**DATA SCIENCE PROJECT 2 (MAST90107)   
Semester 2 2021**

**GROUP 17**

Meeting No. 4

**Date & Time:** 13th August 2021, 10AM - 11:05AM

**Agenda:**

1. Going through XG Boost progress
2. Going through NN/Autoencoders progress

**Attendees:** Roberto Bonelli, Brenden Ansell, Vrinda Raj, Kartika Waluyo, Ziad Al Bkhetan

**Discussion:**

1. XGB:

* To train blood data to lung data - use all of the whole blood data and try to predict atleast 20 genes in lungs. We start with a big predictor set (X) and end up with a small predicted set (Y).
* The function will be required to loop the dataset to predict each lung gene.
* Scatter plot can be used to visualise the predictions.
* Find methods in XGBoost that include matrices
* Try to calculate correlation - to understand accuracy and relationship

1. NN:

* If autoencoders are not working out and are to complicated - we plan to shift to using neural networks.
* Find hyperparams by tuning the existing NN model.

1. To find ways to be able to impute missing values.
2. To implement data partitions instead of using random sampling.