Homework 5

Due: Sunday, May 10, 2020, 11:55pm

Guidelines:
All homeworks will be submitted via Courseworks. For a given assignment, all files should be put in a folder called uni-hw5, all lowercase, which is then zipped and submitted on courseworks. The folder should contain a single readme.txt which has your name, uni, the homework number, and a brief summary of your work for each part (if necessary).

Nearest Neighbor

In this assignment we will complete our machine learning library. In addition to the functionality we have already written, we will implement the following features:

- Provide some statistics on the dataset using pandas:
  - Skewness and kurtosis of each column in the dataset
  - Mean/Std Deviation of each column, grouped into benign/malignant
- Plot some more advanced information using seaboard
  - Pairplot of each column against other columns
  - Heatmap of the column correlations
- Implement the K nearest neighbors algorithm from class
- Use scikit-learn to implement:
  - K Nearest neighbors (for comparison)
  - Support Vector Machine classifier

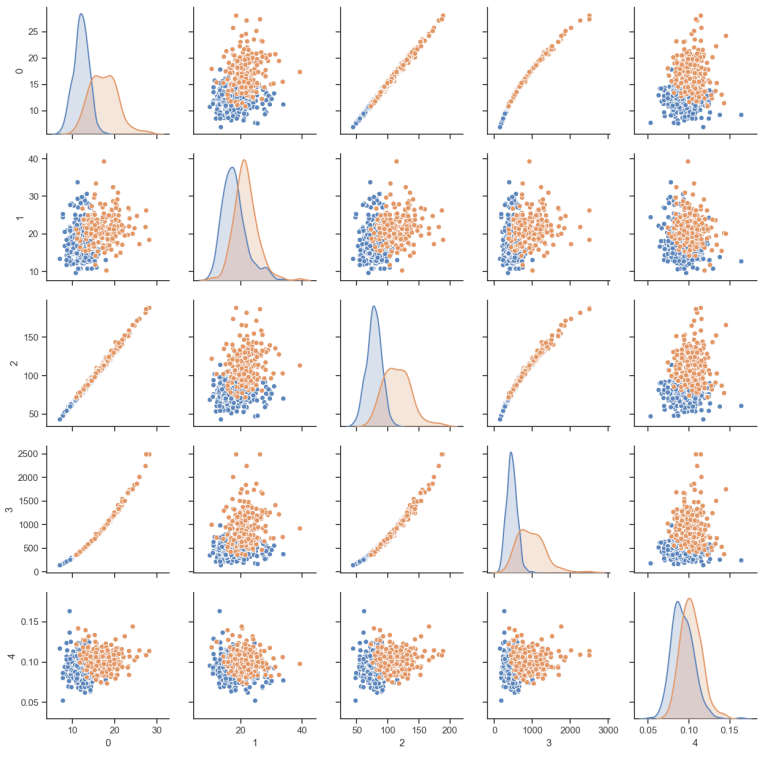
Here is an example output from running %run -m engi1006. Note that the "…" indicate that I have omitted some output (e.g. for other columns).

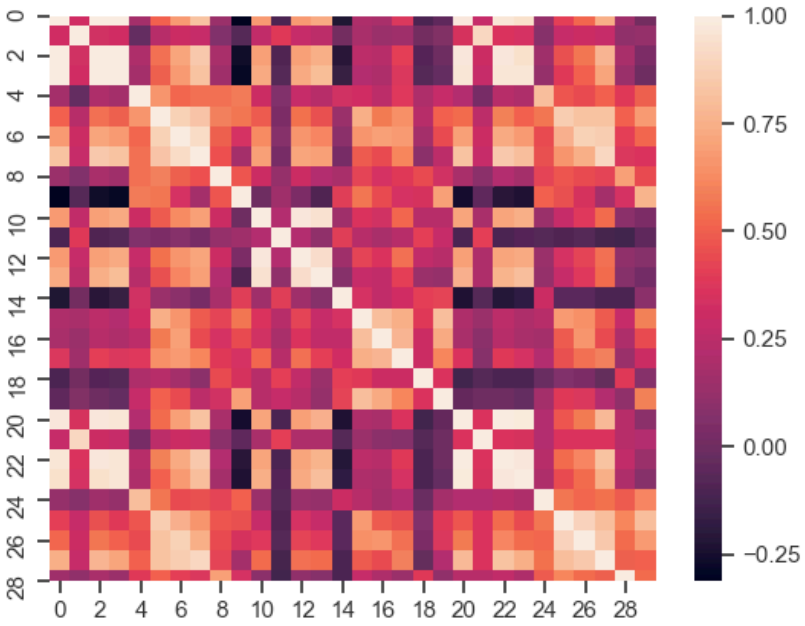
```
Please provide a filename: wdbc.csv
What percentage of the dataset would you like to put aside for testing? 20
What model would you like to run? (knn, sklearn-knn, svm) knn
Reading filename: wdbc.csv
Dataset size: 569 x 30
Number of benign: 357
Number of malignant: 212


Would you like to see more advanced stats? (y/n)y
Column 0 statistics:
        Skewness:0.9423795716731008         Kurtosis:0.8455216229065408
…
Column 29 statistics:
        Skewness:1.6625792663955172         Kurtosis:5.2446105558150125
```

Dataframe statistics
Benign Stats:
Mean:
0 12.146524
1 17.914762
…
29 0.079442
Name: B, dtype: float64
Std:
0 1.780512
1 3.995125
…
29 0.013804
Name: B, dtype: float64

Malignant Stats:
Mean:
0 17.462830
1 21.604906
…
29 0.091530
Name: M, dtype: float64
Std:
0 3.203971
1 3.779470
…
29 0.021553
Name: M, dtype: float64
Would you like to plot the data? (y/n) y
How many columns should we plot in the scatter matrix? (5) 5






```
Splitting dataset into 80% for training and 20% for testing
Test dataset has 113 entries
Train dataset has 456 entries


Hit enter to run algorithm


How many nearest neighbors? 5
Running knn classifier...
Accuracy: 90.3%
Run again? (y/n) n
```

Details

All functions have details about their implementation in their docstrings.