Questions

A. Download the dataset from here. It's a standard dataset, where you have to predict whether the income of a person is above or below 50k$. Implement a classifier using logistic regression from scratch (you are allowed to use only NumPy and Pandas). Also implement L1 and L2 regularization. Report the accuracy on the train, val and test set while using L1 and L2 regularization separately. You may use scikit-learn only for encoding categorical data. You can play around with the data - it may or may not be necessary to use all the features. Can you come up with a reason as to why L1 regularization works better than L2 regularization, or vice versa? Also plot a loss vs iteration and accuracy vs iteration graphs for both models (L1 and L2).

B. Download the MNIST dataset from here. Implement an L1 and L2 regularized logistic regression model using the scikit-learn library. Compute and report the accuracy obtained using one-vs-rest approach for each of the 10 classes, for both the training and test sets. Report the train and test accuracy for both L1 and L2 regularized logistic regression. Comment on whether or not it is a good fit, i.e, underfitting or overfitting.

C. For the MNIST dataset, plot the Receiver Operating Curve (ROC) curve for each class (Do this just for L2 regularized Regression). Plot all ROC curves on the same graph.