COMS 3007: Machine Learning Assignment 2023
Part 2: Solving ML Problems

You have now studied many different machine learning algorithms, with an emphasis on classification. In this assignment, you will be testing out your ability to tackle classification problems.

You have been given a dataset spread over two files. *traindata.txt* contains a set of input data (one data point per row), and *trainlabels.txt* are the corresponding labels (from 0 to 9).

Your task is simple: you are required to build a classifier to classify unseen data drawn from the same distribution as the training data. You can use any method you want, but your goal is to perform as well as possible on the unseen data. This will be auto-marked, with marks assigned based on your accuracy on this data. Specifically, the marking will be competitive, so use everything you have learned to do as well as possible.

Your method must be able to load a file of input data points, and produce an output file of the predicted labels, using your model that you have already trained. Your file must be called *classifyall.py* which reads from a file called *testdata.txt* and writes to a file called *testlabels.txt*.

You are allowed to use these libraries only: *sklearn*, *tensorflow*, *pytorch*, and *pandas*. You will be provided with the same singularity environment that will be used to mark your submission. If your model does not run on the marker, you will receive 0.

You must make two submissions (each just by one person in the group): your code to be auto-marked, and a short report.

Your report should be *at most TWO pages* submitted as a **PDF document** to Moodle. This should include the following points, but should be kept very brief:

(1) All your names and student numbers.
(2) The algorithm you used, with any design decisions, and hyperparameters. Briefly justify what you did, and reference any external papers/websites that you relied on heavily. This document needs to convince me that you understand and can explain exactly how your algorithm works.
(3) Any clever tricks that you found to be helpful, including data pre-processing.
(4) A measure of your expected accuracy (based on any test data you separated out).

Important:

- The due date for submissions is the end of **Tuesday, 6 June**.

- You must submit and work in groups of **between three and four people** (having more or fewer people will automatically halve your marks). Make sure all your names AND student numbers are on the submission, otherwise you will receive 0.