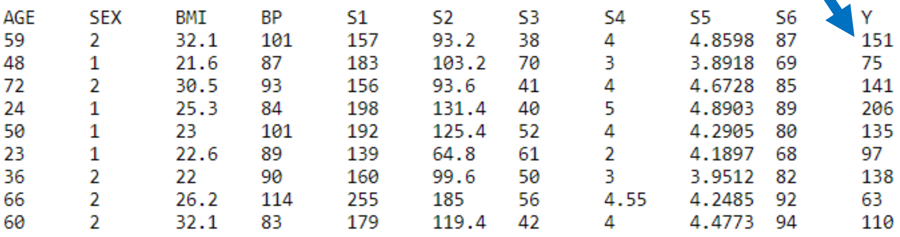
Computer Science Department
CS675 – Introduction to Data Science (CRN: 76747)
Fall 2021

Project #1 / Due 06-Oct-2021

The goal of this assignment is to understand the logic and methods of exploratory data analysis (EDA). The mode of analysis concerned with discovery, exploration, and empirically detecting phenomena in data. EDA has become the default pre-modeling step for every Machine Learning project engagement. Exploratory Data Analysis (EDA) is a way to investigate datasets and find preliminary information, insights, or uncover underlying patterns in the data. Instead of making assumptions, data can be processed in a systematic method to gain insights and make informed decisions.

Investigate the data by utilizing **NumPy**, **Pandas,** any **graph** library (MatPlotlib, Seaborn, Plotly), and Python's **Statsmodel** modules.
The analysis of the data should be focus on predicting the progression of a disease (diabetes in our case).
Get the data from **Stanford U's** Machine Learning Repository:
https://web.stanford.edu/~hastie/Papers/LARS/diabetes.data

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

Load the dataset by using **NumPy's genfromtxt** function (you are allowed to use others…)
https://numpy.org/devdocs/user/basics.io.genfromtxt.html

NOTE: You do NOT build/select a model, you only perform deep-dive analysis on the data.

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

1- Prep the data in order to be ready to be fed to a model.
 Look for missing, null, NaN records.
 Find outliers.
 Transform data – all entries should be numeric.
2- List all types of data, numeric, categorical,…
3- Perform EDA on data.
 Present dependencies and correlations among the various features in the data.
 List the most variables (Feature Importance) that will affect the target label.
4- State limitations/issues (if any) with the given dataset.