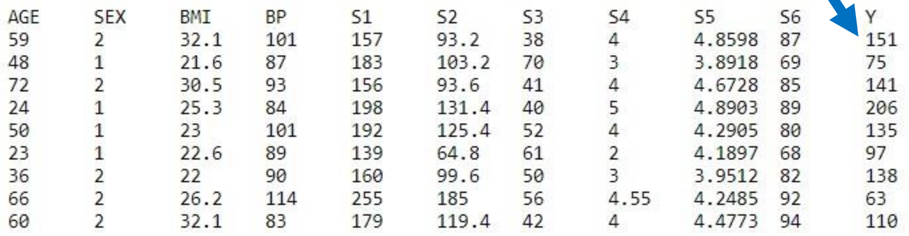
Computer Science Department
CS675 – Introduction to Data Science (CRN: 73405)
Fall 2024

Project #2 / Due 12-Nov-2024

This is a continuation of project #1 (EDA).
Implement a Linear Regression algorithm (model) in Python, by using the Scikit-learn module. The **regression model** should be able to predict the progression of a disease (diabetes in our case) by using the least-squares regression.
The modeling of the data should be focus on predicting the progression of a disease.

Get the data from **Stanford U's** Machine Learning Repository:
https://web.stanford.edu/~hastie/Papers/LARS/diabetes.dat a

Here is a sample of the dataset (out of 442 records):



AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
59	2	32.1	101	157	93.2	38	4	4.8598	87	151
48	1	21.6	87	183	103.2	70	3	3.8918	69	75
72	2	30.5	93	156	93.6	41	4	4.6728	85	141
24	1	25.3	84	198	131.4	40	5	4.8903	89	206
50	1	23	101	192	125.4	52	4	4.2905	80	135
23	1	22.6	89	139	64.8	61	2	4.1897	68	97
36	2	22	90	160	99.6	50	3	3.9512	82	138
66	2	26.2	114	255	185	56	4.55	4.2485	92	63
60	2	32.1	83	179	119.4	42	4	4.4773	94	110

For some background information on the data, see this seminal paper:
Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani (2004) "Least Angle Regression," Annals of Statistics (with discussion), 407-499.
https://projecteuclid.org/euclid.aos/1083178935

Write **Python** scripts in order to complete the following tasks along with their output. All work should be done and submitted in a single **Notebook** (Jupyter or Colab).

1) Predict the feature 'y' using a single feature of 'X' (in the entire dataset) Find out which feature from 'X' should be used for the best prediction of 'y'.

<< Output >>:

- Model's coefficients (slope, y-intercept)
- The Linear Regressor Model (graph) plotting
- The MSE (Mean Square Error)

2) Predict the feature 'y' using a pair feature of 'X' (in the entire dataset) Find out which pair feature from 'X' should be used for the best prediction of 'y'.

<< Output >>:

- Model's coefficients (slope, y-intercept)
- The Linear Regressor Model (graph) plotting.
- The MSE (Mean Square Error)

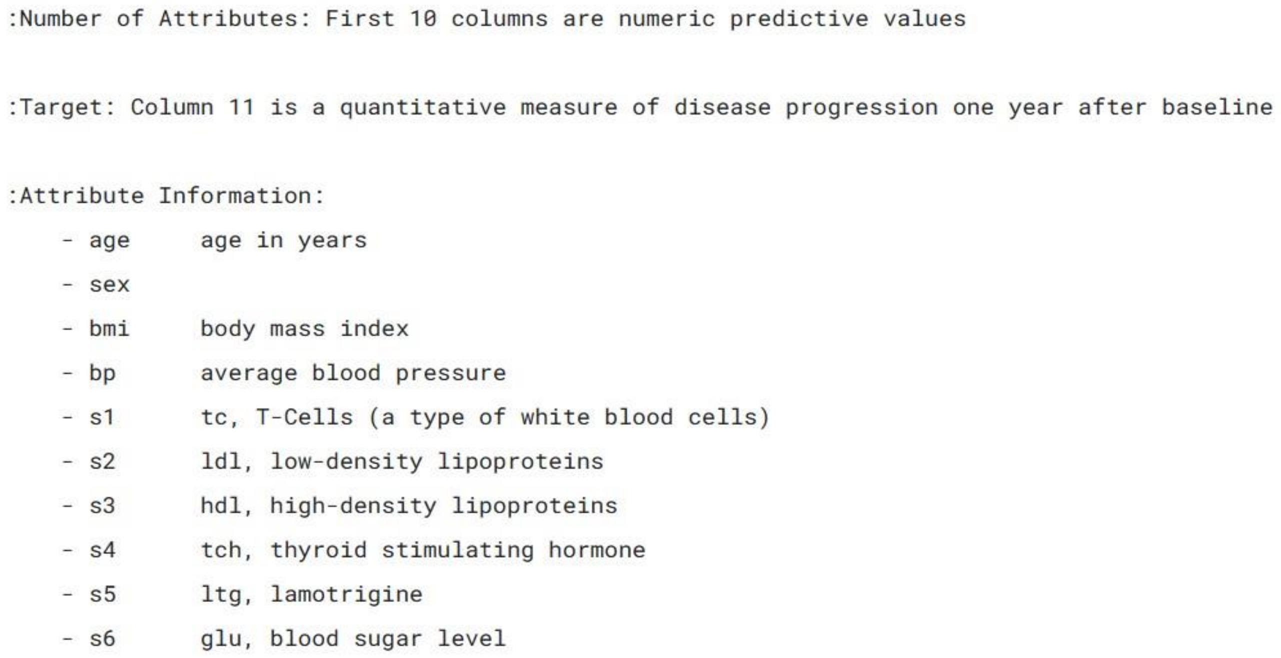
3) Predict the feature 'y' using all (10) features of 'X' (in the entire dataset)

<< Output >>: Model's coefficients & The MSE (Mean Square Error)

4) Compute the training MSE and validation MSE when fitting the regressor in all features, for the following training set sizes: n_train = 20 ,n_train = 50, n_train = 100, n_train = 200

Extra Credit: Create another model by using the **XGBoost** library. Compare and contrast its results with the linear regressor (tasks 1-4).

Useful **metadata** information for each of the features of the dataset, listed below. Make sure you review it:



```
:Number of Attributes: First 10 columns are numeric predictive values

:Target: Column 11 is a quantitative measure of disease progression one year after baseline

:Attribute Information:
    - age     age in years
    - sex
    - bmi     body mass index
    - bp      average blood pressure
    - s1      tc, T-Cells (a type of white blood cells)
    - s2      ldl, low-density lipoproteins
    - s3      hdl, high-density lipoproteins
    - s4      tch, thyroid stimulating hormone
    - s5      ltg, lamotrigine
    - s6      glu, blood sugar level
```

Based on Project #1