**DATA SCIENCE PROJECT 1 (MAST90106)   
Semester 1 2021**

**GROUP 17**

Meeting No. 14

**Date & Time:** 19th May 2021, 10:00AM

**Agenda:** Discussing proportions

**Attendees:** Kartika Waluyo, Ziad Al Bkhetan, Vrinda Rajendar Rajanahally

**Discussion:**

1. Final report - progress
2. Training, test and validation sets

- there is no scientific reasoning to determine proportion

- the training set must be the biggest to avoid overfitting

- the second largest could be assigned to the validation set

- **TRY** cross validation: all data will be assigned to all 3 sets to bring out the best accuracy

1. Multidimensional correlated input to multidimensional correlated output

- gene x, y and z in one sample is used to predict gene x in another sample

- neural network

- traditional machine learning: using linear regression and creating one model per gene in every tissue (not an accurate solution)

- Decision tree/Random Forests

**Next steps:**

1. Send Ziad the address of CSL and this week’s meeting details
2. Research cross validation and bootstrapping
3. Think about how you can model **multidimensional** input data?
4. Keywords to search: feature selection engineering, feature selection

**Resources:**

1. <https://machinelearningmastery.com/difference-test-validation-datasets/>