Assignment-1 Problem Statement (ONLY for AKB and KM TA group):

Deadline: October 3

Attached are two datasets -- Boston House Pricing and Wisconsin Breast Cancer. Analyze the data and create models for the given prediction tasks: For Boston, predict MEDV using the remaining columns. For Wisconsin, predict Class using the remaining columns.

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 + Newton's method, logistic regression using gradient descent + Newton's method, 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 accuracy, but the accuracy on test data** that the model has not seen during training. Take appropriate measures for the same.
8. Metrics to be used to validate the model:
    a. For classification, sklearn.metrics.accuracy_score
    b. For regression
        i. sklearn.metrics.mean_squared_error
        ii. sklearn.metrics.mean_absolute_error
        iii. The following snippet of code (using epsilon = 0.01) for the mean absolute relative error (DO NOT USE THE SKLEARN IMPLEMENTATION)

```
mean_absolute_relative_error = lambda y, yhat, epsilon:
(100/y.size)*np.sum((np.abs(y-yhat)/(np.abs(y) + epsilon)))
```

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 dataset in the same directory so that I can just open-and-run without modifying.
3. A short report of 2-3 pages detailing:
    a. Your assumptions.

b. The methods used.
c. The results obtained.
d. Your observations.

The aim of the report is to demonstrate that you have spent sufficient time and effort on the problem. Obviously, someone who demonstrates that they have tried various different methods, experimented with hyperparameters, etc will earn a higher score in the evaluation than someone who has just run the models once and left it at that.

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.

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