# 1. Introduction

# 2. Background Information

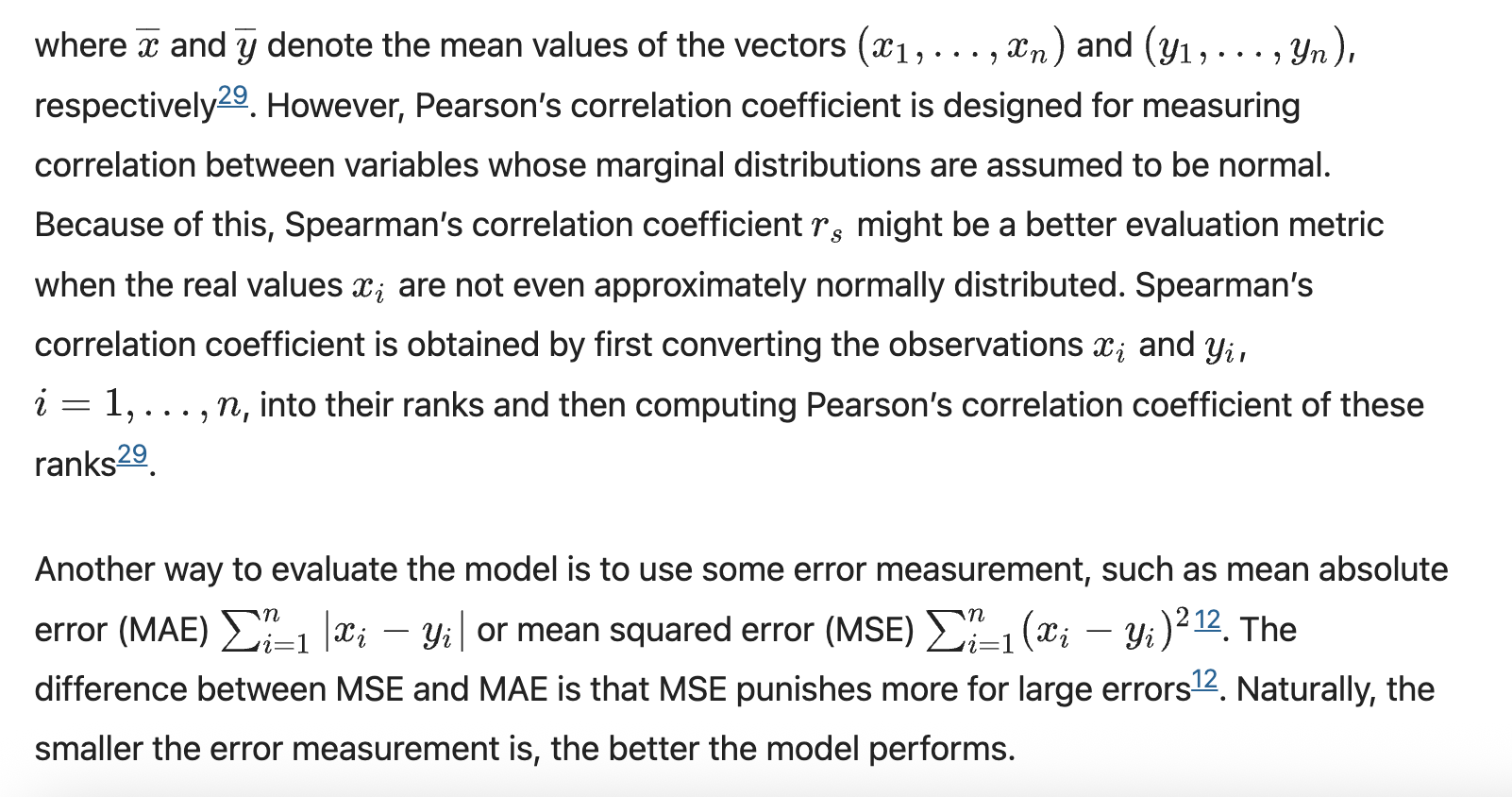
Here I will explain the difference between national and modelled mmr

The national estimates reflect a compilation of country data sourced from vital statistics and civil registration, population surveys and censuses, studies on maternal mortality, and various other data repositories, including data collected from surveillance sites (1). It was compiled the Maternal Mortality Estimation Inter-Agency Group.

**Statistical concept and methodology:**

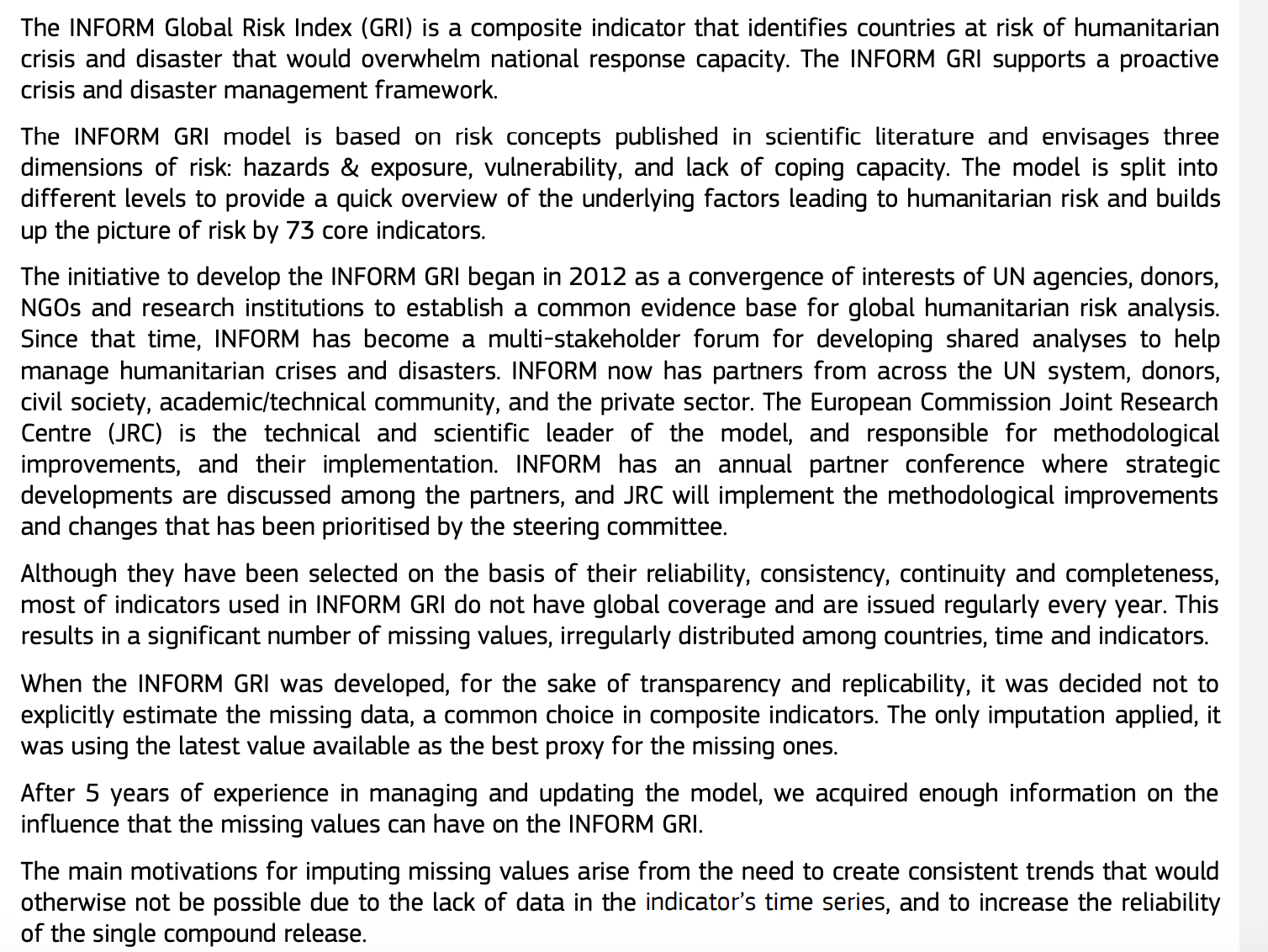
Maternal mortality ratios are generally of unknown reliability, as are many other cause-specific mortality indicators. Household surveys such as Demographic and Health Surveys attempt to measure maternal mortality by asking respondents about survivorship of sisters. The main disadvantage of this method is that the estimates of maternal mortality that it produces pertain to any time within the past few years before the survey, making them unsuitable for monitoring recent changes or observing the impact of interventions. In addition, measurement of maternal mortality is subject to many types of errors. Even in high-income countries with reliable vital registration systems, misclassification of maternal deaths has been found to lead to serious underestimation. The national estimates of maternal mortality ratios are based on national surveys, vital registration records, and surveillance data or are derived from community and hospital records.

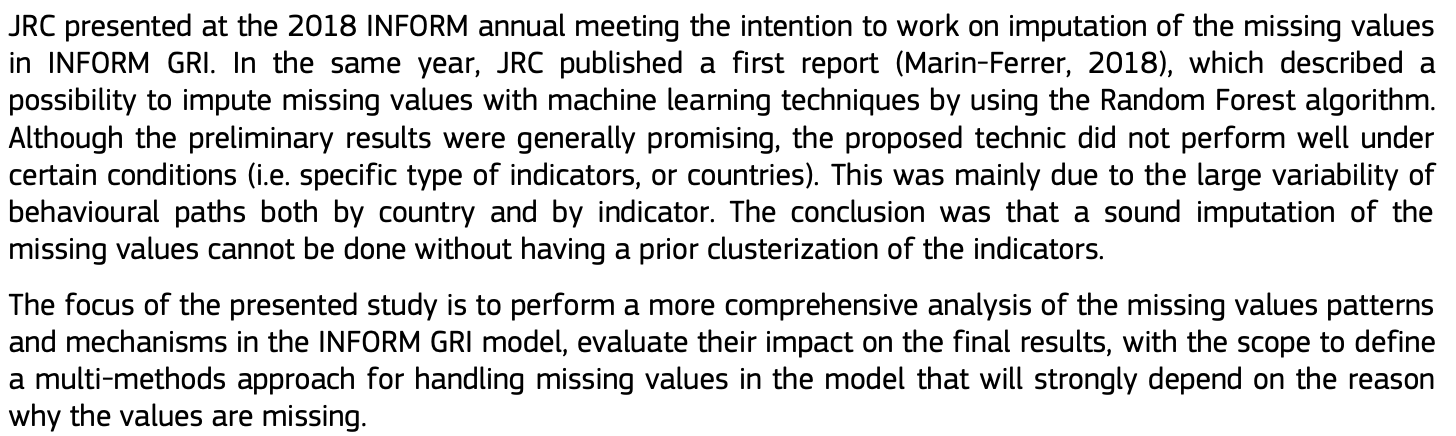
Loss vs performance metrics

* <https://link.springer.com/article/10.1007/s10462-025-11198-7>
* <https://www.nature.com/articles/s41598-024-56706-x>
  + 

# 3. Related Work

<https://drmkc.jrc.ec.europa.eu/inform-index/Portals/0/InfoRM/Publications/EUR%2030037%20-%20Imputation%20of%20missing%20values%20in%20the%20INFORM%20Global%20Risk%20Index%20-%20print.pdf>





Unlike in my study, where modelled results were available

# 4. Materials and Methodology

All code was written using Python3 and run in Visual Code Studio or on the Gadi supercomputer at the National Computational Infrastructure. Where appropriate, the random seed was set to 42 for reproducibility.

**Figure 1**: Overview Flowchart of Experimental Workflow

## 4.1 Data Sources and Merging

### 4.11 Data Sources

Data was sourced from a variety of World Health Organisation (WHO) and World Bank Group (WB) data repositories. The final, merged dataset socioeconomic, health-related, and environmental indicators. Information about the specific datasets used in this study was summarised in Table 1, below. The specific variables gathered from each data source are listed in the Appendix, section 9.1. The specific features included in each dataset were originally gathered from a diverse range of sources. See the webpages for more information.

**Table 1:** Summary information about the datasets used in this study.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type of Dataset** | **Number of Features** | **Date Range** | **Number of Areas Covered** | **Demographic Subsets Chosen from Dataset** | **Source** |
| MMR modelled and estimates | 2 | 1985-2023 | 242 | NA | World Bank Group’s Gender Data Portal |
| Overall health literacy, metrics, agency | 198 | 1960-2023 | 265 | NA | World Bank Group’s Gender Data Portal |
| Illness incidence and prevalence | 193 | 2000-2019 | 194 | Sex | WHO Health Inequality Data Repository & IHME |
| Empowerment | 9 | 1991-2023 | 120 | Economic status (quintiles 1, 5) | WHO Health inequality Data Repository after re-analysis by the WHO Collaborating Center for Health Equity Monitoring |
| Socioeconomic, education, environmental | 64 | 1970-2023 | 195 | Sex, economic status (quintiles 1, 5), residence (urban, rural) | WHO Health Inequality Data Repository, sourced from The World Bank Data Catalogue |
| Income level | 1 | 2024 |  | NA | WHO Health Inequality Data Repository, produced by the WHO’s Global Health Observatory |

The national and modelled estimates for the maternal mortality ratio (MMR) were sourced from the World Bank Group’s Gender Data Portal (1). This dataset contains MMR estimates collected yearly for 242 regions, countries, territories, and areas. The national estimates were collected between 1985 and 2018 while the modelled estimates were collected between 1985 and 2023. See Section 3 for discussion of the WHO’s calculation of the modelled MMR estimates, and how they differ from the national estimates.

Some of the datasets used in this study contained disaggregated data. For example, features were sex or economic status specific. However, the MMR estimations were not disaggregated. Including disaggregations as a separate column would therefore produce a missing value in the MMR estimates columns when merging the datasets. To prevent this label variable from having missing values, I did not include the disaggregated demographic groups as a separate variable. Instead, the indicator was replicated for each included subgroup. For example, rather than a single feature, ‘Feature 1’, multiple versions of the feature – one for each subgroup – were included. See Table 2 for an illustrative example.

**Table 2**: Illustrative example of a feature being presented with demographic specific data. The bolded text represents the subgroup being represented.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Country | Date | Indicator 1 **Female** | Indicator 1 **Male** | Indicator 1**Rural** | Indicator 1 **Urban** | Indicator 1 **Quintile 1** | Indicator 1 **Quintile 5** |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

A dataset containing 198 health-specific indicators, including health literacy and health-related decision-making ability, as well as mortality and disease rates, was collected from the World Bank Group’s Gender Data Portal (2). Its data was collected annually between 1960 and 2023, and covers 265 regions, countries, territories, and areas. This dataset was not disaggregated. For brevity, this dataset will be referred to as the ‘overall health’ dataset from this point forward.

Similarly, a dataset containing 193 indicators detailing the incidence and prevalence of a greater variety of illnesses was downloaded from the WHO Health Inequality Data Repository (3). The estimates were produced by the Institute for Health Metrics and Evaluation (IHME) Global Burden of Diseases, Injuries and Risk Factors Study (GBE). The specific data was sourced from the Global Burden of Disease Study 2019 on the Global Health Data Exchange website. This dataset contained nation-specific information collected annually between 2000 and 2019 for 194 countries, territories or areas. I used features from this dataset that were disaggregated by sex. For ease of reference, this dataset will be referred to as the ‘illness dataset’ for the rest of this paper.

A dataset containing 9 indicators describing women’s social independence, decision-making agency, and attitude to violence was downloaded from the WHO Health inequality Data Repository after re-analysis by the WHO Collaborating Center for Health Equity Monitoring (4). The data was collected between 1991 and 2023, describing 120 countries, territories, and areas. I included this dataset’s economic status specific disaggregations (quintiles 1 (poorest) and 5 (richest)). This dataset will be referenced as the ‘empowerment dataset’ from this point onwards.

A dataset containing 64 features related to economic status, education level, and location was gathered from the WHO Health Inequality Data Repository (5) and sourced from The World Bank Data Catalogue. It contains nation-specific information recorded between 1970 and 2023 for 195 countries, territories or areas. I used this dataset’s sex (female, male), place of residence (urban, rural), and economic status (quintiles 1 (poorest) and 5 (richest)) disaggregations. This dataset will be referred to as the ‘socioeconomic dataset’ for the rest of this paper.

Finally, the World Bank’s 2024 categorisation of an area’s income level (1 to 4, with 4 being the highest income level) was sourced from the WHO Health Inequality Data Repository and was produced by the WHO’s Global Health Observatory (GHO) (6).

1. <https://genderdata.worldbank.org/en/indicator/sh-sta-mmrt>
2. <https://genderdata.worldbank.org/en/topics/health#idAllIndicators>
3. Illness prevalence
4. empowerment
5. <https://www.who.int/data/inequality-monitor/data> check source
6. income level

### 4.12 Merging Data

All datasets used in this report contained a column for the country/region and a column with the associated ISO3 country codes, as described in the ISO 3166 international standard. Given that different versions of the same country’s name could be used in different datasets (e.g. United States versus United States of America), I joined the datasets on ISO3 code and year. Thus, the final merged dataset contained a row per area, year unique ID. The rest of the columns were features extracted from the original dataset. This merged dataset contained 16,948 samples and 733 features.

As described in the Section 3 and Table 1, the WHO recorded national and modelled MMR estimates are only available from 1985. Therefore, I excluded all data collected before 1985 and after 2018 from my analysis and excluded any features for which there was no data collected during these years. This restricted my dataset to 9,018 rows and 727 columns.

The modelled MMR estimates were held aside and used to evaluate the model, as described in Section 4.34.

### 4.13 Data Exploration

To better understand my data and choose the most effective data pre-processing techniques and machine learning models, I plotted the proportion of missing data per year across all countries between 1985 and 2018 (see Figures 2 and 3).

#### 4.131 MMR Estimates

A graph with blue dots

AI-generated content may be incorrect.

**Figure 2:** Proportion of missing national MMR estimates per year between 1985 and 2018.

The proportion of missing national MMR estimate data decreases from approximately 75% in 1985 to roughly 58% in 2011 before increasing quickly to nearly 95% in 2018. See Section 4.241 for a discussion of how this affects the train/test split.

#### 4.132 Feature Data Availability

A graph of the number of years

AI-generated content may be incorrect.

**Figure 3:** Proportion of missing data across all countries and features per year between 1985 and 2018.

Before 2000, the dataset had close to or greater than 90% missing data. In contrast, between 2000 and 2018, the dataset had been 40 and 80% missing data. For 4 years, the proportion of missing data was less than 50%. The specific figures are shown below in Table 3.

**Table 3:** Years with the lowest proportion of missing data across all countries and features, rounded to two decimal places.

|  |  |
| --- | --- |
| **Date** | **Proportion of Non-Missing Data** |
| 2000 | 0.44 |
| 2005 | 0.49 |
| 2010 | 0.42 |
| 2015 | 0.43 |

This pattern was likely due to a group of indicators being reported with a periodicity of 5 years. It was taken into account when splitting the data into train/test subsets (see Section 4.241).

#### 4.133 Missing Data Pattern

As discussed in the background information, past studies have stated that it is nearly impossible to definitively state or test whether a dataset is missing at random (MAR), missing completely at random (MCAR), or missing not at random (MNAR) [1, 2]. As a result, researchers have found that treating all data as MAR is a reasonable approach between it lies in the middle of the MCAR to MNAR spectrum [1]. Thus, it is reasonable to treat this dataset as MAR for future data processing and imputation.

Moreover, the pattern of missingness may be MAR because the data is more likely to be missing if it is collected in a year other than 2000, 2005, 2010, 2015, 2019. Thus, the probability of a datapoint being missing is related to observed data (e.g. year). While there is the possibility that the data is MNAR, as potentially a country may have reduced willingness data that reflects negatively, this may be unlikely due to international reporting obligations [citation]

Nevertheless, in this paper I assume that the data is MAR. Therefore, as discussed in Section 2, assuming that the data was MCAR and proceeding to remove all rows and columns with missing data could bias the data towards those countries reporting greater amounts of data (1). As a result, I solely chose to use machine learning models that could work missing data.

1. <https://link.springer.com/article/10.1186/s40537-021-00516-9>
2. <https://link.springer.com/article/10.1007/s40273-023-01297-0>

## 4.2 Data Cleaning and Pre-Processing

### 4.21 Initial Cleaning and Removal

#### 4.211 MMR Estimate Data Removal

All year/region combinations that were missing national MMR estimate values were removed to prevent the estimate from needing to be imputed. This imputation may have caused the machine learning models to be trained on incorrect feature/MMR estimate pairing, introducing inaccuracy into the models.

As a result of this pre-processing step, the number of rows in the dataset decreased from 9, 018 to 2,816.

#### 4.212 Initial Feature Removal

I removed the modelled MMR estimate from the dataset, as keeping it as a feature would help the machine learning models ‘cheat’ as they use the modelled estimate to predict the national estimate. Instead, this modelled estimate was used later to evaluate the model [section?].

Additionally, I removed the following features:

* 'Number of maternal deaths’
* ‘Lifetime risk of maternal death (1 in: rate varies by country)’
* ‘Lifetime risk of maternal death (%)’

According to the World Bank, ‘A maternal death refers to the death of a woman while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes.’ (1). This is very similar to the definition of the MMR estimate, with the major difference being that the ‘number of maternal deaths’ is a count while the MMR is measured per 100,000 live births. Thus, the ‘number of maternal deaths’ was excluded from my feature dataset, as the model could simply ‘cheat’ and use this feature instead of learning how to predict the MMR from socioeconomic and health-related data.

Similarly, the two features measuring the ‘lifetime risk of maternal death’, as a rate or percentage, are similar statistics to the MMR. More specifically, the Pearson’s correlation coefficient between the national MMR estimate and the ‘lifetime risk of maternal death (%) was 0.93, indicating their strong positive correlation. As before, to prevent the model from ‘cheating’ by using these features to predict the MMR, I excluded these features from my dataset.

1. <https://genderdata.worldbank.org/en/indicator/sh-mmr-dths>

#### 4.213 Initial Data Removal Summary

As a result of rows with missing national MMR estimates and removing the three features discussed above, the final, merged dataset consisted of *2,816 samples and 723 columns*.

### 4.22 Splitting into Train/Test Sets

The cleaned and pre-processed dataset was split into train and test subsets in a 90:10 ratio. This split was performed in two different ways, each for a specific purpose.

Citation for 90:10

#### 4.221 Train/Test Split – Split by Region

Independence between the train and test set was guaranteed by mandating that a specific region could not be present in both the train and test set. For example, data about Australia could be in either the train or test set, not both. This independence preserving measure meant that the train/test split was not exactly 90:10, but instead within one or two percentage points, as the number of rows per region did not always allow for this proportional split. The list of regions in the dataset was shuffled being split into train and test sets to prevent alphabetical ordering from affecting the region’s allocation.

Additionally, the World Bank categorised regions as being high income, upper middle income, lower middle income, or low income. The 2024 categorisations were used to label all countries in this dataset. The 90:10 split was applied to the subset of countries in each specific income level to ensure the same proportion of regions from each broad income level were represented in the train and test sets. These train test subsets were then merged into a single train set and test set.

This split was used to evaluate whether the machine learning models could predict the national MMR estimate for a region within the current study period to inform policy makers about their population’s current health status.

#### 4.222 Train/Test Split – Split by Year

The second method for splitting the dataset into train/test subsets was to divide the data by year. All data collected between 1985 and 2014 was placed in the train set and all data collected between 2015 and 2018 was placed in the test set. This division produced an 88:12 train/test split. The proportions were not cleanly 90:10 because I wanted to ensure that the test set contained sufficient non-missing data to be useful. Thus, I included data from 2015 in the test set, as it was one of the four years containing less than 50% missing data (see Section 4.132).

The purpose of this split was to determine whether the trained model could be used to predict future MMR estimates given values for specific socioeconomic and health-related indicators. In this way, policy makers can predict the effect of a program targeting one of these indicators to determine its effect on MMR.

### 4.23 Cross-Validation

The two train datasets (split either by region or by year) were further split into train/validation pairs in an 80:20 ratio. Each train dataset was split into 5 cross-validation folds, each of which being a specific permutation of the train/validation 80:20 split.

When splitting the train dataset split by region, a specific region could only be in the train or validation set. When splitting by year, a specific year could only be in the train or validation set. The choice of set could change between folds. These conditions resulted in slight deviations from the 80:20 ratio, but only within a percentage point.

By performing cross-validation, I could assess how slight changes in the composition of the training dataset affected model performance, thus commenting on the models’ generalisability. Citation and motivation for this approach, and check independence of folds

Splitting into cross-validation folds was performed by scikit-learn’s GroupKFold method.

### 4.24 Iteratively Removing Rows and Columns Containing More Missing Data Than a Specific Threshold

- graphs

- cv

Many studies published in the public health domain use a missing data threshold of 60 to 90%, where variables with more than 90% missing values are excluded from the dataset [1]. This 90% threshold has been justified with the use of simulation studies, which have shown that principled imputation methods can produced unbiased results for up to 90% missingness assuming the data is missing at random [1].

To test this approach, I trained my machine learning models on three versions of my dataset, each with a different proportion of missing data. This experimental approach to choosing an appropriate missing data threshold was particularly important for my dataset, as only 161 features out of 722 (excluding the national MMR estimate) had a missing data proportion of less than 90%.

I experimented with training my machine learning models on three versions of my dataset:

* No missing data removed
* All columns and rows missing more than 95% of their data
* All columns and rows missing more than 85% of their data

Stricter missing data limits were not employed due to the possibility of introducing bias via removing samples of missing data when data is not missing completely at random, as discussed in Section 4.133.

To produce these different versions of my dataset, I iterated through each row and column of the data and removed any that had a higher proportion of missing data than the relevant threshold. I repeated this process until no new rows or columns were removed, as removing a row could change the proportion of missing data in a column, and vice versa.

This process was performed for each individual fold of training data, as performing iterative data removal on the initial, whole dataset could cause data leakage given that the proportions of missing data would have been calculated using both the train and test data.

Given that the machine learning models could handle missing data, imputation was not deemed necessary.

1. <https://www.jclinepi.com/article/S0895-4356(18)30871-0/pdf>

### 4.25 Encoding Categorical Variables

My dataset contains one categorical feature: region name. Unfortunately, Scikit-Learn’s random forest regression model cannot work with non-numeric data. I created a one-to-one pairing between region name and an integer, then replaced the region name by the integer in all feature data used to train the random forest regressor. Given that there are over 200 regions present in the data, I used this approach instead of one-hot encoding, which would meaningfully increase the data’s dimensionality, increasing computational complexity and challenge [1].

This encoding was not applied to the data used to train the other machine learning models, as they could work with categorical data and using this type of ordered numeric encoding is that it could introduce ordinal relationships between countries, potentially causing the machine learning model to learn a non-existent pattern [1].

1. <https://arxiv.org/html/2312.16930v1>

### 4.26 Summary of Datasets Produced Via Pre-Processing

The cleaned data was split into two train/test subsets. One subset was split by region while the other was split by year. These subsets were each further split into 5 train/validation set permutations, again split by region or year depending on which initial train/test split was being used. 3 versions of each of these folds were then produced by iteratively removing rows and columns with more missing data than a threshold proportion.

This produced 15 train/validation datasets per train/test split.

### 4.27 Normalisation

S

## 4.3 Machine Learning Models

### 4.31 Choice of Model

This thesis trained three machine learning models:

* Scikit-Learn’s Random Forest Regressor
* XGBoost’s XGBRegressor
* LightGBM’s LGBMRegressor

Information about the models’ specifics was discussed in section ??.

Each model was trained on each train/validation fold, with their results discussed in the results section.

### 4.32 Hyperparameter Tuning

The Optuna hyperparameter optimisation framework was used to determine the best hyperparameters for each model. See Tables 4, 5, and 6 for the specific hyperparameters and their ranges for each model. 300 trials were run for each model/cross-validation fold combination. Each trial represents a choice of hyperparameter values for the subset of hyperparameters selected for tuning. All other hyperparameters were set to their default values. During each trial, the machine learning model was trained on the specific train data fold with the chosen hyperparameter values. Its performance was evaluated by calculating the mean squared error for the model’s predictions on data from the associated validation fold. The model trained with the choice of hyperparameter values that produced the lowest mean squared error across all trials was saved and evaluated on the test data.

**Table 4:** Hyperparameter Tuning for Scikit-Learn’s Random Forest Regressor

|  |  |  |
| --- | --- | --- |
| **Hyperparameter Name in Scikit-Learn** | **Hyperparameter Function** | **Range of Potential Values** |
| n\_estimators |  | 10 to 300 |
| max\_depth |  | 3 to 25 |
| min\_samples\_split |  | 2 to 10 |
| bootstrap |  | True or False |
| max\_samples |  | 0.01 to 1.0 |

n\_trees = trial.suggest\_int("n\_estimators", 10, 300)

max\_depth = trial.suggest\_int("max\_depth", 3, 25)

min\_sample\_split = trial.suggest\_int("min\_samples\_split", 2, 10)

bootstrapping = trial.suggest\_categorical("bootstrap", [True, False])

if bootstrapping == False:

max\_samples = None

else:

max\_samples = trial.suggest\_float("max\_samples", 0.01, 1.0)

**Table 5:** Hyperparameter Tuning for XGBoost’s XGBRegressor

|  |  |  |
| --- | --- | --- |
| **Hyperparameter Name in XGBoost** | **Hyperparameter Function** | **Range of Potential Values** |
| n\_estimators |  | 10 to 300 |
| max\_depth |  | 3 to 25 |
| learning\_rate |  | 0 to 1 |
| reg\_alpha |  | 0 to 0.001 |
| reg\_lambda |  | 0 to 0.001 |
| booster |  | ‘gbtree’ or ‘dart’ |
| sampling\_method |  | Always set to ‘uniform’ |
| subsample |  | 0.1 to q |

**Table 6:** Hyperparameter Tuning for LightGBM’s LGBMRegressor

|  |  |  |
| --- | --- | --- |
| **Hyperparameter Name in LightGBM** | **Hyperparameter Function** | **Range of Potential Values** |
| n\_estimators |  | 10 to 300 |
| max\_depth |  | 3 to 25 |
| learning\_rate |  | 0 to 1 |
| reg\_alpha |  | 0 to 0.001 |
| reg\_lambda |  | 0 to 0.001 |
| boosting |  | ‘gbdt’ or ‘dart’ |
| bagging\_fraction |  | 0.1 to 1.0 |
| bagging\_freq |  | 0 to 10 |

- <https://optuna.org/>

### 4.33 Comparison of Cross-Validation Folds? Model Evaluation

The fine-tuned model was then evaluated on the test data, with its performance measured by its mean squared error, root mean square error, mean absolute error, R2 score, and a symmetric version of the mean absolute percentage score. Each evaluated metric was averaged across the folds? same type of fine-tuned model (e.g. xgboost)

### 4.34 Comparison with Modelled MMR Values

Whether the model can behave out of sample (other income levels via comparison)

Figure 2 demonstrates the difference between the national and modelled estimates, with the latter having a much lower proportion of missing data per year (8-10% versus 60-100%). As explained in the background information, this is likely due to the challenges involved in collecting national MMR estimates, especially for lower income countries without vital registration systems. This is highlighted by the difference between the median modelled MMR estimate (86) and median national MMR estimate (19), as the lower median national estimate reflects how the majority of data informing the national estimate is obtained in higher income countries, which tend to have lower MMR estimates. Merging these estimates will be discussed in more detail in section 4.24, below.

### 4.35 Sensitivity Analysis

# 5. Results

# 6. Discussion

# 7. Conclusion

# 8. References

# 9. Appendix

### 9.1 Features From Each Data Source