Computer Science 3202/6915
Assignment 3 – Linear models for regression – Baseline

Goal
Apply linear methods for regression.

Due date
Sunday February 28th by 11:30pm.

Specifications
Your program should be called A3_Reg.py and it should run in Linux. You program should take one command-line arguments: A filename specifying a tab-delimited plain-text file containing the training data.

For example, we should be able to execute your program as follows:

```
$python3 A3_Reg.py Traindata.txt
```

where the $ indicates the terminal prompt.

The data you should use is provided in Brigthspace. For generating your model, you will train with the data given in the file A3_TrainData_noDup.tsv. This is a tab-delimited text file containing 2698 observations. The file has a header but not instance IDs. There are 11 numerical attributes, and the last column is the value to predict.

This data was taken from the article: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In *Decision Support Systems*, Elsevier, 47(4):547-553, 2009. If you are interested to take a look, this manuscript is available in Brightspace.

Functionality
Your program should do the following:

1. Read the input data. You can assume the input file is in the working directory.
2. Use k-fold CV to assess the performance of a linear model using least squares linear regression with the scikit-learn function LinearRegression (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html). This is your baseline model.

Submission
Submit through Brightspace the following (one submission per team):

a) Your python code in a single file called A3_Reg.py
b) A PDF file containing:
    1. the cross-validation performance (R^2 and RSS) of your baseline model,
    2. a table with the model coefficients and a brief interpretation of the coefficients (e.g., which wine characteristic(s) lower(s) wine quality? which wine characteristic(s) increase(s) wine quality?)
    3. an acknowledgement section listing your collaborations and online sources, and

4. a program specification section listing the Python version and libraries you used.

Common pitfalls to avoid (i.e., DO NOT do the following or points will be deducted):

1. Submit files in a compressed file.
2. Fail to include some of the sections in the PDF file.
3. Forget to acknowledge a collaboration or source.
4. Miss some functionality.
5. Use different data sets.

Online examples/tutorials:

- https://www.kaggle.com/jnikhilsai/cross-validation-with-linear-regression
- https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html
- https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/