Classify handwritten digits using the famous MNIST data (Kaggle)

# Goal:

The goal of your extra credit project is to take an image of a handwritten single digit, and determine what that digit is. You are expected to implement the project using an algorithm of your choice.

# Description of the project:

The data for your project is taken from the MNIST dataset. The MNIST ("Modified National Institute of Standards and Technology") dataset is a classic within the Machine Learning community that has been extensively studied.  More detail about the dataset, including Machine Learning algorithms that have been tried on it and their levels of success, can be found at <http://yann.lecun.com/exdb/mnist/index.html>.

## What’s in the Data!!!

The data files train.csv and test.csv contain gray-scale images of hand-drawn digits, from zero through nine.

Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive.

The training data set, (train.csv), has 785 columns. The first column, called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image.

Each pixel column in the training set has a name like pixelx, where x is an integer between 0 and 783, inclusive. To locate this pixel on the image, suppose that we have decomposed x as x = i \* 28 + j, where i and j are integers between 0 and 27, inclusive. Then pixelx is located on row i and column j of a 28 x 28 matrix, (indexing by zero).

For example, pixel31 indicates the pixel that is in the fourth column from the left, and the second row from the top, as in the ascii-diagram below.

Visually, if we omit the "pixel" prefix, the pixels make up the image like this:

000 001 002 003 ... 026 027

028 029 030 031 ... 054 055

056 057 058 059 ... 082 083

| | | | ... | |

728 729 730 731 ... 754 755

756 757 758 759 ... 782 783

The test data set, (test.csv), is the same as the training set, except that it does not contain the "label" column.

Your submission file should be in the following format: For each of the 28000 images in the test set, output a single line with the digit you predict. For example, if you predict that the first image is of a 3, the second image is of a 7, and the third image is of a 8, then your submission file would look like:

3

7

8

(27997 more lines)

See benchmark code and a sample submission with the [Random Forest Benchmark](https://www.kaggle.com/users/993/ben-hamner/digit-recognizer/random-forest-benchmark).

The evaluation metric for this contest is the categorization accuracy, or the proportion of test images that are correctly classified. For example, a categorization accuracy of 0.97 indicates that you have correctly classified all but 3% of the images.

# What to turn in:

You are expected to create a final project report which should have the following:

1. Description of machine learning method and approach
2. Detailed description of experimental setup (validation, training, testing etc.)
3. Report of results (accuracy, F1, and maybe a confusion matrix)

# Due Date

You will be required to submit all of your documents and source code by **Last lecture**

Submission has to be done through Blackboard. No email submissions will be considered!