The project is based on the NIH Chest-X-ray dataset found on Kaggle: https://www.kaggle.com/nih-chest-xrays/data. The dataset is composed of 112,120 jpg images of chest x-rays from 30,805 unique patients labeled with 14 different disease classes.

My partner and I decided to approach separately the data loading, data cleaning and model building portion of the task. We each had our own methods and models to utilize. My partner used a pre-trained CNN and I used an MLP. Our goal, in the end, is to form a consensus about which approach works best and which we would present.

My contribution to the project consists of all of the code on the MLP side of the project, as my partner's contribution consists of all of the code on the pre-trained CNN side. I had difficulty in every step of the project. Importing the data from Kaggle into my VM instance proved to be difficult. I had to create a hidden folder in my instance and download my Kaggle API key into the folder. I then had to pip install the Kaggle API in my instance. Only then I was then able to down all 46 GB of data into my VM instance. After unzipping the data and locating the 12 folders they were unzipped to, I was able to compile the data. I relied heavily on a Kaggle notebook for importing and mapping the labels to the data: https://www.kaggle.com/adamjgoren/nih-chest-x-ray-multi-classification. The compiled dataset consists of 112,120 images mapped to 15 unique labels. There was an over representation of the "No Finding" class which was dropped to avoid bias in the model. The labels were then one-hot encoded in preparation for an MLP model. The data was split into 20 percent testing and 80 percent training, and the images were converted to grayscale and resized to 128 by 128. The MLP model consists of two layers. The first layer consists of 12 neurons with a relu activation function. The output layer consists of 14 neurons with a softmax activation function.

I was unable to produce results with my model. I had issues with TensorFLow and my Cloud instance, and we decided to abandon MLP for my partner's CNN and MobileNet models.

I found 48 lines of code from the internet, modified 10 of those lines and added 17 lines of my own code. The percentage is 58.5.