# Summary of Module 5 - Week A Discussion

The group discussion for DATA 6550 focused on the selection of a paper and dataset for the reproducibility project. Key topics included:

1. \*\*Paper & Dataset Selection:\*\*

- The team reviewed multiple papers, prioritizing those with well-documented code and recent publication dates.

- Chance found several promising papers but noted that some were too old or lacked a cohesive coding structure.

- A Toronto-based study was considered, but it primarily used Python in a manual, object-oriented approach, which differed from the group’s usual methods with Jupyter Notebooks.

- Eventually, the group leaned toward selecting a MOOC dropout prediction paper that used machine learning and deep learning models to predict student dropouts in Massive Open Online Courses (MOOCs).

2. \*\*Discussion of Code Structure:\*\*

- Chance explained the difference between traditional Python scripts (.py files) and Jupyter Notebooks, emphasizing that the Toronto paper relied heavily on a script-based approach that would be harder to follow and debug.

- He highlighted the challenges of working with multiple Python files, emphasizing the need for a consistent structure to avoid errors during execution.

3. \*\*Reproducibility Concerns:\*\*

- The group discussed the criteria for selecting a paper, emphasizing the need for publicly available datasets and clear methodologies.

- Chance mentioned concerns about hyperparameter tuning and model accuracy for the chosen paper and emphasized that understanding the model’s parameters would be key to achieving successful reproducibility.

4. \*\*Proposed Division of Work:\*\*

- Chance suggested dividing the work by method/model:

- One person to handle Keras with TensorFlow.

- Another to work with PyTorch.

- Others to run models for the different datasets.

- Each team member would be responsible for replicating results from the paper using different techniques.

5. \*\*Exploration of Data Sources:\*\*

- Two primary datasets were identified:

- KDD Cup Data – used for predicting MOOC dropouts.

- CEGEP Academic Performance Data – a smaller dataset used to predict enrollment in science courses.

- The team confirmed that both datasets met the project’s requirements.

6. \*\*Code and Repository Review:\*\*

- Chance identified the GitHub repository associated with the MOOC dropout prediction paper, which included well-commented code and accessible datasets.

- The paper's methodology used Scikit-learn for classification and included detailed comparisons of models such as Keras, TensorFlow, and FastAI.