General Assembly DC - DAT5
Predicting HIV Progression in Patients
By: Nana Adu-Krow

1.0 Problem Statement and Hypothesis

By applying our knowledge of viral load and cd4+ counts found in patients with HIV can we actively predict whether that patient would respond well to a 16 week therapy program?

Another possibility I'd like to explore is whether we can effectively use the amino acid sequence in our predictions to see whether patients would respond positively or negatively.

I believe through the ensembling method and combining different models emphasizing various features I should be able to come up with an effective algorithm. With the conceptual approach of combining logistic regression, single tree decision analysis, as well as naive bayes I should be able to accurately predict whether a patient will respond well to treatment or not.

2.0 Description of the Dataset

The data set was provided through one of the older Kaggle Competitions. It highlights the different features given on both the train and test data set. The Dataset is organized as follows:
Col-1: Patient ID
Col-2: Responder status ("1" for patients who improved and "0" otherwise)
Col-3: Protease nucleotide sequence (if available)
Col-4: Reverse Transcriptase nucleotide sequence (if available)
Col-5: Viral load at the beginning of therapy (log-10 units)
Col-6: CD4 count at the beginning of therapy

3.0 Preprocessing

The data was collected from the Kaggle Competition website and appropriately gave two csv files. The first was the training dataset and the second smaller data was the testing dataset.

4.0 Exploring the Data

The first thing that I was interested in doing was looking at the training data set and seeing the visual relationship between the viral load and CD4+ count of the patients who responded well and the patients who didn't respond well. There are definitely more patients who didn't respond well to patients who did, and you can see the negative relationship between viral load and cd4 count in patients with responder status 0. Patients with responder status 1 had a bit more of positive slope in comparison.

However with what I want to do with prediction, looking at linear regression relationships wouldn't help. Classification methods would be the method I use moving forward.