Assignment-1 Problem Statement:

Deadline: 4th September EOD

Attached are two datasets.

Classification

- File(s): 'Dry_Bean_Dataset.xlsx', 'Dry_Bean_Dataset.txt' (this text file contains the metadata, i.e. the description of each of the columns in the dataset)
- Description: Given a total of 16 features; 12 dimensions and 4 shape forms, predict the class of a given bean. The 7 classes are - Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira.

Regression

- File(s): 'Insurance.csv'
- Description: Given a number of metrics, predict the individual medical charges billed by health insurance.

Metrics to be used to validate the model:

- **For classification: F1-score**
  Read more about the F1-score here.
  PS: Open in incognito mode if it asks for a subscription.
- **For regression: Mean Squared Error**

1. You may have to perform basic preprocessing as taught in the last week.
2. Write the code for the models and training on your own using Numpy/Scipy**. The use of Scikit-Learn or any automatic differentiation package is forbidden**. Note: You can't use numpy.gradient either.
3. You will be expected to explore univariate + multivariate linear regression in closed-form + gradient descent , logistic regression using gradient descent, and Naive Bayes models.
4. Investigate whether selecting a few columns instead of selecting all features yields a better result.
5. For classification, explore if there are any columns due to which Multivariate Gaussian models would be more suitable. (No need to implement Multivariate Gaussian, just mention which columns and how you figured this out).
6. For the purposes of debugging, you may check your implementations on any randomly generated data.
7. Keep in mind that **the aim is not to maximize the training metric, but the metric on test data** that the model has not seen during training. Take appropriate measures for the same.

What you need to submit is a ZIP file with your name as the roll number:

1. A Jupyter Notebook with well-documented code. The code needs to be in working condition without any modifications to be done by the TA to get the results.
2. The datasets in the same directory so that we can just open-and-run without modifying.
3. A PDF version of the notebook with all cell outputs printed.

Note that different groups have different datasets and will get different results. Do not compare with them -- that will only lead to more anxiety.

Plagiarism from your friends' work or from online will not be tolerated and will invite harsh penalties. Discussion is permitted -- sharing of code is not.
We will be conducting a viva-voce for all students so that we can accurately gauge your sincerity during the assignment as well as your competence.