Assignment 3 – Linear Regression

**Description:**

In statistics, linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). In machine learning, a linear regression model is widely used to predict the real-valued output. For this assignment, you need to build a univariate linear regression model using the *scikit-learn* *diabetes* dataset.

**Purpose:**

* Get familiar with Python programming language and the scikit-learn library.
* Make use of feature selection method.
* Develop a univariate linear regression model for a given task.

**Directions:**

For this assignment, you need to build a linear regression model from scratch. Below is a detailed instruction of what you may need to do.

* Dataset Preparation
  + You need to load the dataset using *sklearn.datasets.* *load\_diabetes*.
    - More information about the function can be found at: <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html#sklearn.datasets.load_diabetes>
  + After loading the dataset, randomly shuffle the dataset to split the dataset to train/dev/test sets.
    - Use the 70% of data for the train set, 15% for the dev set, and 15% for the test set
    - You need to make sure that the labels and features are still matching after shuffling the data.
    - You may want to use the random shuffle function provided by Numpy.
* Univariate Linear Regression Development
  + You need to implement a univariate linear regression model from scratch.
    - You need to use a gradient descent algorithm to solve the optimal parameters for the univariate linear regression model.
    - You need to implement the gradient descent algorithm from scratch.
  + The dataset contains 10 features; however for a univariate linear regression model, you may only use one feature. Thus, to build the linear regression model, you need to decide which feature to you. There are three approaches you may use:
    - You may train 10 univariate linear regression models, one model for each feature, and select the one with the best performance on the dev set.
    - You may run a feature selection algorithm to select a feature. There are plenty of feature selection algorithms available. Feel free to search on the internet and use a method at your choice.
    - You may use PCA for dimension reduction to reduce the number of features to 1. There are many python libraries that provide PCA algorithms you may use. In fact, scikit-learn also has a function. Feel free to search on the internet and choose one at your choice.
* Test the Model
  + Test the model using the test set.
* Submission
  + You need to submit a written report for this assignment.
    - For this report, you need to:
      * Explain what you have done
      * Report the best performance on the test set (in terms of MSE)
      * Include your code as an appendix
        + You could save your Colab code as a PDF file and attach it to your report, or you could copy and paste your code into the report.

If you want to copy/paste your code, make sure to maintain the appropriate indentation and make the code readable.

**Evaluation Criteria:**

This assignment is worth 100 points.

* (50 pts) The code must be runnable and works as expected
* (50 pts) The report must contain the three components