# ENG EK 125 Summer 2020
## Final Project

**Project Intro:**

Machine Learning! Big Data! Artificial Intelligence!

It is everywhere. Large, publicly available, data sets are ubiquitous, and will be even more so with the advent of 5G. You have also seen how to scrape your own data from web sites using API keys. Everyone is impacted by ML and AI. It is more and more important for everyone (not just engineers) to understand ML and AI. The two foundations of Machine Learning are coding and statistics. You now have the necessary coding chops. You know a little bit about statistics, and will learn even more when you take EK 381. This project is intended to get you started on ML. You will do this using the Machine Learning Toolbox in MATLAB, somewhat as a black box since you do not yet have all of the background to fully understand all of the ML algorithms that are in the Toolbox.

It is important to understand that there are biases in many data sets and algorithms. A well-known example is facial recognition software that was trained on images of mostly white men, and therefore had difficulty recognizing faces that were not white males. As a societal engineer, it is your duty to avoid such biases. Therefore, to begin your project, find an article that talks about these biases and write a paragraph synopsis of the article.

**Group Hand-in Details:**

You will work in the same group of two or three from the Project Proposal submission. This project is due on **Monday, August 10th.** Please note that this is a Final Project in lieu of a Final Exam, so expect minimal help – this is YOURS to do!

You will submit one copy of your project per group.

**Project Description:**

You are to, as a team, create a project and the solution. The project must be based on the Machine Learning Toolbox. It must include at the very least the following:

- Reading and analyzing at least one data set of substantial size (at least 10,000 data points) into your own data structure(s)
- Plot meaningful data points relevant to the problem you are trying to solve

**Be creative! Make it interesting! Your project must be arguably useful to society! Embrace the concept of Boston University creating the Societal Engineer ™. You will be in competition with the other project groups in terms of usefulness and creativity.**

What you are to do is similar to Homework 4, only using your own data set to try to solve (or assist in solving) a societal problem. It is strongly recommended that you play around with the ML Toolbox using one of the built-in data sets (the Fisher irises data is a good one) first, before trying your large data set.

You are to document your process, showing all results and explaining your decisions. Do not include your entire data set (!), but representative rows from it.

Usually, with ML, scrubbing the initial data set is the most important and time-consuming aspect. It may seem a little boring, but it is critical to the success of your ML algorithm(s). Start by considering rows that have missing data or outliers. Are the outliers errors or potentially valuable information? For the features, decide whether to delete any that are redundant. Consider one-hot encoding. Consider normalizing the data. Show the code that you use to modify your data set. Note that you may need to go back and do this again once you try the ML Toolbox. For example, some algorithms allow categorical data, but some require only numbers, so you may need to revisit your Feature Analysis.

It is recommended that you import your data set as a .csv file. That way, you can easily read it into MATLAB as a table. Begin by randomizing your data set. Run some statistical analyses.

Get your randomized data set as a table variable in the base workspace, and then get into the ML Toolbox. In MATLAB, click on the APPS tab (not the HOME tab). You will see two supervised learning apps, "Classification Learner" and "Regression Learner". *If you are trying to classify data, choose the former and if you are trying to predict a real number, choose the latter*. Click on New Session, then From Workspace, and then choose your table variable. Some features will be the "predictors" and one will be the "response". You can change that. Start the session. You will see a lot of algorithms listed. If any are greyed out, it is because your data set is not in the correct format for that particular algorithm. Choosing "All Quick to Train" and then clicking on Train will run through all of the available algorithms.

Play around with this! Part of the project is for you to explore, and to learn how to learn. You will not understand some things. Look them up! Use the Discussion Forum on edge! If you want to save your work, choose Export Model.

**Project Deliverables:**

1. **Cite your article** on biases and write a synopsis of the article.
2. Explain your project at a high level. Give some **background on the problem** you are trying to solve, the group of people for whom the product solves a problem and why you believe this solution solves that problem. Include the sources/links to the datasets you used here.
3. The remainder should be a description of your process as you make sense of all of your data and make inferences. Include *everything* you try, whether it works or not!