**Agenda:**

* Discussion of data split
* Discussion of Gadi usage
* Discussion of how to best present my results
* Project timeline
* Discussion of what should be included in the introduction

**Discussion of the Data Split: 80/10/10**

As discussed in our previous meeting, I divided the raw merged data into a training, validation, and testing set.

*Split by countries:*

For each of high, upper middle-, lower middle-, and low-income countries (as categorised by the World Bank), I allocated roughly 80% of the rows into training, validation and testing sets. I had the caveat that the same country could only be in one of the training, validation, and testing sets to preserve independence between sets. I then concatenated the training/validation/testing data from each of the income brackets together to form the full training/validation/testing sets.

*Split by year:*

I can have the same structure when splitting by year. However, I had a few critical questions that I needed answered before I could proceed:

* There are different proportions of missing data per year.
  + Should I try to distribute the proportion of non-missing data according to the 80/10/10 split (likely then having non-consecutive years in the same split, e.g. 2003 and 2012), or just do the split using consecutive years?

**Discussion of Gadi usage:**

I’ve had a range of issues (from how to write the scripts effectively to debugging to how to load the necessary modules) and a steep learning curve with Gadi but am slowly but surely getting there!

*Current major roadblock*: my code does not appear to be making use of the GPUs on the supercomputer, causing it to run extremely slowly. I’ve done multiple attempts, with wall times ranging from 5 to 10 hours where the script only progresses through ~40 to 70 Optuna trial runs before timing out and all progress is lost.

* How do you ensure that your script is making use of the supercomputing resources?

**Discussion of How to Best Present my Results:**

I have the following ‘experimental variables’:

* Whether 95% thresholding has been applied
* Whether correlation-based imputation was applied
* Whether standardisation was applied
* Type of imputation method
* Type of machine learning model

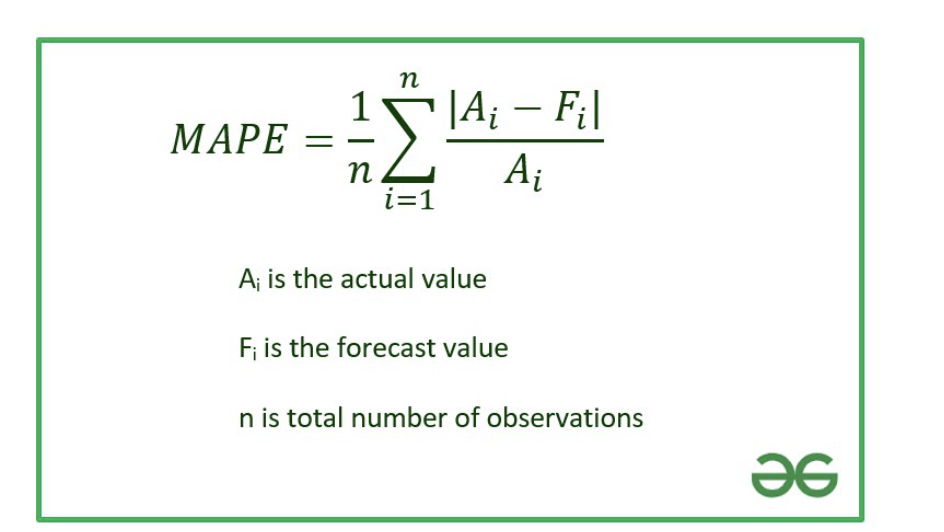
*Proposed method:*

* Average over all datasets and do a heatmap of average MAPE for each machine learning model / imputation method combination
* Then choose best combination/few combinations and represent as heatmap with combination on the x-axis and type of dataset on the y-axis, showing the MAPE score
* Choose the best combination and test possibility of ensemble-based model

**Project Timeline:**

*Poster deadline:*10 October, 2025

*Thesis deadline:* 24 October, 2025

- swapping the A and F will change the value of the metric

Making it asymmetric

* Absolute difference / max(abs(A), abs(F))
  + Change to this
* Also use average absolute error/difference (do not divide by anything)
  + Mae
* Root mean square error
* Expecting a high correlation

*Proposed timeline:*

Between now and 21 July (first day of the new semester):

* Finish running the code on Gadi for both possible splits
* Represent all data neatly
* Write introduction and method
  + Methodology should be well-laid out
    - Explore, compare techniques, and argue why I continue with one method than others
    - Peer-reviewed review giving pros and cons of different imputation methods to show that I am following the standard
      * Try to prove that you are not introducing bias in this process
      * Measure potential bias
    - Baseline is original data without imputation (can use 95% threshold or 100%), can also normalise (similarity between columns is a good reasoning for normalisation)
      * Apply ml models that work on such data to have a baseline metric to use as a comparison
      * Once finished this, explore further with imputing data
  + Back up methodological choices from papers in the literature
    - Need to use this in the methodology
    - Find papers that have done imputation for public health data

Expect errors to be higher for testing that for validation

21 July to 3 August:

* Find best combination of models and test ensemble-based combination, then compare with single best model

3 August to 31 August:

* Finish all coding and experimentation, including:
  + Using SHAP to find relative predictive power of different variables

31 August to end of semester/due dates:

* Write thesis (results, discussion, etc.)
* Produce poster
* Fine-tuning

Marking process:

* Externally marked and moderated

**Discussion of What Should be Included in the Introduction**

* What is maternal mortality and why is it an important issue to tackle
* Why reducing maternal mortality is dependent on effective data collection
* Machine learning in health economics
  + Previous work and explanation of what is possible
* Machine learning methods used in this paper