 'pot': Blood potassium level test result (numeric: from 2.5 to 47)

- ‘hemo’: Hemoglobin blood test result (numeric: from 3.1 to 17.8)
- ‘pcv’: Packed cell volume test result (numeric: from 9 to 54)
- ‘wc’: White blood cell test result (numeric: from 2200 to 26400)
- ‘rc’: Red blood cell test result (numeric: from 2.1 to 8)
- ‘CKD’: Chronic kidney disease (binary: 0 for healthy individuals and 1 for kidney patients)

The Email Data includes the text of the numerous emails labeled as spam or not spam.
This data includes the following columns:

- ‘email’: Text of the email (string)
- ‘label’: Label of the email (binary: 1 for spam and 0 for not-spam)

Note. You must put the CSV files in the same folder as your code file. If you use Jupyter notebook it should be in the address: ‘C:/Users/YOUR-USER-NAME’.
You can also read the file by its address; for example:

```
f = open('C:/files/sample-file.txt')
```

Question 1

Write a code to learn a simple regression model to predict the ratio of violent crimes based on (i) percentage of unemployed people and (ii) percentage of people with a bachelor’s degree or higher education. Then explain the impact of each of these two factors on violent crimes by interpreting the regression coefficients.

Question 2

Write a code to learn a multiple regression model to predict the ratio of violent crimes based on all the other variables. Report the most influential factor in violent crimes.

Question 3

Use `LogisticRegression` class of `sklearn` package to learn a logistic regression model that predicts chronic kidney disease based on other variables in Kidney Disease Data.

Question 4

Split Kidney Disease Data into two parts of training data (70%) and testing data (30%), and train the model in Question 3 using the training data. Then predict the chronic kidney disease for the testing data samples and report the accuracy and f1 score of the predictions.

Question 5

The following function takes a text as input and returns a dictionary that includes the frequency of each word in the text. Change this function to return the frequency ratio of the most frequent word to the length of the text.

```python
def get_frequency(input_string):

    list_of_words = input_string.split(' ')
    dict_of_frequencies = {}

    for word in list_of_words:

        if word in dict_of_frequencies.keys():
            dict_of_frequencies[word] = dict_of_frequencies[word] + 1
        else:
            dict_of_frequencies[word] = 1

    return(dict_of_frequencies)
```

Question 6

The following code is to extract useful features from the email texts included in the Email Data and train a model to predict if the email is spam. Complete the code to:

- Extract four binary features representing the presence of the words ‘hyperlink’, ‘free’, ‘click’, and ‘business’ in email texts,
- Use the `get_frequency` function in Question 5 to extract the ratio of the most frequent word of the text as a numeric feature,
- Train a logistic regression model on 70% of the data to classify the email as spam or not spam based on the five extracted features,
- Predict the label of the remaining 30% of the data and report the accuracy of the predictions.

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

data = pd.read_csv('email.csv')

# adding empty columns
data['hyperlink'] = None
data['free'] = None
data['click'] = None
data['business'] = None
data['frequency'] = None

################# your code here ##################
## you need to
## 1. for each row
## 1-1. check if the mail text includes the words
## 'hyperlink', 'free', 'click', and 'business' and
## fill the corresponding columns with 0 or 1
## 1-2. Use the get_frequency function to get the ratio of
## the most frequent word and fill the frequency column
##
## 2. split the data into the training (70%) and testing
## (30%) data
##
## 3. Use LogisticRegression class of sklearn package
## to train a model to predict the label of emails
## based on the extracted features


#######################################################
```

References

[1] Communities and crime data set. https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime.

[2] Chronic kidney disease dataset. https://www.kaggle.com/datasets/mansoordaku/ckdisease?resource=download.

[3] Spam or not spam dataset. https://www.kaggle.com/datasets/ozlerhakan/spam-or-not-spam-dataset.