The Australian National University
2600 ACT | Canberra | Australia

Australian National University

School of Computing

College of Systems and Society

Using Machine Learning Methods to Estimate Maternal Mortality Ratios

— Honours project (S1/S2 2025)

A thesis submitted for the degree
Bachelor of Philosophy (Honours) - Science

By:
Rosalita Rosenberg

Supervisors:
Dr. Minh Bui
Dr. Nhung Nghiem

October 2025

Declaration:

I declare that this work:

- upholds the principles of academic integrity, as defined in the Academic Integrity Rule;
- is original, except where collaboration (for example group work) has been authorised in writing by the course convener in the class summary and/or LMS course site;
- is produced for the purposes of this assessment task and has not been submitted for assessment in any other context, except where authorised in writing by the course convener;
- gives appropriate acknowledgement of the ideas, scholarship and intellectual property of others insofar as these have been used;
- in no part involves copying, cheating, collusion, fabrication, plagiarism or recycling.

I acknowledge that I am expected to have undertaken Academic Integrity training through the Epigeum Academic Integrity modules prior to submitting an assessment, and so acknowledge that ignorance of the rules around academic integrity cannot be an excuse for any breach.

October (2025), Rosalita Rosenberg

Acknowledgements

I would like to thank Dr. Minh Bui and Dr. Nhung Nghiem for their invaluable support throughout this project. I would also like to thank Dr. Trong Nhan Ly for his help in using the NCI's GADI supercomputer.

Abstract

Maternal mortality is an ongoing, critical international health challenge, especially in low- and middle-income countries. In 2023, the global maternal mortality ratio (MMR), or the number of maternal deaths per 100,000 live births, was 197. Approximately 95% of maternal deaths occur in low and lower-middle income countries and fragile settings. Many of these deaths could have been prevented by using existing, effective interventions. The World Health Organisation has highlighted how low-quality, sparse data about maternal mortality hinders effective intervention, especially as countries with the highest MMRs tend to have the most missing data. The primary contribution of my thesis to the literature was its development of decision-tree based machine learning models to predict the MMRs of 172 countries between 1985 and 2018 using data from the World Health Organisation and World Bank. In contrast to existing approaches, my proposed models estimated MMR without needing to make potentially invalid assumptions about the underlying distribution of data. My models also used a wider range of socio-economic and health-related features than existing methods. This produced an alternative set of MMR estimates that can be used provide consensus about the true MMR values as well as encourage debate about the validity of different MMR modelling techniques. The best-performing Random Forest Stacking Ensemble achieved a test mean relative error of 0.07 when predicting all MMR data for a specific country and a test mean relative error of 0.37 when forecasting MMR values. Despite being limited by low-quality and sparse input data, my models' MMR predictions were similar to those produced by the most widely used models in the literature, reinforcing their validity. The socio-economic and health-related variables with highest predictive power for MMR in my models were established risk factors. My analysis highlighted the importance targeting socio-economic drivers of maternal mortality, such as women's employment prospects, to successfully reduce maternal deaths.

Table of Contents

Chapter 1

Introduction

1.1 Problem Motivation

The United Nations and other international organisations have recognised that high rates of death due to complications from pregnancy and childbirth is an ongoing, critical global health challenge (World Health Organization, 2025). As a result, they have set numerous goals and resolutions to encourage countries to take substantive action to reduce maternal mortality. Despite the number of maternal deaths decreasing by 40% between 2000 and 2023, maternal mortality remains unacceptably high (World Health Organization, 2025). In 2023, one woman was estimated to die from complications due to pregnancy and childbirth every two minutes. Many of these deaths were avoidable, with almost 3 million women predicted to have died from preventable, maternity-related causes between 2010 and 2020 (Souza et al., 2023). The vast majority of deaths attributed to complications from pregnancy and childbirth occur in low and lower-middle income countries due to substantial country-level inequities (Cresswell et al., 2025). For example, a woman in Australia or New Zealand is 400 times less likely to die from giving birth than a woman in sub-Saharan Africa (World Health Organization, 2025).

Further reduction in the rate of maternal mortality has stalled, with only two regions (central and south Asia, and Australia and New Zealand) showing continued decrease in the rate of maternal mortality between 2016 and 2023 (Souza et al., 2023). All other regions either experienced no change or increases in the rate of maternal mortality. Researchers and international organisations have emphasised how sparse, low-quality data about maternal mortality has hindered effective interventions, as maternal mortality is often substantially underestimated in official statistics (World Health Organization, 2025; Cresswell et al., 2025; Koblinsky et al., 2016; Peterson et al., 2022, 2024; Ahmed et al., 2023). As a result, in 2015, the World Health Organisation highlighted the need to improve measurement of maternal mortality in its Strategies toward Ending Maternal

Mortality report (Cresswell et al., 2025; Ahmed et al., 2023).

1.2 Existing Maternal Mortality Models

To address this data gap, the United Nation's Maternal Mortality Inter-Agency Group and the Institute of Health Metrics and Evaluation formulated models that estimate maternal mortality ratios (MMR), or the number of maternal deaths per 100,000 live births, on a global scale (World Health Organization, 2025; Alkema et al., 2017; GBD 2021 Causes of Death Collaborators, 2024). These models use classical machine learning techniques that are heavily informed by statistics (Alkema et al., 2017). In contrast, the more recently published Global Maternal Health Microsimulation model estimates MMR by simulating the reproductive lifecycles of thousands of individual women (Ward et al., 2023a). These models were developed using domain specific knowledge (World Health Organization, 2025; Alkema et al., 2017; GBD 2021 Causes of Death Collaborators, 2024; Ward et al., 2023a).

All three of these models make assumptions about the underlying data distribution, which may bias their MMR predictions (World Health Organization, 2025; Alkema et al., 2017; GBD 2021 Causes of Death Collaborators, 2024; Ward et al., 2023a). For example, the models assume a certain degree of regional homogeneity, especially when estimating the MMR of countries with sparse data. Additionally, the most widely used models only consider a small subset of variables that impact maternal mortality and assume that the relationships between their chosen covariates and maternal mortality are globally applicable (World Health Organization, 2025; Alkema et al., 2017; GBD 2021 Causes of Death Collaborators, 2024). These assumptions may reduce the accuracy of their MMR predictions.

Due to their differing methodologies, the three models sometimes produce dissimilar MMR estimates (Ward et al., 2025). At times, their MMR estimates differ by hundreds of deaths per 100,000 live births (Ward et al., 2025). These contradictory results can cause confusion about which set of estimates to use and reduce trust in the modelling process, hampering policy makers' ability to effectively use the models' MMR predictions (AbouZahr, 2011).

This motivates the central question addressed in my thesis: *"Can an alternative modelling technique that does not make assumptions about the underlying data distribution and that considers a wide variety of socio-economic and health-related variables be used to estimate maternal mortality ratios?"*

1.3 Contributions of My Research to the Literature

This question motivated the primary aims of my research:

1. To use interpretable machine learning methods to estimate countries' maternal mortality ratios and assist in global MMR monitoring.
2. To identify important socio-economic and health-related features to inform targeted policies that will most reduce MMR.

To address these aims, I developed, tested, and compared a series of alternative machine learning models to estimate MMR. I based all proposed models on the decision-tree architecture because decision trees can effectively handle high-dimensional, sparse data without making assumptions about the underlying data distribution (Costa and Pedreira, 2023). These properties were essential given the high proportion of missing data and large number of feature variables in my data. While deep learning methods also have strong performance on high-dimensional data, they cannot natively handle sparse data and I did not want to risk introducing bias into my data by imputing the missing values (LeCun et al., 2015; Kia et al., 2022). Therefore, no deep learning was used in this research.

The main contributions of my research were that:

- **I developed decision-tree based Random Forest, XGBoost, and LightGBM models that can effectively deal with sparse data to estimate and forecast MMR.** I found that the specific training data used to fit the models had a greater impact on their predictive accuracy than the choice of model type, feature selection strategy, or proportion of missing data included in the dataset.

- **I used stacking and voting ensemble methods to combine predictions from 300 Random Forest, LightGBM, and XGBoost models fit on different training data to further improve predictive accuracy.** The best-performing ensemble leveraged patterns learned by each component model on the various training datasets. I found that the Random Forest Stacking Ensemble had the highest overall predictive performance, with higher performance gains observed when the performance of the models being combined was less uniform.

- **I examined the performance of the best-performing ensemble when it was trained on data from all income levels versus a specific income level.** Estimates of past MMR values were more accurate when informed by trends across all income levels while MMR forecasts were more accurate when based on income-specific data. Generally, the lowest mean-squared error was achieved when predicting the MMR of higher-income countries.

- **I benchmarked my models' MMR predictions against estimates from existing maternal mortality models in the literature.** While my predictions were broadly similar to the literature's estimates, they tended to predict lower

MMR values due to methodological differences, variation in model variables, and possible underestimation of MMR in my ground truth data.

- **I designed the Python code** used to implement and evaluate these models. The code is **freely available on GitHub** at https://github.com/R0sle/health_economics_honours. Model training and evaluation was performed on the National Computational Infrastructure's **Gadi Supercomputer**.

- **I determined the socio-economic and health-related features with the highest predictive power for MMR**, many of which were established risk factors. I used these results and existing causal research to suggest that investment in women's education, incentives for skilled medical personnel to practice in rural areas, and increased provision of family planning services would reduce MMR by addressing important drivers of maternal mortality.

- Using my models, **I provided alternative MMR estimates for 172 countries between 1985 and 2018**. These estimates can be used to **resolve existing disagreement** about the true maternal mortality ratios and **inform scientific debate** about the relative merits of different MMR modelling approaches.

- **I showed that my models achieved comparable MMR predictive accuracy to existing models in the literature without a similarly heavy reliance on domain knowledge.** Therefore, my models have wider applicability in low resource countries where domain knowledge in this field is still developing.

Chapter 2

Background

2.1 Maternal Mortality

In 2015, the United Nations committed to achieving 17 Sustainable Development Goals by 2030 to fuel progress toward eliminating global poverty and protecting the planet (Miranda et al., 2023). Specific Sustainable Development Goals outline important targets for improving global health and environmental outcomes as well as reducing inequality and conflict. Progress toward the Sustainable Development Goals is monitored by a panel of independent scientists. In 2023, this panel issued warnings that the international community would fail to meet many of the Sustainable Development Goals, as progress has stalled and, for some countries and goals, regressed (Miranda et al., 2023). The panel attributed this to a mixture of factors, including limited government resources devoted toward the goals, lack of available data for monitoring the goals, and unequal global distribution of infrastructure and innovation. The effects of these trends combine with, and amplify, crises like the COVID-19 pandemic to further hinder progress.

The report emphasised the lack of progress toward maternal and child mortality goals (Miranda et al., 2023). In response, in 2024, the 77th World Health Assembly passed an additional resolution to increase progress toward decreasing maternal mortality (World Health Organization, 2025). This resolution targeted Sustainable Development Goal 3.1, which aims to reduce the global maternal mortality ratio (MMR), or the number of maternal deaths per 100,000 live births, to below 70 by 2030, with no single country having an MMR of greater than 140. In this context, a maternal death is defined by the International Classification of Diseases as (World Health Organization, 2022):

> *"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."*

International concerns about trends in maternal mortality were driven by recent MMR estimates (World Health Organization, 2025). More specifically, in 2023, the global MMR was 197 deaths per 100,000 live births (uncertainty interval 174 to 234), notably higher than the Sustainable Development Goal's target of 70 (World Health Organization, 2025). Concerningly, substantial country-level inequity means that many countries have even higher national MMRs, as approximately 95% of maternal deaths occur in low and lower-middle income countries and fragile settings (Cresswell et al., 2025). For example, in 2023, Nigeria had an MMR of 993 (uncertainty interval 718 to 1540) while Australia had an MMR of 3 (uncertainty interval 2 to 4) (World Health Organization, 2025). As a result of this inequality, only a small subset of countries is projected to meet Sustainable Development Goal 3.1 (Cresswell et al., 2025).

The leading global cause of maternal deaths between 2009 and 2020 was haemorrhage, which refers to a large loss of blood due to excessive internal or external bleeding (Cresswell et al., 2025). Studies estimate it caused 27% of maternal deaths globally, with a disproportionate incidence in lower income countries. Effective haemorrhage treatments exist, meaning that many of these deaths were preventable. Indirect obstetric deaths, or deaths due to a condition tangential to pregnancy that was aggravated by the pregnancy, caused 23% of global maternal deaths between 2009 and 2020. The second and third most common causes of death during this time period were hypertensive disorders (16% of deaths), abortion (8%) and pregnancy-related infection (7%). Experts predict that, over time, MMRs will decrease and the majority of maternal deaths will be caused by indirect, non-communicable conditions instead of direct complications of pregnancy and childbirth (Souza et al., 2014). A country's position within this 'obstetric transition' has important implications for the choice of strategies used to reduce its MMR.

2.2 Monitoring Maternal Mortality

Reports published by both the World Health Organisation (WHO) and academic researchers highlight how lack of access to accurate, complete data about maternal mortality hinders effective interventions (World Health Organization, 2025; Cresswell et al., 2025; Ahmed et al., 2023). The missing data would help policymakers identify regions with high burden of maternal deaths as well as possible region-specific causes of maternal mortality (World Health Organization, 2025; Ahmed et al., 2023). This would allow them to implement timely, targeted, and useful programs to reduce maternal mortality. However, data collected about maternal deaths is known to substantially underestimate true maternal mortality due to a mixture of underreporting and misclassification of maternal deaths (Mgawadere et al., 2017; Ahmed et al., 2023; Peterson et al., 2024).

MMR is estimated from one or more of a diverse range of data sources, with a large sample size and/or complete records needed for stable MMR estimates given the relative rarity of maternal deaths (Peterson et al., 2024). Where possible, MMR estimates are informed by civil registration and vital statistics (CRVS) systems, which are national data collection systems that continuously record births and medically certified deaths.

Cause of death is recorded in line with the International statistical classification of diseases and related health problems. Thus, CRVS systems generate vital information for mortality monitoring and policy development, as in a perfect world they record all deaths in a country with their associated causes (Mgawadere et al., 2017; Peterson et al., 2022). However, in 2017, less than 40% of countries had CRVS systems that enabled continuous and accurate maternal mortality monitoring (Mgawadere et al., 2017). Unfortunately, this prevents monitoring of trends in maternal mortality, especially in the lowest income countries that have the highest MMR burdens, as they tend to have the most missing data (Cresswell et al., 2025; Peterson et al., 2024). For example, in 2017, only 2 of the 49 least developed countries had greater than 50% death registration coverage (Mgawadere et al., 2017).

Even when CRVS systems are in place, they are limited by their national coverage and can be subject to a myriad of underreporting and misclassification errors, reducing the quality of the reported data (Cresswell et al., 2025; Mgawadere et al., 2017; Ahmed et al., 2023; Peterson et al., 2024). More specifically, underreporting occurs when a maternal death is not registered, while misclassification occurs when the incorrect cause of death is recorded (Peterson et al., 2022). While maternal mortality is underreported at all stages of pregnancy, it is more frequent at the earliest phases when signs of pregnancy may be missed (Ahmed et al., 2023; Peterson et al., 2024). Underreporting also increases when the maternal death occurs at home or when it occurs as a result of abortion or extramarital pregnancy due to social stigma or legal barriers (Ahmed et al., 2023). Maternal mortality is also often misclassified due to the complexity of isolating the exact cause of death, especially when the death is caused by an underlying health condition (Ahmed et al., 2023; Peterson et al., 2024). Due to misclassification and underreporting, studies predict that maternal mortality is underestimated by at least 40%, with large differences between countries (Ahmed et al., 2023). Thus, reliability of CRVS data must be confirmed before use (World Health Organization, 2025).

MMR estimates can also be informed by specialised studies, which determine the MMR within a specific geographic region using police and medical records, national registries, administrative reviews, medical autopsies, and censuses (World Health Organization, 2025). They are often considered the gold-standard.

In addition to CRVS systems and specialised studies, MMR estimates are informed by broader national and household surveys, censuses, national surveillance data, and data collected from health providers (World Health Organization, 2025; Mgawadere et al., 2017). These sources are particularly useful in low and middle-income countries that lack CRVS systems (Ahmed et al., 2023). Unfortunately, surveys may not provide adequate coverage, especially of rural areas that are difficult and/or expensive to reach. Additionally, the relative rarity of maternal mortality means these surveys require a large sample size to be statistically significant, which can make them prohibitively expensive to conduct (Mgawadere et al., 2017). Alternatively, maternal deaths can be monitored using surveys based on the sisterhood method, where adult respondents detail how many of their sisters have died from a pregnancy-related cause. This is the WHO recommended

method for countries without other reliable sources of data, as asking respondents about the health of others immediately increases sample size. However, the survey does not provide current data for monitoring purposes.

As a result of these limitations, maternal mortality data can be sparse and low-quality, motivating use of modelling techniques that can use global data to fill in the gaps.

2.3 Machine Learning

In the past few decades, improvements in communication, data storage, and processing power, such as through the development of the internet of things and data centres, have allowed large quantities of data to be collected and shared at scale (Jordan and Mitchell, 2015). For the first time, researchers can analyse massive datasets from a wide variety of sources, such as health records like those discussed above. As a result, researchers have the opportunity to identify complex, insightful, data-driven patterns (Jordan and Mitchell, 2015; Costa and Pedreira, 2023). Increasingly, researchers are detecting and then analysing these patterns using machine learning (ML), where they train models to learn relationships within the data (Jordan and Mitchell, 2015; Greener et al., 2022). This approach differs from the traditional strategy of designing the model using hand-crafted rules that are informed by prior domain knowledge. ML is particularly useful when applied to datasets with many datapoints and/or variables, as the technique can find hidden patterns that may be missed by humans (Greener et al., 2022). ML models can then take these patterns and use them to make predictions in the absence of empirical data. Thus, it could be a useful technique to employ when working with missing epidemiological data (Zuhair et al., 2024). In this section, I overview the main machine learning methods, giving particular attention to the decision-tree based ML techniques used in this research.

Conventionally, the input dataset to an ML model consists of a number of samples/ observations, where each sample is referred to as a 'datapoint' (Greener et al., 2022). Each datapoint is defined by a certain number of variables, which are referred to as 'features'. Features with discrete values are called 'categorical' (or discrete) and features with continuous numerical values are called 'continuous'. For example, if a feature describes 'risk' and its values were 'high', 'medium', or 'low', it would be considered categorical. In contrast, if its values were a risk score between 0 and 5, it would be continuous. Generally, each row of the input dataset corresponds to a datapoint, and each feature corresponds to a column.

2.3.1 Types of Machine Learning

ML models can be broadly classified as supervised or unsupervised methods depending on whether the rows of the input dataset are associated with an output value (Greener et al., 2022).

Unsupervised Learning

Unsupervised learning models act on input datasets whose datapoints are not associated with a specific categorical or continuous output value (Greener et al., 2022). For example, datapoints in unsupervised learning may consist of a series of observations about feature variables 'temperature', 'day of the week', and 'location'. However, the observations would not be associated with an output variable, like 'quantity of ice cream sold'. The aim of unsupervised learning is to uncover hidden patterns and learn the data's structure. By not providing output values, the model is not explicitly guided toward learning a specific type of pattern in the data.

A common application of unsupervised learning is clustering, where the model learns relationships in the data that allow it to group similar datapoints together (Greener et al., 2022). For example, clustering can be used to place patients with similar attributes in the same group. Another popular use of unsupervised learning is dimensionality reduction, which transforms a dataset with many variables into a dataset with fewer variables while retaining as much of the data's original variation as possible. The transformed dataset may contain linear and non-linear transformations of the original variables. One widely used dimensionality technique is called principal component analysis and is often applied to be able to represent a dataset with many variables using only two variables, making it easier to visualise patterns in the data.

Supervised Learning

In contrast, supervised learning occurs when a model is fit to a labelled dataset, where each input datapoint is associated with one or more output categories or values (Greener et al., 2022). The true values of the output variables are referred to as the 'ground truth', which the model is trained to predict by learning relationships between input datapoints and the output values. Supervised learning can be applied to classification problems, where the ground truth is two or more specific categories, or regression problems, where the ground truth is a continuous numerical output. This thesis will focus on supervised machine learning for regression analysis, as models can be trained on data labelled with ground truth MMR values, which are continuous.

More formally, a dataset containing n samples is denoted as $D = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$ refers to a d-dimensional input feature vector and $y_i \in \mathbb{R}$ refers to the corresponding continuous, numeric output value (Terven et al., 2025). When solving a regression problem, the model's goal is to learn a mapping between the input data and the associated ground truth, $f : \mathbb{R}^d \to \mathbb{R}$ (Jordan and Mitchell, 2015). For a new input datapoint, x^*, the model uses the learned function to predict the associated output, $\hat{y} = f(x^*)$. The type of mapping used defines the ML model being implemented. At its core, this mapping is a mathematical function defined by a series of parameters, where the function takes the dataset as input and gives its prediction as the output (Greener et al., 2022). To produce accurate predictions, the mapping must approximate the true, underlying relationships between features and the variable being predicted.

2.3.2 Model Development

Model performance depends on whether the model's parameters are well-suited to the model's purpose and dataset Greener et al. (2022). Model performance is defined by a loss function, which quantifies the difference between the model's predictions and the ground truth. The process of optimising the model's parameters involves minimising this loss function, which commonly involves a technique called gradient descent (Terven et al., 2025; LeCun et al., 2015). Intuitively, gradient descent takes advantage of the observation that the gradient quantifies the direction of greatest increase. Thus, taking the negative gradient of the loss function with respect to each parameter gives the direction that the parameter's value would need to move to produce the greatest decrease in loss (LeCun et al., 2015). As a result, to minimise the loss function, L, the gradient of the loss with respect to a specific parameter, θ, can be subtracted from the parameter's current value, as shown in (Eq.1) (Baldi, 1995). The symbol η is the learning rate, which determines the degree to which the negative gradient is used to adjust the parameters' value. This gradient descent algorithm is applied to all model parameters to minimise the model's loss function through optimising its parameter values. There are many different implementations of gradient descent, such as sample gradient descent, which calculates the gradient using a subset of the dataset to reduce computational complexity (Ganie and Dadvandipour, 2023).

$$\theta_{\text{new}} = \theta_{\text{current}} - \eta \frac{\partial \mathcal{L}}{\partial \theta_{\text{current}}} \tag{Eq.1}$$

Model development must be done with care, as the model's parameters are optimised with respect to a specific input dataset (Greener et al., 2022). This can produce overfitting, where the model has high performance on the input dataset but low performance on out-of-sample data. Overfitting can occur due to noise in the input dataset, where the model learns the noise as a true pattern in the data. This prevents the model from learning the true, underlying patterns in the data that would allow it to extrapolate to out-of-sample data, which may have a different noise pattern. Generally, more complex models have a higher risk of overfitting, as they have more parameters that can be configured to the exact, noisy patterns in the input dataset. The risk of overfitting must be balanced with the risk of underfitting, which occurs when the model is too simple to accurately capture the underlying relationships in the data (Greener et al., 2022). Overfitting and underfitting are related to the bias-variance trade-off, where bias refers to errors in the model's predictions while variance refers to change in the model's predictions based on the training data used. The goal of model development is to produce a model with low bias and low variance. However, to reduce bias, the model generally must become more complex, which can cause overfitting and increase variance, necessitating a trade-off.

To balance the goal of low bias while avoiding overfitting, the dataset is split into non-overlapping training and testing subsets, generally in a ratio between 75:25 and 90:10 (Greener et al., 2022). The model's parameters are fit to the training dataset through

minimising the loss function. Then, the model's performance is evaluated on the previously unseen test data to determine whether the model is generalisable or is overfit to the training data.

However, the model should not be adjusted based on its test performance to prevent overfitting to the test data, which would prevent the test set from being able to measure out-of-sample performance (Greener et al., 2022). This is a problem when using the test set to compare the performance of different hyperparameter specifications, where hyperparameters govern the architecture of a model and the training process, but are not themselves fine-tuned during training. They define the structure of the mapping used by the model, not the mapping itself. For example, the learning rate η, or the rate at which parameter values are changed during training, is a hyperparameter. To address this problem, the training data can be further split into non-overlapping training, validation subsets (Greener et al., 2022). Model parameters are fit using the training data, and different model architectures and hyperparameter specifications are tested on the previously unseen validation set. The ability of the best performing model to generalise to out-of-sample data is then evaluated using the unseen test set. Thus, the model's performance on the test set is often considered a measure of its real-world performance. As a result, the test set should only be used once.

Training data is often split into training, validation subsets through a process called K-fold cross-validation (Greener et al., 2022). In this process, the training data is split into $K-1$ equally sized, non-overlapping subsets. For each of K iterations, the training data consists of $K-1$ folds while the validation data consists of the single, remaining fold. One version of the model is trained per iteration on the $K-1$ training folds, with its performance tested on the validation fold. By having K iterations, each individual fold has a turn to be the validation fold, testing the model's ability to generalise on all parts of the training-validation set. The performance of the K models (one per iteration) is then compared and/or combined.

Loss Functions and Parameter Tuning

The model's performance is measured using a loss function, as described above (Terven et al., 2025). There are a variety of possible loss functions that can be used for regression problems, with one of the most common being the mean squared error (MSE), or L2 loss. The MSE is the averaged squared difference between the ground truth output, y_i, and the model's predicted output, $\hat{y}_i$, across n datapoints. The MSE is defined in (Eq.2). A limitation of MSE is its sensitivity to outliers, as squaring the difference between the true and predicted outputs places high importance on large errors.

$$MSE = \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) \tag{Eq.2}$$

A widely used variation of MSE is mean absolute error (MAE), also referred to as the L1 loss (Terven et al., 2025). MAE measures the average absolute difference between the

true and predicted outputs, and is defined more formally in (Eq.3). Taking the absolute difference instead of the squared difference means MAE is less affected by outliers than MSE. However, unlike MSE, MAE is not differentiable everywhere due to the absolute value, presenting difficulties when using gradient based optimisation techniques.

$$MAE = \frac{1}{n}\left(\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid\right) \tag{Eq.3}$$

Another common variation of MSE is root mean square error (RMSE), or the square root of the MSE (Terven et al., 2025). The RMSE is defined formally in (Eq.4). Like MSE, the squared function in RMSE heavily penalises outliers. However, unlike MSE but similar to MAE, RMSE is in the scale of the original data, making it easier to interpret.

$$RMSE = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)} \tag{Eq.4}$$

An alternative, widely used metric is the mean absolute percentage error (MAPE), which calculates the average prediction error as a percentage of the ground truth value (Terven et al., 2025). It is defined in (Eq.5). MAPE is criticised for being asymmetrical, as always dividing by the true output, y_i, can produce different errors depending on whether the predicted value underestimates or overestimates the true value. For example, predicting a value of 50 if the true value is 100 gives a MAPE of 50% while predicting a value of 100 if the true value is 50 gives a MAPE of 100%. Thus, the same absolute error produces different MAPE scores depending on whether the under- or over-estimate is used as the denominator. Another limitation of using MAPE is that it can become very large or undefined if y_i is close to zero (Terven et al., 2025).

$$MAPE = \frac{1}{n}\left(\sum_{i=1}^{n} \mid \frac{y_i - \hat{y}_i}{y_i} \mid *100\%\right) \tag{Eq.5}$$

The coefficient of determination, also called the R^2 score, is another commonly used performance metric (Terven et al., 2025). It determines the proportion of variation in the output variable explained by the model. R^2 is defined in (Eq.6), where $\bar{y}$ is the mean true value. R^2 is equal to 1 if the model explains all variation in the output. However, a high R^2 score can sometimes reflect overfitting in the model. R^2 is negative if the model performs more poorly than if it simply predicted $\bar{y}$. Unfortuantely, the R^2 is known to be sensitive to bias and can arbitrarily increase with the number of features.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{Eq.6}$$

2.4 Machine Learning Models

As described above, the type of mapping from the input feature data to the output defines the machine learning model.

Linear Regression

One of the most well-known, basic machine learning models is linear regression, which is often described as a 'line of best fit' through the data (Zou and Hastie, 2005). Model development focuses on minimising the distance between the true values and the line produced by the model's predictions (Greener et al., 2022). More formally, linear regression is often used to predict output $\hat{y}$ using a linear combination of d-dimensional input feature vectors, $x \in \mathbb{R}^d$(Zou and Hastie, 2005). The model is described in (Eq.7), with the d-dimensional weights denoted by $\theta \in \mathbb{R}^d$.

$$\hat{y} = \left(\sum_{d=1}^{D} \theta_d * x_d \right) \tag{Eq.7}$$

One of the symptoms of overfitting is large parameter weights on feature dimensions, as this signals that the model has found a complex pattern in the dataset, which is more likely to be noise and thus less generalisable (Zou and Hastie, 2005). As a result, many linear regression implementations incorporate a regularisation term, which is added to the loss function to penalise model complexity. More specifically, the regulariser increases the loss by some function of the model's parameters (Greener et al., 2022). To minimise loss, training generally involves actions to reduce the regularisation term and thus prevent the parameter values from becoming too large. The L1 norm, or the sum of the parameters' absolute values, is a commonly used regularisation function (Zou and Hastie, 2005). By penalising parameters' absolute values, it encourages the model to use zero feature weights, thus performing automatic feature selection. Another widely used regularisation function is the L2 norm, which is the sum of the squared parameter values, and thus severely penalises large parameter values. Elastic Net is a special version of the linear regression model that combines the L1 and L2 norms. Elastic Net model's regularisation term is defined in (Eq.8), where α is a hyperparameter that controls the influence of the L1 versus L2 norm.

$$regulariser = (1 - \alpha) * L_1 + \alpha * L_2 \tag{Eq.8}$$

Linear regression is solely linear in the parameters, meaning the feature variables do not need to be linear (Greener et al., 2022). However, the model can still underfit if the relationship between feature variables is non-linear. Thus, more complex models have been developed.

Support Vector Machines

Support vector regression fits a model based on the most informative data points (Smola and Schölkopf, 2004). More specifically, only predictions that were incorrect by at least epsilon contribute to the model's loss during training. Epsilon is a hyperparameter that defines the model's error tolerance. Data points associated with a predictive error of at least epsilon are referred to as "support vectors". This procedure allows the model to focus on correcting larger errors. Model predictions are generated from a linear combination of support vectors to be able to capture the most complex relationships in the data. Often, input data is transformed into a higher dimensionality feature space to more effectively model non-linear relationships (Smola and Schölkopf, 2004).

Decision Tree Based Methods

Since their original proposal in the 1960s, decision trees have become an important part of the most widely used ML models (Costa and Pedreira, 2023). Intuitively, decision tree models function like flowcharts. A regression decision tree is visualised below 2.1, with the tree's internal nodes given by circles and its terminal nodes given by squares. When predicting the output value for a specific datapoint, the model starts at the root node and applies a logical test to the values of one or more feature dimensions. For regression, this test is usually in the form $feature <= value$, and defines a split (Costa and Pedreira, 2023). Based on the test's Boolean result, the model moves to the right or left child node. This process repeats until the model reaches a terminal leaf node, which is a node with no children. The terminal node's value determines the model's prediction. An alternative way to conceptualise decision trees is as a specific partitioning of the input space, where each node partitions the feature space and each new partition is passed down to the node's children. The tree's prediction then corresponds to a specific area of the feature input space.

The decision tree's structure is developed during training, where the logical tests that best predict the outcome variable are chosen (Costa and Pedreira, 2023). For regression problems, the logical test at each node is determined through finding the split that minimises the mean squared error in the associated child nodes (Loh, 2014). Traditionally, this is done in greedily, where specific splits are evaluated solely by their effect on their children's error (Costa and Pedreira, 2023).

One of the primary advantages of decision trees is their ability to work with data that has missing values, with specific implementations having different methods (Costa and Pedreira, 2023; Loh, 2014). CART (Classification and Regression Trees) is one of the classic decision tree implementations and uses 'surrogate' splits to deal with missing data. When a datapoint is missing a value in a specific feature dimension, nodes that partition the input space using that feature instead split using an alternate, related variable (Loh, 2014).

Another advantage of decision trees is their interpretability, which is due to their flowchart-like structure (Costa and Pedreira, 2023). As a result, they are valued in disciplines

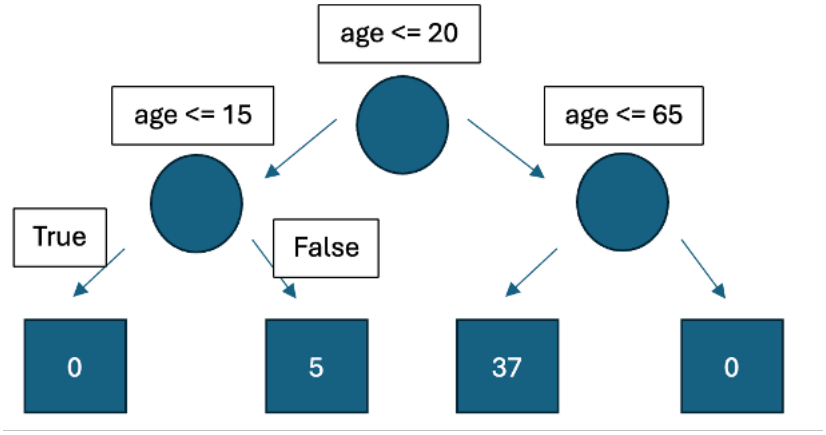


Figure 2.1: Regression decision tree visualisation, where splits are defined in terms of the feature 'age' and the model is trying to predict the number of hours worked per week. If the result of the test is True, the model moves to the left child node, but if it is False, it moves to the right child node. The values inside the terminal nodes are the predictions.

that place more emphasis on understanding why ML models have made a specific prediction, such as in drug development. A further benefit of decision trees is their relatively low computation cost when compared to other ML models.

However, a major limitation of decision trees is their propensity to overfit, where the input space is partitioned by overly complex rules based on the specific training examples and noise in the training data (Costa and Pedreira, 2023). As a result, shallower trees with fewer partitions tend to generalise better, but may have lower performance due to their lower complexity.

There are many variations of decision trees (Costa and Pedreira, 2023). For example, to better represent more complex functions, studies have explored basing the splits in internal nodes on multiple feature variables and/or having predictive models in the terminal nodes instead of a constant. Additionally, research has explored replacing the greedy approach used to determine splits with look-ahead algorithms to avoid suboptimality.

Deep Learning

As detailed above, models based on the decision tree architecture make predictions by working directly with the features in the given input data, which are often handpicked through feature engineering. In contrast, deep learning (DL) models are trained to transform the given features into representations of the input data that most effectively enable regression and classification (LeCun et al., 2015). This is called representation learning.

As part of representation learning, DL models have a hierarchy of representations, with the first layer containing input data. At each layer, the model uses non-linear transformations to combine and transform the representation from the previous layer into a more useful representation of the input (LeCun et al., 2015). For example, in deep learning model used to interpret images, the early layers often detect edges and corners in the image, while later layers combine these early representations to be able to identify shapes. By learning increasingly nuanced representations of the (potentially high-dimensional) input data, DL models can approximate complex functions.

These stacked, hierarchical layers are referred to as a "neural network" (Serghiou and Rough, 2023). The first layer contains the raw, input data and the final layer contains the model's predictions. The layers in between are called 'hidden layers' and contain processing units called neurons". There are a wide variety of architectures used for supervised deep learning, but generally data from the previous layer (either the raw data or a previous representation of the data) input to each neuron is combined in a weighted sum. A bias term is typically added to the sum, then a non-linear transformer is applied to the entire sum. The Rectified Linear Unit (ReLU) is a popular non-linear transformation, which maps any negative value to zero but does not affect positive values. This transformed sum is the output of a neuron. Intuitively, the neuron is transforming the previous representation of the data into a more complex representation.

The weights and biases used to define the connections between layers of the network are learned during training (Serghiou and Rough, 2023). DL models can use hundreds of millions of these adjustable parameters to make their final predictions, with training performed using hundreds of millions of examples (LeCun et al., 2015). In deep learning, adjustment to model parameters is performed using backpropagation, where gradient descent is used to determined how the model's loss changes with respect to the parameters in each layer.

Variables like the number of layers in the network and the number of neurons per layer are hyperparameters (Serghiou and Rough, 2023). The larger the number of layers and the more neurons per layer, the more complex the network (Serghiou and Rough, 2023). According to the universal approximation theory, a neural network with sufficient layers and neurons can approximate any continuous mathematical function, making DL techniques extremely useful for modelling complex relationships (Serghiou and Rough, 2023).

How the layers of a neural network are combined defines the architecture of a neural network (Serghiou and Rough, 2023). When every neuron from one layer is connected to every neuron in the following layer, and there is at least one hidden layer, the network is called a "feed-forward neural network". Recurrent neural networks (RNNs) are a variation of the feed-forward neural network (FFNN) often used to process sequential data (LeCun et al., 2015; Serghiou and Rough, 2023). The first component of the sequence is fed into the network, which predicts the next component of the sequence. This prediction is then used as input and fed back into the network to get the prediction

for the next component of the sequence (Serghiou and Rough, 2023). At its most basic form, the recurrent neural network combines the representation of the current component of the sequence with the hidden layer representation of the previous component, which is used to represent the history of the sequence.

Despite the success of different DL architectures, DL models are associated with important limitations. For example, they are limited by their need to be trained on large datasets, which can be less available in domains like epidemiology (Serghiou and Rough, 2023). Additionally, DL models tend to overfit to their training data, as the complexity of their underlying architecture allows them to effectively capture and model noise. Thus, they are less useful in settings where the model is needed to generalise to new settings, such as new epidemiological populations. This complexity also reduces the interpretability of DL models. It can be challenging for humans to understand the feature representations used by the models to make decisions, as these hidden representations are based on non-linear combinations of potentially millions of parameters (LeCun et al., 2015; Serghiou and Rough, 2023).

A final limitation of DL models is their need to be trained on high-quality, complete data, as data is passed through the neural network using matrix operations, which cannot work with missing values (Kia et al., 2022). Therefore, missing data is generally removed before being given to the model, with newer methods exploring how to ignore the missing feature dimensions for specific samples to avoid needing to remove the entire datapoint. Alternatively, the missing data is imputed.

2.4.1 Ensemble Machine Learning Models

Studies have found that combining predictions from multiple models can have better predictive performance than solely using predictions from a single model (Ganaie et al., 2022). This is called 'ensemble learning'. Ensemble methods can reduce generalisation error when the models being combined, called base estimators or weak learners, are independent and diverse (Mahajan et al., 2023). This allows them to cover a wider range of possible outcomes. Additionally, a single model may become stuck in a local optimum, but if each base estimator in an ensemble model starts in a different place and/or has a different formulation or training trajectory, it is unlikely that all base estimators will become stuck in the same local optimum (Ganaie et al., 2022). Ensemble methods also perform well when complex relationships in the data can be approximated better by a combination of base estimators than by a single base estimator.

Ensemble models can generally be categorised as bagging, boosting, voting, or stacking algorithms (Mahajan et al., 2023).

Bagging

During bootstrap aggregation, or 'bagging', predictions from multiple versions of the same type of base estimator are combined (Breiman, 1996). Different versions of the

same base estimator are produced by training each estimator on a bootstrap replicate of the training set. In other words, datapoints are drawn at random and with replacement from the training set to form independent, bootstrapped datasets of the same size. Then, each base estimator is trained on one of the bootstrap replicates of the dataset, producing an ensemble of base learners whose predictions are combined. For regression tasks, the predictions are generally averaged (Breiman, 1996). Bagging works particularly well when models trained on different versions of the training set are substantially different, allowing the ensemble model to cover a wider variety of outcomes. Additionally, bagging can reduce variability and overfitting by cancelling out noise in the dataset (Mahajan et al., 2023). An example of bagging is combining the predictions of multiple decision trees made on separate, bootstrapped versions of the training dataset.

The Random Forest model is a widely used variation of the basic decision tree-based bagging ensemble (Ganaie et al., 2022). In the Random Forest algorithm, each split in the base decision trees is created using a random subset of features. This modification further reduces overfitting by forcing the model to learn patterns in the data based on different combinations of features. However, there is no guarantee that an important feature will be used for splitting, potentially causing important information to be lost (Mahajan et al., 2023).

Boosting

While bagging trains base estimators independently, boosting ensemble methods train base estimators sequentially (Mahajan et al., 2023). During boosting, each base estimator in the sequence tries to correct the errors of the previous estimator, giving the ensemble model higher prediction accuracy and lower bias (Mahajan et al., 2023; Rizkallah, 2025). The base estimators in boosting ensembles are often decision trees due to their empirically demonstrated prediction accuracy (Rizkallah, 2025).

There are many implementations of boosting algorithms, with gradient boosting being one of the most popular (Rizkallah, 2025). In gradient boosting, the first base estimator predicts the output variable. Then, each new base estimator is trained to minimise the current model's loss (Ke et al., 2017). To do so, the new base estimator predicts the negative gradient of the previous estimator's loss function. This negative gradient indicates the direction of greatest decrease in loss. By learning this direction, the new base estimator can move the ensemble's prediction in a direction that most reduces its prediction loss. This is expressed more formally in (Eq.9), where $F_m(x)$ is the mth base estimator in the sequence, $h_m(x)$ is the base estimator trained on the negative gradient of $F_(m-1)(x)$, and $_m$ is the weight attached to $h_m(x)$, quantifying its importance (Rizkallah, 2025). The final prediction from a boosting ensemble is the sum of predictions from its base learners (Chen and Guestrin, 2016).

$$F_m(x) = F_{m-1}(x) + \rho_m * h_m(x) \tag{Eq.9}$$

When the base learners are decision trees, gradient boosting is referred to as the gradient

boosting decision tree algorithm (GBDT) (Rizkallah, 2025). Studies have shown that GBDT is accurate, efficient, and interpretable, precipitating its use in a wide variety of disciplines (Ke et al., 2017). Two of the most common GBDT methods are Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting (LightGBM).

Extreme Gradient Boosting (XGBoost)

The XGBoost algorithm is a high-performance, scalable GBDT method (Rizkallah, 2025; Chen and Guestrin, 2016).

In the base gradient boosting method described above, the m_{th} base estimator, $h_m(x)$, predicts the negative first-order gradient of the $(m-1)_{th}$ base estimator to move the ensemble's predictions in the direction that most reduces loss (Rizkallah, 2025). The XGBoost model takes this a step further. It constructs a 2nd-order Taylor approximation of the current model's loss function using the loss function's first and second-order derivatives (Chen and Guestrin, 2016). This Taylor approximation is minimised to find the optimal leaf node weights, which are the tree's predictions. This method provides a more controlled error correction mechanism, as the second-order derivative indicates how quickly the gradient is changing, guiding how much change should be made in response to the gradient.

When building a new base estimator, the algorithm must decide which feature value to use as the 'value' part of the $feature <= value$ logical tests on the internal nodes (Chen and Guestrin, 2016). When evaluating a candidate logical test, the algorithm separates the input data into two groups – the data that would push the model to the left child and the data that would push it to the right child. The model then calculates the approximated loss function for each group using the current predictions from the previous base estimators. By minimising this loss, it determines the optimal node weight for the children nodes and can determine the potential reduction in loss produced by this specific split.

A major advantage of XGBoost is its ability to work with missing data (Chen and Guestrin, 2016). As described above, the logical tests at a base estimator's internal nodes are determined during training. This process is completed with non-missing data only. Then, the model determines the 'default direction' for each internal node. This is the direction taken when the feature dimension used in the node's logical test has a missing value. The default direction is set to left or right, depending on whether moving to the left or right child node produced lower predictive error during training. As an aside, this is the same method used to handle missing data in the Random Forest model (Pedregosa et al., 2011).

Unlike the base GBDT algorithm, the loss function used in the XGBoost model has an additional regularisation term (Chen and Guestrin, 2016). This regulariser is a function of the number of leaves in the base estimator's decision tree and the squared absolute values of the leaf node scores. Adding this regulariser to the loss function increases loss when the number of leaf nodes increase. Consequently, the regularisation term penalises

model complexity, as it encourages the model to have fewer internal nodes and input space partitions to reduce the number of terminal nodes. XGBoost also supports feature subsampling, like the Random Forest model, to further reduce overfitting.

One of the key hurdles to constructing GBDT ensembles is the need to trial all possible feature values in the 'value' part of the $feature <= value$ logical test when determining the optimal structure for the base estimator (Chen and Guestrin, 2016). In the exact greedy algorithm approach, all possible splits for all features must be tested, with split performance quantified by how much it reduces the loss. While this has strong performance, it is computationally demanding, especially when evaluating all possible splits for continuous feature variables and when the input data does not fit into memory. The approximate algorithm was introduced to address this problem. This algorithm splits the distributions of continuous features into percentiles, using the differences between percentiles as candidate split points, thus reducing the number of possible splits needed to be evaluated.

Light Gradient Boosting Machine (LightGBM)

LightGBM is another commonly used GBDT algorithm that also places a strong emphasis on maximising computational efficiency (Ke et al., 2017). One of the main modifications proposed by the LightGBM algorithm is gradient-based one-side sampling (GOSS). The GOSS algorithm reduces the number of samples used to determine internal node splits. Instead of using all data points to determine each split, GOSS uses the most informative data points and samples a subset of less informative points to maintain the same general data distribution. Its choice of samples is derived from the observation that datapoints associated with small gradients offer smaller potential reduction in error and are thus less useful for increasing model performance. Using this observation, GOSS takes all datapoints with gradient greater than a certain threshold and samples randomly among the remaining datapoints with smaller gradients. It uses this subsampled dataset to determine the internal node split, increasing computational efficiency.

Another innovation used in the LightGBM model is exclusive feature bundling (Ke et al., 2017). This approach can be applied to sparse feature spaces, which generally have mutually exclusive feature variables, or groups of features where no more than one feature takes a non-zero value at the same time. Groups of mutually exclusive features can be 'bundled' together into a single feature, further increasing computational efficiency.

Voting

The voting ensemble model is another method of aggregating predictions from multiple base estimators (Mahajan et al., 2023). In contrast to bagging, base estimators in the voting ensemble model can have different model architectures, and all models are trained on the same dataset. In regression, the final prediction from a voting ensemble model is the unweighted or weighted average of the base estimators' predictions. Using a weighted average allows more importance to be placed on specific base estimators. The voting

strategy benefits from combining the strengths of each model class in the ensemble but can show lower performance if the base estimators are too similar.

Stacking

In a stacking ensemble model, predictions from base estimators serve as inputs to a meta-learning model, which combines the inputs to produce a single, final output (Ganaie et al., 2022). In other words, the predictions from each base estimator serve as the input dataset for the meta-estimator, which learns patterns within these predictions to output a final, low-error prediction. The meta-estimator can learn which base estimators are the most important and how to most effectively combine predictions from base estimators (Mahajan et al., 2023). The meta-learner can have a different structure from the base estimators, with examples of meta-learners being Random Forests, support vector machines, linear regressors, and neural networks. While stacking ensembles can improve performance in similar ways to those discussed above, they can be computationally expensive to train, as all the base estimators and meta-learning model must be fit to the data.

Chapter 3

Related Work

A thorough literature review revealed three methods that used domain knowledge and data modelling techniques to estimate global maternal mortality ratios. Section 3.1 discusses these approaches in detail. My proposed models' MMR estimates will compared against estimates from these existing models later in this thesis. Section 3.2 introduces an alternative algorithmic modelling approach, with Section 3.3 providing examples of this approach in public and maternal health research.

3.1 Existing Methods for Estimating Maternal Mortality Using Domain Knowledge and Data Modeling

The Model Produced by the United Nations' Maternal Mortality Estimation Inter-Agency Group

The Maternal Mortality Estimation Inter-Agency Group (MMEIG) is a collaboration between United Nations (UN) Member States, the WHO, the World Bank Group, and various UN agencies (World Health Organization, 2025). This collaboration models regional and country-specific MMRs between 2000 and 2023 for women 15 to 49 years old in 195 countries and territories. The consistent methodology used to determine these estimates enables global monitoring of maternal mortality trends and progress towards the UN's Sustainable Development Goal. The group's modeling approach has evolved over this time frame, with each report updating the model used to estimate the MMR.

The MMEIG uses a combination of two models to estimate MMR (World Health Organization, 2025; Peterson et al., 2024). First, the Bayesian maternal mortality misclassification (BMis) model calculates adjustment factors for the provided civil registration and vital statistics (CRVS) data to account for underreporting and misclassification of maternal deaths, as discussed in the Background Information (Section 2.2). This adjustment

was applied specifically to CRVS data because it was the largest input to the MMEIG estimates (Peterson et al., 2024). Global and country-specific adjustment factors were calculated using high-quality specialised studies, which provided a benchmark for the accuracy of CRVS data. Adjustment factors for countries with no specialised studies were calculated using global estimates. Data from non-CRVS or specialised study sources was increased by 10% to account for underestimation of maternal mortality. This explicit adjustment for poor-quality data was a strength of the UN MMEIG model, especially given the widespread underreporting and misclassification of maternal deaths reported in the literature.

After using the BMis model to correct CRVS data errors, the MMEIG estimated the MMR per country per year using the Bayesian maternal mortality estimation (BMat) model (World Health Organization, 2025). The BMat calculates MMR as the sum of non-HIV-related maternal deaths and HIV-related maternal deaths, where death was due to pregnancy aggravating an existing HIV/AIDs condition. The BMat estimates non-HIV-related MMR using a Bayesian hierarchical regression model (World Health Organization, 2025). Briefly, a hierarchical model determines a general trend and individual-specific deviations from the trend, which are referred to as random effects (Veenman et al., 2024). In BMat model, non-HIV related maternal deaths were estimated using global covariate trends with region and country specific effects, reflecting the belief that countries in the same geographic region have similar MMR trends (Alkema et al., 2017). This enables the model to estimate MMR for countries with sparse data, as it can use information from countries in a similar geographic region, which is an advantage of the modelling approach. The model parameterises the general trend and individual effects using prior knowledge, with each of the parameters initially drawn from a prior probability distribution (Veenman et al., 2024). The parameters were updated via Bayes' rule upon observation of data, with less data causing the final parameter values to be closer to the parameters drawn from the prior distribution.

The covariates used to calculate non-HIV MMR were gross-domestic product (GDP) per capita, general fertility rate, and presence of a skilled birth attendant (World Health Organization, 2025). An autoregressive integrated moving average (ARIMA) model was then used to determine whether empirical country-maternal mortality tracked with the covariates. For example, if the data indicated that non-HIV MMR decreased more slowly than predicted by the covariates, the non-HIV MMR estimate would be reduced. This incorporation of a data-driven factor is a strength of the BMat model, as it enables the model to reflect local conditions and be more responsive to sudden shocks that cause the country's MMR to deviate from the trend predicted by the covariates (Alkema et al., 2017). The less data available for a specific country-year, the less the associated non-HIV MMR estimate could be adjusted, making it more strongly influenced by the covariates (World Health Organization, 2025).

HIV-related MMR was estimated separately because evidence indicates that HIV/AIDS is a prominent cause of maternal mortality in countries with ongoing HIV/AID epidemics, with studies showing that women infected with HIV have approximately 8 times

higher risk of pregnancy-related death (World Health Organization, 2025; Zaba et al., 2013). Calculation of HIV-related MMR involves a constant that defines the relative risk of dying from HIV/AIDS for a pregnant versus non-pregnant women, which is estimated in conjunction with experts (World Health Organization, 2025; Alkema et al., 2017). The small subjectivity in the value of this constant is demonstrated through the change in its value between the 2010 and most recent BMat models (Wilmoth et al., 2012; Alkema et al., 2017).

While the model was fit to all data provided by the country, it placed higher weight on values with lower error variances, which were derived from calculating the random error in the data collection processes (World Health Organization, 2025). As a result of incorporating error, the final BMat estimates had smaller uncertainty intervals for countries with higher-quality data. By incorporating these uncertainty intervals, the BMat model could warn stakeholders about which statistics are more reliable, thus providing more nuanced information about maternal health trends.

The Global Burden of Disease Study Model

Coordinated by the Institute of Health Metrics and Evaluation (IHME), the Global Burden of Disease Study (GBD) is an international scientific initiative that has benchmarked major diseases and risk factors since 1993 (Murray, 2022). It encourages scientific debate by providing an alternate set of population health estimates from the UN (AbouZahr, 2011). Additionally, unlike the UN MMEIG estimates which are produced under consultation with Member States, GBD estimates are produced independently, removing potential bias (Mathers, 2020). Many GBD Studies have been published in high-impact, peer-reviewed journals like Lancet (Murray, 2022). The GBD Study's strong reputation is apparent in its use in national planning by a variety of governments, such as the United Kingdom, Norway, and China.

The 2021 GBD Study produced estimates of maternal mortality for 204 countries between 1990 and 2021 (Naghavi, 2024). Similar to the UN MMEIG estimates, using a consistent methodological approach across countries enables GBD estimates to be used to monitor global trends in MMR and track progress toward the Sustainable Development Goals (Murray, 2022). Unlike the UN MMEIG estimates, the GBD estimates were informed by sub-national data as well as national data, allowing it to incorporate local realities into its predictions (Naghavi, 2024). The data was then cleaned, standardised, and any deaths reported with an unclear or incorrect cause-of-death were probabilistically redistributed to a more likely cause of death (Johnson et al., 2021). This reduced the effect of misclassification errors on the model's final outputs, improving the accuracy of its predictions.

GBD maternal mortality estimates were produced using cause of death ensemble modelling (CODEm) (GBD 2021 Causes of Death Collaborators, 2024). The ensemble consisted of linear mixed effects regression (LMER) and spatiotemporal Gaussian process regression (ST-GPR) models. Like the Bayesian hierarchical models discussed above,

LMER models use random effects to quantify super-region, region, country, and age-level trends (Foreman et al., 2012). The ST-GPR model first uses a LMER model without country-specific effects to give a basic maternal mortality prediction. Then, it performs additional regression on the initial prediction errors using smoothing, where each datapoint's individual error was replaced by the weighted average of errors in its neighbourhood. The neighbourhood was defined over time, geographic space, and age. The smoothing operation reduces noise, with the extent of smoothing increasing for countries with less data.

This technique is a strength of the GBD Study, as it enables CODEm to estimate the MMR of countries with sparse data using information from similar regions. Additionally, by adding the computed residuals to the model's initial predictions, the model can capture trends in the data not represented by the covariates (Foreman et al., 2012).

The relationship between all chosen covariates and maternal mortality was statistically significant (GBD 2021 Causes of Death Collaborators, 2024). Additionally, each covariate had a causal link with maternal mortality, as established by existing scientific literature and expert analysis. The GBD 2021 Study used more covariates than the MMEIG, with the former using 19 covariates while the latter only used three (GBD 2021 Causes of Death Collaborators, 2024; World Health Organization, 2025). More specifically, the GBD study estimated maternal mortality using covariates including, but not limited to, age-specific fertility, maternal education, neonatal mortality ratio, skilled birth attendance, age-specific HIV mortality in females 10 to 54 years old, and age-standardised wasting (GBD 2021 Causes of Death Collaborators, 2024).

The final maternal mortality prediction was the mean of 1000 CODEm ensemble predictions, with each prediction generated by one individual, component model. The likelihood of each model being chosen was determined by its weight, which was calculated from the base model's out-of-sample predictive performance (GBD 2021 Causes of Death Collaborators, 2024). The 1000 draws allowed the construction of a 95% uncertainty interval. This approach was another strength of the CODEm model, as it leverages patterns learned by a variety of models and increases generalisation, as discussed in the Background Information (Section 2.4.1) (Mahajan et al., 2023).

Limitations of the UN MMEIG and GBD Study's Maternal Mortality Models

While the UN MMEIG's BMat model and the GBD Study's CODEm model are widely respected, and their estimates are used to inform policy, they have some limitations (World Health Organization, 2025; Murray, 2022).

The MMR estimates produced by the BMat and CODEm models were based on sparse, low-quality data (Alkema et al., 2017; Peterson et al., 2024; GBD 2021 Causes of Death Collaborators, 2024). Inaccurate data can cause the models to produce misleading MMR estimates, reducing their ability to inform policy. Unfortunately, limited data about the extent of underreporting and misclassification errors in different countries and systems make the errors difficult to correct, reducing the efficacy of the UN's BMis model and the

GBD Study's data cleaning mechanisms. This was especially true in countries without CRVS systems, which have lower data quality and quantity (Cresswell et al., 2025). As a result, MMR estimates for countries with less developed data collection systems tend to be predicted with wide uncertainty bounds, reducing their informativeness (World Health Organization, 2025).

Similarly, lack of data about the trends captured by the models' parameters can force researchers to use wide prior distributions for the parameters (Veenman et al., 2024). When there is little empirical data available to update the parameters' values, the parameters are largely informed by this wide prior distribution. This can result in incorrect results with large uncertainty, again reducing the utility of the models' outputs.

Additionally, when there is little to no data for a country, BMat would estimate non-HIV related MMR solely using regional effects and covariates (Alkema et al., 2017). Similarly, the CODEm LMER models would estimate MMR using the covariates, super-region and region effects (Foreman et al., 2012). CODEm ST-GPR models would increase smoothing in their residual regression analysis and use the region and super-region effects. While this enables the models to estimate MMR for data-sparse countries, if there is regional heterogeneity, the country's MMR estimates would be pulled towards an unrepresentative region-level estimate, resulting in inaccurate country-level predictions. Additionally, the lack of data may be due to an abrupt change or crisis, which smoothing may obscure, causing the model to lose important information about maternal mortality.

Prediction error may also be due to the assumption of a global relationship between the covariates and MMR. However, research has found that skilled birth attendance (SBA), a covariate used in both BMat and CODEm, only significantly reduces MMR when SBA coverage across the country is at least 40% (McClure et al., 2007). Thus, countries with very low SBA coverage have a different relationship to MMR than countries with high coverage. If these countries are also missing data, the UN ARIMA model or GBD ST-GPR models could not adjust the covariate-driven estimates, contributing to inaccurate model predictions.

Another major limitation of BMat and CODEm is their consideration of only a small subset of relevant covariates. There are a wide variety of factors that influence maternal health. For example, non-communicable disease (NCDs) are a leading cause of maternal mortality, with cardiovascular disease being one of the primary causes of maternal mortality between 2018 and 2020 in Australia (Ramson et al., 2024; Akselrod et al., 2023). Other NCDs like diabetes, asthma and mental health conditions also commonly affect pregnant women, with anemia increasing probability of postpartum hemorrhage, the primary cause of global maternal mortality. The literature also gives evidence for how excess maternal mortality is linked to quality of medical care, the incidence of infectious diseases like malaria, climate-related hazards, availability of contraception, financial constraints, violence in the woman's region, the woman's geographic remoteness and education, racism in the health system, and gender inequities that influence a woman's ability to make decisions about childbearing and medical care (Ramson et al.,

2024; Akselrod et al., 2023; Koblinsky et al., 2016; Conway et al., 2024; Tunçalp et al., 2014).

Thus, BMat's consideration of only 3 covariates limits its accuracy, as it did not account for other important socio-economic and health-related trends. While CODEm uses 19 covariates, only two of the covariates were unrelated to quality of care, fertility or mortality rates, limiting its consideration of socio-economic variables (GBD 2021 Causes of Death Collaborators, 2024). Consequently, BMat and CODEm did not model many of the factors that affect MMR, limiting their accuracy and reducing their ability to inform policymakers about which socio-economic factors should be targeted to reduce MMR.

Specific Limitations of the UN MMEIG's Maternal Mortality Model

It may be challenging to use BMat to model different candidate policies. More specifically, it can be difficult to determine exactly how a policy would impact aggregate measures like GDP per capita and general fertility rate, which would be used to model the new MMR estimates produced as a result of the policy (Ward et al., 2023a). Additionally, while GDP per capita is a powerful predictor of MMR, the mechanism through which it impacts maternal mortality involves a variety of other factors. Therefore, using BMat to determine that increasing GDP per capita would reduce MMR would not provide politicians with enough information to craft policies that can effectively reduce MMR.

Unfortunately, researchers have observed that, when tested on out-of-sample data from more recent years, BMat can overestimate decreases in maternal mortality in low-income countries (Alkema et al., 2017). This overestimation may be due to the limitations in the modelling process discussed above. Consequently, BMat's authors indicated the need for further exploration of possible modelling techniques.

Specific Limitations of the GBD Study's Maternal Mortality Model

CODEm's use of a combination of complex models could make it difficult to interpret the underlying associations between covariates and maternal mortality. This could hinder the estimates' ability to effectively inform health policy.

Researchers have also noted that the GBD Study's authors do not have access to all available national and sub-national data due to data privacy restrictions (Rommel et al., 2018). This has produced discrepancies between GBD Study and government estimates. For example, researchers found that the estimates of the number of diabetes-related deaths from the German federal health reporting system were outside the uncertainty intervals of the GBD's estimates (Rommel et al., 2018). This restricted access to data affects the accuracy of GBD estimates, and thus their ability to inform national health policy.

The Global Maternal Health Microsimulation Model

The Global Maternal Health Microsimulation Model (GMatH) was first proposed in 2023 (Ward et al., 2023a). The authors of the GMatH model motivated their approach by describing how the models produced by the UN MMEIG and GBD Study may inadequately describe intra-country trend. More specifically, they described how the MMEIG and GBD estimates were based on statistical relationships between aggregate country-level factors and MMR, preventing them from modelling variation within a specific country. In contrast, the GMatH model simulates individual women's reproductive lifecycles to determine estimates of maternal mortality, with differences in how those lifecycles are simulated used to reflect country-level heterogeneity. These estimates were produced for 200 countries and territories for every year since 1990. MMR estimates were also forecasted to 2050. Additionally, the calibrated model was used to make projections for each year up to 2050.

The GMatH model used monthly cycles to simulate each stage of pregnancy and childbirth (Ward et al., 2023a). At each stage, the model estimates the probabilities of pregnancy, termination, and complications as a result of individual-level, social, and institutional risk factors. Parameters governed the transition probabilities to different states within the model. These parameters were estimated from probability distributions based on empirical data where possible, and on expert opinion when data was unavailable. Relationships between parameters were similarly derived through a mixture of empirical data and expert opinion. Parameters' prior probability distributions were based on a hierarchical model with up to five levels (global, country income group, continent, region, and country). This allowed the model to determine parameter values for regions with sparse data, which was a strength of the GMatH model. GMatH was then fit to empirical data.

GMatH used 5 sets of parameters, categorised into biological parameters, family planning parameters, health system parameters, obstetrical complications, and clinical interventions (Ward et al., 2023a). Examples of biological parameters include age-specific probability of pregnancy and anaemia status, while examples of family planning parameters include contraceptive preferences. Health system parameters include the type of care available at birth and underreporting of maternal deaths. Parameters representing obstetrical complications include the risk of postpartum haemorrhage and parameters representing clinical interventions include the use of elective interventions, such as caesareans. The model's use of a wide variety of parameters causally related to maternal mortality was a strength, enabling the model to produce robust MMR estimates. Additionally, GMatH's use of parameters specific to demographic groups allowed model estimates to more clearly represent local conditions.

To test the model's predictive accuracy, the authors calibrated the model's maternal death estimates using CRVS data collected between 1990 and 2015, then compared the model's estimates for 2016 to 2020 to the CRVS estimates for the same time period (Ward et al., 2023a). The mean absolute error for the total number of maternal deaths

in test set was 47.5.

GMatH's authors argue that, by simulating causal relationships between risks and the stage of a woman's reproductive lifecycle, their model can use causal-inference to predict maternal outcomes more robustly than the MMEIG and GBD correlation-based approaches (Ward et al., 2023a). Additionally, and in contrast to BMat and CODEm, GMatH's breadth of parameters allows a wide variety of policies and health system barries to be modelled. For instance, GMatH has been used to investigate differences in maternal mortality between women in rural versus urban areas, as well as for women with different education levels (Warda et al., 2024). This analysis showed the importance of addressing women's education as an avenue for reducing maternal mortality. In contrast, it is difficult to produce an effective policy to reduce MMR from observing that BMat's MMR estimates are primarily predicted by GDP, which is a difficult outcome for politicians to change (Ward et al., 2023a).

Limitations of the Global Maternal Health Microsimulation Model

Similar to the limitations of BMat and CODEm discussed above, a primary limitation of the GmatH model was its use of sparse and low-quality data, again reducing its prediction accuracy (Ward et al., 2023a). For example, there were multiple instances of parameters for high-income countries using the prior distribution calculated for upper-middle income countries (**?**). These parameters were generally informed by Demographic and Health Surveys, which collected data solely from lower-income countries, preventing informative priors from being generated for high-income countries. Other parameters that lacked supporting empirical evidence were instead informed by expert opinion, which may not reflect the local reality. Additionally, these parameters were often estimated using hierarchical models, with the associated limitations discussed in Section 3.1. These uninformative or unrepresentative priors could reduce the model's accuracy, thus decreasing its ability to inform policy.

There were also a variety of limitations unique to the GMatH model. For example, any misspecification of the causal relationships between maternal mortality and feature variables would reduce accuracy (van Imhoff and Post, 1998). Additionally, while the model can consider a wide variety of explanatory variables, each variable is associated with uncertainty, especially in the case of countries with little empirical data for parameter-tuning. By progressively adding variables, the model may become overly influenced by uncertainty, with its estimates becoming dominated by the parameters' uncertainty. As a result, the variation in the estimates may increase beyond the point at which the estimates themselves are informative, as they cover too wide a range of outcomes. Small inaccuracies in each of the parameters' values may also accumulate, further decreasing accuracy (Spielauer, 2011). Moreover, the more parameters included in the model, the greater its complexity, and thus the greater the chance of overfitting, reducing the model's ability to generalise (Li and O'Donoghue, 2013). This is particularly relevant for low-income countries with little data.

GmatH's accuracy may also be affected by how it orders simulated events. According to the original paper, the model "progresses in monthly intervals", indicating that the simulated women's states are updated at discrete timesteps (Ward et al., 2023a). Many of the parameters in the model are inter-related, and as a result the order they are updated can affect the model's final estimates (van Imhoff and Post, 1998). For example, if a woman experiences a severe complication, her chance of mortality would chance substantially depending on whether she was treated before or after occurrence of a secondary infection (Li and O'Donoghue, 2013).

The validity of GmatH's estimates is also affected by the starting state of its simulated population, which is a sampled population and thus could result in the model's final estimates being unrepresentative (Ward et al., 2023a; van Imhoff and Post, 1998).

Comparison of MMR Estimates from the Bayesian Maternal Mortality Model, Cause of Death Ensemble Model, and Global Maternal Health Microsimulation Model

On average, GMatH's country-level MMR estimates were 22% higher than CODEm's estimates and 19% than BMat's estimates (Ward et al., 2025). 85.8% of CODEm's MMR estimates were contained within GMatH's 95% confidence intervals compared to 88.1% of estimates from BMat. The correlation between GMatH and CODEm's MMR estimates was 0.828 versus 0.879 between GMatH and BMat.

Despite these similarities, Ward et al. (2025) found large variation across the models' estimates for certain countries, such as Nigeria and Afghanistan. The inter-model variation is likely due to their different methodologies and input datasets. For example, Ward et al. (2025) noticed that inter-model variation was often greatest when the only data available about a country was survey-based data about pregnancy-related mortality. In contrast to maternal mortality, the cause of pregnancy-related death does not need to be related to pregnancy, childbirth, or termination (Ward et al., 2025). To be used in the BMat and CODEm models, pregnancy-related mortality must be converted into a model-recognisable metric using a series of calculations and assumptions. In contrast, the pregnancy-related mortality data can be inputted directly into GMatH, preventing the model from needing to make potentially invalid assumptions that bias the model's estimates. This difference in pre-processing may explain the high inter-model variation for these cases.

The variation in MMR estimates can produce confusion and uncertainty about the type of policy that should be implemented (Ward et al., 2025). Consequently, the authors of the GMatH model describe their hope that their intrinsically different modelling approach could provide further insight into the reason why the models' estimates have diverged (Ward et al., 2023a).

3.2 Existing Maternal Mortality Models versus Decision-Tree Based Machine Learning Models

The BMis/BMat, CODEm, and GMatH models are part of the 'data modelling culture', as they estimate MMR by modelling the processes that generate the input data (Breiman, 2001). To do so, they make assumptions about the data's structure and the relationships that exist within the data, which inform their choice of priors and covariates. In contrast, decision-tree (DT) based machine learning techniques are part of the 'algorithmic modelling culture', where the model focuses on predicting the outcome of interest instead of trying to learn how the data is generated,. Using an algorithmic modelling approach may avoid error from uninformative priors, regional heterogeneity, and misspecification of causal relationships between MMR and feature variables.

Unlike DT models, BMat and CODEm must consider the effect of multicollinearity when selecting their feature variables. This process is described explicitly in CODEm's documentation (Breiman, 2001). Multicollinearity occurs when features are linearly dependent, which makes it difficult to attribute change in MMR to a specific feature (Chan et al., 2022). Linearly dependent variables contain similar information, making the model more likely to learn noise in the data, overfit and generalise less easily. However, use of a small subset of features can cause the models to ignore valuable information about how other health-related and socio-economic variables affect MMR, reducing their predictive accuracy.

In contrast, DT models are particularly suited to working with high-dimensional data, as splits in the individual DTs are determined by the feature partition that best reduces error (Scornet et al., 2015; Loh, 2014). Thus, if three feature variables are highly correlated and one of the features is already used in a split, the others are less likely to be chosen for future splits because they would not add additional information. Therefore, DT methods can include all features that could influence MMR with a lower risk of overfitting, unlike the previously described approaches. Additionally, if a variable does not help reduce predictive error because it is not correlated with MMR, it will not be used in any splits and thus will be ignored (Scornet et al., 2015; Loh, 2014). Therefore, error in redundant and uninformative variables will contribute less to uncertainty in the final predictions from DT models than GMatH, which relies on all variables. Consequently, DT models can work well with high-dimensional datasets.

Another strength of DT models is their treatment of missing values. They use surrogate splits or default directions to handle sparse data, as described in the Background Information (Section 2.4.1) (Costa and Pedreira, 2023; Chen and Guestrin, 2016). This prevents them from needing to use imputation methods that may introduce error or Bayesian hierarchical models that may overly smooth regional heterogeneity. As the number of dimensions increase, the likelihood that a specific section of the input space contains data decreases, thus increasing sparsity (Naghavi, 2024). As a result, DT models' ability to work with sparsity may help them avoid further limitations of high-

dimensional data (Krzywinski, 2018).

As a consequence of these benefits, researchers have stated the importance of exploring how algorithmic modelling approaches can be used to improve prediction of maternal health outcomes (Hu et al., 2025). Thus, I propose the use of a decision-tree based machine learning model to estimate MMR. This model will provide an alternate set of estimates to help form a consensus out of the three existing estimates presented in the literature.

3.3 Algorithmic Machine Learning Techniques in Public Health

To further motivate my use of algorithmic machine learning (ML) models, I present examples of how these approaches have been used in public health research. In this section, I refer to all algorithmic models as ML models to use their common name despite the data modelling approaches discussed above technically also being ML models.

3.3.1 General Overview

Machine learning models are being applied to a wide variety of public health research problems to take advantage of the large quantities of health data being generated by wearable devices, clinical records, and social media (Sadr et al., 2025). For example, they can be used for image-based medical diagnostic tasks, improving operational efficiency, predicting patient-specific risks, and drug discovery. ML models can be categorised as 'white box', 'grey box', and 'black box' depending on their level of interpretability, where 'white box' models are the most interpretable and 'black box' are the least (Neha Margret et al., 2024). 'Black box' models include deep learning and neural networks and are frequently used in image-based diagnosis. Unfortunately, their complexity and lack of interpretability make it difficult to identify the feature variables with the highest predictive power, reducing their ability to inform policy targets. In contrast, 'white-box' models like decision tree-based algorithms can be used to predict medical risk factors and complications in an interpretable manner, allowing the model user to understand the factors used by the model to produce its results.

In their review, Mahajan et al. (2023) discussed the use of ensemble-based models in public health. They found that bagging and boosting algorithms were the most popular in the surveyed literature, as they were used in 41 and 37 of 45 studies, respectively Mahajan et al. (2023). However, they were only the most accurate algorithms evaluated in the study in 26.8 and 40.5% of instances. In contrast, stacking and voting were less frequently used (23 and 7 out of 45 studies), but they had the highest accuracies 82.6% and 71.4% of the time, respectively. Stacking models' high performance was attributed to their ability to learn the best base estimators. As a result, stacking models have been used to predict incidence of diabetes, heart disease, liver disease, and skin cancer.

3.3.2 Estimations of Cause Specific Maternal Mortality and Risk of Mortality

Many studies that predict maternal health outcomes using DT based methods focus on estimating patient risk and cause-specific maternal mortality. As a result, much of the ML research in this domain uses classification models, which can categorise a woman's mortality risk as 'high', 'medium', or 'low'. While I use a regression model in my thesis, I include examples of classification models to contextualise how DT-based ML is being used in maternal health research. As described below, DT and boosting models generally had the highest, or among the highest, performance, motivating their use in my thesis.

Akazawa et al. (2021) used ML models to classify a woman's risk of postpartum haemorrhage, a leading cause of maternal mortality, to inform treatment. They compared the performance of logistic regression, DT, random forest (RF), boosted tree, and deep learning models trained on 11 clinical variables. The boosted tree model had the highest accuracy. However, the model was trained and tested on data from the same institution, potentially reducing its generalisability.

Similarly, researchers have used ML techniques to predict a woman's risk of pre-eclampsia, another leading cause of maternal mortality, to improve identification and treatment of high-risk pregnancies (Marić et al., 2020). The study trained Elastic Net and gradient boosting (GB) models on a range of medical and socio-demographic covariates, with similar performance between the two models. However, the model was trained and tested on data from a single referral hospital, which had a higher proportion of high-risk patients, potentially reducing the model's generalisability.

Sylvain et al. (2025) predicted more general, adverse pregnancy outcomes in Rwanda using logistic regression, DT, RF, GB models, support vector machines, and neural networks. The RF and GB models had the highest accuracy (90.6% and 88.49%, respectively). The study also determined the most predictive variables. However, the study predicted occurrence of adverse outcomes as a binary variable, with a negative outcome encompassing a wide variety of possible maternal and neonatal health events. This could reduce nuance in the model's predictions and thus its ability to inform treatment. This model may also not generalise to rural regions, as it was only trained on data from district hospitals, which are only responsible for only roughly 35% of births in Rwanda.

As a final example, Khadidos et al. (2024) used a stacking-ensemble model to classify maternal health risk in Bangladesh, where they trained GB, RF, DT, and k-Nearest Neighbours models as the base estimator, with each base estimator trialled as the meta-learner. Using the GB model as the meta-learner had the highest precision (0.86), with all DT based stacking ensemble outperforming sole use of bagging or boosting.

Machine learning classifiers are also used to predict health system attributes. For instance, Taye et al. (2025) used a RF classifier to predict whether a birth was attended by a skilled birth attendant using a mixture of socio-economic and health system quality variables. The RF model was also used to indicate the most predictive variables. The

model achieved 92% accuracy despite being trained and tested on survey data, which is known to be of lower quality. This data was imputed, which may introduce bias into the model's estimates. However, this study shows how DT based ensembles can achieve high performance with low-quality data. Similarly, Fredriksson et al. (2022) compared the performances of a RF model and artificial neural network to more classical statistical models when classifying the likelihood of a woman delivering her baby in a health facility. The RF had the highest classification accuracy (74%), with the paper also reporting the most predictive variables. The lower accuracy may be related to the authors' use of imputation.

3.3.3 Estimation of Maternal Mortality

There is a severe lack of studies that use algorithmic modelling techniques to estimate MMR and the number of maternal deaths. The only relevant study found in my literature review was published in 2025 and only estimated MMR for Bangladesh (Molla et al., 2025). The authors compared the performance of a Bidirectional Recurrent Neural Network and an Elastic Neural Network to predict MMR, with the associated root mean square errors being 3.30 and 3.44 per 100,000 live births, respectively. This study was severely weakened by its dataset size, as it reported having only 21 observations of MMR, which is insufficient to train a robust model. No critique can be made of its feature variables, as these did not appear to be reported. Despite this study's limitations, it serves as proof of concept for using algorithmic modelling techniques to estimate MMR. The lack of robust studies doing this type of analysis highlights a gap in the literature that my thesis aims to fill.

3.4 Summary

In conclusion, the UN MMEIG and Global Burden of Disease models are the primary modelling techniques for estimating global maternal mortality ratios. The GMatH is a newer method for MMR prediction. These three methods are limited by their low-quality, sparse input data. As a result, estimates for countries with less data may be estimated with parameters generated by uninformative priors. The parameter values and overall MMR estimates may also be oversmoothed, with the modelling techniques ignoring potential regional heterogeneity. Given that countries with less data tend to have higher MMR, this can reduce the estimates' ability to inform national health policy. The models are also weakened by their consideration of only a small subset of the socio-economic and health-related variables that impact maternal mortality. Using DT based models would circumvent the need to model the data generating process, eliminating assumptions about data distribution and the need for priors. Additionally, DT models can handle a wide range of feature variables, allowing them to make more comprehensive estimates, with evaluation of feature importance covering a wider range of domains. In maternal health research, ML methods are generally used to classify a woman's overall and cause-specific maternal mortality risk. I found only one algorithmic modelling approach to

estimating MMR, with the study solely predicting MMR in Bangladesh. Thus, there is a gap in the literature about how a decision-tree based ML model can be used to estimate MMR at a global level.

As a final note, building a model with an entirely new methodology will produce another set of MMR estimates. My model's estimates can be compared to the literature and contribute to resolving some of the lack of consensus around current MMR estimates.

Chapter 4

Materials and Methods

This thesis used a variety of socio-economic and health-related features to predict the MMR for each (country, year) sample. Separate models were trained to perform country-level prediction and forecasting. In the following chapter, I present information about the datasets used to develop my models, as well as the data processing and exploratory analysis pipeline (section 4.0-4.3) and computational workflow (section 4.4-4.8). Development of the final, highest performing model involved 4 major steps.

1. Applying cross-fold validation, feature selection, and missing data removal to generate different train datasets.
2. Training Random Forest, XGBoost, and LightGBM regressors on different versions of the training data to explore various pre-processing techniques.
3. Training stacking and voting ensemble models on different combinations of base estimators to reduce predictive error.
4. Evaluate best-performing model by analysing feature importance and sensitivity to input data as well as comparing its predictions to MMR estimates in the literature.

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. All data and code were uploaded to a public GitHub repository (https://github.com/R0sle/health_economics_honours). All data is available upon request.

4.1 Data Pre-Processing and Exploratory Analysis

Figure 4.1 gives a high-level overview of the data cleaning and exploratory analysis applied to the raw data, as well as the pre-processing steps used to generate different versions of the training data. I first merged various World Health Organisation and World Bank datasets into a single, raw dataset (Figure 4.1a). This raw data was cleaned

and split into train/test sets (Figure 4.1b). Separate train/test sets were used to train models for country-level prediction and forecasting. Each train set was further split into 5 cross-validation folds (Figure 4.1c). I applied five different feature selection techniques to each training fold, creating five versions of each fold (Figure 4.1d). Finally, I removed rows and columns from each training fold/feature subset combination that had more than a threshold proportion of missing data (Figure 4.1e). I tested four missing data thresholds, creating four versions of each training fold/feature subset. This process generated 100 versions of the training data for both country-level prediction and forecasting. These steps are explained in more detail in the following subsections.

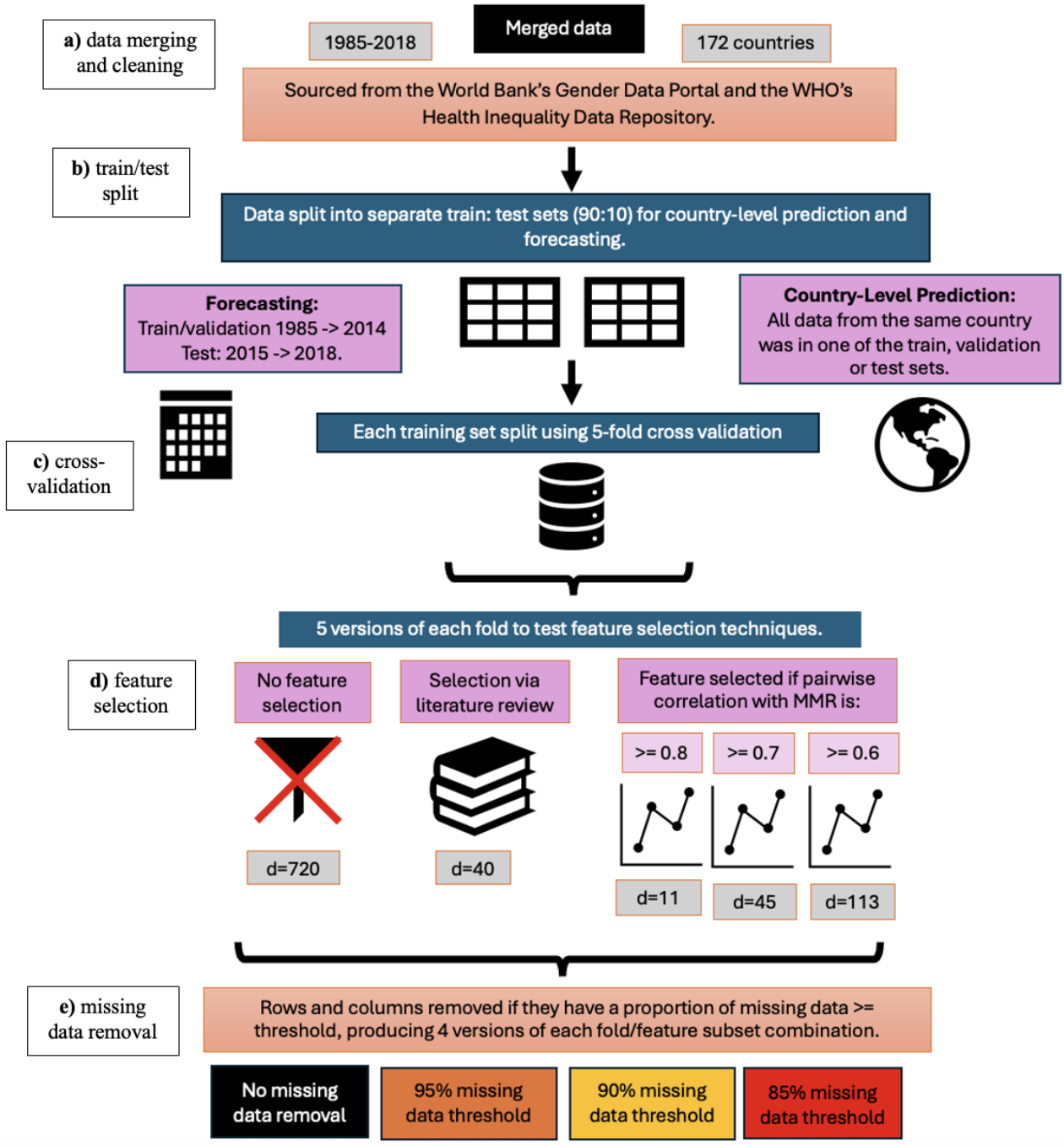


Figure 4.1: High-level overview of the data merging and pre-processing pipeline. I merged data from the WHO and World Bank into a single dataset (a). I then split this data into train/test sets, with separate splits used for country-level prediction and forecasting (b). I further divided each training set into 5 cross-validation folds (c). I then produced 5 versions of each fold using different feature selection methods (d). Finally, I generated 4 versions of each feature subset dataset by removing missing data with different levels of flexibility (e). This process generated 100 versions of the training data.

4.2 Data Sources and Merging

In this subsection, I describe the data merging and cleaning steps (Figure 4.1a).

4.2.1 Data Sources

National MMR data for 242 regions, countries, territories, and areas between 1985 and 2018 was sourced from the World Bank Group's Gender Data Portal (Maternal Mortality Estimation Inter-Agency Group, 2025). The data was derived from information provided by countries' national data collection systems, such as from national surveys, hospital records, and civil registration and vital statistics systems. These MMR values served as the 'ground truth' for training my model. They were different from the outputs of the BMat, CODEm, and GMatH models discussed in the literature review.

Feature data was retrieved from a variety of World Health Organisation (WHO) and World Bank Group repositories. More specifically, 5 datasets were downloaded from these repositories, with each component dataset itself a compilation of variables, sometimes provided by a range of sources. Information about each component dataset was summarised in Table 4.1. See the GitHub repository for the specific variables gathered from each data source.

Briefly, the feature dataset sourced from the World Bank Group Gender Data Portal (2025) describes various health and socio-economic outcomes, which were each estimated by UN/WHO divisions or other partner organisations. For example, one of the indicators included in this dataset was the 'Probability of Survival to Age 5', which was calculated by the UNESCO Institute for Statistics. Similarly, data describing health determinants related to the environment, employment, education and social protection was compiled by various agencies, with many of the monitored indicators used to track progress toward the UN's Sustainable Development Goals (The World Bank Data Catalog, 2024).

The dataset describing illness incidence and prevalence was compiled by the Institute for Health Metrics and Evaluation (2023), which publishes the Global Burden of Disease Study. Thus, illness incidence and prevalence would be determined using disease-specific versions of the CODEm framework.

The WHO Collaborating Center for Health Equity Monitoring (2024) re-analysed data from Demographic and Health Surveys to provide information about women's empowerment. More specifically, this re-analysis measured women's social independence, such as women's ability to complete schooling and achieve their goals, women's ability to make household decisions, and women's attitudes to violence.

This use of data from a variety of data sources was motivated by Onambele et al. (2023), who recommended combining multiple sources to take advantage of the different datasets offered by the WHO.

Each country was categorised as low, lower-middle, upper-middle, or high-income by the World Bank (World Health Organization's (WHO) Global Health Observatory (GHO),

2024). These categories were converted into numbers using ordinal encoding, where low-income was denoted as '1' and high-income as '4', to preserve their implicit order. No other features were modified from their raw format using methods like one-hot encoding, ordinal encoding, or log transformation.

Table 4.1: Summary information about the datasets used in this study.

Type of Dataset	Number of Features	Date Range	Number of Areas Covered	Demographic Subset Used	Source
National MMR estimates	1	1985-2018	242	NA	Maternal Mortality Estimation Inter-Agency Group (2025)
Health outcomes & literacy, agency	198	1960-2023	265	NA	World Bank Group Gender Data Portal (2025)
Illness incidence and prevalence	193	2000-2019	194	Sex	Institute for Health Metrics and Evaluation (2023)
Empowerment	9	1991-2023	120	Wealth quintiles (1, 5)	WHO Collaborating Center for Health Equity Monitoring (2024)
Socioeconomic, education, environmental data	64	1970-2023	195	Sex, wealth quintiles (1, 5), residence (urban, rural)	The World Bank Data Catalog (2024)
Categorisation of a Country's Income level	1	2024	198	NA	World Health Organization's (WHO) Global Health Observatory (GHO) (2024)

Some of the datasets contained disaggregated data. For example, features were sex or

Table 4.2: Illustrative example of subgroup specific versions of a single feature, 'Feature 1', with the bolded text defining the demographic being represented.

Country	Date	Feature 1 **Female**	Feature 1 **Male**	Feature 1 **Urban**	Feature 1 **Rural**	Feature 1 **Wealth Quintile 1**	Feature 1 **Wealth Quintile 5**

economic status specific. However, the ground truth MMR values were not disaggregated. Including the disaggregation as its own feature column would therefore produce a missing value in the associated MMR estimate column when merging the datasets. To prevent this, I created subgroup specific versions of the feature. See Table 4.2 for an illustrative example. If the data was disaggregated on a scale (e.g. Feature A was reported for wealth quintiles 1 through 5), I only used values from the most extreme subgroups (e.g. quintiles 1 and 5) to prevent the number of features, and thus the dimensionality of the dataset, from becoming too large.

As a note, my proposed model uses different input data than the BMat, CODEm, and GMatH models discussed in the literature review. This was due to variation across the models' covariates/features/parameters and methodologies. While the BMat and CODEm models use some pre-compiled datasets, they mainly rely on processed mortality data collected from a mixture of sources, including surveys and official reporting mechanisms World Health Organization (2025); GBD 2021 Causes of Death Collaborators (2024). GMatH relies both on mortality databases, Demographic and Health Surveys, the WHO database, and the medical literature (Ward et al., 2023a).

4.2.2 Merging Data

All datasets used in this report contained columns specifying the country and its associated ISO3 country code, as described in the ISO 3166 international standard (IBAN.com, 2025). However, different datasets sometimes used a different version of the same country's name (e.g. United States versus United States of America). Therefore, datasets were merged using the unique ISO3 code and year.

The national MMR estimates were collected between 1985 and 2018. Therefore, all data collected before 1985 and after 2018 was excluded. Features with no data between 1985 and 2018 were also excluded.

4.2.3 Data Cleaning

All (country, year) samples that were missing an associated MMR estimate were removed from the dataset. This avoided needing to impute the ground truth variable, which may

have caused the models to be trained on incorrect feature/MMR estimate pairings, introducing inaccuracy.

Additionally, I removed the following feature variables:

- 'Numbe maternal deaths'
- 'Lifetime risk of maternal death (1 in: rate varies by country)'
- 'Lifetime risk of maternal death (%)'

The 'number of maternal deaths' is the numerator of the MMR. Similarly, the two features measuring the 'lifetime risk of maternal death', as a rate or percentage, are calculated using the MMR (World Health Organization, 2025). The Pearson's correlation coefficient between the MMR estimate and the 'lifetime risk of maternal death (%) was 0.93, showing their strong positive correlation. Therefore, these three variables were excluded from the feature dataset to prevent the model from using them to predict the MMR instead of learning the relationship between MMR and socio-economic and health-related data.

The final data cleaning step involved removing the 'country' and 'year' as feature variables, instead using them as unique sample identifiers. This step was performed because 'country' and 'year' could introduce bias in the data, where the model learns the typical MMR for a country or year instead of learning to use the relationships between MMR and the features.

4.3 Exploratory Data Analysis

An initial exploratory data analysis was conducted to gain a better understanding of the dataset and motivate choice of pre-processing techniques.

4.3.1 Trends in Missing Input Data

No data imputation was used despite the high level of missing data. This was done to prevent imputation from introducing bias into the dataset, especially given the high level of missing data. More specifically, the pattern of missing data in this dataset would typically be categorised as 'missing not at random' (MNAR). In other words, the probability of data being missing relies on both observed and missing data, or on other, unobserved variables (Emmanuel et al., 2021; Mukherjee et al., 2023). This dataset would be considered MNAR because the probability of missing data is heavily related to the robustness of the country's data collection systems, which is an unseen variable, but which may be related to a country's MMR. Additionally, some years have a much lower proportion of missing data than others (20 to 35% missing data versus 80 to 90%). This indicates structural differences in data collection due to periodic data reporting, which is an unobserved variable.

Given that data MNAR is dependent on unseen data, it is extremely difficult to remove or impute the missing data without ignoring the important unseen variables and introducing bias (Emmanuel et al., 2021). For example, removing all rows and columns with missing data or imputing the missing values based on the observed datapoints would bias the data toward countries reporting greater amounts of data. As a result, I only used ML models that could work with missing data.

4.3.2 Key Statistics

To give deeper insight into the feature dataset, I presented key summary statistics about some of the features that the literature describes as having a particularly meaningful relationship with MMR. The proportion of missing data was calculated after data cleaning, explaining why the MMR estimates have a missing data proportion of 0%. Both mean and median were included to give an indication of outlier occurrence.

4.3.3 Principal Component Analysis

Principal component analysis (PCA) was employed for dimensionality reduction purposes. Rather than trying to visualise patterns by plotting all 721 feature dimensions, PCA was used to project samples into the 10 principal components that captured the maximum amount of total variance across the dataset.

PCA cannot be performed on a sparse dataset. Thus, Scikit Learn's k-Nearest Neighbours imputation method was used to impute missing data before applying PCA (Pedregosa et al., 2011). This method imputed missing values using datapoints that were most similar to the sparse datapoint in their non-missing dimensions. The data was standardised to a zero mean and unit standard deviation before applying the Scikit Learn PCA method (Pedregosa et al., 2011). This prevented variance calculations from being skewed by features with high magnitudes. Note that imputation was not used at any other point in this research.

The input feature data was plotted across its two most important principal component axes, with samples being coloured according to their income level, MMR estimate, and year to better identify patterns and clusters in the data.

4.3.4 Correlation Analysis

I calculated Pearson's pairwise correlation coefficients between each feature and the maternal mortality ratio. Analysis of these pairwise correlations were used to inform feature selection strategy.

4.4 Processing for Machine Learning Pipeline

Figure 4.1 visualises the flowchart overview of the data pre-processing process. Different versions of the train dataset were produced to explore the effect of various pre-processing

techniques. The distribution of data across the train and test sets was presented in Chapter 5 to give insight into model performance and generalisation.

4.4.1 Splitting Input Data into Train/Test Sets

The cleaned dataset was split into train/test subsets in two different ways (Figure 4.1b). Separate train/test datasets were used for country-level prediction (Figure 4.2a) and forecasting (Figure 4.2b).

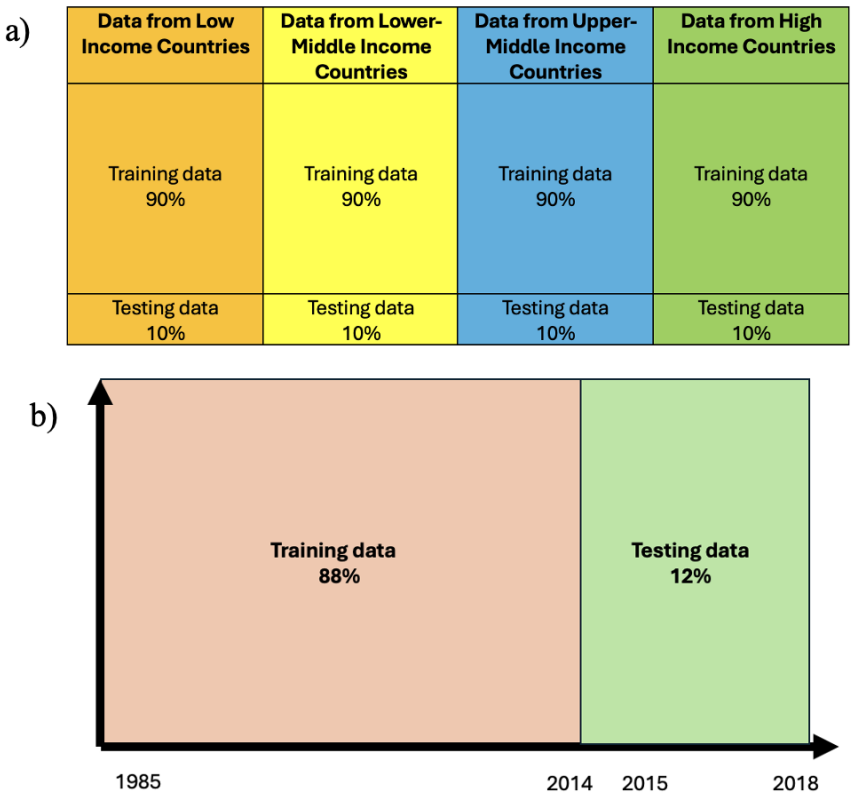


Figure 4.2: Train/test split visualisation. a) When split for country-level prediction, 90% of the data from each income level was placed into the train set and 10% into the test set, with all data from the same country in only one set. b) When split for forecasting, all data between 1985 and 2014 was placed in the training set (88% of available data) and all data between 2015 and 2018 was put in the test set (12%).

Country-Level Prediction (CLP)

I split the input dataset so all data from a specific country was in either the train or test set. Splitting the input data by country prevented data leakage by preserving

independence between the train and test sets. Models trained in this way could estimate missing MMR values and thus inform policy makers about their country's maternal health status.

Each country in the input dataset was classified as high-income, upper-middle income, lower-middle income, or low-income by the World Bank. I split the original dataset into 4 subsets, each containing all countries from a specific income level. Each of these subsets were split into train/test datasets in a 90:10 ratio. The 4 train sets were merged into one complete train set, and the 4 test sets were merged into a complete test set (Figure 4.2a). This ensured that samples from each income level were in both the train and test sets to allow the test set to effectively evaluate whether the model could generalise to different income levels. However, ensuring all data from the same country was in either the train or test set meant the train/test split was not exactly 90:10. Instead, the true ratio was within one or two percent of 90:10, as the number of rows per country did not allow an exact 90:10 split.

Forecasting

MMR forecasts can give policymakers information about future MMR trends as well as allow researchers to predict the effects of candidate policies. I simulated this scenario by placing all data from 1985-2014 in the train set and all data from 2015-2018 in the test set (Figure 4.2b). The goal of this strategy was for the model to learn patterns in the historical data to use to predict the future.

This division produced an 88:12 train/test split. I included data from 2015 in the test set to ensure the test set contained sufficient non-missing data to be useful, as 2015 contained more than 50% less missing data than average.

4.4.2 Cross-Validation

Each of the CLP and forecasting train datasets were further divided into 5 cross-validation folds, each of which being a specific permutation of the train/validation 80:20 split (Figure 4.1c).

If the data was being split for CLP, all data for the same country was placed in either the train or validation set to prevent data leakage and evaluate the model's ability to predict sparse country data. If the data was split to perform forecasting, all data from the same year was placed in either the train or validation set to prevent data leakage and assess the model's ability to predict for an unknown year. These conditions resulted in slight deviations from the 80:20 ratio, but only within a percentage point.

The train data was split into cross-validation folds using Scikit Learn's GroupKFold method, which ensures that entries of the same group only appear in one validation set (Pedregosa et al., 2011). Members of the groups were countries for CLP and years for forecasting. This method ensured that the train and validation sets varied across the

different folds, allowing me to assess how changes in the composition of the training dataset affected model performance.

4.4.3 Feature Selection

As discussed in the literature review, decision-tree based models can work with high-dimensional data. I tested whether this ability meant they achieved the greatest performance when using high-dimensional data, or if they achieved greater performance when working with a subset of features. This difference may be due to a higher number of features introducing noise, sparsity, and additional computational complexity, as discussed in Section 3.2.

I created 5 versions of each fold, each with a different selection of features (see Table 4.3 and Figure 4.1d for a summary and the GitHub repository for a spreadsheet giving the specific features used in each subset). In the first case, no features were removed to evaluate model performance on the full feature dataset, which contained 720 features. To create the second feature subset, I surveyed a number of papers about maternal mortality to learn which features researchers believed most strongly influence MMR. This search allowed me to derive a subset of 40 biological and socio-economic feature variables. While there were many other relevant features I could have chosen from the available dataset, I believed these covered the major MMR determinants.

The final three feature subsets were produced using the correlation between feature variables and MMR. More, specifically, I computed the pairwise Pearson's correlation coefficient of all feature columns with the MMR estimate. I used the Pandas correlation method, which ignores rows where either feature pair has a missing value. I created a dataset containing features whose absolute pairwise correlation coefficient with MMR was at least 0.8. I produced two additional datasets containing features whose absolute correlation coefficient with MMR was at least 0.7 and 0.6, respectively. This allowed me to test the strength of correlation needed for the features to improve model performance.

As a note, feature selection was performed after cross-fold validation to ensure that the training samples used in each fold were consistent across the feature subsets.

4.4.4 Iterative Removal of Rows and Columns with a Higher Proportion of Missing Data Than a Specific Threshold

As discussed in Section 4.3.1, the input data contained information missing not at random. In this case, imputation and/or data removal can introduce bias, as the presence of missing data may signal important information about the country's health system dynamics. However, if the specific feature has a very high proportion of missing data, such as over 85%, the model may overfit to its small dataset. Therefore, I experimented with removing rows and columns with very high proportions of data. More specifically, I iteratively removed columns and rows that contained a higher proportion of missing values than a pre-defined threshold. Higher thresholds meant rows and columns with a

Table 4.3: The 5 feature selection methods used to create 5 versions of each train fold.

Feature Selection Method	**Number of Features**
No feature selection employed.	720
Features that the literature describes as having a strong influence on MMR (Zaman et al., 2024; Patel, 2023; World Health Organization, 2025; Ramson et al., 2024; Akselrod et al., 2023; Koblinsky et al., 2016; Conway et al., 2024; Warda et al., 2024)	40
All features whose absolute Pearson's pairwise correlation coefficient with MMR >= 0.8. Called 'Correlation 0.8' from this point forward.	11
All features whose absolute Pearson's pairwise correlation coefficient with MMR >= 0.7. Called 'Correlation 0.7' from this point forward.	45
All features whose absolute Pearson's pairwise correlation coefficient with MMR >= 0.6. Called 'Correlation 0.6' from this point forward.	113

greater proportion of missing values were kept in the dataset.

Given that removal of a sparse row could affect the proportion of missing data in a column, and vice versa, the removal of rows and columns was conducted iteratively until the dataset stabilised. Similarly, since each feature subset had a different number of columns, the iterative removal of missing data had to be performed per feature subset/fold pair, as the number of columns influenced the proportion of missing data per row.

Iterative data removal was performed per fold to prevent data leakage between the folds (Figure 4.1e). This procedure was only applied to the training data, not the validation or testing sets. This allowed evaluation results from different training datasets to be compared. However, to allow the models to function, columns dropped from a model's train set due to feature selection or missing data removal were also dropped from the validation and test sets.

I produced 4 versions of each feature subset/fold combination using missing data thresholds of 85%, 90%, 95%, and 100% (no missing data removed). I included missing data thresholds above and below 90% because researchers have hypothesised that even princi-

pled imputation methods can introduce bias into datasets with greater than 90% missing data (where data is missing at random) (Memon et al., 2022). I used this range of missing data thresholds to determine whether a similar cutoff point occurs with data missing not at random when using decision-tree based methods like default directions to handle missing data.

4.4.5 Summary of Datasets Produced Via Pre-Processing

The cleaned data was split into train/test subsets in a 90:10 ratio, with different versions of the split implemented for CLP and forecasting (Figure 4.1b). These subsets were each further split into 5 cross-validation folds (Figure 4.1c). Five versions of each fold were created by applying different feature selection mechanisms (Figure 4.1d). Finally, 4 versions of each of each feature subset/fold combination was produced by applying iterative missing data removal with different thresholds of missing data allowed (Figure 4.1e).

This produced 100 datasets (5 folds x 5 feature selections x 4 missing data thresholds) for each of country-level prediction and forecasting analyses.

4.5 Computational Workflow

Figure 4.3 visualises the development, testing, and investigation of a variety of decision-tree based machine learning models, with each of the following steps again done separately for country-level prediction and forecasting. I first trained a Random Forest, XGBoost, and LightGBM model on each version of the training data, creating 300 models (Figure 4.3a). I fine-tuned each of the models' parameters over 1,000 Optuna trials (Figure 4.3b). I compared the performance of models trained on each version of the training data to determine which model type and combination of pre-processing techniques produced the best-performing model. I then experimented with combining the models' predictions using a voting or stacking ensemble (Figure 4.3c). I trialled each of the Elastic Net, Random Forest, and Support Vector Machine architectures as the stacking ensemble's meta-learner. Each of the voting and stacking ensembles' hyperparameters were tuned over 1,000 Optuna trials. I then investigated whether using a subset or randomly permuted ordering of base estimators' predictions reduced the predictive error of the best-performing voting or stacking ensemble (Figure 4.3d). Next, I conducted a feature importance analysis for the best-performing model to find the variables with the highest predictive power for MMR (Figure 4.3e). I then determined whether the predictive error of the best-performing model changed when it was trained on data from only one income level, thus investigating its sensitivity to its input data (Figure 4.3f). Finally, I compared my best-performing model's MMR predictions to estimates from the BMat, CODEm, and GMatH models described in the literature review (Figure 4.3g). These steps will be described in detail in the following subsections.

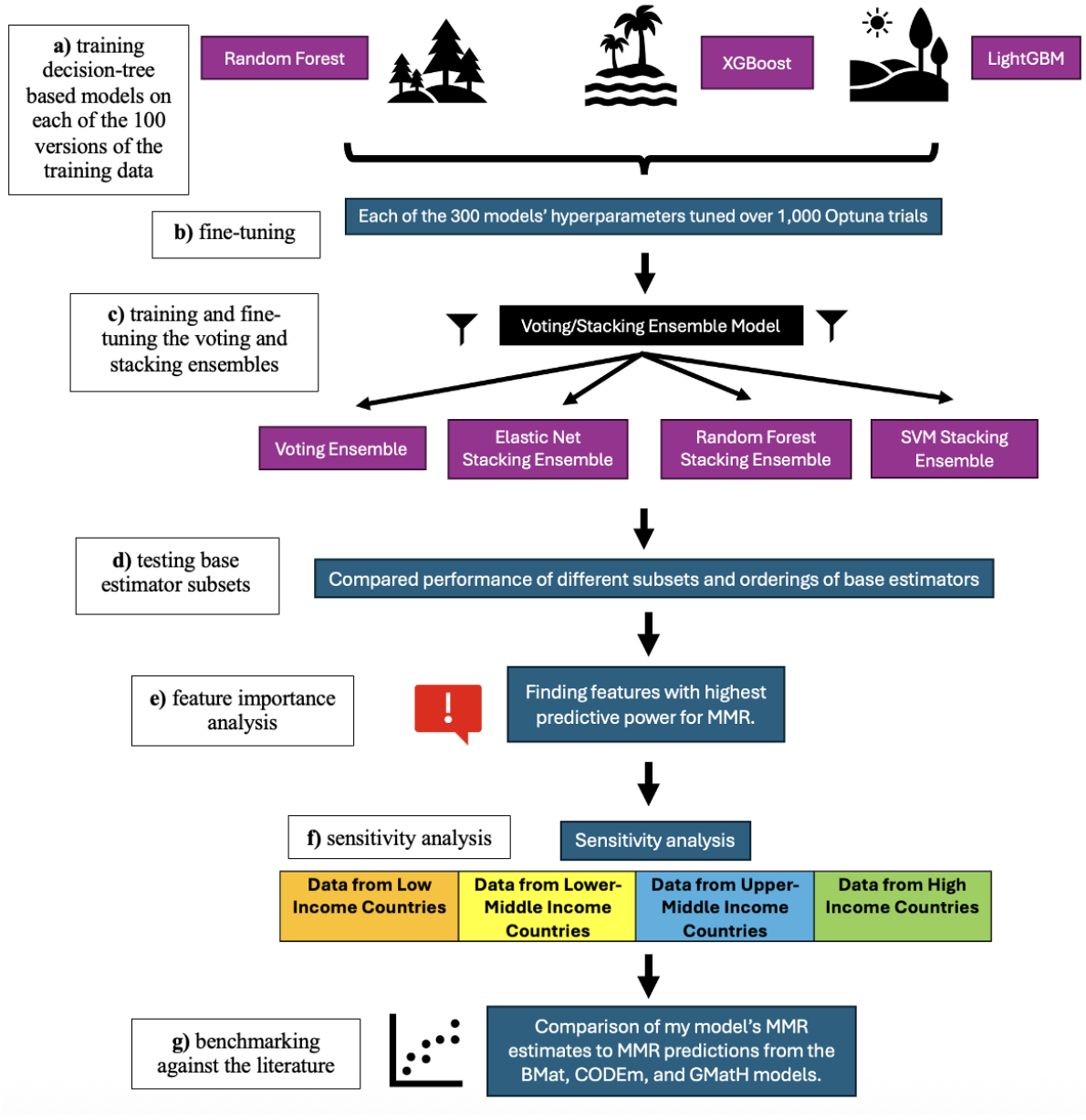


Figure 4.3: Experimental overview of the development of decision-tree based machine learning models used to estimate MMR. Development was done separately for models used to perform country-level prediction and forecasting. Each of the 100 versions the training data was used to fit a single Random Forest, XGBoost, and LightGBM model (a), with each model's hyperparameters fine-tuned over 1000 Optuna trials (b). Different combinations of predictions from each of these base estimators were used to train voting and stacking ensembles (c, d). The features with the highest predictive power for MMR in the best-performing model were identified (e). Finally, I tested the best-performing model's sensitivity to the input data and similarity to MMR predictions from existing models (f, g).

4.6 Base Model Training and Fine-Tuning

This section details the methodology used to train and fine-tune the Random Forest, XGBoost, and LightGBM base models (Figure 4.3a and 4.3b).

I did not use deep learning methods in this research because deep learning models cannot natively handle missing data (Kia et al., 2022). To use deep learning models with my sparse data, I would need to remove, ignore, or impute the missing values. However, this could introduce bias, especially given that information was missing not at random in my dataset. Therefore, I only used decision-tree based models, which can natively handle missing data by learning a default direction to move through the tree when they encounter missing data (Chen and Guestrin, 2016).

4.6.1 Base Model Training and Fine-Tuning

Scikit Learn's Random Forest Regressor, XGBoost's XGBRegressor and LightGBM's LGBMRegressor were trained to predict the MMR for a specific country, year datapoint (Pedregosa et al., 2011; Chen and Guestrin, 2016; Ke et al., 2017). 100 versions of each model type were trained for country-level prediction (CLP), with each version corresponding to one of the 100 versions of the train dataset produced during pre-processing. Another 100 versions of each model type were trained to perform forecasting.

The models' hyperparameters were fine-tuned using the Optuna hyperparameter optimisation framework (Akiba et al., 2019). Finetuning occurred over 1,000 Optuna trials, where each trial represented a choice of values for the subset of hyperparameters being tuned. Tables 4.4, 4.5, and 4.6 show the specific hyperparameters fine-tuned for Scikit Learn's Random Forest, XGBoost and LightGBM, respectively. All other hyperparameters were set to their default values. During each trial, the model being fine-tuned was fit to its associated train data using the chosen hyperparameter values. Its performance was evaluated by calculating the mean squared error (MSE) of its predictions on the associated validation fold. The set of hyperparameters with the lowest validation MSE across the 1,000 trials was saved and used to define the highest performing model. These best-performing hyperparameters can be accessed via the Optuna trial objects in the linked GitHub repository.

This method produced 300 fine-tuned models (100 each of XGBoost, LightGBM, and Random Forest models) for CLP and 300 models for forecasting. Thus, 600 models were fine-tuned in total, which took 1 to 2 days. Due to this computational demand, fine-tuning was conducted with only one metric as opposed to multiple metrics capturing different information. MSE was used because it heavily penalises outliers, which in this context would most likely be errors in the high MMR estimates for lower-income countries.

Table 4.4: Hyperparameter Tuning for Scikit-Learn's Random Forest Regressor (Pedregosa et al., 2011)

Hyperparameter Name in Scikit-Learn	**Hyperparameter Function**	**Range of Potential Values**
n_estimators	The number of trees.	10 to 300
max_depth	The maximum depth of trees.	3 to 25
min_samples_split	The minimum number of samples/rows for which an internal node can be split.	2 to 10
bootstrap	Whether each tree was trained on a random subset of samples.	True or False
max_samples	Proportion of the full dataset used to train each base estimator. This parameter was not used when bootstrap was set to False.	0.01 to 1.0

Table 4.5: Hyperparameter Tuning for XGBoost's XGBRegressor (Chen and Guestrin, 2016)

Hyperparameter Name in XGBoost	**Hyperparameter Function**	**Range of Potential Values**
n_estimators	Number of trees/boosting iterations.	10 to 300
max_depth	The maximum depth of trees.	3 to 25
learning_rate	Controls the extent to which each new tree influenced the model's predictions.	0 to 1
reg_alpha	Constant used for L1 regularisation	0 to 0.001
reg_lambda	Constant used for L2 regularisation	0 to 0.001
booster	'gbtree' was the XGBoost gradient boosting method. 'dart' modified 'gbtree' to randomly drop trees to reduce overfitting.	'gbtree' or 'dart'
subsample	Proportion of dataset randomly chosen for each boosting iteration during training.	0.1 to 1

Table 4.6: Hyperparameter Tuning for LightGBM's LGBMRegressor (Ke et al., 2017)

Hyperparameter Name in XGBoost	**Hyperparameter Function**	**Range of Potential Values**
n_estimators	Number of trees/boosting iterations.	10 to 300
max_depth	The maximum depth of trees.	3 to 25
learning_rate	Controls the extent to which each new tree influenced the model's predictions.	0 to 1
reg_alpha	Constant used for L1 regularisation	0 to 0.001
reg_lambda	Constant used for L2 regularisation	0 to 0.001
boosting	'gbdt' was the LightGBM gradient boosting method. 'dart' modified 'gbdt' to randomly drop trees to reduce overfitting.	'gbdt' or 'dart'
bagging_freq	Every k-th iteration, a random subset of data was used for the next k iterations of training.	0 to 10
bagging_fraction	Proportion of input data randomly chosen for training. Used if bagging_freq was not zero.	0.1 to 10

4.6.2 Testing and Comparison

The best performing hyperparameter values for each model were saved at the end of the Optuna fine-tuning process. When being evaluated, each model was re-fit on its associated training data using these hyperparameter values.

The 300 fine-tuned models for CLP were evaluated on the same test set, which had no missing data removed. The only difference between the test sets used for the various models was that each test set contained only the features included in the model's training data.

The fine-tuned models were evaluated on the accuracy of their test set predictions. Specifically, the MSE, root mean-squared error (RMSE), mean absolute error (MAE), R^2, and mean relative error of its test set predictions were calculated. Using a wide range of metrics enabled a more nuanced evaluation of the model's performance, as the metrics placed different emphasis on outliers. The mean relative error was a symmetrical version of the mean absolute percentage error (MAPE) described in Section 2.3.2. The formula for the mean relative error (Eq.10) modified the base MAPE formula Eq.5 to adjust for its asymmetrically.

$$MeanRelativeError = \frac{1}{n}\left(\sum_{i=1}^{n} \mid \frac{y_i - \hat{y}_i}{max(|y_i|, |\hat{y}_i|)} \mid *100\%\right) \quad \text{(Eq.10)}$$

Each combination of model type, feature subset and missing data threshold had 5 associated training folds. The combination's performance was the average test performance of the models trained on these folds. The performances of different combinations were then compared.

The 300 fine-tuned models that performed forecasting were similarly evaluated and compared.

4.6.3 Feature Importance Analysis

Each feature's importance was calculated using the models' in-built methods. Scikit Learn's Random Forest Regressor used the mean decrease in impurity/MSE across the feature's splits (Pedregosa et al., 2011). XGBoost used the mean decrease in loss across the feature's splits taking into account information from the loss function's derivatives (Chen and Guestrin, 2016). LightGBM calculated feature importance in a similar way to XGBoost but used the total decrease in loss rather than the average decrease (Ke et al., 2017). Feature importance was used to determine which variables had the highest predictive power for MMR, which was one of the primary aims of this thesis. Additionally, variation in feature importance across models was used to investigate differences in model performance.

4.7 Development of Voting and Stacking Ensembles

This section discusses how the 300 MMR predictions from the fine-tuned Random Forest, XGBoost, and LightGBM models were combined using voting and stacking ensemble methods (Figures 4.3c, 4.3d, and 4.3e).

While the fine-tuned XGBoost, LightGBM and Random Forest models discussed above are bagging and boosting ensembles, they were referred to as "base estimators" for the remainder of this thesis, with the stacking and voting ensembles used to combine them referred to as "ensemble models". Ensemble modelling was performed for both the CLP and forecasting analyses. Figure 4.4 visualises the ensemble models' inputs.

4.7.1 Development of Different Types of Voting and Stacking Ensembles

Each (country, year) sample in the input dataset was associated with a single ground truth MMR estimate. Using the previously described method, the 300 fine-tuned models would each predict sample's MMR, giving 300 predictions per sample. The ensemble model treated each prediction as a feature, learning how to combine the 300 features to produce a final, accurate MMR estimate (Figure 4.4a).

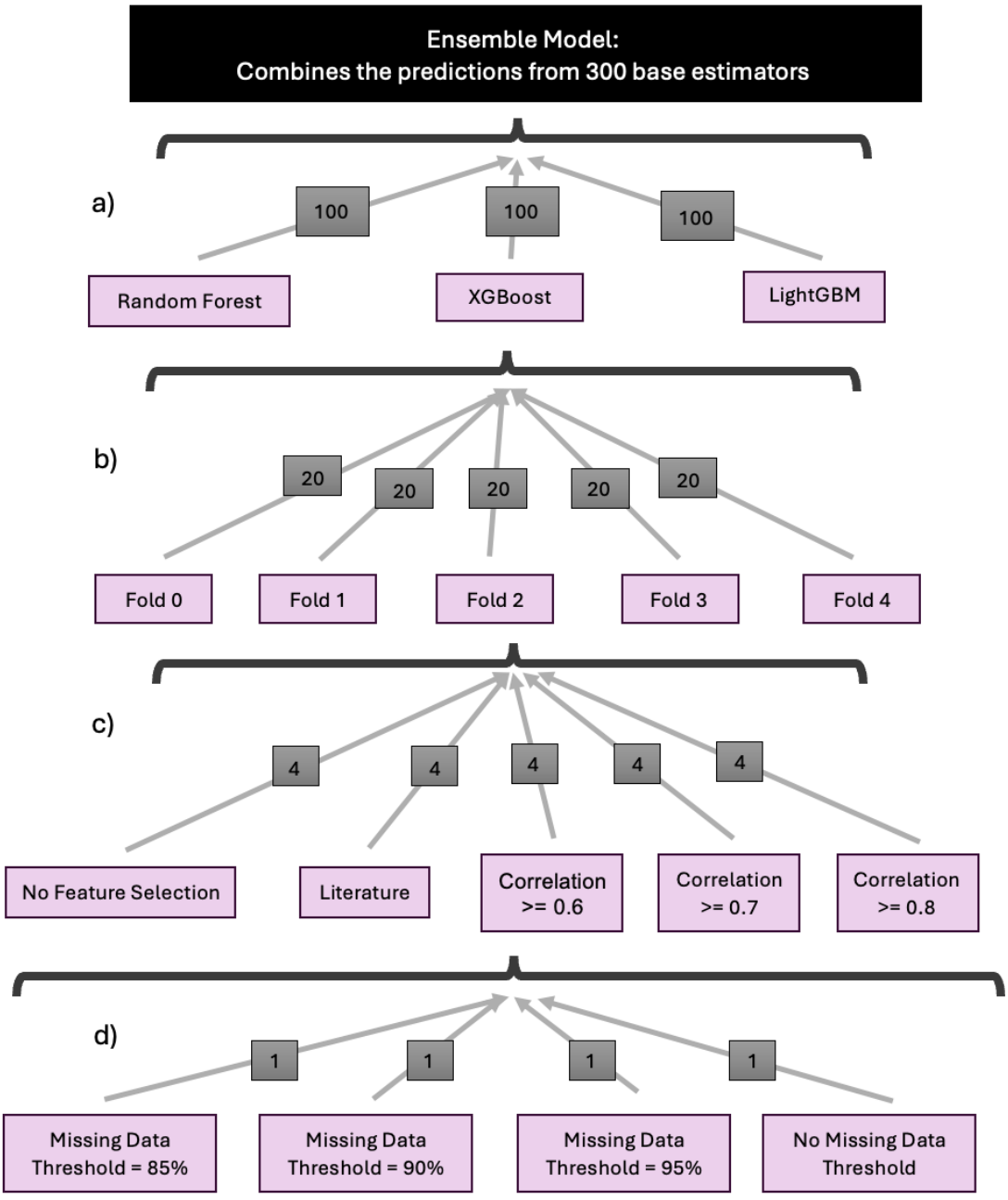


Figure 4.4: Visualisation of where the 300 base estimators used as input into the ensemble models came from. The grey boxes gave the number of base estimators trained per method. The 300 base estimators consisted of 100 Random Forest, XGBoost, and LightGBM models (a). 20 versions of each model-type were trained on each cross-validation fold (b), with the 4 models from the same fold trained on different feature subsets (c). Each of these 4 models were trained on a dataset with a different missing data threshold (d).

More specifically, the base estimators were fit on their associated training data using their best hyperparameter settings, which were determined through the fine-tuning process described above. Then, each of the fine-tuned models predicted the MMR for the full, concatenated training and validation sets. In other words, the fine-tuned models predicted on the training data before it was exposed to cross-fold validation, feature selection, and missing data removal. These predictions served as the training dataset for the ensemble models.

I used voting and stacking ensemble models to combine the base estimators' predictions, as, in their review of ensemble methods, Mahajan et al. (2023) found that, of all models tested within a study, voting and stacking ensembles most frequently had the highest performance.

Voting Ensemble

The voting ensemble model produced a weighted average of the base estimators' predictions for each country, year sample. The weighting given to each base model was determined through 1,000 Optuna fine-tuning trials, with weights ranging from 0 to 1. The weights that produced the lowest MSE on the ensemble training data were chosen as the optimal hyperparameters. No validation set was used because the voting ensemble was not 'trained', it was simply given different sets of weights to use to combine the various base estimators. Thus, the 'training data' served the same function as 'out of sample validation data', as it was 'unseen' by the ensemble.

Stacking Ensemble

The stacking ensemble method uses a meta-learner, which itself was an ML model. This meta-estimator learned how to combine the predictions from the 300 base estimators to produce the lowest predictive error. I created three versions of the stacking ensemble to compare the performance of different meta-estimators. More specifically, I used the Elastic Net linear regression model, Random Forest regressor, and Support Vector Regressor as candidate meta-learners. All models were used with the Scikit Learn's implementation (Pedregosa et al., 2011). Elastic Net was used as a progression of the voting ensemble model, where the combination of L1 and L2 regression could both perform feature selection and reduce the possibility of overfitting, as described in the background (Zou and Hastie, 2005). Additionally, as described in the literature review, decision-tree based stacking ensembles outperform base estimators solely based on bagging and boosting, thus motivating use of the Random Forest regressor as a meta-learner (Khadidos et al., 2024). Support vector regression was used as a meta-learner because its approach of only using datapoints outside its error tolerance margin could have interesting effects on how it uses the predictions from different base learners (Smola and Schölkopf, 2004).

Each meta-estimator had internal parameters that needed to be tuned, rather than just hyperparameters like in the case of the voting ensemble. Table 4.7 describes the hyperparameters tuned for each meta-learner. As a result, the training dataset had to be split

into train/validation sets. This allowed the stacking ensemble models to fit their internal parameters on the train set and fine-tune their hyperparameters on the validation set over 1,000 Optuna trials. The ensemble training data was split into train/validation sets in an 80:20 ratio using Scikit Learn's 'train_test_split' method (Pedregosa et al., 2011). The hyperparameter values that produced the lowest MSE on the validation set were used in the final stacking ensemble models. These best-performing hyperparameters can be accessed via the Optuna trial objects in the linked GitHub repository.

Table 4.7: Hyperparameters Tuned in Each Stacking Ensemble Meta-Estimator

Ensemble Model	Hyperparameter Name	Hyperparameter Function	Range of Values
Elastic Net Stacking Ensemble	Alpha	Specifies the extent of regularisation.	0.1 to 1
	L1_ratio	Controls the weighting of the L1 versus L2 norm. Higher values push the regulariser closer to the L1 norm.	0 to 1
Random Forest Stacking Ensemble	Same parameters as described in Table 4.4.		
Support Vector Machine Stacking Ensemble	kernel	Type of kernel used to transform input into a non-linear space. If 'poly', degrees tested were 2–5.	polynomial or radial basis function
	C	Strength of regularisation term.	0.1 to 1
	epsilon	Error tolerance, used to determine support vectors.	0.05 to 1

4.7.2 Evaluating the Voting and Stacking Ensemble Models

A test set was generated to be able to evaluate the voting and stacking ensembles' predictive performance on out-of-sample data. The ensembles' input test data consisted of the base estimators' predictions on their test sets. The ensemble models used these predictions to give final MMR estimates, which were compared to the test ground truth. The ensembles' models' test performance was used to assess their generalisability and determine the best ensemble.

4.7.3 Analysis of Base Estimator Importance

To better understand the variation in the ensemble models' predictive performance, I explored how each ensemble valued the contribution of the various base estimators. More specifically, I investigated whether different ensemble models placed the most importance on predictions from the same set of base estimators or on different base estimators.

As described earlier, the different base estimators can be thought of as "features" in the ensemble model. Thus, the importance placed on each base estimator by the ensemble model was quantified using the model's built-in feature importance methods. Model importance in the Random Forest ensemble was determined using the same Scikit Learn feature importance calculation described in Section 4.6.3 (Pedregosa et al., 2011). The weighting of each base estimator in the voting ensemble's final prediction was used as a proxy for model importance. Similarly, base estimator importance in the Elastic Net model was determined using the coefficient attached to each base estimator's predictions.

4.7.4 Testing the Performance of the Best Voting/Stacking Ensemble with Different Subsets of Base Estimators

After establishing the best performing voting or stacking ensemble, I tested whether its performance could be improved by using a different combination of base estimators (Figure 4.3d). I compared its test performance when its input dataset only consisted of predictions from the following subsets of base estimators for both CLP and forecasting.

- 300 base estimators consisting of XGBoost, LightGBM, and Random Forest regressors (original ensemble model).
- 100 base estimators consisting of just XGBoost regressors.
- 100 base estimators consisting of just LightGBM regressors.
- 100 base estimators consisting of just Random Forest regressors.

4.7.5 Investigating Base Estimator Selection in the Best Performing Voting/Stacking Ensemble

To gain a deeper understanding of the best performing voting/stacking ensemble, I investigated possible reasons for its choice of base estimators. More specifically, I explored whether the chosen base estimators had the lowest mean-squared error, as this was the metric used to train and fine-tune the ensembles. I compared the predictive performance of all base estimators given an importance score of at least 0.03. I also tested whether choice of base estimator was arbitrarily determined by its position in the input dataset used for the stacking/voting ensemble. I did this by permuting the base estimators' positions and re-estimating base model importance in the ensemble.

4.7.6 Feature Importance Analysis in the Best Performing Ensemble

Determining predictive power of various socio-economic and health-related variables was a primary aim of this thesis. Thus, I calculated the importance scores of the feature variables used by the most important base estimators in the highest performing ensemble (Figure 4.3e).

4.7.7 Analysis of the Best Performing Ensemble's Prediction Error by Income Level

The distribution of prediction errors per income level made by the best performing voting/stacking ensemble was visualised and analysed. This experiment gave greater insight into how the model performed in different settings.

4.7.8 Analysis of the Best Performing Ensemble's Uncertainty

There were 300 base estimator predictions for each ground truth MMR estimate. I calculated the standard deviation among the 300 predictions for every datapoint in the test set to explore the base estimators' agreement on the true MMR prediction. Lack of consensus among the base predictors would likely make the ensemble's prediction less stable. As a result, this analysis provides an approximation for uncertainty in the ensemble's predictions.

4.8 Sensitivity Analysis

I conducted a sensitivity analysis to gain a deeper understanding of how the input dataset affected the quality of the best performing ensemble's predictions (Figures 4.3f and 4.5). The sensitivity analysis was conducted in the same way both country-level prediction and forecasting.

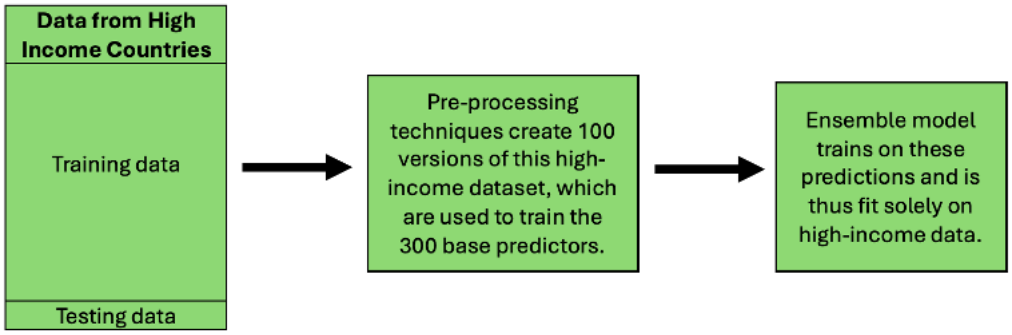


Figure 4.5: Visualisation of the sensitivity analysis procedure, where the ensemble's base estimators were fit on data from a single income level (in this case, high-income). The sensitivity analysis created a separate ensemble model for each subset of the input data corresponding to a specific income level.

To perform the sensitivity analysis, I created 4 new versions of each fold/feature subset dataset by filtering the input dataset by income level. I then conducted missing data thresholding on each filtered dataset, as above. The filtered datasets were referred to

as "sensitivity datasets" from this point forward, with the four filtered datasets characterised in the following list. This process generated 300 base estimators trained on each sensitivity dataset.

- Data from low-income countries only.
- Data from lower-middle income countries only.
- Data from upper-middle income countries only.
- Data from high-income countries only.

Each sensitivity analysis ensemble was fit on the base estimators' predictions for a filtered version of the concatenated train, validation data, which only contained data from the relevant income level. Fine-tuning followed the same procedure described above.

The sensitivity analysis ensemble models were then evaluated using a version of the test dataset filtered to only contain data from the relevant income level. Each sensitivity analysis ensemble was compared to the original best performing ensemble, which was trained on data from all income levels but only used to predict on the filtered test set. For example, the sensitivity analysis ensemble trained and tested on data from low-income countries was compared to the original ensemble trained on all data but tested on data from low-income countries. Using the same test set ensured comparability between the original and sensitivity analysis models.

4.9 Comparison to Literature

I compared my best performing ensemble, trained on data from all income levels, to the latest versions of the GMatH simulation model, BMat model, and CODEm model (Figure 4.3g). More specifically, I compared my model's MMR predictions for each country/year sample to the predictions made by models described in the literature (World Health Organization, 2025; GBD 2021 Causes of Death Collaborators, 2024; Ward et al., 2023a). The MMR estimates from the literature were given with their 95% confidence intervals.

First, I calculated the percentage difference between my test set MMR estimates and the corresponding estimates from each literature model using Eq.11. This gave an indication of the similarity between my MMR estimates and the literature's MMR predictions.

$$Percentage\ Difference = 100\% * \frac{my_{estimate} - literature_{estimate}}{literature_{estimate}} \quad \text{(Eq.11)}$$

Then, I determined the percentage of test set MMR estimates from my best performing ensemble that fell in the 95% confidence interval of the corresponding estimates from the literature. This analysis was performed for all test set estimates and per income level, where I calculated the proportion of test set estimates for countries from a particular income that fell within the 95% confidence intervals of the associated estimates from

the literature. I also calculated the proportion of ground truth MMR estimates used to test my model that fell within the 95% confidence interval of the literature models' estimates. This compared the MMR estimates that my model was trained to predict with the literature's MMR estimates to provide more information about whether differences between my model and the literature were due to poor performance or training data.

Finally, I visualised the difference between my model's MMR estimates and the literature's estimates for an exemplar country from each income level. I visualised the model's MMR estimates for country-level prediction and forecasting separately. I attempted to maximise geographic coverage by comparing estimates for at least one country in the Americas, Africa, Europe, Oceania, and Asia.

I performed this comparison for MMR estimates produced by both the country-level and forecasting models.

4.10 Note About Limited Computational Resources

Further experiments that involved training different base estimators and/or ensemble combinations of base estimators were not conducted due to a lack of computational resources. Many of the experiments described in this chapter were performed on the Gadi supercomputer, as the model training process ran too slowly to be feasible on my personal computer. However, after completing the experiments in this chapter, my team had used all the computational resources allocated to them on the supercomputer this quarter, preventing further experiments from being run. In total, 527 kilo-Service Units were used this quarter by the team, with my research using 123.22 KSUs.

Chapter 5

Analysis

In this chapter, I first present the effects of my data cleaning and pre-processing strategies (5.1) as a result of data processing. I then show the results of my exploratory data analysis (5.2) and investigation into the distribution of data between training/validation and test sets (5.3).

I next present how Random Forest, XGBoost, and LightGBM models performed when fit on training data from different cross-validation folds and generated with different feature subsets and missing data thresholds (5.4). I then explore the result of incorporating these base estimators into different stacking and voting ensembles (5.5). Finally, the architecture (5.6) and performance (5.7) of the best-performing stacking ensemble was analysed, before the MMR estimates from this best-performing model were compared to MMR estimates from the existing BMat, CODEm, and GMatH models in the literature 5.8). Models' MMR estimates for the country-level prediction and forecasting tasks were analysed separately. Section 6 discussed the results presented in this chapter.

5.1 Effect of Data Cleaning and Pre-Processing on the Input Data

The effect of the merging and cleaning processes described in the methods and overviewed in Figure 4.1a were described in this section. The raw, merged data set had 731 features and 16,948 samples uniquely identified by their country and year.

5.1.1 Effect of Data Cleaning

Removing all (country, year) samples from before 1985 and after 2018 reduced my input dataset to 725 features and 9,018 samples. The number of samples decreased further

when all samples missing an associated MMR estimate were removed, with greater proportional decreases observed for lower-income countries (Table 5.1).

Table 5.1: Number of samples per income level before and after rows with missing MMR data were removed. The proportion of samples remaining after cleaning was given as a percentage.

Income Level	**Number of Samples**		**Proportion of Samples Remaining (%)**
	Before Removing Samples with Missing MMR	**After Removing Samples with Missing MMR**	
Low	884	78	8.8
Lower-Middle	1734	310	17.9
Upper-Middle	1802	996	55
High	2176	1405	65

As a result of this pre-processing, **the final, merged dataset consisted of 720 features and 2,789 samples**.

5.1.2 Effect of Missing Data Removal During Pre-Processing

As described in Section 4.4.4, rows and columns with greater than a threshold proportion of missing data were removed to generate different versions of each fold/feature subset combination. Figure 5.1 shows how iterative data removal affected the size of the entire input dataset, which gives a rough idea of the effects of data removal per fold. My lowest missing data threshold of 85% still preserved a fair amount of missing data (61%), staying true to the data sparse conditions of countries without robust data collection systems. Higher missing data thresholds retained a larger number of rows and columns. Decreasing the missing data threshold from 95% to 90% had a large impact on the number of rows (2568 to 2070) but only a small effect on the number of columns (611 to 610), indicating that missing data was more likely due to a data-sparse sample than a data-sparse feature.

Figure 5.1 explains why I did not use stricter missing data thresholds, as they would have reduced the size of the dataset to less than 700 rows, which is relatively small and may not be enough for model training.

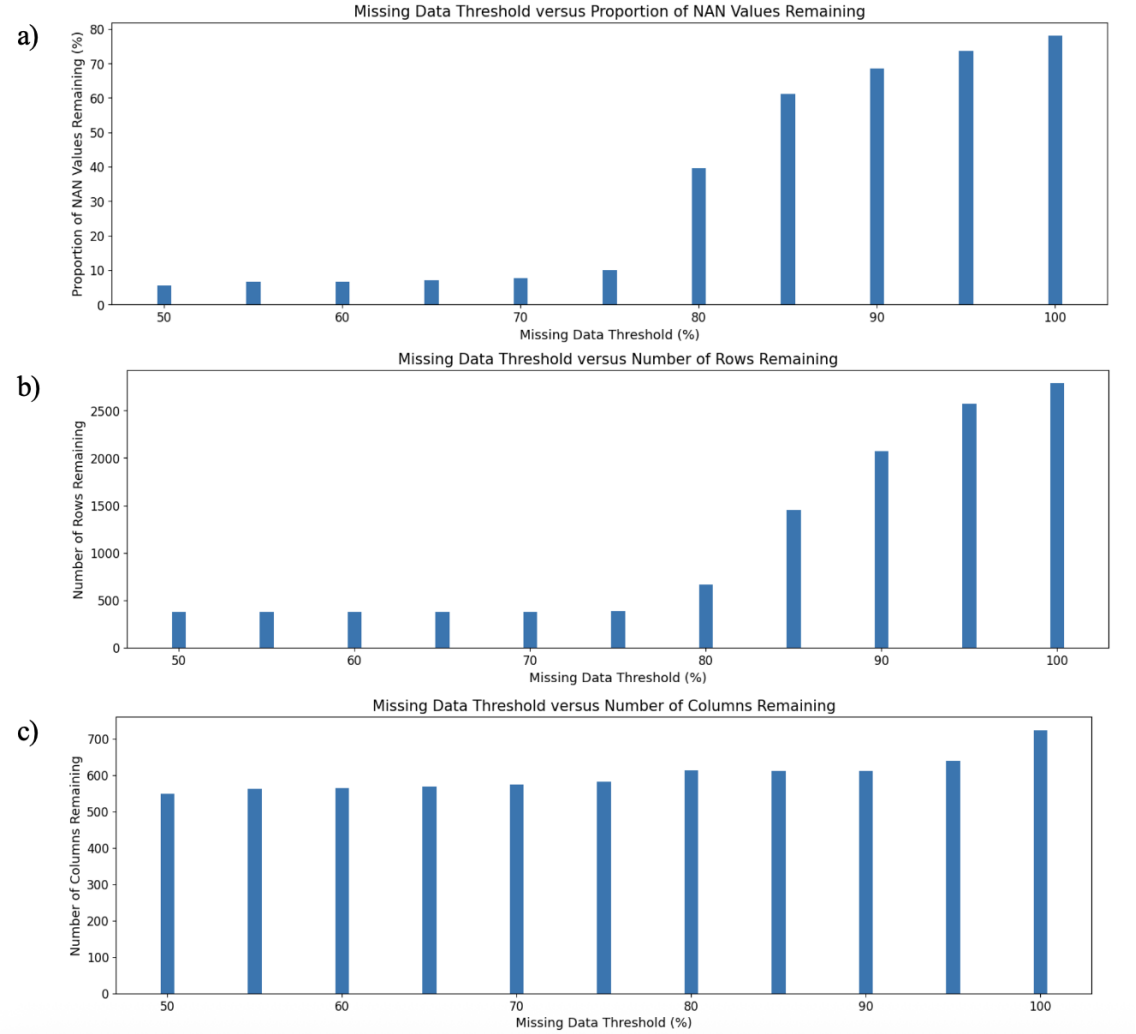


Figure 5.1: The a) proportion of missing data, b) number of rows and c) number of columns remaining in the full input dataset (not split into folds or feature subsets) after missing data removal for thresholds between 50% and 100% (no missing data removed).

5.2 Exploratory Data Analysis

The results of my exploratory data analysis presented this section contextualised model performance, as discussed in Chapter 6.

5.2.1 Analysis of Trends in Missing Data

The proportion of (country, year samples) missing an associated MMR estimate out of all samples from the same income level was referred to as "the proportion of missing

estimates" in the following analysis. This proportion varied widely across income levels, with the greatest difference observed between lower-middle and upper-middle income countries (Figure 5.2). The proportion of missing estimates decreased as income level increased. Additionally, this proportion decreased between 1985 and 2010 for each income level. For example, between 1985 and 2010, the proportion of missing estimates decreased from 50% to 35% in the low-income data, 45% to 35% in the lower-middle income data, 31% to 16% in the upper-middle data, and 19% to 14% in the high-income data. The proportion of missing estimates started increasing for all income levels post-2011, with the greatest increases observed in high and upper-middle income countries (38 and 30 percentage points, respectively).

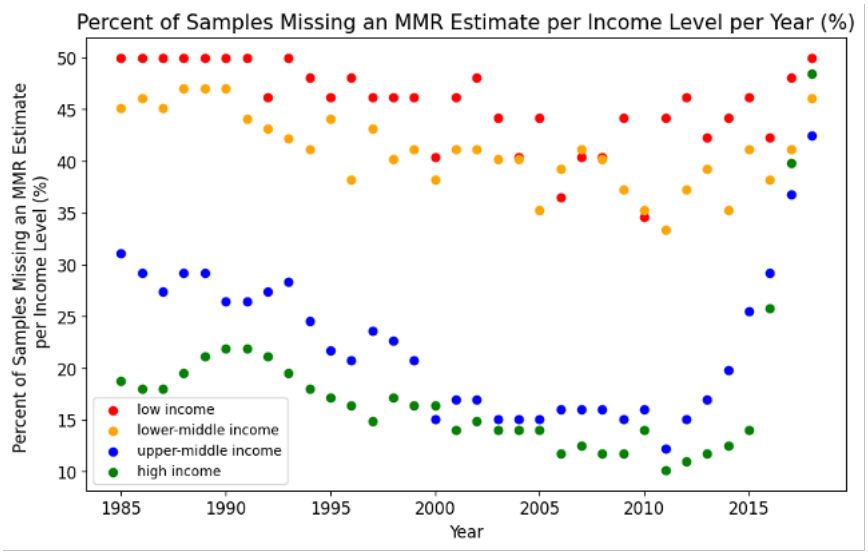


Figure 5.2: The percent of samples in the input dataset missing MMR estimates before cleaning or pre-processing. Results were presented per year between 1985 and 2018 and per income-level (red for low-income, orange for lower-middle, blue for upper-middle, and green for high).

Figure 5.2 visualises the proportion of missing feature data per year for each income level between 1985 and 2018.

Before 2000, the dataset had close to or greater than 90% missing data. Between 2000 and 2018, the dataset generally had 80 to 90% missing data. For 4 years (2000, 2005, 2010, 2015), the proportion of missing data was only between 22 and 35% (Figure 5.2). There was little difference between the proportion of missing feature data across the different income levels.

Having a few years with substantially less missing data than the norm was likely due to a group of indicators being reported with a periodicity of 5 years. This pattern was considered when splitting the data into train/test subsets, where at least one year of low missing data was used in the test set (see Section 4.4.1).

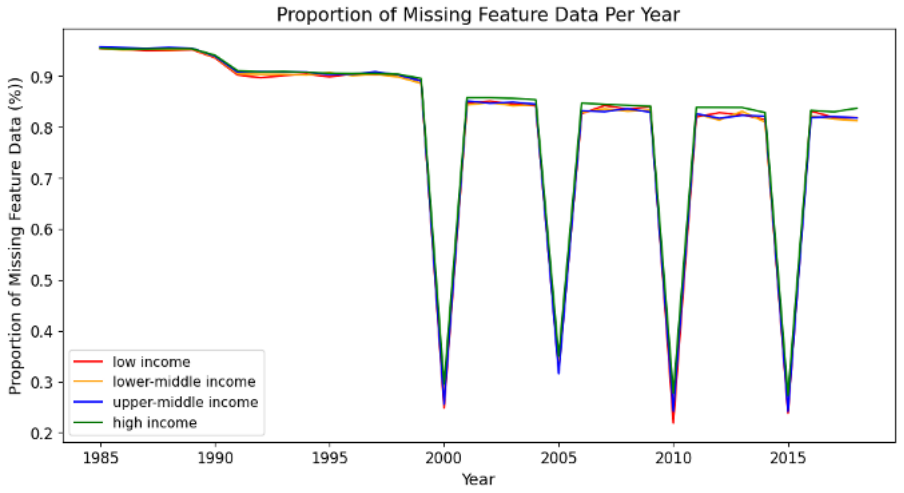


Figure 5.3: Proportion of missing feature data across all countries per year from 1985 to 2018.

5.2.2 Key Statistics in the Merged Input Data Before Pre-Processing

To better understand the input data, I calculated key summary statistics about a few of the features indicated by the literature to have a meaningful relationship with MMR (Table 5.2). Generally, health outcomes improved as income level increased. Standard deviation in the feature decreased as income level decreased. While many of the important variables had low rates of missing data, some of the socio-economic and quality of care features had increasing proportions of missing data for higher income levels. For example, the dataset for the lowest income countries was missing 58% of measurements for 'women participating in own health care decisions (% of women age 15-49)' while the highest income dataset was missing 99.9%.

The national, ground truth MMR estimates were subject to large outliers, as the mean values were larger than the median values for all income levels (Table 5.2). Additionally, standard deviation for the MMR estimates was large. The difference between mean and median, as well as the magnitude of standard deviation, decreased as income level increased.

5.2.3 Principal Component Analysis

Figure 5.4 visualises the variance captured by the feature dataset's top ten principal components. The first principal component captured 31% of total variance in the dataset. The proportion of variance captured decreased sharply to 9 and 6% for the second and third principal components before levelling out at 1.7 to 3% for the remaining top ten principal components. Thus, using the first two principal components to represent the feature data would capture approximately 40% of the data's total variance, providing an adequate representation for the purposes of exploratory data analysis.

Table 5.2: Mean, median, standard deviation and proportion of missing data of features with a meaningful relationship to MMR. The key summary statistics were presented per income level.

Feature	Income Level	Mean	Median	Std Dev	Missing Data (%)
WHO national MMR estimates (ground truth)	Low	657	617	453	0
	Lower-Middle	197	55	260	0
	Upper-Middle	51	38	55	0
	High	15	8	21	0
Infant mortality rate (per 1,000 live births)	Low	63	65	29	0
	Upper-Middle	43	39	23	0
	Lower-Middle	24	19	15	0
	High	9	7	7	2
Pregnant women receiving prenatal care (%)	Low	74	85	23	28
	Upper-Middle	81	86	18	65
	Lower-Middle	92	96	10	78
	High	93	97	8	95
Women participating in own health care decisions (% of women age 15-49)	Low	55	60	22	58
	Upper-Middle	65	67	22	86
	Lower-Middle	84	84	8.7	97
	High	91	91	NaN	99.9
Communicable, maternal, neonatal, & nutritional diseases prevalence in females (age standardized, per 100,000 population)	Low	79,399	84,661	14,140	77
	Upper-Middle	73,030	73,279	9,389	84
	Lower-Middle	62,248	63,092	10,658	86
	High	38,835	36,807	11,821	87
Survival to age 65, female (% of cohort)	Low	59	58	13	0
	Lower-Middle	71	73	12	0
	Upper-Middle	79	80	8	0
	High	87	88	5	0
Unmet need for contraception	Low	27	28	7	33
	Upper-Middle	22	23	8	72
	Lower-Middle	13	12	6	88
	High	13	10	10	97

The feature data was projected onto its first two principal components to better visualise clusters in the data. The dense cluster in the bottom center-left of Figures 5.5a and 5.5b corresponded to higher income countries with low MMRs. This cluster extended upwards to the centre-left. A country's income level tended to decrease and its MMR generally increased travelling up and to the right of Figures 5.5a and 5.5b. Yet, despite this general trend, there were datapoints belonging to upper-middle, lower-middle, and low-income countries throughout this strip of points. There were no similarly large clusters when the datapoints were coloured by year, potentially due to heterogeneity in countries' MMR values at every time point (Figure 5.5c). However, there was slight clustering at the

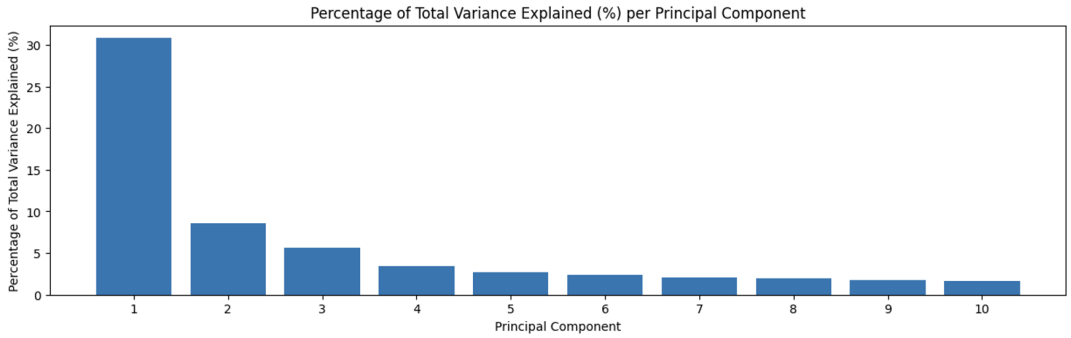


Figure 5.4: Percent of total variance in the dataset captured by the top 10 principal components.

U-shape's leftmost and rightmost edges, which corresponded to more recent years. The small cluster above the U's valley represented years further in the past.

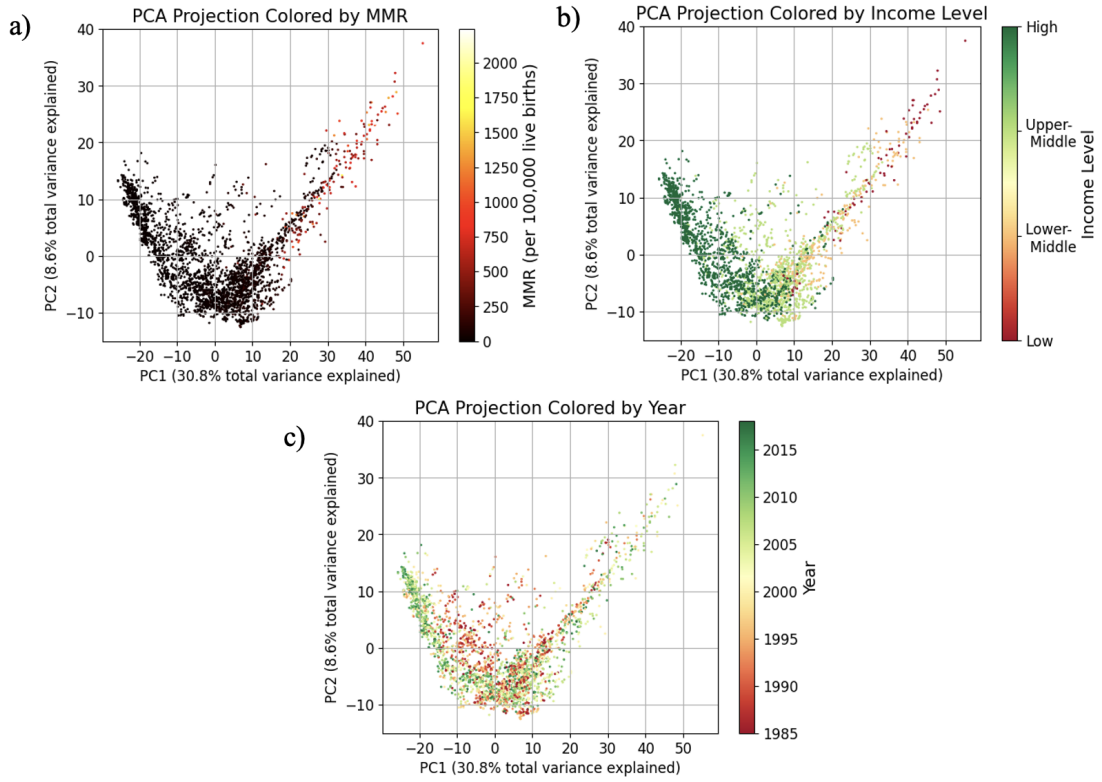


Figure 5.5: PCA projection of the feature data onto its first 2 principal components, which captured 30.8 and 8.6% of the data's total variation, respectively. The projection was coloured by the samples' a) MMR, b) income level, and c) year.

5.2.4 Correlation Analysis

While there were a broad range of correlations between features and MMR, the frequency of correlations was not uniformly distributed (Figure 5.6). More specifically, over 50% of the pairwise correlation coefficients were between -0.25 and 0.25, i.e., weak or no correlation. In contrast, approximately 2% of pairwise coefficients were less than -0.75 or greater than 0.75. The low frequency of high magnitude correlations motivated the use of feature selection methods as a possible way to reduce overfitting to noise. Use of feature selection was also motivated by the observation that there were 482 features with an absolute pairwise correlation greater than 0.9 with another feature, indicating they contained similar information. Thus, including all features may be redundant.

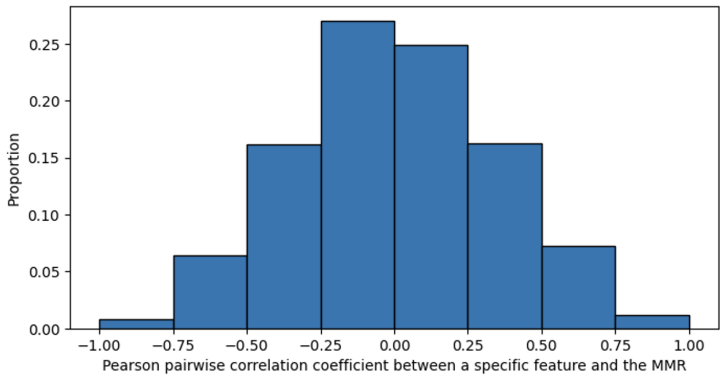


Figure 5.6: Pearson's pairwise correlation coefficient between a specific feature and MMR plotted against the proportion of features in the cleaned dataset with this correlation coefficient.

5.3 Data Distribution Between the Train/Validation and Test Sets

This sub-section compared the ground truth MMR distribution in the train/validation and test sets to provide foundation for later discussions about model performance. This comparison was performed separately for the datasets used to build country-level prediction and forecasting models.

5.3.1 Train/Test Split When Training Models to Perform Country-Level Prediction

To train models for country-level prediction (CLP), the input data was split into train/validation and test sets by country. As described in Section 4.4.1, this meant that all data from a specific country was placed in either the train/validation or test set. This split was performed for each income level, with the income-level specific train/validation and test sets merged into a single, unified train/validation and test set.

When the input data was split for CLP, the distribution of ground truth MMR values for lower-middle, upper-middle, and high-income countries was generally similar across the train and test sets (Figure 5.7). In fact, the data for lower-middle countries in the train/validation and test sets had the same Q2 MMR (52). Similarly, the Q2 MMRs for upper-middle and high-income data only differed by 3 and 1, respectively, between the train/validation and test sets. The Q1 values for these countries' test datasets were marginally greater than for their train/validation sets (e.g. 41 versus 33 for lower-middle countries). In contrast, the Q3 and maximum MMR values in these countries' test data were smaller than in their train/validation data. Consequently, the Q1 to Q3 MMR values of the test datasets were completely within the train/validation sets' Q1 to Q3 range. The train/validation set for these three income levels also contained outliers with higher MMR values than the associated test sets.

The differences between train/validation and test MMR data were greater for lower-middle income countries than upper-middle and high-income. For example, data for lower-middle income countries had a train/validation Q3 MMR 223 points higher than the associated test Q3 MMR.

While the distribution of ground truth MMR values in the low-income data mostly overlapped between the train/validation and test sets, the train/validation distribution was smaller than the test distribution. More specifically, the train/validation data had a Q2 MMR value of 610 while the test set had a Q2 of 772. The test set's Q3 and Q1 similarly exceeded the train/validation set's Q3 and Q1 by 126 and 103, respectively. As a result, the test set's MMR distribution was shifted higher than the train set's distribution. Given the small number of low-income samples, which cover a wide range of MMR values (Table 5.2), samples with high MMRs may have been included in the test set by chance. However, the maximum MMR value in the low-income train/validation data was greater the maximum in the corresponding test set.

5.3.2 Train/Test Split When Training Models to Perform Forecasting

To train models to perform forecasting, the input data was split into train/validation and test sets by year. As detailed in Section 4.4.1, the train/validation data consisted of all samples from 1985 to 2014, while the test data consisted of samples from 2015 to 2018.

When splitting the input data into train/validation and test sets by year, the income-level specific test MMR distributions were similar to the corresponding train/validation distributions (Figure 5.8). For example, the low-income data's train/validation set had only slightly smaller Q1 and Q3 values than its associated test set (404 vs 466 and 860 vs 887, respectively). The train/validation set filtered for lower-middle income data had larger Q1 and Q3 values than its associated test set (36 vs 34 and 316 vs 237, respectively). Similarly, the high-income data's Q1 and Q3 values were marginally larger in the train/validation set (4 vs 2 and 17 vs 10, respectively). The test set filtered for upper-middle income data had a slightly higher Q3 and lower Q1 than the associated

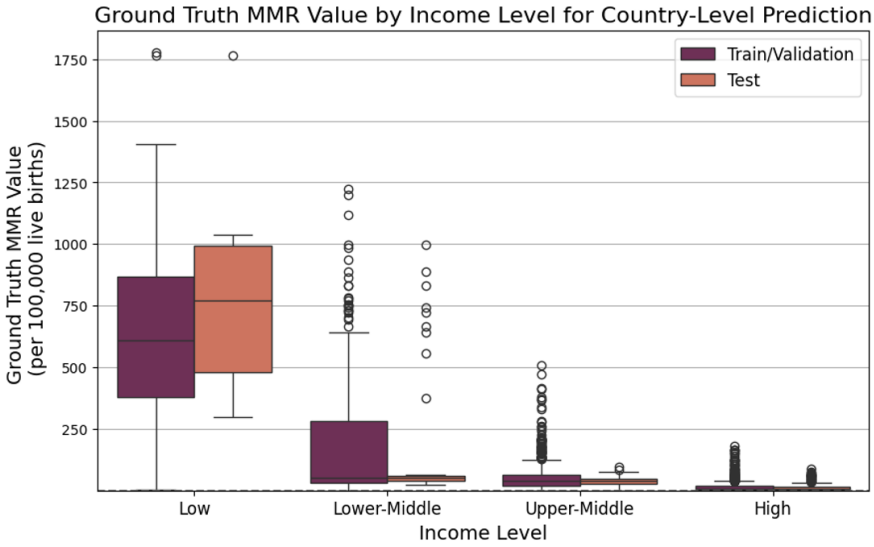


Figure 5.7: Boxplots of the distribution of ground truth MMR values across the train/validation and test datasets for different income levels. These datasets were used to train models for country-level prediction.

train/validation set, meaning the test set completely encompassed the train/validation set (63 vs 61 and 15 vs 19, respectively).

The test set's Q2 MMR value was lower than the train/validation set's Q2 MMR value for all income levels. The difference was greatest for the low-income data, where the test set's Q2 was 104 points larger than the train/validation set's Q2. In contrast, this difference ranged between 3 and 15 for lower-middle, upper-middle, and high-income countries.

The test set filtered for low-income data had median MMR values within the spread of the associated train/validation set (Figure 5.9a). However, it did not have examples of the train/validation set's sporadic decreases in MMR, potentially explaining why the test set's Q1 was higher than the train/validation set's Q1. The test set for lower-middle, upper-middle, and high-income countries all contained an outlier year (2017 or 2018) with a much higher MMR value than typically observed in the train/validation set (Figures 5.9b, 5.9c, and 5.9d). While the lower-middle income test set's outlier was contained within the train/validation distribution, the outlier years from the upper-middle and high-income test sets were not. Nevertheless, the other years in the upper-middle and high-income test sets were generally lower the train/validation set, especially for the high-income test set. This explains why the Q1 and Q2 metrics for the upper-middle and high-income test sets were lower than the Q1 and Q2 values in the associated train/validation sets, despite the presence of the outlier year.

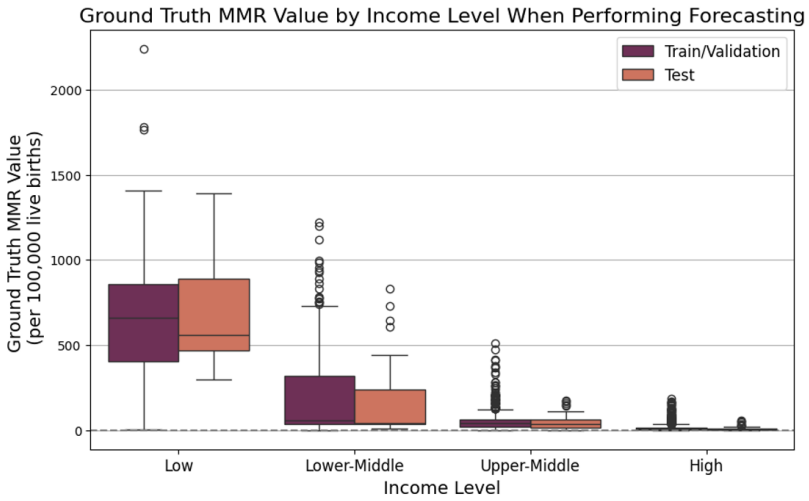


Figure 5.8: Boxplots of the distribution of ground truth MMR values across the train/validation and test datasets for different income levels. These datasets were used to train models to perform forecasting.

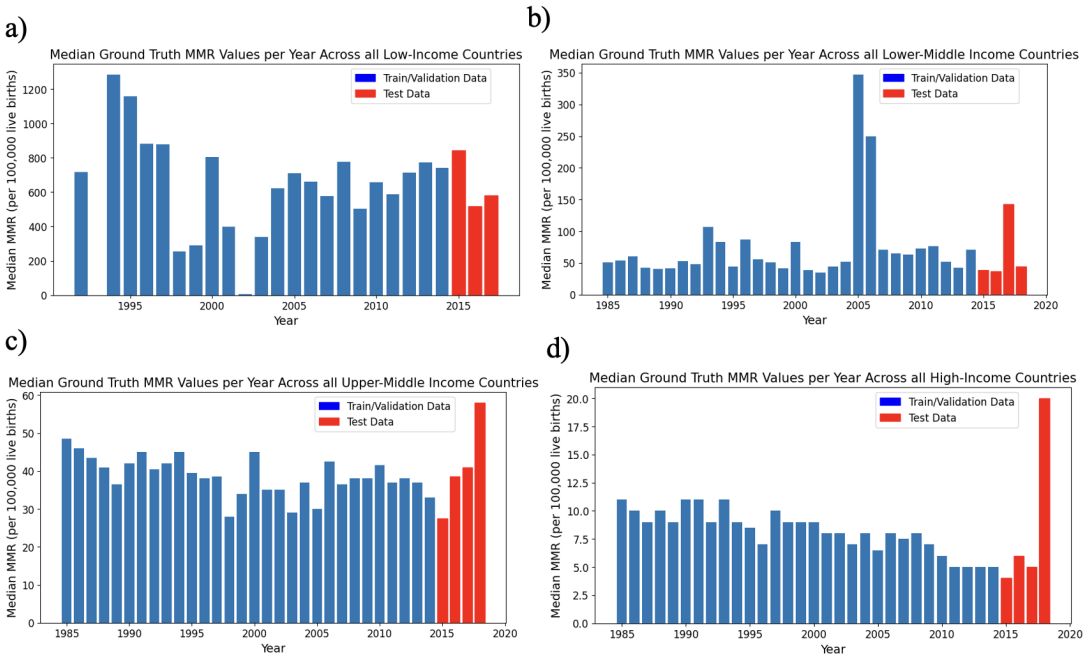


Figure 5.9: Median ground truth MMR per year for a) low, b) lower-middle, c) upper-middle, and d) high-income countries in the train/validation (blue) and test (red) sets used for forecasting.

I further investigated the outlier year (2018) in the high-income data due to the large discrepancy between the train/validation and test sets. Only two countries (Oman and Uruguay) had non-NAN test MMRs for 2018. These countries' ground truth MMRs were larger than the norm for high-income countries, consistently ranging between 14 and 25 throughout the test set (Table 5.3). In the non-outlier test years, the median ground truth MMR did not reflect these high values because it was brought down by low MMRs in other high-income countries. For example, in 2015 Australia and Norway had ground truth MMRs of 3 and 0, respectively. Thus, the large median ground truth MMR observed in 2018 for the high-income countries was not due to a change in circumstances within a certain country. It was solely due to data only being reported from countries with MMR values on the higher end of the spectrum.

Table 5.3: The ground truth MMR values for Oman and Uruguay between 2015 and 2018. These were the only two high-income countries with non-NAN data for 2018.

Country	**Ground Truth MMR Value (per 100,000 live births)**			
	2015	**2016**	**2017**	**2018**
Oman	21	20	14	23
Uruguay	19	25	19	17

5.4 Performance of Single Random Forest, XGBoost and LightGBM Models

This section discusses the performance of Random Forest, XGBoost, and LightGBM models trained on cross-validation data curated with different feature subsets and missing data removal techniques (as described in Figure 4.3). As explained in Section **??**, performance was measured using mean relative error (MRE), MSE, RMSE, MAE, and R^2. However, to keep this chapter concise, only MRE and MSE were presented, with the other metrics given in Appendix A.1. While MRE described the model's error as a proportion of the ground truth and prediction values, MSE provided information about the model's performance when predicting for outliers. Combining the two metrics provided a comprehensive measure of model performance.

5.4.1 Base Estimator Performance on Different Feature Subsets and Missing Data Removal Thresholds for Country-Level Prediction

Random Forest

Random Forest models trained on different feature subsets generally had similar predictive error on the test set, especially when accounting for standard deviation in their performance (Figure 5.10). The models' MRE and MSE typically ranged from 0.25 to

0.32 and 5,000 to 10,000 across the different feature subsets, respectively. The exception was models trained only on the 'Correlation 0.8' feature subset. In this case, models had notably lower predictive performance, with mean relative error at least 0.5 and MSE at least 28,000. Models trained on the 'Correlation 0.7' feature subset generally had the second largest errors across both metrics.

The Random Forest models with the lowest MRE (0.25) were trained on datasets formed without feature selection and with a missing data threshold of 90% (Figure 5.10a). However, the models with the lowest MSE (4,986) were trained with a missing data threshold of 85% and on the subset of features chosen based on descriptions of their significant relationships with MMR in the literature (Figure 5.10b). The set of literature-based features was most consistently associated with low MSE scores, and to a lesser degree low MRE scores. Thus, models trained on this feature subset may have had less outlier-induced error.

The relative ordering of best to worst performing missing data threshold changed when considering MSE versus MRE. For example, Random Forest models trained on a missing data threshold of 85% typically had higher MRE than models trained on higher missing data thresholds. Additionally, models trained with no missing data removal had the lowest, or tied for the lowest, MRE. In contrast, models trained with a missing data threshold of 85% had both the highest and lowest MSE scores, depending on the feature selection method. The same applied for no missing data removal. However, these comparisons must be taken with caution, as the standard deviation in both error metrics always overlapped with that of models trained on other thresholds.

XGBoost

The XGBoost models had similar trends in their performance as the Random Forest models (Figure 5.11). For example, the XGBoost models also incurred their highest MRE and MSE when trained on the 'Correlation 0.8' feature subset. Additionally, they generally had the second highest error in both metrics when trained on the 'Correlation 0.7' feature subset and tended to have lower MSE when trained with the literature-based feature subset. Similar to the Random Forest models, XGBoost models trained on a missing data threshold of 85% had the highest MRE across most feature subsets with no consistent trend observed for MSE.

Excluding the high-error models trained on the 'Correlation 0.8' feature subset, the MRE and MSE of XGBoost models ranged from 0.27 to 0.43, and 4,000 to 10,000, respectively. The MRE range's lower and upper bounds were higher than for the Random Forest models. The MSE range's lower bound was slightly smaller than for the Random Forest models. Three XGBoost models had the same lowest MRE score (0.27). They were trained on the 'Correlation 0.6' feature subset (missing data thresholds 95% and 100%) and no feature selection (missing data threshold 95%). The models with the lowest MSE (4,185) were trained with the literature-based feature subset and a 90% missing data threshold.

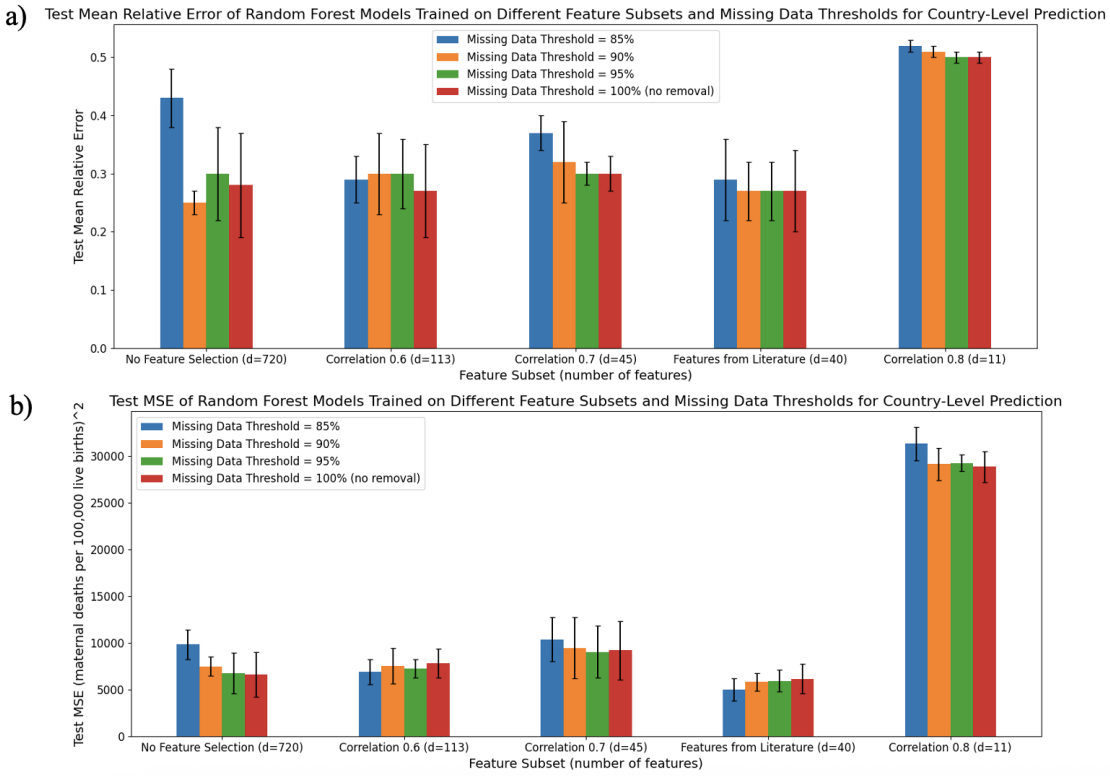


Figure 5.10: a) Mean relative error and b) mean-squared error for Random Forest base estimators fit on different feature subsets and missing data thresholds for country-level prediction.

One of the major differences between the XGBoost and Random Forest models was the magnitude of their standard deviation, with XGBoost models showing larger differences between their performance on different cross-validation folds. For example, standard deviation in the MSE of XGBoost models trained with no feature selection ranged from 2,271 to 5,037. In contrast, this standard deviation varied between 1,021 and 2,379 for Random Forest models.

As observed for the Random Forest models, there was no universally best performing feature subset or missing data threshold, especially given XGBoost models' wide standard deviations.

LightGBM

The LightGBM models had similar performance trends as XGBoost and Random Forest (Figure 5.12). For example, they had their worst performance on the 'Correlation 0.8' feature subset and among their worst performance on the 'Correlation 0.7' subset. Additionally, models trained on a missing data threshold of 85% and no feature selection had

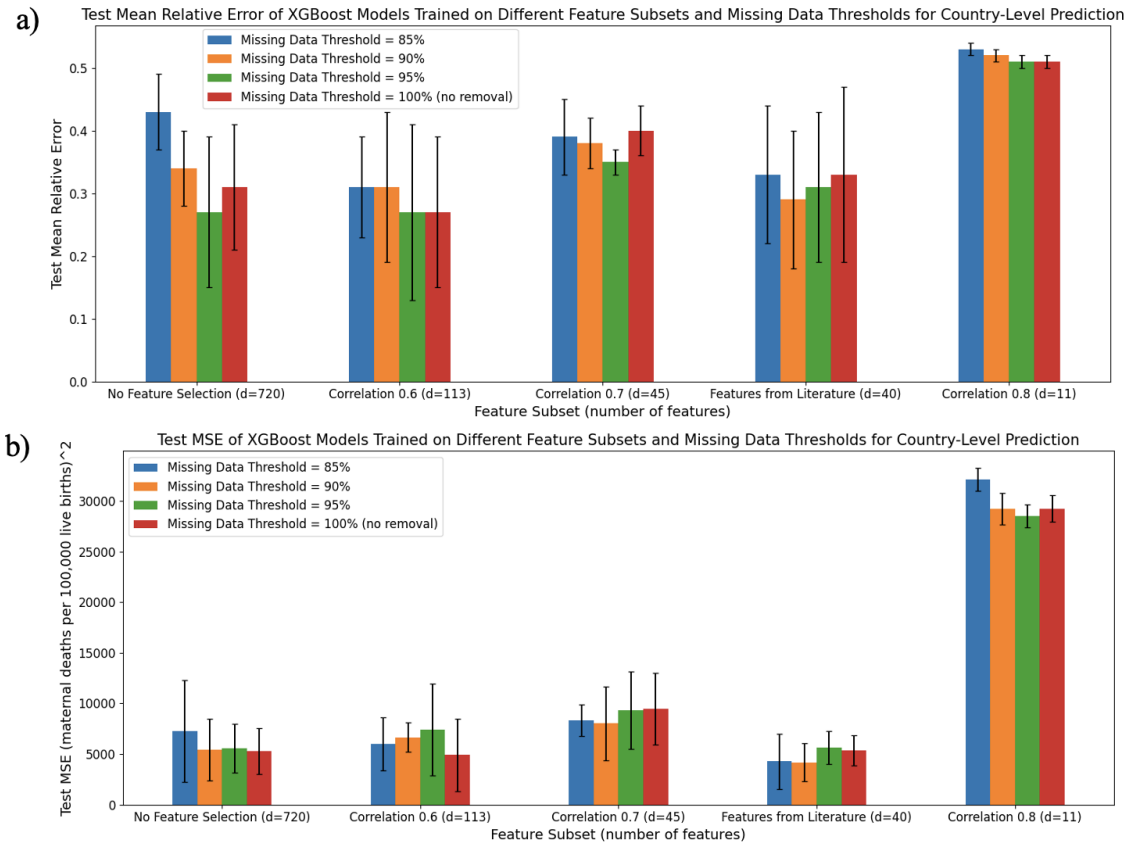


Figure 5.11: a) Mean relative error and b) mean-squared error for XGBoost base estimators fit on different feature subsets and missing data thresholds for country-level prediction.

the worst MRE performance across all three model types. As with the Random Forest and XGBoost models, LightGBM models did not have a consistently best performing missing data threshold or feature subset. However, similar to the XGBoost and Random Forest models, LightGBM models experienced higher performance more consistently on the literature-based feature subset.

Excluding performance on 'Correlation 0.8' feature subsets, LightGBM models had MRE between 0.27 and 0.49 and MSE between 6,000 and 11,000. Both ranges were higher than for the Random Forest and XGBoost models, although the lower bound of the MRE range was the same as for XGBoost. The standard deviation in LightGBM's performance was smaller than for the XGBoost models but higher than for the Random Forest models. For instance, the standard deviation in MSE for LightGBM models trained with no feature selection ranged from 777 to 3,989, compared to 2,271 to 5,037 for XGBoost and 1,021 to 2,379 for Random Forest.

The LightGBM models with the lowest MRE were trained on the literature-based feature subset with no missing data threshold (0.27). In contrast, the LightGBM models with the lowest MSE were trained with no feature selection and with a missing data threshold of 95%. These combinations of pre-processing techniques also produced the best performing XGBoost models. However, wide standard deviations in error prevented these techniques from being conclusively designated as the highest performing combination, especially given they did not produce the best performing Random Forest models.

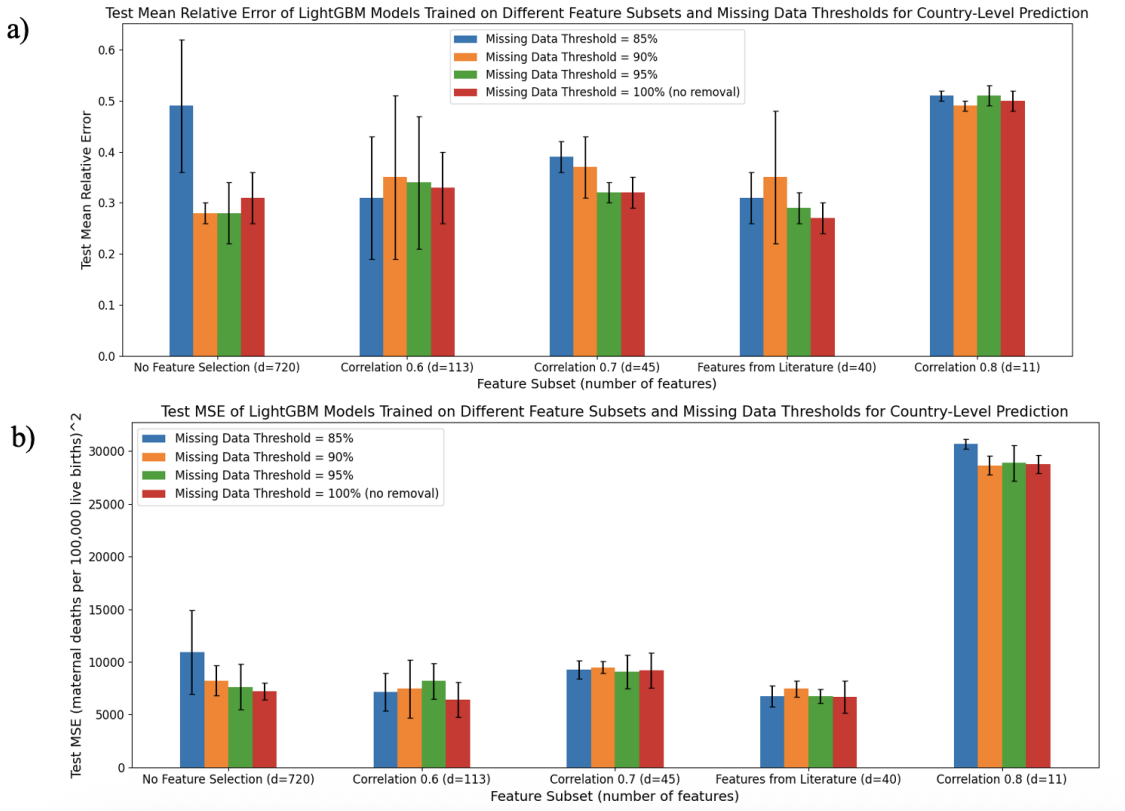


Figure 5.12: a) Mean relative error and b) mean-squared error for LightGBM base estimators fit on different feature subsets and missing data thresholds for country-level prediction.

5.4.2 Base Estimator Performance on Different Feature Subsets and Missing Data Removal Thresholds for Forecasting

Random Forest

When trained to perform forecasting, Random Forest models fit on different feature subsets had very similar MREs, with their MRE ranging from 0.37 to 0.40 (excluding models trained with the 'Correlation 0.8' feature subset) (Figure 5.13a). The Random

Forest models with the lowest MRE (0.37) were trained with no feature selection and a missing data threshold of 95%.

There was more variation in the models' MSE, indicating variation in the effect of outliers on the different pre-processing methods (Figure 5.13b). The Random Forest models had MSEs between 4,900 and 9,500, excluding errors from models trained on the 'Correlation 0.8' feature subset. The 'Correlation 0.6' subset generally produced the lowest MSE scores (all below 6,000). For example, the Random Forest models with the lowest MSE were trained on the 'Correlation 0.6' subset with no missing data removal (MSE=4,917). The 'Correlation 0.6' subset's stronger performance was more consistent when measured with MSE than MRE. Thus, it may handle outliers more effectively.

The Random Forest models trained to perform forecasting had the highest MRE and MSE scores on the 'Correlation 0.8' feature subset, like the models used for country-level prediction. As previously observed, the standard deviation in the error metrics prevented one missing data threshold from consistently producing the highest model performance.

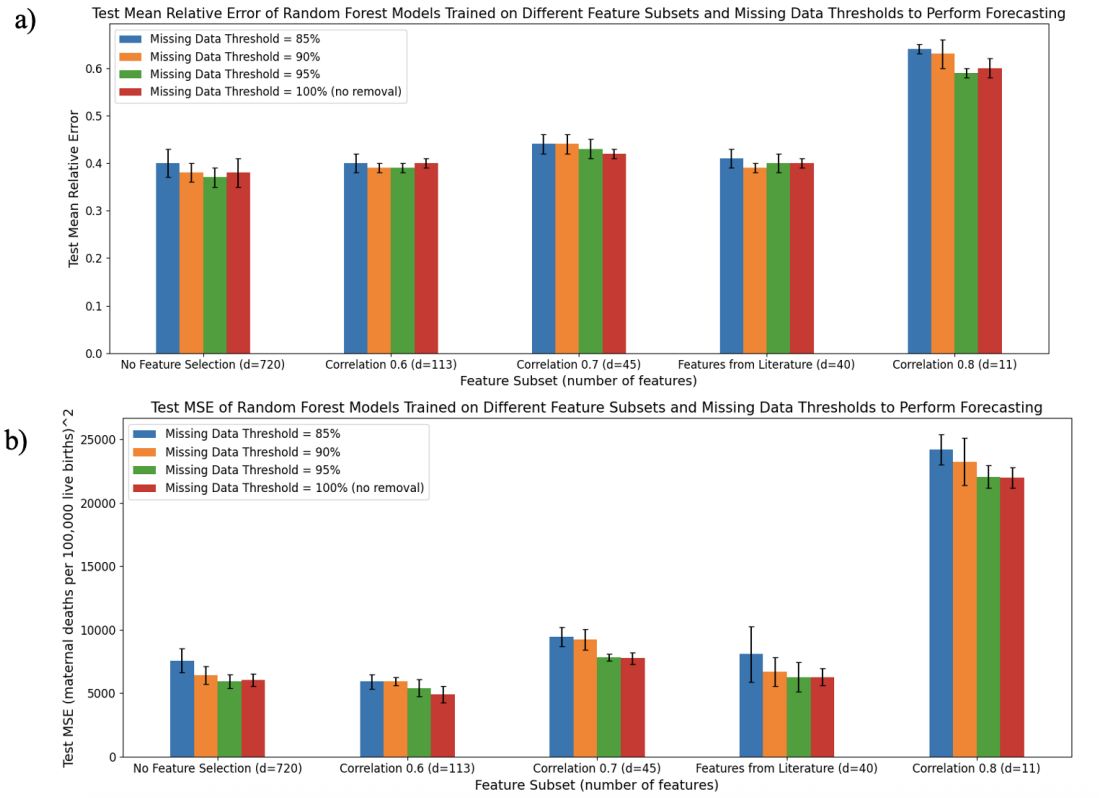


Figure 5.13: a) Mean relative error and b) mean-squared error for Random Forest base estimators fit on different feature subsets and missing data thresholds for forecasting.

XGBoost

Similar to the Random Forest models, the XGBoost models trained to perform forecasting had very similar MRE scores across the different feature subsets (Figure 5.14a). More specifically, when excluding the low performance 'Correlation 0.8' feature subset, their MRE ranged from 0.42 to 0.49, which was higher than that observed for the Random Forest models used for forecasting. This lack of variation made it difficult to identify a feature subset and missing data threshold that consistently had the lowest MRE, especially when taking into account the standard deviation in each error estimate across the cross-validation folds.

The MSE score for XGBoost models ranged from 6,100 to 11,200 (excluding models trained on the 'Correlation 0.8' feature subset) (Figure 5.14b). This range had larger lower and upper bounds than the MSE range for the Random Forest models. In general, XGBoost models also had higher standard deviation in their MSE scores than the analogously trained Random Forest models (497 to 3,734 versus 270 to 2,188). As observed for these Random Forest models, the XGBoost models trained on the 'Correlation 0.6' feature subset generally had lower MSE scores (all less than 7,000). Models trained on the literature-based subset also had lower error.

Three XGBoost models tied for the lowest MRE (0.42). They were trained on datasets with no feature selection (missing data thresholds 85% and 95%) and the 'Correlation 0.6' feature subset (missing data threshold 85%). The XGBoost model with the lowest MSE (6,163) was trained on data with the 'Correlation 0.6' feature subset and a missing data threshold of 95%.

LightGBM

As observed for the Random Forest and XGBoost models, the LightGBM models had relatively uniform MRE scores across different versions of the input dataset (Figure 5.15a). Excluding models trained on the 'Correlation 0.8' feature subset, the LightGBM models had MRE scores between 0.44 and 0.54. This range was higher than for the Random Forest models and had greater upper and lower bounds than the XGBoost models' MRE. While the LightGBM MRE range was wider than the XGBoost and Random Forest MRE ranges, it was still relatively small. In combination with the large standard deviations in the error metrics, this meant that no single feature subset or missing data threshold consistently had the highest performance.

The LightGBM models' MSE ranged from 4,773 to 9,156, excluding models trained on the 'Correlation 0.8' feature subset (Figure 5.15b). This was similar to the Random Forest models and lower than the XGBoost models. The standard deviation in the LightGBM models' MSE ranged from 546 to 1,336. This range was fully contained within the analogous ranges for the XGBoost and Random Forest models. The 'Correlation 0.6' feature subset produced low MSE scores more consistently than any other feature subset, as observed for the XGBoost and Random Forest models. LightGBM models also had more consistently low error when trained on the literature-based feature subset, like

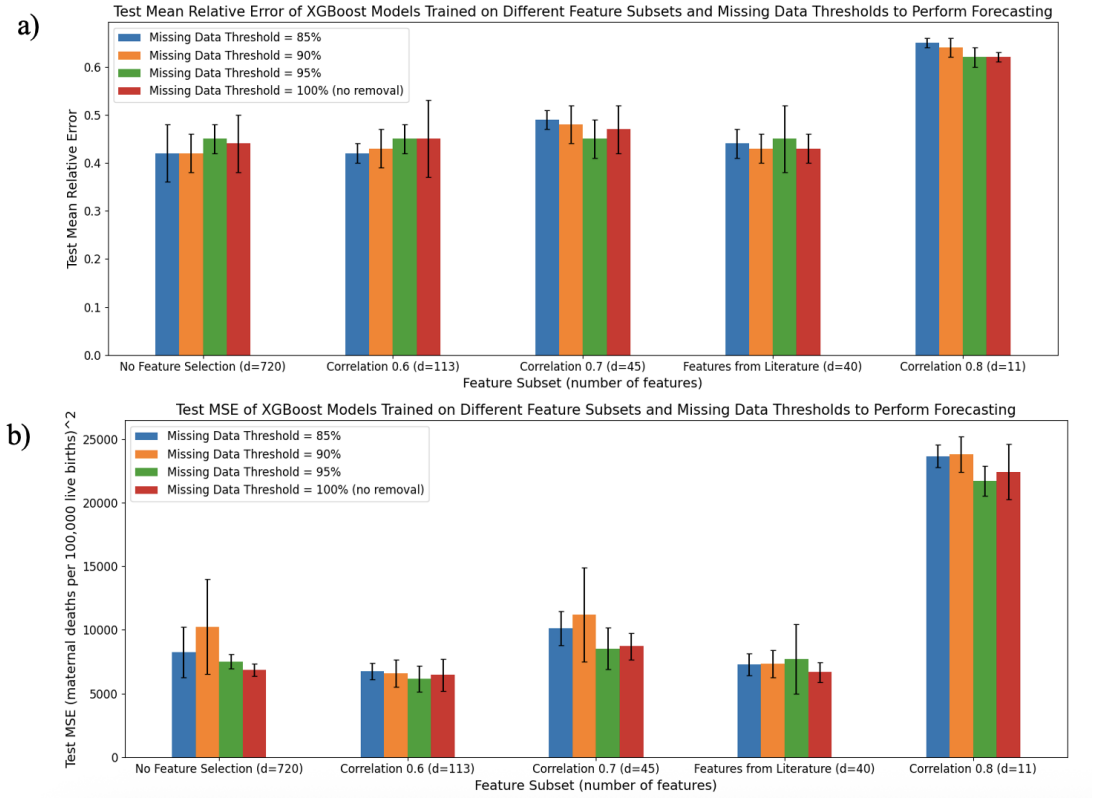


Figure 5.14: a) Mean relative error and b) mean-squared error for XGBoost base estimators fit on different feature subsets and missing data thresholds for forecasting.

the XGBoost models but unlike the Random Forest. No single missing data threshold consistently produced the lowest MSE for the LightGBM models.

The LightGBM models with the lowest MRE were trained on the 'Correlation 0.6' feature subset (missing data threshold 95%) and 'Correlation 0.7' feature subset (no missing data threshold). The lowest MSE scores were also observed in LightGBM models trained on the 'Correlation 0.6' feature subset and 95% missing data threshold.

5.4.3 Comparisons Between Random Forest, XGBoost, and LightGBM Performance on Different Feature Subsets and Missing Data Removal Thresholds

In this section, I compared the Random Forest, XGBoost, and LightGBM models directly. While the following plots contained a lot of detail, the most salient information was the difference between the various models (plotted in different colours). See Appendix A.2 for comparisons between the models' MAE, RMSE, and R^2 scores.

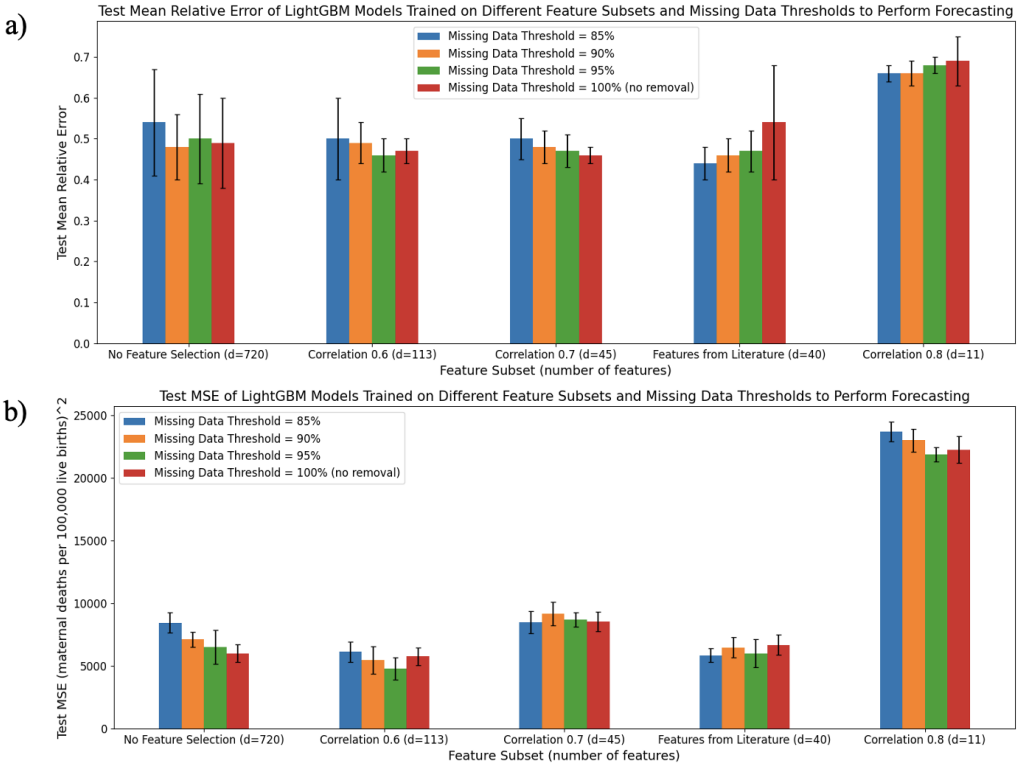


Figure 5.15: a) Mean relative error and b) mean-squared error for LightGBM base estimators fit on different feature subsets and missing data thresholds for forecasting.

Country-Level Prediction

The LightGBM and XGBoost models had the highest MRE in almost every scenario (Figure 5.16). The Random Forest models thus often had the lowest MRE across the 5 cross-validation folds. However, the standard deviation in the XGBoost models' MRE indicated they achieved lower MRE scores on specific folds when trained with no feature selection or on the 'Correlation 0.6' and literature-based feature subsets.

The XGBoost models with the lowest MSE were trained with no feature selection or on the 'Correlation 0.6' and literature-based feature subsets. While the standard deviations for the different model types overlapped, the XGBoost models' MSE standard deviation indicated higher performance on specific cross-validation folds. Generally, when XGBoost did not have the highest performance, LightGBM and Random Forest performed similarly. While the LightGBM models had the highest MSE on when trained with no feature selection or on the literature-based feature subset, they rotated with the Random Forest models for the worst MSE performance on the other feature subsets.

Despite XGBoost's high fold-specific performance, no single model type had consistently

superior performance across all data pre-processing technique combinations, especially given the overlapping standard deviation in the models' error metrics (Figure 5.16).

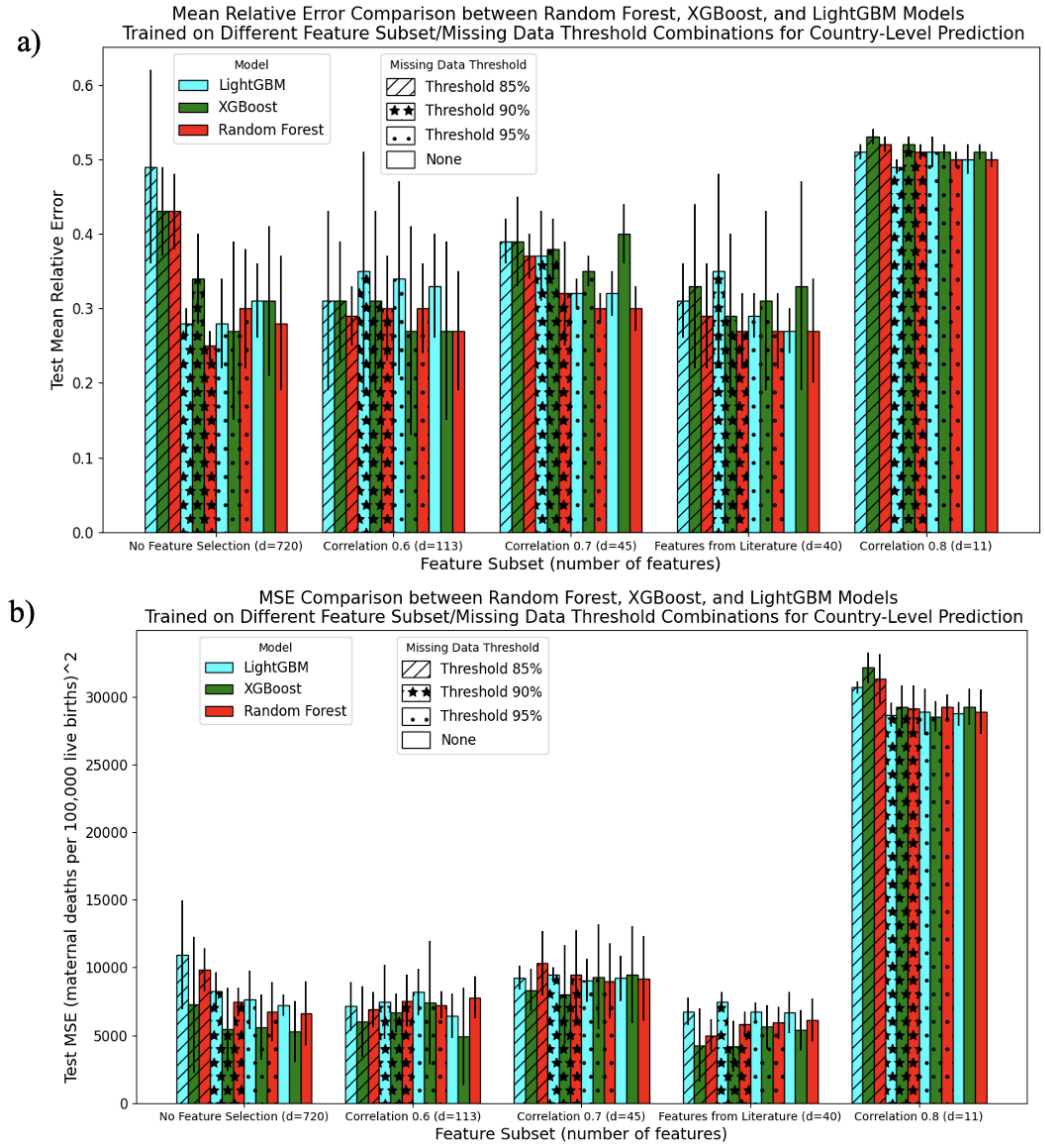


Figure 5.16: a) Mean relative error and b) mean-squared error for Random Forest (red), XGBoost (green), and LightGBM (blue) models fit on different feature subsets and missing data thresholds to perform country-level prediction.

Forecasting

LightGBM models often had the highest MRE and Random Forest models often had the lowest (Figure 5.17). Unlike the scenario described above, the XGBoost models' MRE

standard deviation did not indicate consistent fold-specific high performance.

XGBoost models trained for forecasting had either the highest or second-highest MSE, with the former occurring more consistently. The LightGBM and Random Forest models performed similarly, with strong overlap in their standard deviations.

Thus, no single model type consistently had the lowest error across both MRE and MSE (Figure 5.17). While Random Forest models had more consistently low MREs, they had similar MSEs to the LightGBM models, indicating a potential susceptibility to outliers.

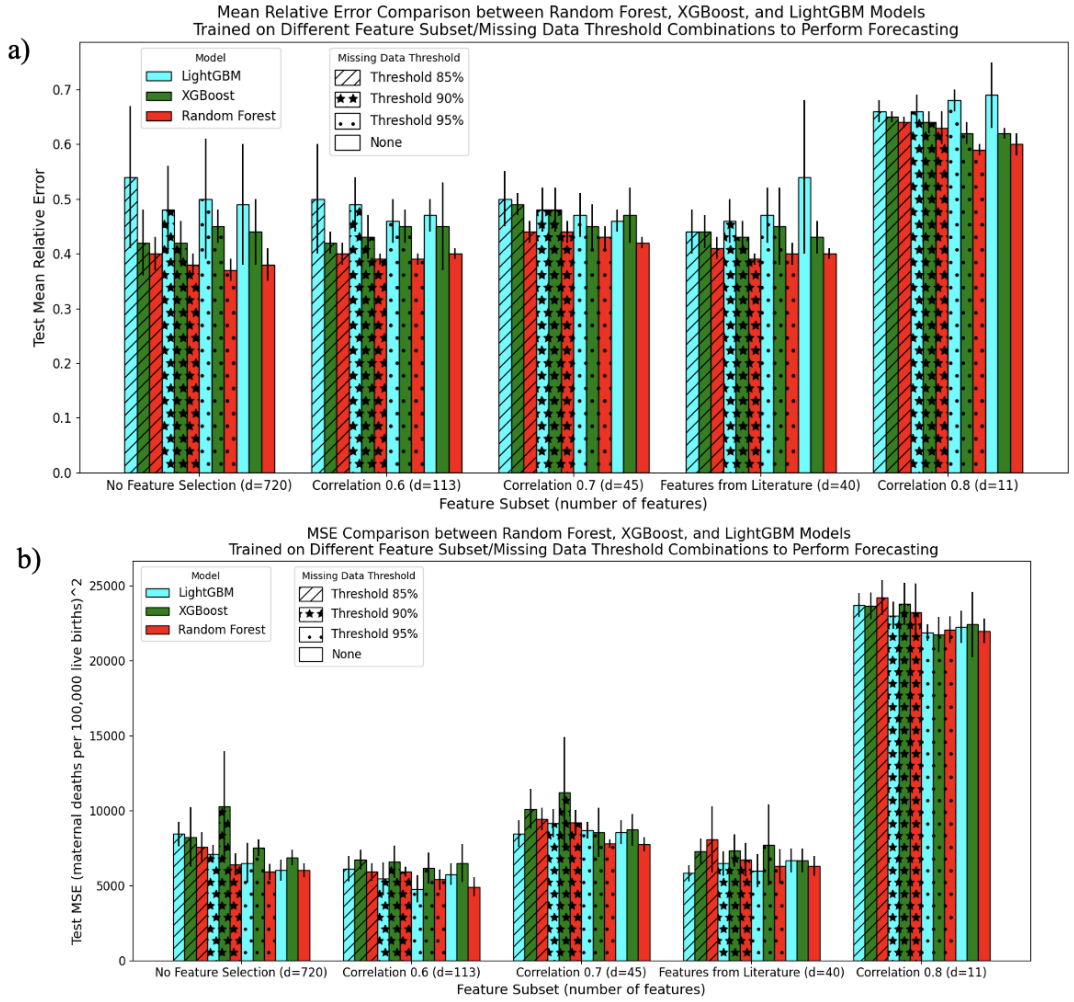


Figure 5.17: a) Mean relative error and b) mean-squared error for Random Forest (red), XGBoost (green), and LightGBM (blue) models fit on different feature subsets and missing data thresholds to perform forecasting.

5.5 Performance of Stacking and Voting Ensembles Used to Combine Base Estimator Predictions

The observation that no single model type consistently had the highest performance motivated experimentation into use of a stacking or voting ensemble to combine predictions from Random Forest, XGBoost, and LightGBM models trained on the various input datasets (overviewed in Figure 4.3c). These models were referred to as "base estimators" from this point forward. As a note, the Random Forest Stacking Ensemble was fit on the predictions of 300 base estimators. In contrast, the Random Forest models detailed above were base estimators fit on feature data.

5.5.1 Stacking and Voting Ensemble Performance When Trained on All Base Estimators

Voting and stacking ensemble performance was measured according to Section 4.7.2, with their RMSE, MAE, and R^2 scores reported in Appendix A.3.

Country-Level Prediction

The stacking and voting ensemble models trained for country-level prediction achieved MRE scores between 0.07 and 0.33 (Figure A.9a). The Random Forest Stacking Ensemble achieved the lowest MRE score while the SVM Stacking Ensemble had the highest. The stacking and voting ensembles had MSE scores between 2,161 and 7,100 (Figure A.9b), where the Elastic Net Stacking Ensemble had the lowest MSE while the Voting Ensemble had the highest.

The Random Forest Stacking Ensemble's MSE was approximately 1.3 times greater than the Elastic Net Stacking Ensemble's MSE (1,689 versus 2,161). In contrast, the Random Forest Stacking Ensemble's MRE was roughly 2.8 times smaller than the Elastic Net Stacking Ensemble's MRE (0.07 versus 0.19). Thus, the benefit of using the Random Forest Stacking Ensemble to reduce MRE was greater than the benefit of using the Elastic Net Stacking Ensemble to reduce MSE. Additionally, MRE provides a better holistic understanding model of model performance, while MSE tends to exaggerate outliers, indicating the Random Forest Stacking Ensemble had better performance on the dataset as a whole. Thus, **the Random Forest Stacking Ensemble was chosen as the best-performing ensemble.**

Forecasting

The stacking and voting ensembles trained to perform forecasting achieved MRE scores ranging from 0.37 to 0.56 and MSE scores between 5,100 and 8,000 (Figure A.10). **The Random Forest Stacking Ensemble was the best-performing model** and the SVM Stacking Ensemble was the worst-performing model in terms of both MRE and MSE.

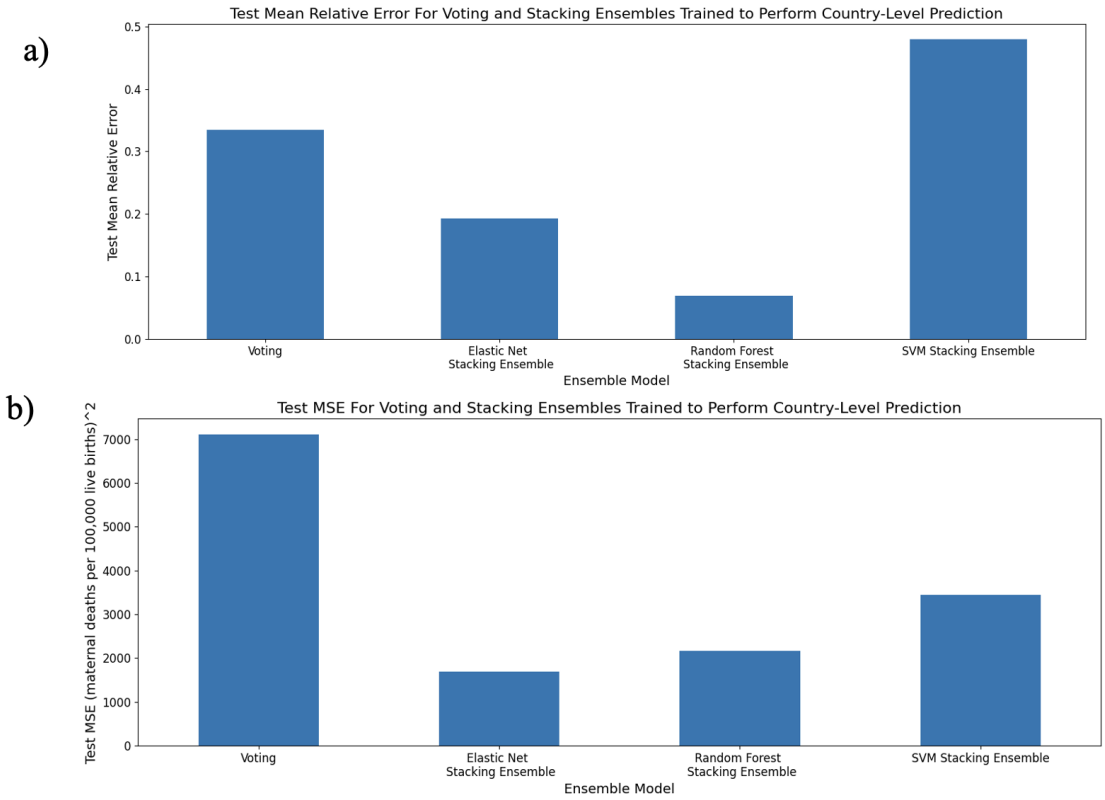


Figure 5.18: a) Mean relative error and b) mean-squared error for voting and stacking ensembles trained on all base models to perform country-level prediction.

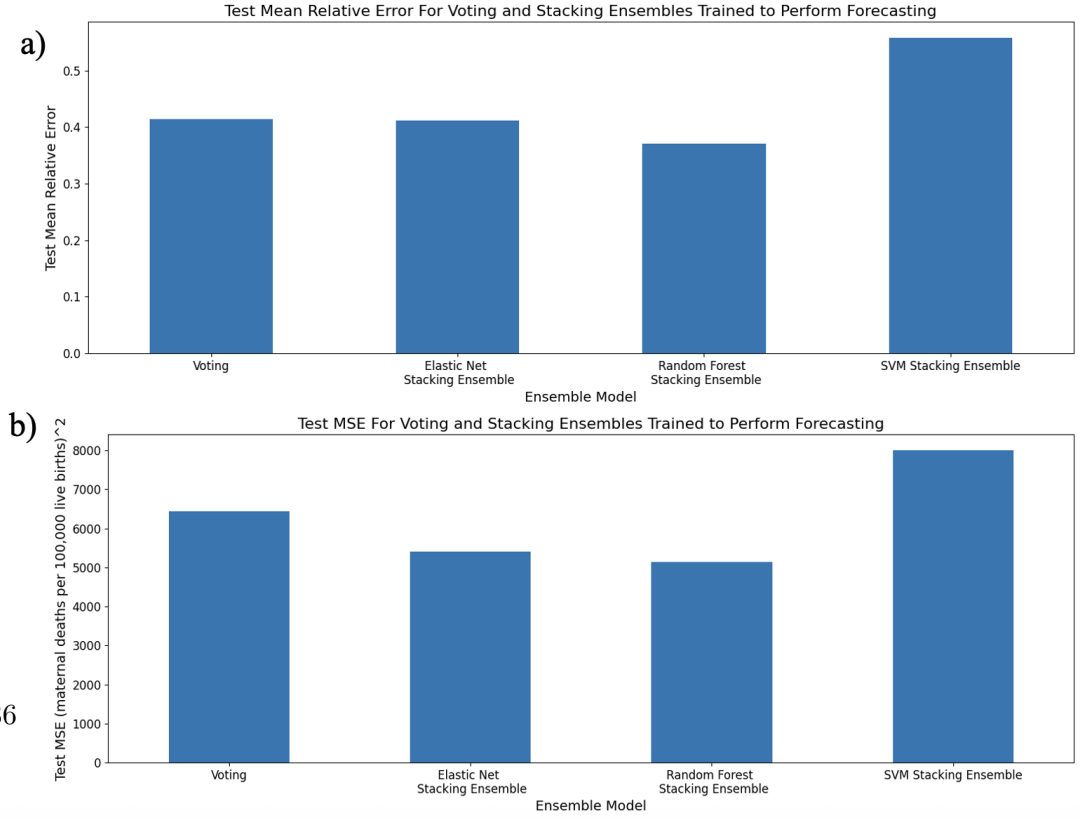


Figure 5.19: a) Mean relative error and b) mean-squared error for voting and stacking ensembles trained on all base models to perform forecasting.

5.5.2 Weighting Given to Each Base Estimator in the Stacking and Voting Ensembles

To better understand the performance differences between the various ensembles, I explored which base estimators were weighted most heavily by each ensemble. I did not further investigate the SVM Stacking Ensemble because the Scikit Learn implementation lacked a 'feature importance' method. Each of the 300 base estimators were referenced using a number between 0 and 299. LightGBM base estimators were numbered 0 to 99, Random Forest base estimators numbered 100 to 199, and XGBoost base estimators numbered 200 to 299.

The Random Forest Stacking Ensemble (RFSE) only placed importance on a subset of base estimators when it was trained for both country-level prediction and forecasting (Figure 5.20). It primarily drew strength from the XGBoost base estimators, with some support from LightGBM models. It placed very little importance on Random Forest base estimators. The RFSE used a greater number of base estimators to perform forecasting than country-level prediction.

Unlike the RFSE, the Elastic Net Stacking Ensemble derived support from most base estimators, with importance placed on all model types (Figure 5.21). This difference was shown clearly by how the Elastic Net Stacking Ensemble placed high importance on some Random Forest base estimators. However, like the RFSE, the Elastic Net Stacking Ensemble placed only a small amount of importance on a subset of base estimators. In contrast, the Voting Ensemble placed a very small, but relatively equal, amount of importance on all base estimators, with only a few base estimators contributing little to the final prediction (Figure 5.22).

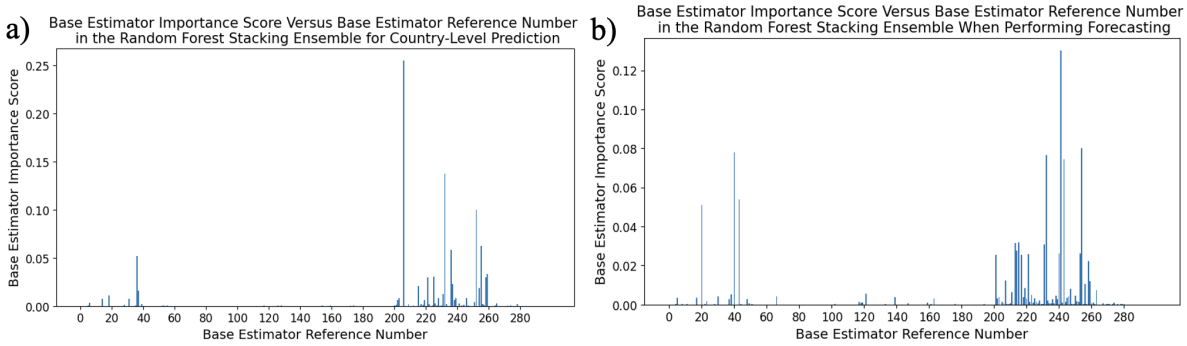


Figure 5.20: Importance score for each of the 300 base estimators used in the Random Forest Stacking Ensemble trained for a) country-level prediction and b) forecasting.

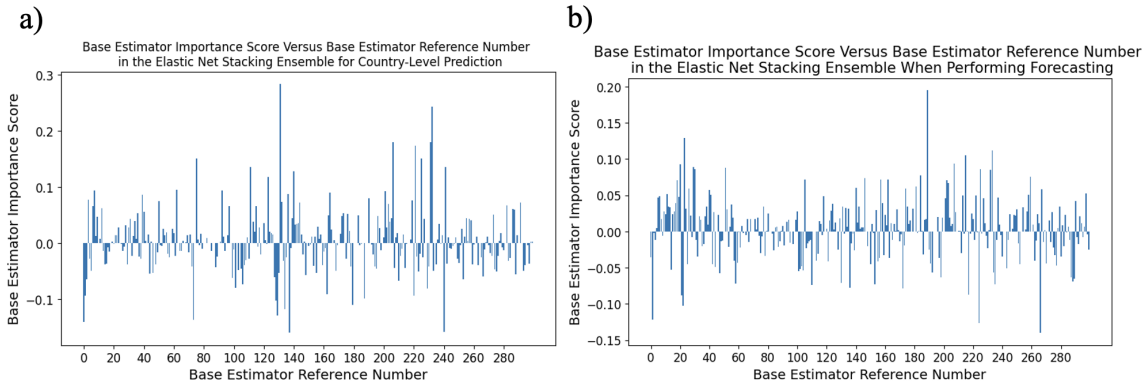


Figure 5.21: Importance score for each of the 300 base estimators used in the Elastic Net Stacking Ensemble trained for a) country-level prediction and b) forecasting.

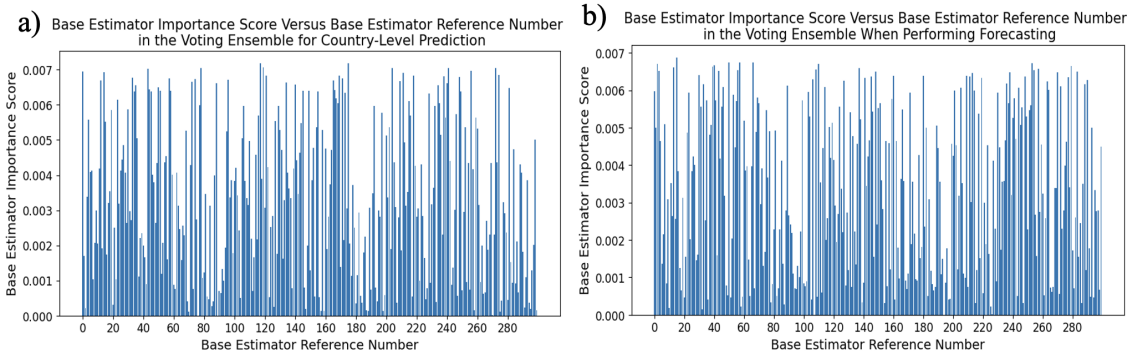


Figure 5.22: Importance score for each of the 300 base estimators used in the Voting Ensemble trained for a) country-level prediction and b) forecasting.

5.5.3 Comparison of the Best Performing Stacking/Voting Ensemble and the Single Base Estimators

As described in Section 5.5.1, the best performing stacking and voting ensemble was the Random Forest Stacking Ensemble (RFSE). Its predictive error was compared with that of its base estimators to establish whether stacking reduced error. While the following plots contained a lot of detail, the most important information conveyed was the difference between RFSE and its base estimators (light purple versus red, green, and light blue). See Appendix A.4 for comparisons using MAE, RMSE, and R^2.

Country-Level Prediction

When trained for country-level prediction, the Random Forest Stacking Ensemble had substantially lower MSE and MRE than the analogously trained base estimators (Figure

5.23). Specifically, it achieved an MRE of 0.07 compared to the best MRE achieved by a base estimator of 0.25. Similarly, the RFSE had an MSE of 2,161 while the lowest MSE produced by a base estimator was 4,185. **Thus, the Random Forest Stacking Ensemble was superior to the base estimators for country-level prediction.**

Forecasting

The Random Forest Stacking Ensemble trained for forecasting did not always produce greatly lower error than the base estimators (Figure 5.24). The RFSE's best MRE of 0.37 was the same as the MRE produced by the Random Forest base estimator trained with no feature selection and a missing data threshold of 95%. The RFSE's lowest MSE of 5,134 was larger than the MSE produced by the LightGBM base estimator (4,773) trained on the 'Correlation 0.6' feature subset with a missing data threshold of 95%.

Despite these less promising results, **the Random Forest Stacking Ensemble was still considered the 'best performing model'** because of the substantial improvement it produced for country-level prediction. The RFSE was used for forecasting as well **for consistency and because it did not increase MRE or MSE by a notable amount**, as its MSE was only 1.08 times greater than the best base estimator's MSE.

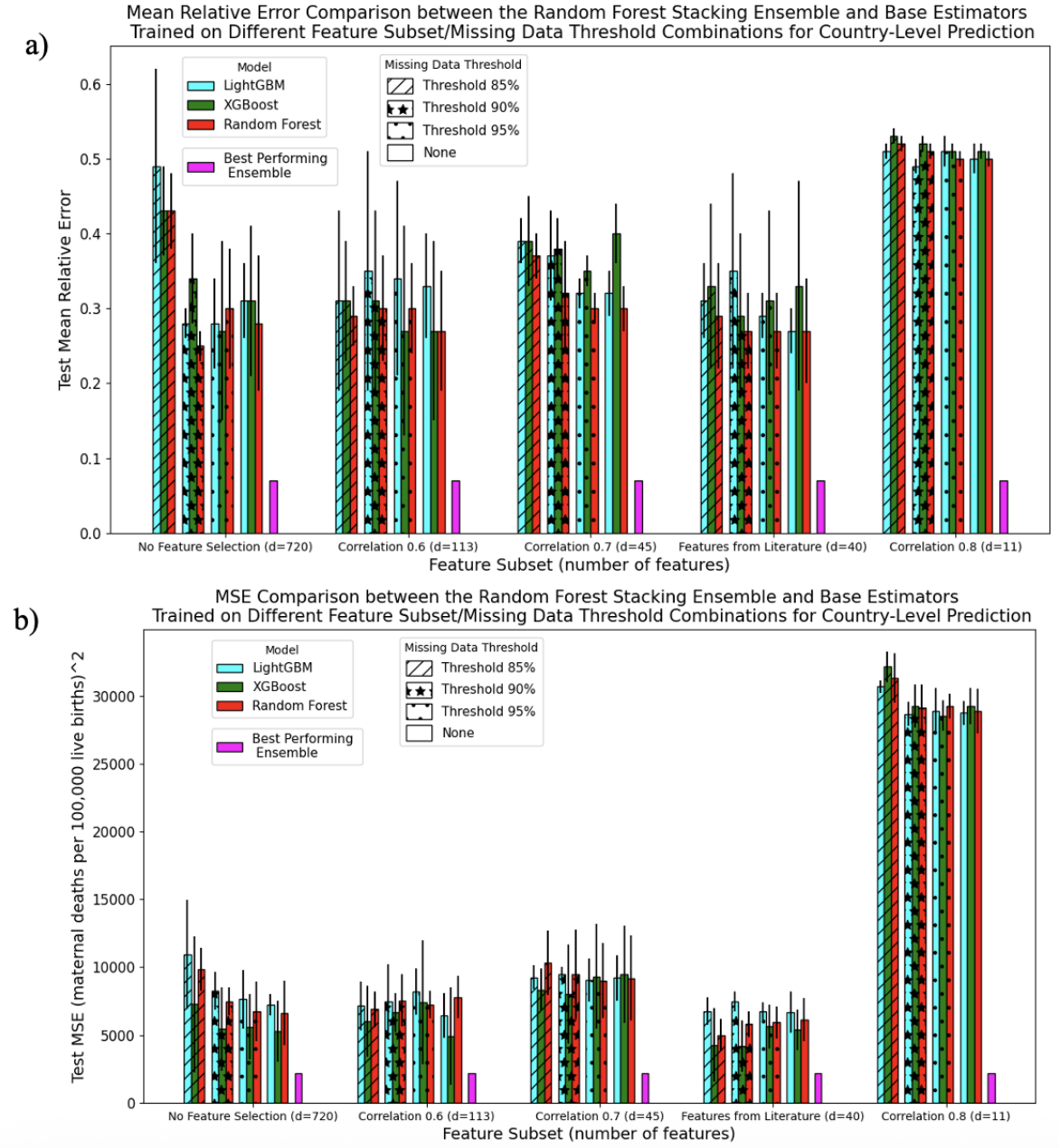


Figure 5.23: a) Mean relative error and b) mean-squared error for the Random Forest Stacking Ensemble (purple) and the Random Forest (red), XGBoost (green), and LightGBM (blue) base estimators trained for country-level prediction.

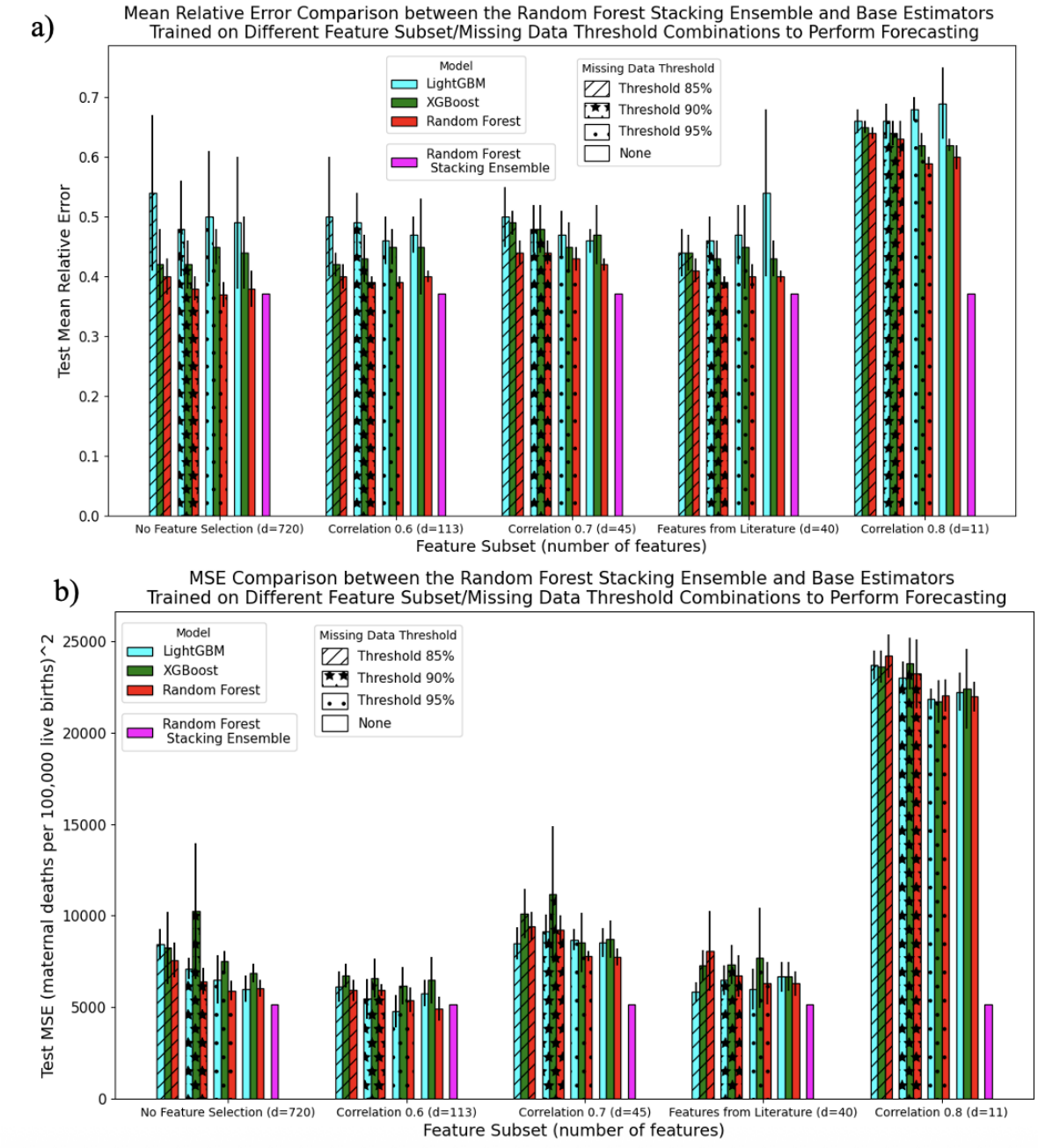


Figure 5.24: a) Mean relative error and b) mean-squared error of the Random Forest Stacking Ensemble (purple) and the Random Forest (red), XGBoost (green), and LightGBM (blue) base estimators trained for forecasting.

5.6 Investigation into the Random Forest Stacking Ensemble's Architecture

The Random Forest Stacking Ensemble's architecture was investigated further to better understand its performance and determine if it could be improved (overviewed in Figure 4.3d and 4.3e).

5.6.1 Random Forest Stacking Ensemble with Different Combinations of Base Estimators

Training the Random Forest Stacking Ensemble (RFSE) on a subset of available base estimators rather than all base estimators, as above, did not greatly improve performance of either country-level prediction or forecasting (Figures 5.25, 5.26, Appenix A.5).

When training the RFSE for country-level prediction, only using predictions from XGBoost models reduced MRE by roughly 0.3% (from 0.0695 to 0.0667) and decreased MSE by 87 (from 2,161 to 2,074). When training the RFSE for forecasting, only using predictions from Random Forest base estimators reduced MRE by about 0.19% (from 0.3708 to 0.3689) and only using predictions from LightGBM base estimators decreased MSE by 405 (from 5,134 to 4,729). All other combinations of base estimator inputs reduced performance. For example, using a different base estimator subset as input to the RFSE trained for country-level prediction increased its MRE by at least 5% and its MSE by at least 1,200.

The improvements due to using a different subset of base estimators were extremely small, especially when the improvement was put in terms of MRE. Additionally, the best subset of base estimators to use changed for each metric and type of analysis. Choosing to remain with the original RFSE model trained on all base estimators prevented the need to conduct all future experiments on three different stacking ensemble formulations. Given the lack of compute resources at the tail-end of this project, the decision was made to use all base estimators. Additionally, using all available base estimators in an ensemble model more closely follows convention. **Thus, the Random Forest Stacking Ensemble trained on predictions from all base estimators was considered the 'best-performing' model from this point forward.**

5.6.2 Importance Analysis of the Base Estimators in the Best-Performing Random Forest Stacking Ensemble

In Section 5.5.2, I discussed how the RFSE only placed high importance on a small subset of mostly XGBoost and LightGBM models. Given the decision to continue using all base estimators as input to the ensemble, I tested various hypotheses for why the specific subset of base estimators was chosen by the RFSE using the procedure outlined in Section 4.7.5.

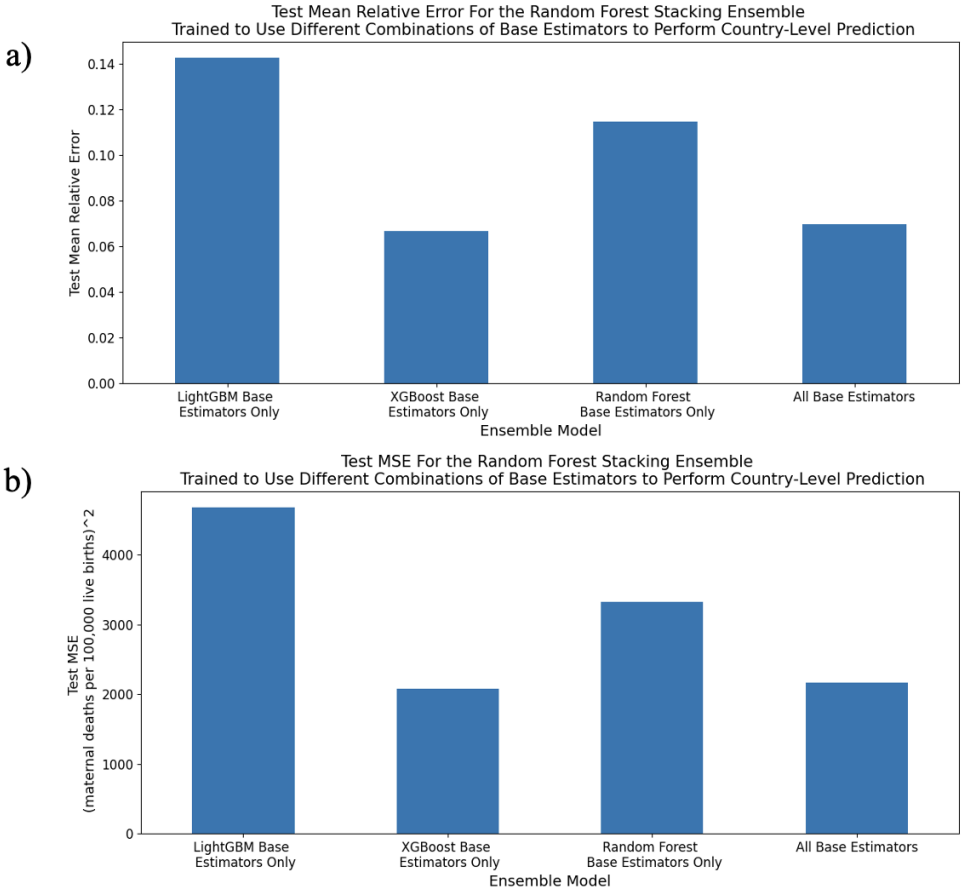


Figure 5.25: a) Mean relative error and b) mean-squared error for the Random Forest Stacking Ensemble trained to perform country-level prediction using different combinations of base estimators.

Differences in the Predictive Performance of Most Important Base Estimators

I first determined whether mean predictive error over the test set was the sole predictor of the RFSE's choice of base estimator. I compared MSE because this was the metric used to train and fine-tune the stacking ensemble. Of the 8 base estimators trained for country-level prediction that were given an importance score of at least 0.03, 7 were XGBoost models and 1 was LightGBM. While many of the MSE scores of the chosen base estimators were low, two had MSE scores of greater than 6,000 and one had an MSE of almost 10,000, which was at the higher end of the range of observed MSE scores (Figure 5.27a). Of the 10 base estimators trained to perform forecasting that were given an importance score of at least 0.03, 7 were XGBoost models and 3 were LightGBM. Again, while most of these estimators produced MSE at the bottom of the observed

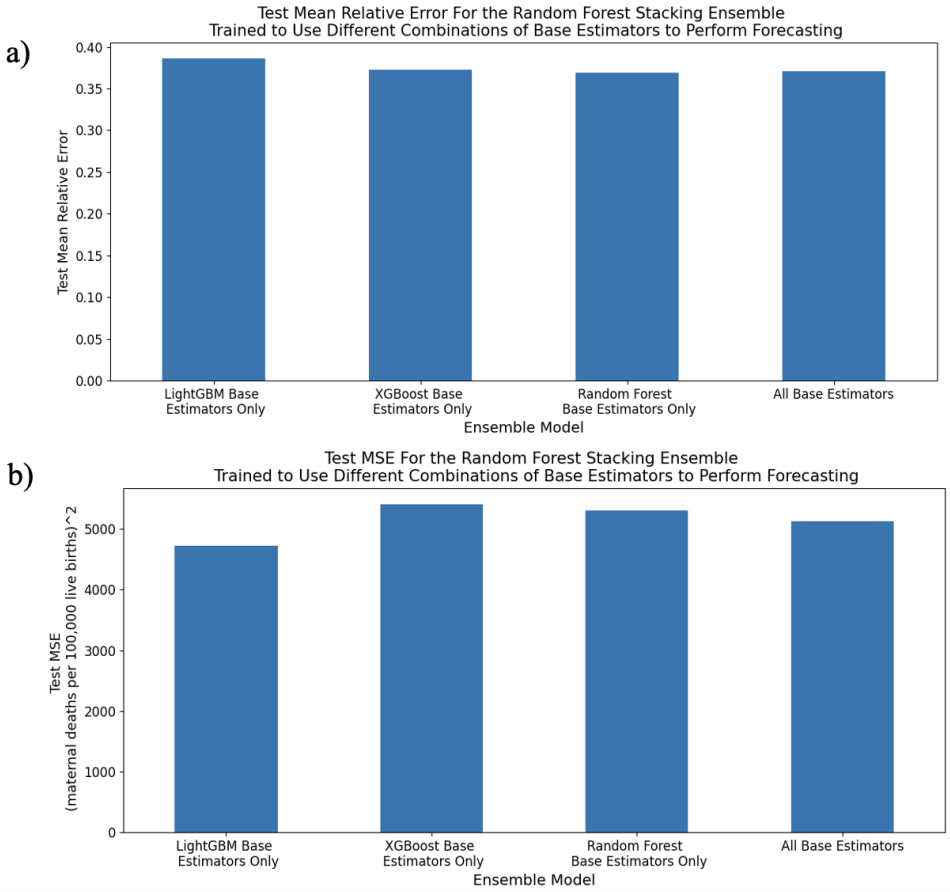


Figure 5.26: a) Mean relative error and b) mean-squared error for the Random Forest Stacking Ensemble trained to perform forecasting using different combinations of base estimators.

range, there was one XGBoost base estimator with an MSE score towards the top end of the range (close to 10,000) (Figure 5.27b). Thus, **MSE did not completely explain how the RFSE gave importance to its base estimators**.

Effect of Permutating the Order of Base Estimators in the RFSE's Input Data

I next tested whether the Random Forest Stacking Ensemble was biased in its choice of estimator. For example, by default, the first 'features' it used to create splits in its decision trees may have had specific positions in its input dataset. Practically, this would mean it first tried to create splits using predictions from base estimators located at default positions in its input data. If none of the base estimators it subsequently trialled produced a split with a lower predictive error, it would remain with the default,

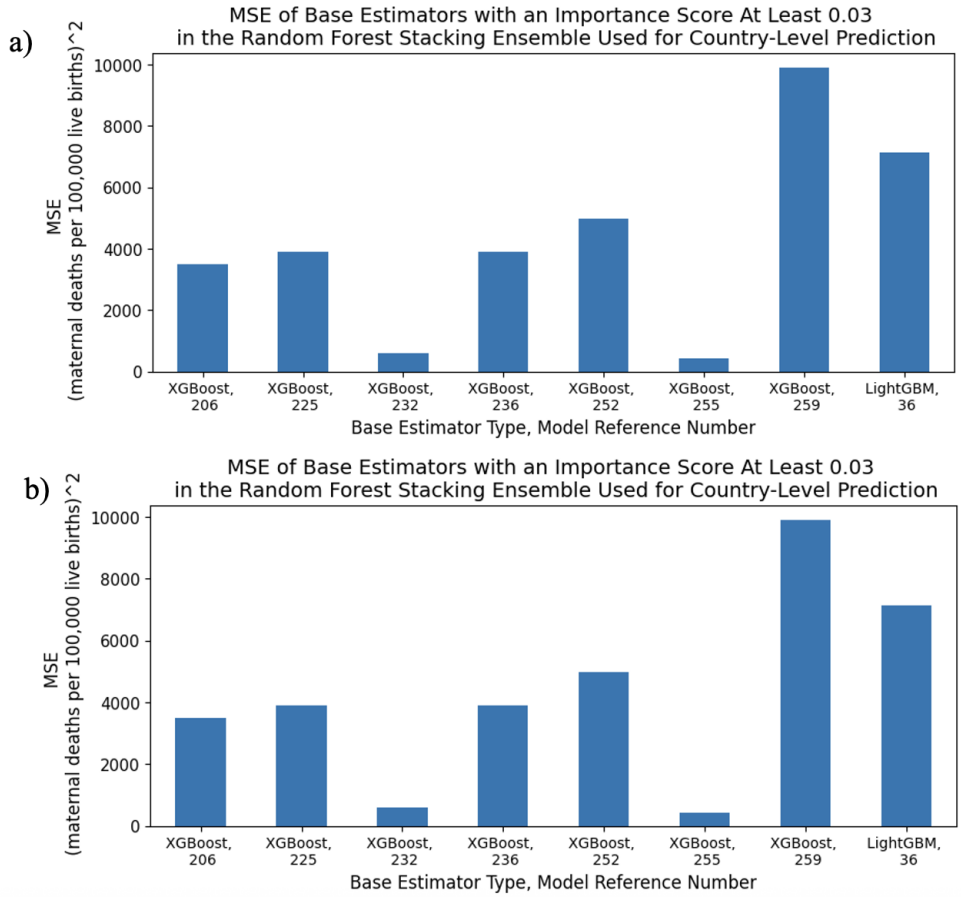


Figure 5.27: MSE for the base estimators given an importance score of at least 0.03 by the Random Forest Stacking Ensemble trained to perform a) country-level prediction and b) forecasting. Each base estimator was identified with its model type and a reference number that specified its position in the RFSE's input data.

biased base estimator selection. This was somewhat likely given the relatively similar performance between base estimators. To test this hypothesis, I randomly permuted the positions of base estimators in the RFSE's input dataset.

Country-Level Prediction

Nine of the ten base estimators given importance scores of at least 0.03 in the original RFSE were also given importance scores of at least 0.03 when base estimator order was permuted (Table 5.4). After permutation, their importance score magnitudes generally did not change by a large amount. The largest change was in the model given the highest importance, which lost 0.06 importance points after permutation. The RFSE's predictive accuracy did not change greatly after permutation, indicating that the subset

of base estimators used was not a random decision (MRE=0.07, MSE=2,188 versus the original MRE=0.07, MSE=2,161). However, the weighting given to each base estimator may be unstable and subject to change via retraining.

Forecasting

Randomising base estimator order had a greater effect of the RFSE trained for forecasting (Table 5.4).While 10 base estimators in the original RFSE had importance scores at least 0.03, only 4 base estimators had a sufficiently high importance score in the permuted RFSE. Only 2 of these 4 base estimators were also in the list of 10 given high importance under the original ordering, with both of these base estimators' importance scores increasing by 0.26 to 0.30 points. The RFSE's predictive accuracy decreased after permutation, with its MSE increasing by roughly 860 points (MRE=0.39, MSE=6,063 versus the original MSE=0.37, MSE=5,134). These changes showed that the RFSE's choice of base estimators was more affected by ordering and/or was instable and subject to change through retraining.

Conclusions of Base Estimator Importance Experiments

I did not find a robust explanation for why the Random Forest Stacking Ensemble placed high importance on a specific subset of base estimators. While some of the subset had low predictive error, others had very high MSE scores. Potentially, even the base estimators with high predictive error performed well for a specific subset of estimates, making them useful for the stacking meta-estimator, which can learn when and how to use them. While the permutation analysis showed the base estimators trained for country-level prediction were not chosen at random, it did reveal instability in their importance scores over different training instances. This was shown more explicitly for the RFSE trained for forecasting. This instability may be due to many of the base estimators having similar performance, allowing them to substitute for each other and/or take some of each other's importance weighting.

5.6.3 Feature Importance Analysis for Chosen Base Estimators

One of the primary aims of this research was to determine the socio-economic and health-related variables with the highest predictive power for MMR. The following section presents the features given the highest importance by the two base estimators with the highest weightings in the RFSE. For comparison, I also presented the features given the highest importance by 2 base estimators with low importance in the RFSE and relatively higher MSEs.

Country-Level Prediction

The two base estimators given the highest importance scores in the Random Forest Stacking Ensemble placed the greatest importance on features detailing women's level and type of employment, women's knowledge of contraceptive options, the percentage of

Table 5.4: The base estimators given an importance score of at least 0.03 by the RFSE when present in the RFSE's input data in the original and permuted orders. The model reference numbers were given in terms of the original ordering to allow comparison. For example, if the base estimator originally in the 206th position in the input data was moved to the 2nd position in the permuted order, it was still presented below as the 206th model.

Country-Level Prediction			
RFSE with Original Base Estimator Order		**RFSE with Permuted Order of Base Estimators**	
Model Reference	**Importance Score**	**Model Reference**	**Importance Score**
206	0.26	206	0.20
232	0.14	232	0.11
252	0.10	236	0.11
255	0.06	252	0.08
236	0.06	255	0.06
36	0.05	225	0.04
259	0.03	36	0.04
225	0.03	258	0.04
Forecasting			
RFSE with Original Base Estimator Order		**RFSE with Permuted Order of Base Estimators**	
Model Reference	**Importance Score**	**Model Reference**	**Importance Score**
241	0.13	241	0.43
254	0.08	243	0.33
40	0.08	258	0.09
232	0.08	207	0.05
243	0.07	All remaining models had importance scores < 0.03	
43	0.05		
20	0.05		
215	0.03		
213	0.03		
231	0.03		

women who were teenage mothers, and the country's World Bank defined income level (Table 5.5). Health-related variables monitoring the presence of skilled health staff at births, fertility rates, medical outcomes related to nutritional status, and life expectancy were also highly valued.

While the two base estimators with higher errors and lower importance scores in the RFSE also placed value on variables that monitor contraception prevalence and literacy rates, they focused more on features that monitored the prevalence of different diseases. For example, they placed high importance on features monitoring the rate of still-births and specific nutritional deficiencies as well as infectious disease and maternal disorders prevalence.

Overall, these results indicate that the base estimators given more importance in the RFSE placed more value on socio-economic related variables and aggregate health-trends while base estimators given less importance monitored more specific trends in health and disease.

Forecasting

As above, the two base estimators with the highest importance scores in the Random Forest Stacking Ensemble placed the greatest importance on features that measured the amount and type of female employment as well as knowledge about contraceptive options and nutritional status (Table 5.6). There was also slightly more emphasis on long-term conditions, such as measuring mortality due to non-communicable diseases. While the base estimators that added little value to the RFSE placed highest importance on similar features, these estimators placed slightly more emphasis on mortality measures and contained more information about trends in health outcomes for the whole population and for men rather than focusing on women.

Table 5.5: The 5 features given the highest importance scores in: (blue) the two base estimators given the highest importance scores in the RFSE, (orange) a medium-low accuracy base estimator from the 'Correlation 0.7' feature subset, and (purple) a low-accuracy base estimator from the 'Correlation 0.8' feature subset. All models were used for country-level prediction.

Base Estimator Model & Importance Score in the RFSE	**Feature Name**
XGBoost, fold 2, Missing data threshold 95%, No feature selection Importance score: **0.26**	Vulnerable employment (% of total employment), female;
	Knowledge of any modern method of contraception (% of all women ages 15–49);
	Wage and salaried workers (% of total population), female;
	Knowledge of any modern method of contraception (% of all married women ages 15–49);
	Teenage mothers (% of women ages 15–19 who have had children or are currently pregnant);
XGBoost, fold 4, Missing data threshold 85%, Features derived from the literature Importance score: **0.14**	Country income level;
	Births attended by skilled health staff (% of total);
	Fertility rate, total (births per woman);
	Survival to age 65, female (% of cohort);
	Prevalence of overweight (% of adults)
XGBoost, fold 4, Missing data threshold 90%, 'Correlation 0.7' feature subset Importance score: $\mathbf{1.44 \times 10^{-4}}$	Cause of death by communicable diseases and maternal, prenatal and nutrition conditions (% of male population);
	Contraceptive prevalence, any method (% of married women ages 15–49);
	Stillbirth rate (per 1,000 total births);
	Tetanus prevalence (age-standardised) (per 100,000 population), male;
	Other infectious diseases prevalence (age-standardised) (per 100,000 population), male
XGBoost, fold 4, Missing data threshold 100%, 'Correlation 0.8' feature subset Importance score: $\mathbf{1.56 \times 10^{-6}}$	Literacy rate, youth total (% of people ages 14–24), female;
	Cause of death by communicable diseases and maternal, prenatal and nutrition conditions (% of total);
	Maternal disorders prevalence (age-standardised) (per 100,000 population);
	Vitamin A deficiency prevalence (age-standardised) (per 100,000 population), male;
	Probability of survival to age 5, male

Table 5.6: The 5 features given the highest importance scores in: (blue) the two base estimators given the highest importance scores in the RFSE, (orange) a medium-low accuracy base estimator from the 'Correlation 0.7' feature subset, and (purple) a low-accuracy base estimator from the 'Correlation 0.8' feature subset. All models were used to perform forecasting.

Base Estimator Model & Importance Score in the RFSE	**Feature Name**
XGBoost, fold 1, Missing data threshold 90%, 'Correlation 0.6' feature subset Importance score: **0.13**	Wage and salaried workers (% of total employment), female;
	Vulnerable employment (% of total employment), female;
	Prevalence of stunting, height for age, male (% of children under 5);
	Contraceptive prevalence, any method (% of married women ages 15-49); Self-employed, total (% of total employment), female
XGBoost, fold 3, Missing data threshold 95%, 'Correlation 0.6' feature subset Importance score: **0.08**	Vulnerable employment (% of total employment), female;
	Children in employment (% of children ages 7-14), female;
	Cause of death, by non-communicable diseases, female (% of female population);
	Yellow fever prevalence (age standardised) (per 100,000 population), female;
	Contraceptive prevalence, any modern method (% of married women ages 15-49)
Random Forest base estimator, fold 5, Missing data threshold 95%, 'Correlation 0.7' feature subset Importance score: $\mathbf{1.55 \times 10^{-5}}$	Births attended by skilled health staff (% of total);
	Contraceptive prevalence, any method (% of married women ages 15-49);
	Stillbirth rate (per 1,000 total births);
	Mortality rate, under-5, male (per 1,000);
	Demand for family planning satisfied by any methods (% of married women with demand for family planning)
XGBoost, fold 4, Missing data threshold 100%, 'Correlation 0.8' feature subset Importance score: $\mathbf{5.28 \times 10^{-7}}$	Maternal disorders prevalence (age standardised) (per 100,000 population), female;
	Literacy rate, youth total (% of people ages 15-24), female;
	Vitamin A deficiency prevalence (age standardised (per 100,000 population), male;
	Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of male population);
	Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)

5.7 Performance Analysis of the Random Forest Stacking Ensemble

Building on the previous results, section explores the RFSE's performance, as described in Figure 4.3f.

5.7.1 Random Forest Stacking Ensemble's Predictive Error per Income Level

To gain a deeper understanding of how the RFSE performs in different settings, I analysed how its prediction errors changed when estimating MMR for countries from different income levels.

Country-Level Prediction

Generally, the RFSE's MRE on the test set decreased as income level increased (Figure 5.28a). For example, its test MRE was 0.18 for low-income countries but 0.07 for high-income countries. In contrast, the RFSE achieved its lowest test error for lower-middle income countries (0.02). Train and validation MREs were similar and smaller than the test MRE for all income levels, with the exception again being the lower-middle income subgroup. The difference between the train/validation and test MREs was greatest for low-income countries (approximately 0.14). The test MRE for low-income countries had the greatest standard deviation (0.22).

MSE uniformly decreased as income level increased, with the differences between income levels spanning orders of magnitude (Figure 5.28b). More specifically, the RFSE incurred an MSE of 62,133 for low-income countries versus an MSE of 6 for high-income countries. The highest standard deviation in MSE was observed for low-income countries. The largest difference in MSE between consecutive income levels occurred between low-income and lower-middle income countries (62,133 to 356).

Forecasting

The MRE of the RFSE trained to forecast MMR increased as income level increased from lower middle to high. (Figure 5.29a). For instance, the RFSE had an MMR of 0.25 for lower-middle income countries versus an MRE of 0.47 for high-income countries. In contrast to this trend, the RFSE had a test MRE of 0.25 for both low and lower-middle income countries. The train and validation errors also increased as income level increased from lower middle to high. The RFSE's MRE had a large standard deviation for its validation and test sets, with the large validation deviation indicating considerable differences between cross-validation folds. Generally, train and validation errors for the same income level were similar, with test error always being at least 0.2 greater than train error. The low-income countries had the greatest difference (0.02) between train and validation MRE of any income level.

Test MSE decreased uniformly as income level increased, with decreases between income levels generally spanning an order of magnitude (Figure 5.29b). For instance, the RFSE achieved a test MRE of 85 for high-income countries versus 88,585 for low-income countries.

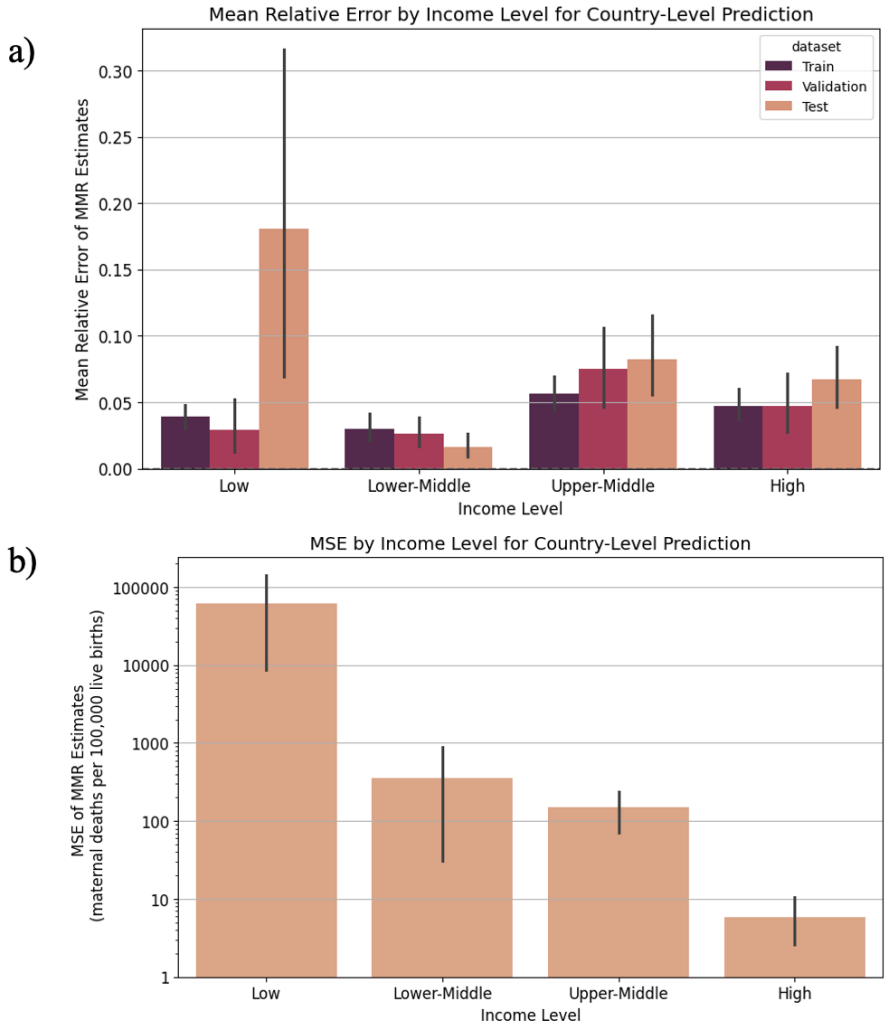


Figure 5.28: a) Mean relative error and b) mean-squared error (log scale) for income-level specific MMR estimates from the Random Forest Stacking Ensemble used for country-level prediction. MRE was given for its performance on the train, validation, and test sets. MSE was only given for the test set.

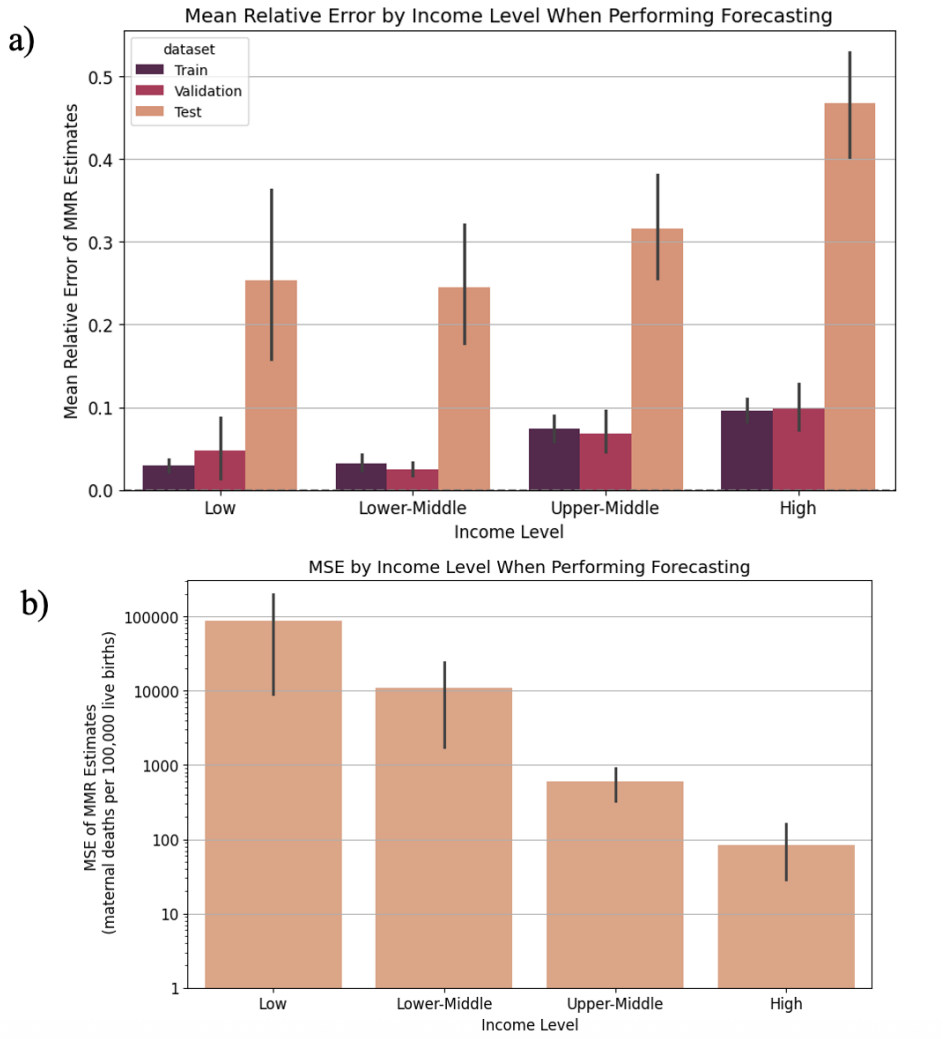


Figure 5.29: a) Mean relative error and b) mean-squared error (log-scale) for income-level specific MMR predictions from the Random Forest Stacking Ensemble trained to perform forecasting. MRE was given for the RFSE's performance on the train, validation, and test sets. MSE was only given for the test set.

5.7.2 Uncertainty Analysis for the Random Forest Stacking Ensemble

To provide a measure of uncertainty about the MMR estimates from the Random Forest Stacking Ensemble, I computed the standard deviation among the MMR estimates of its base estimators. The smaller the standard deviation, the greater the agreement, and thus the more certainty the stacking ensemble had in its final estimate.

As the ground truth MMR increased, standard deviation among base estimators trained

for country-level prediction also increased (Figure 5.30a). For ground-truth MMR values between 0 and 150, standard deviation was generally less than 50. In contrast, for ground truth MMRs between 300 and 1,050, base model predictions ranged from 50 to 350. For the extremely high ground truth MMR of 1,763, standard deviation among base estimators was 441.

Similarly, standard deviation among the MMR predictions of base estimators trained to perform forecasting increased (0 to 300) as the ground truth MMR increased (0 to 900) (Figure 5.30b). However, a slight decrease in standard deviation was observed for ground truth MMRs greater than 1,150. More specifically, the ground truth MMRs 1,194 and 1,389 had standard deviations 216 and 170, respectively.

As a note, these findings must be cautioned, as there were few datapoints for ground truth MMRs greater than 1,050 for country-level prediction or 750 for forecasting.

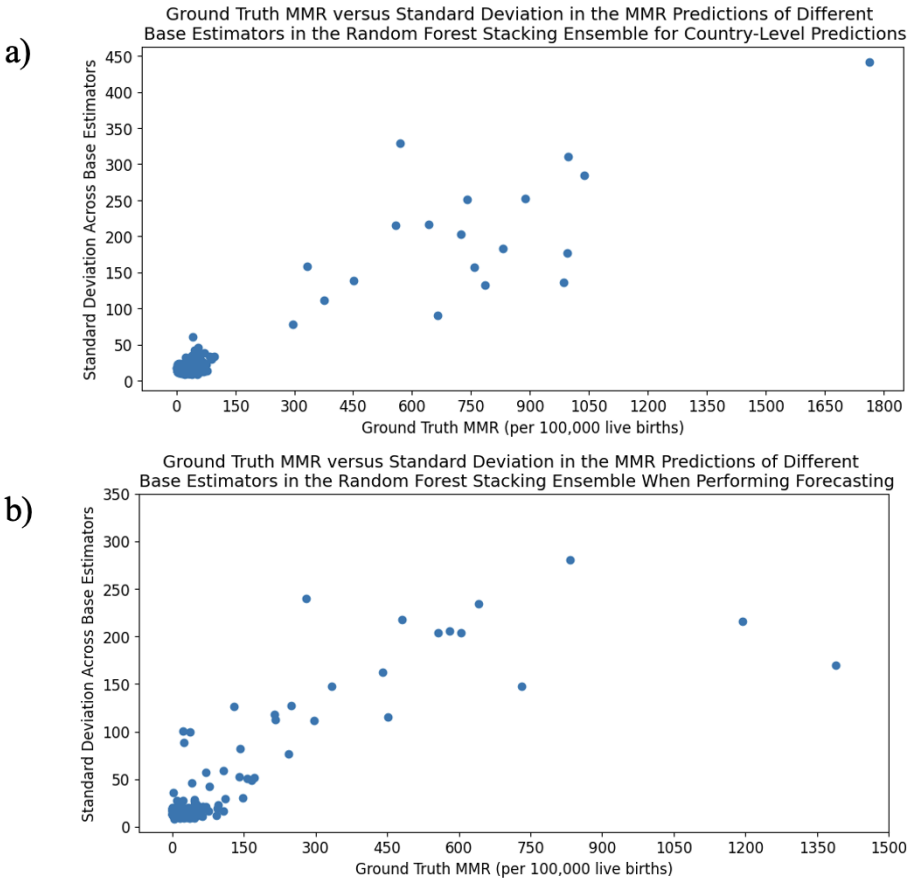


Figure 5.30: Standard deviation among the predictions made by base estimators in the Random Forest Stacking Ensemble versus the ground truth MMR estimate they were trying to predict. This analysis was done for base estimators trained to perform a) country-level prediction and b) forecasting.

5.7.3 Sensitivity Analysis

I conducted a sensitivity analysis to explore how the input data structure affected the RFSE's final MMR predictions (see Appendix A.6 for additional performance metrics.)

Country-Level Prediction

The MRE for the Random Forest Stacking Ensemble trained on data from all income levels was very similar to the MRE for RFSEs trained on data from a specific income level (Figure 5.31a). For example, the RFSE trained on all data had a test MRE of 0.068 on the high-income dataset while the RFSE trained on just high-income data had an MRE of 0.070. Similarly, the RFSE trained on all data had a test MRE of 0.082 on the upper-middle dataset while the RFSE trained on just upper-middle income data had a test MRE of 0.072. No difference in MRE between the models was greater than 1%. Therefore, the MRE differences between the model trained on all data versus the models trained on income-specific data was not significant.

However, there were larger differences between the RFSE trained on all data and RFSEs trained on income-specific data when performance error was calculated in terms of MSE (Figure 5.31b). The RFSE trained on all data had lower MSE scores than the RFSEs solely trained on low, upper-middle, and high-income data by 166,874, 219, and 2.2, respectively. The difference of 166,874 was the largest difference between the original and sensitivity models across all datasets. The difference in the trend observed for MSE and MRE suggests the presence of outliers in the income-specific data. The exception to this trend was the lower-middle income dataset, where the model trained on all data had a higher MSE than the RFSE trained solely on data from this income level. Nevertheless, this difference was just 18, which may not be significant given that standard deviation in lower-middle income countries' ground truth MMR values was 260 (Table 5.2).

Forecasting

Unfortunately, some of the cross-validation folds for the 'Correlation 0.8' feature subset had insufficient non-missing data when filtered for just high-income data. As a result, some of the base estimators for the high-income dataset could not be trained, preventing an RFSE from being fit on the high-income sensitivity data, as it expected a certain number of base estimators. Consequently, only results from the sensitivity models trained on the low, lower-middle, and upper-middle datasets were presented.

The RFSE trained on all data always exceeded the MSE of the RFSEs trained on specific income levels (Figure 5.32). The largest discrepancies occurred for low-income countries (29,884 MSE points) and lower-middle income countries (2,928 MSE points). In contrast, the difference for upper-middle income countries was only 6 MSE points. The MRE of the RFSE trained on all data was also greater than the MRE of RFSEs trained on only low and upper-middle income data (0.25 versus 0.21 and 0.32 versus 0.30, respectively). However, its MRE was smaller than the RFSE solely trained on lower-middle income data (0.25 versus 0.31).

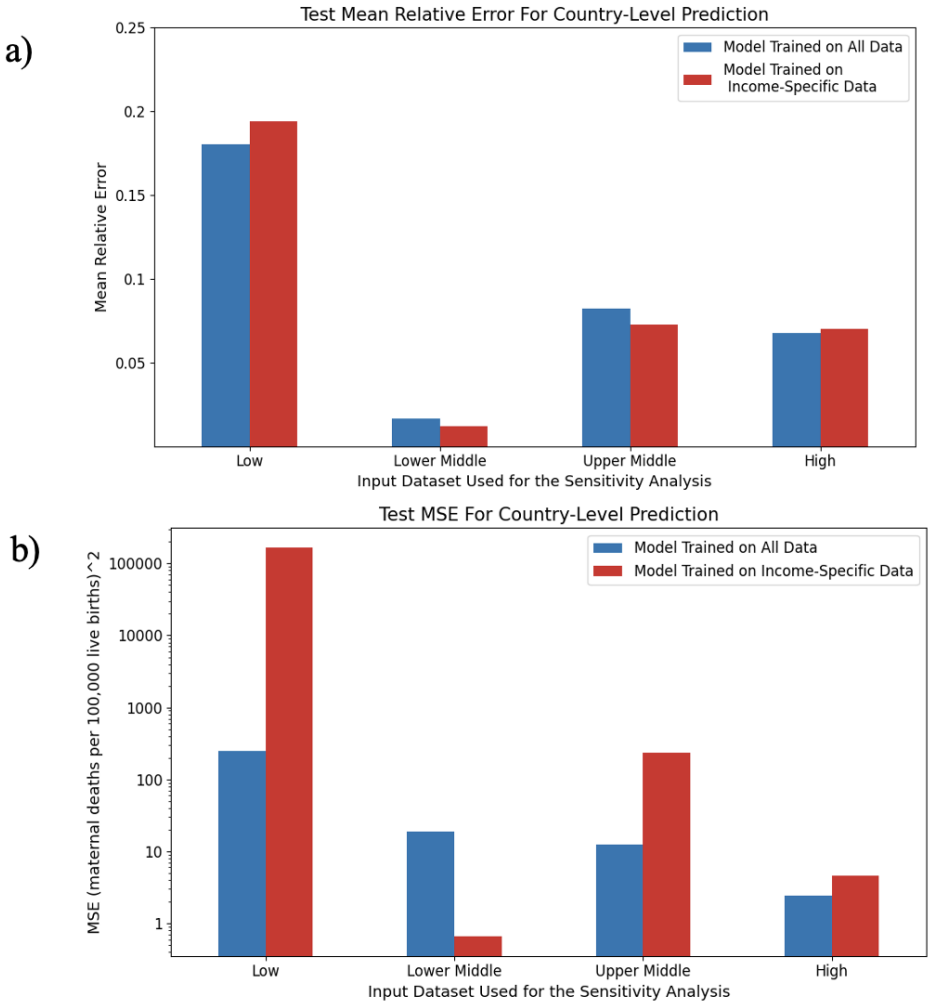


Figure 5.31: a) Mean relative error and b) mean-squared error (log scale) for the RFSE trained on data from all income levels (blue) and RFSEs trained on data from a specific income level (red) for country-level prediction. The models being compared were tested on data from the same income level.

5.8 Comparison of the Random Forest Stacking Ensemble to the Literature

In this section, I compared the MMR estimates of my best-performing Random Forest Stacking Ensemble to the MMR estimates produced by the UN MMEIG's BMat Model, the Global Burden of Disease Study's CODEm model, and the GMatH model. See the literature review for detailed descriptions of these models and their MMR estimation processes, and Figure 4.3f for the overview of this process.

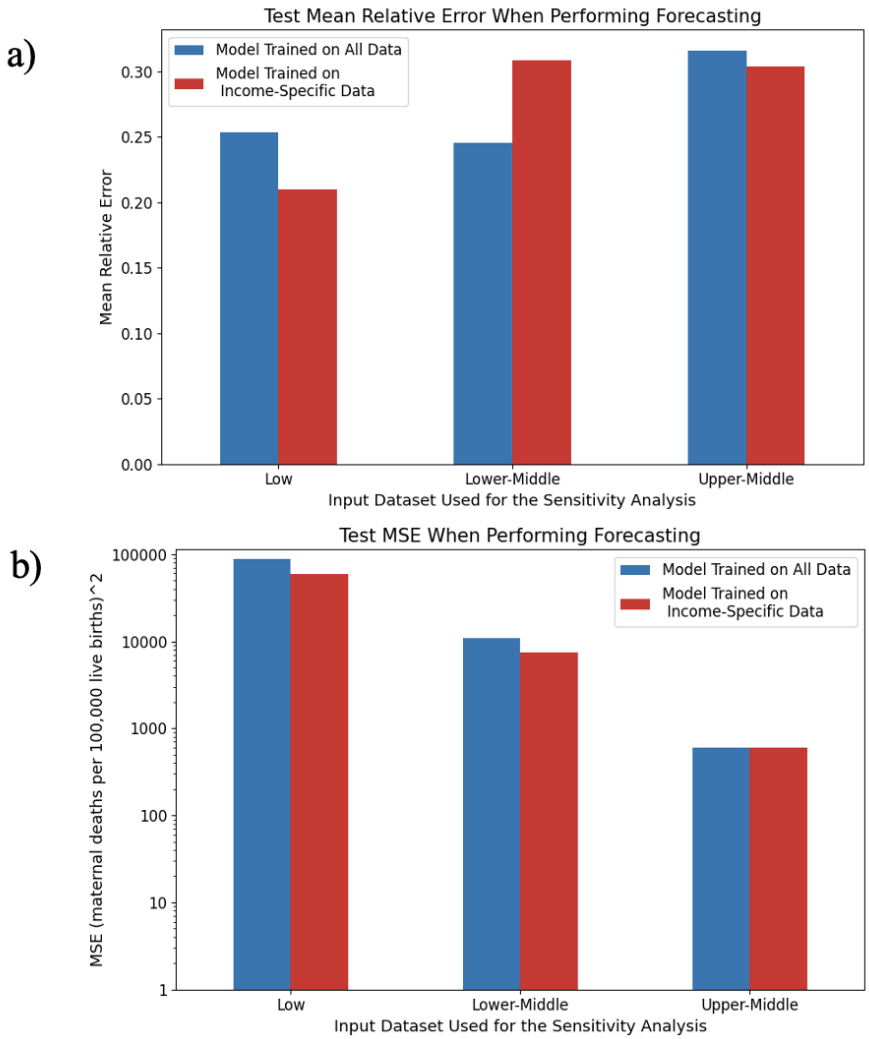


Figure 5.32: a) Mean relative error and b) mean-squared error (on a log scale) for the RFSE trained on data from all income levels (blue) and RFSEs trained on data from a specific income level (red) to perform forecasting. The models being compared were tested on data from the same income level.

5.8.1 Percentage Difference

I took the percentage difference between my MMR estimate for each country, year datapoint and the associated MMR estimate from the literature. A negative percentage difference meant that my estimate was smaller than the literature's estimate.

Country-Level Predictions

Over 70%, 75%, and 80% of the MMR estimates from my best-performing RFSE were smaller than the corresponding estimates from the BMat, CODEm, and GMatH models, respectively (Figure 5.33). Over 40% of my MMR predictions were between 75 and 100% smaller than the corresponding GMatH estimates. In contrast, less than 5% of my MMR estimates were over 75% smaller than the BMat or CODEm estimates. More specifically, approximately 50% of my MMR estimates were between 0 and 39% or 0 and 32% smaller than the associated estimates from the BMat and CODEm models, respectively.

Approximately 15 to 20% of my MMR predictions were larger than the associated BMat and CODEm estimates while only roughly 5% of my predictions were greater than the corresponding GMatH estimates. There were no extreme outlier differences between my MMR predictions and the literature's estimates.

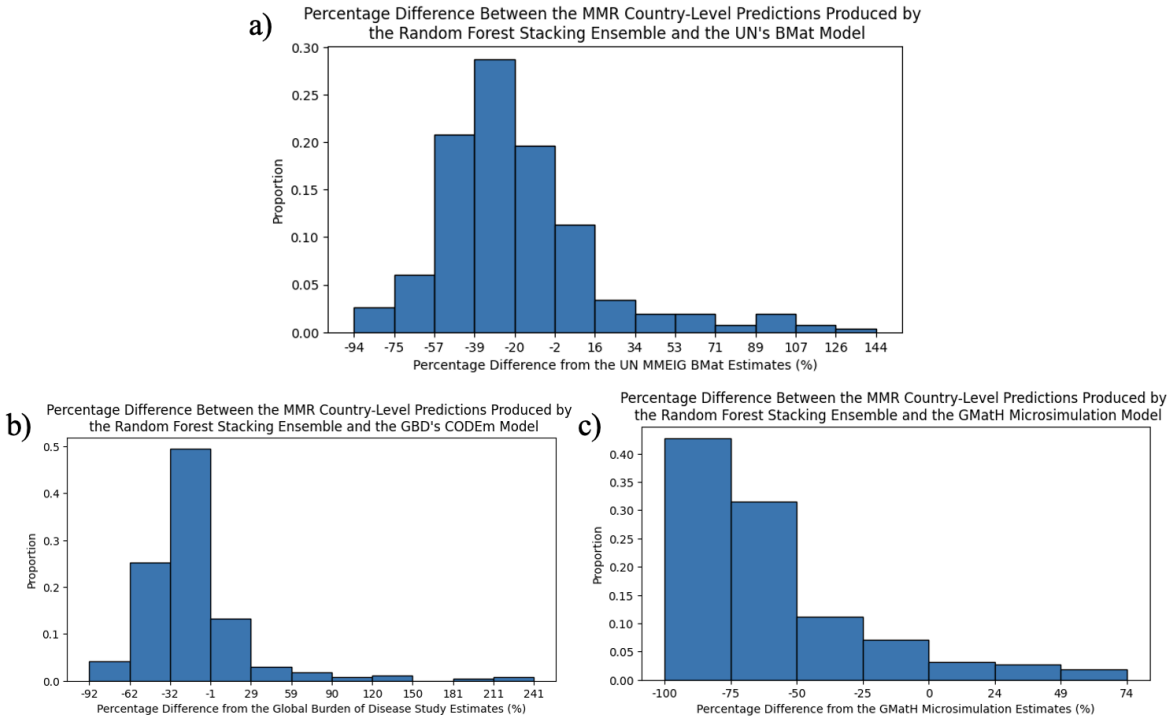


Figure 5.33: Percentage difference between the MMR estimated by my best-performing Random Forest Stacking Ensemble trained for country-level prediction and the MMR estimated by a) BMat, b) CODEm, and c) GMatH.

Forecasting

Almost 70% and 60% of my MMR forecasts were smaller than the associated BMat and CODEm estimates, respectively (Figures 5.34a, 5.34b). In contrast, over 80% of

my MMR estimates were smaller than the associated GMatH estimates (Figure 5.34c). Over 40% of my model's MMR predictions were between 0 and 50% smaller than both the BMat and CODEm estimates. In contrast, over 50% of my forecasts were between 65 and 100% smaller than the GMatH predictions.

Approximately 30% of my MMR forecasts were larger than the BMat and CODEm estimates while only roughly 10% of my forecasts were larger than the associated GMatH predictions. There was a small proportion of instances where my forecasts were over 1300% greater than the corresponding BMat estimates and between 960 and 1,000% greater than the associated CODEm predictions. In comparison, this small proportion of instances was only between 180 and 215% larger than the corresponding GMatH estimates.

Summary of Differences

The magnitude difference between the GMatH predictions and both my MMR country-level predictions and forecasts was generally larger than the differences to BMat and CODEm.

5.8.2 Coverage

Approximately 67.1% of my RFSE's country-level MMR predictions were within the 95% confidence intervals (CI) of GMatH's MMR predictions (Table 5.7). In contrast, only 22.3 and 29.4% were within the 95% CI of the BMat or CODEm's MMR estimates, respectively. Similarly, only 20.4, 33.2, and 67.1 of the ground truth MMR estimates used to train my RFSE were within these models' 95% confidence intervals.

A higher proportion of my RFSE's MMR forecasts were within BMat and GMatH's 95% CI than the proportion of my model's country-level predictions. In contrast, a smaller proportion of my model's MMR forecasts were within the 95% CI of the CODEm model's estimates. More specifically, 30.9, 23.5, and 81.6% of my models' MMR forecasts were within the 95% CI of the BMat, CODEm, and GMatH models, respectively. A similar, but slightly higher, proportion of the ground truth MMR estimates used to train my model were within these 95% confidence intervals (32.6%, 25.1%, and 84.9%).

5.8.3 Per-Country Comparison

To better understand the differences between my best-performing model's predictions and those found in the literature, I compared my model's MMR estimates for a specific country from each income level to the associated BMat, CODEm, and GMatH estimates.

Comparison Between the Literature and My RFSE Trained to Perform Country-Level Prediction

Compared to the other models, my Random Forest Stacking Ensemble underpredicted New Zealand's MMR between 1990 and 2015, with greater underprediction pre-2010

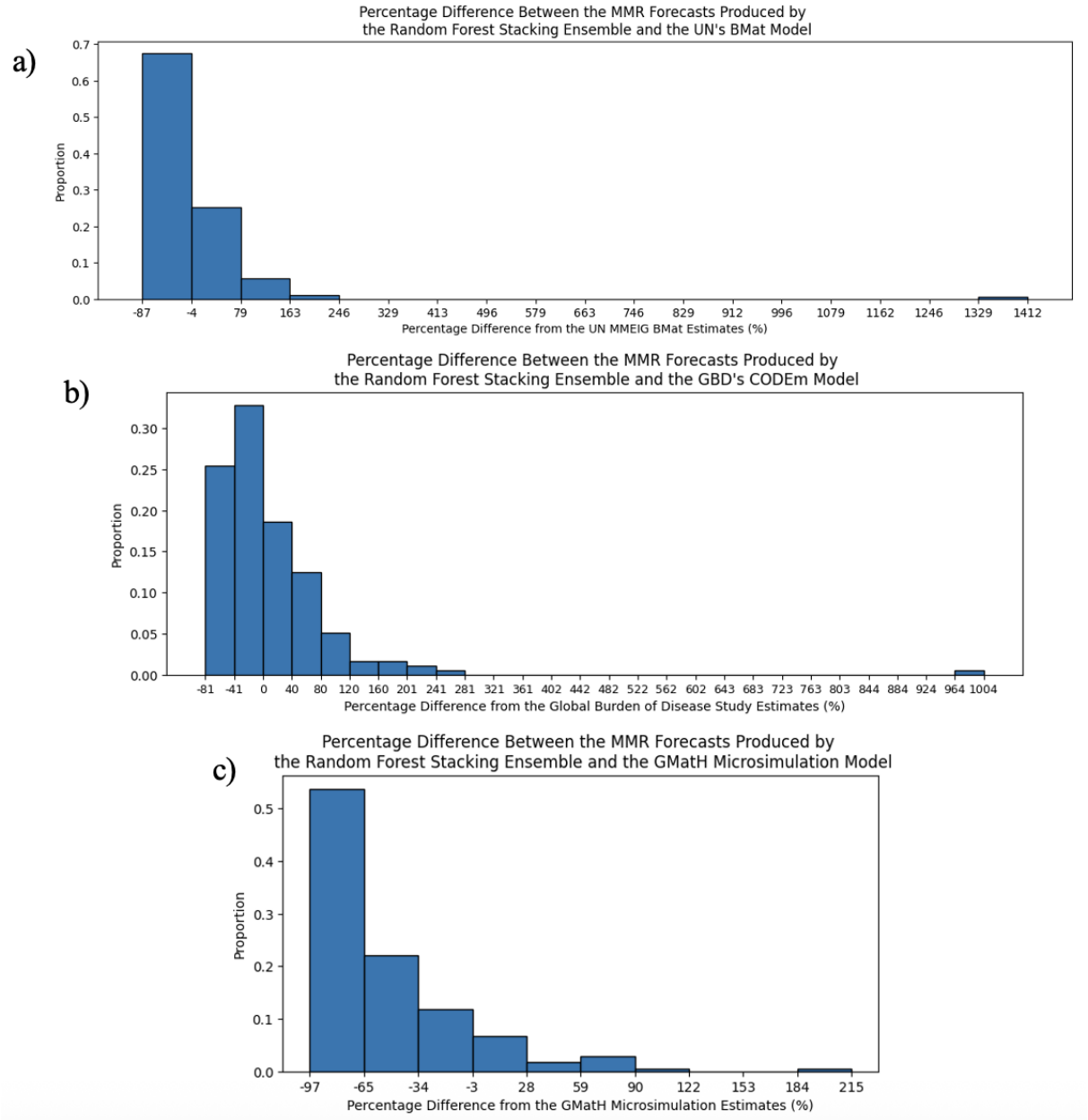


Figure 5.34: Percentage difference between MMR estimated by my best-performing Random Forest Stacking Ensemble trained to perform forecasting and the MMR estimated by a) BMat b) CODEm, and c) GMatH.

(Figure 5.35a). While my model's estimates were outside the 95% confidence intervals (CI) of the BMat and CODEm models, the actual difference in MMR between the estimates was between 5 and 20. The GMatH model strongly overestimated New Zealand's

Table 5.7: The percentage of MMR country-level predictions (blue) and forecasts (pink) from my best-performing Random Forest Stacking Ensemble that fell within the 95% confidence intervals (CI) for BMat, CODEm, and GMatH's predictions. The proportion of ground truth MMR estimates used to train my model that fell within these CI was also presented

Type of Analysis	Percent of RFSE MMR estimates within BMat's 95% CI	Percent of ground truth MMR estimates within BMat's 95% CI	Percent of RFSE MMR estimates within CODEm's 95% CI	Percent of ground truth MMR estimates within CODEm's 95% CI	Percent of RFSE MMR estimates within GMatH's 95% CI	Percent of ground truth MMR estimates within GMatH's 95% CI
Country-Level Prediction	22.3%	20.4%	29.4%	33.2%	67.1%	67.1%
Forecasting	30.9%	32.6%	23.5%	25.1%	81.6%	84.9%

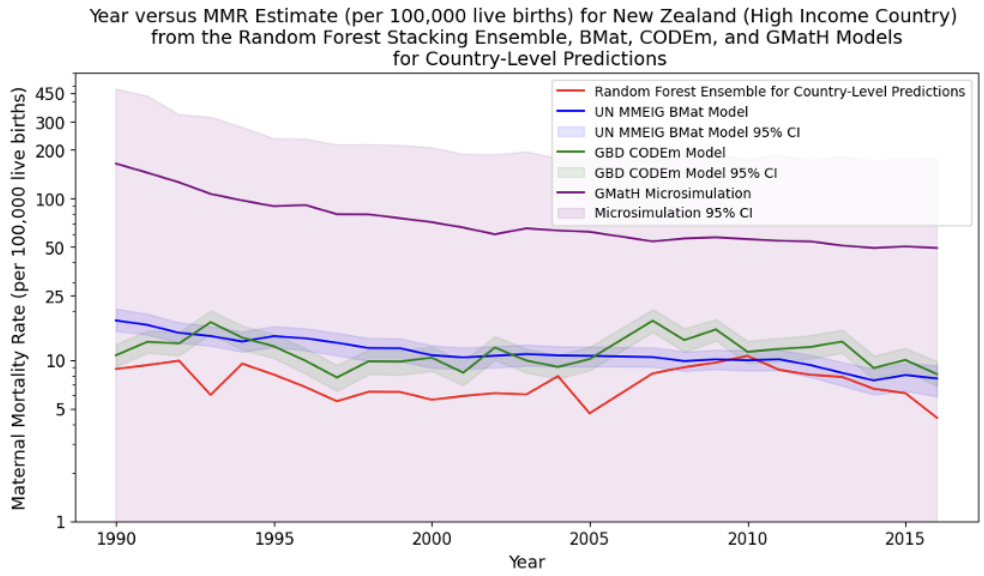
MMR, predicting MMR to be close to 200 in 1990 and fall to roughly 50 by 2015. In contrast, the BMat and CODEm models did not predict an MMR of higher than 25 for New Zealand in this time interval. These larger GMatH estimates came with a wide 95% CI that enveloped my model's MMR estimates.

My RFSE's estimates had greater intersection with the literature's estimates for Colombia, an upper-middle income country (Figure 5.35b). My model's estimates were generally 0 to 40 points off the closest literature estimate. They were generally within the GMatH model's wide 95% confidence interval, with the GMatH predictions again higher than the other literature estimates. At times, my model's estimates were within the 95% CI of either the BMat or CODEm models' predictions. However, my model's estimates fluctuated more strongly between consecutive years, compared to the smoother literature estimates. Many of the sharp peaks predicted by my model corresponded to the years with the greatest amount of non-missing data (see Figure 5.3).

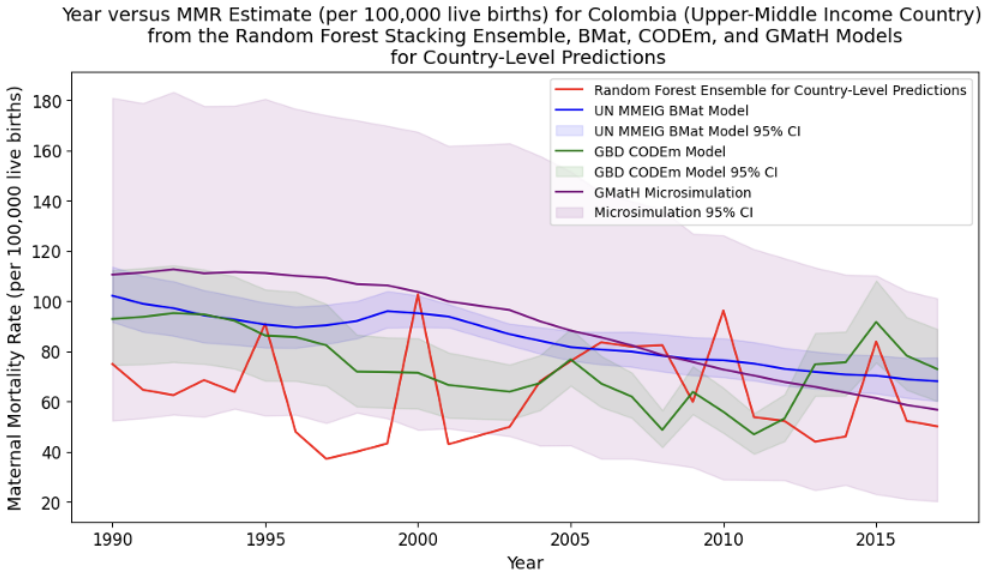
My models' estimates were completely within the 95% CIs of at least one other model when predicting for Kenya and Rwanda, which are lower-middle and low-income countries, respectively (Figures 5.35c, 5.35d). However, the magnitude difference between my estimates and the literature's estimates was in the hundreds, with the greatest difference observed between my estimates and the CODEm predictions. My RFSE's estimates for these countries generally exceeded the BMat and CODEm estimates. They were more often greater than the GMatH estimates for Rwanda than for Kenya. However, these comparisons must be taken with a grain of salt, as there were only 4 datapoints for each of Kenya and Rwanda over this period of time.

Therefore, it appeared that my model's country-level predictions were underestimates for higher-income countries but over-estimates for lower-income countries.

a)


Year versus MMR Estimate (per 100,000 live births) for New Zealand (High Income Country)
from the Random Forest Stacking Ensemble, BMat, CODEm, and GMatH Models
for Country-Level Predictions
Random Forest Ensemble for Country-Level Predictions
UN MMEIG BMat Model
UN MMEIG BMat Model 95% CI
GBD CODEm Model
GBD CODEm Model 95% CI
GMatH Microsimulation
Microsimulation 95% CI
Maternal Mortality Rate (per 100,000 live births)
450
300
200
100
50
25
10
5
1
1990
1995
2000
2005
2010
2015
Year

b)


Year versus MMR Estimate (per 100,000 live births) for Colombia (Upper-Middle Income Country)
from the Random Forest Stacking Ensemble, BMat, CODEm, and GMatH Models
for Country-Level Predictions
Random Forest Ensemble for Country-Level Predictions
UN MMEIG BMat Model
UN MMEIG BMat Model 95% CI
GBD CODEm Model
GBD CODEm Model 95% CI
GMatH Microsimulation
Microsimulation 95% CI
Maternal Mortality Rate (per 100,000 live births)
180
160
140
120
100
80
60
40
20
1990
1995
2000
2005
2010
2015
Year

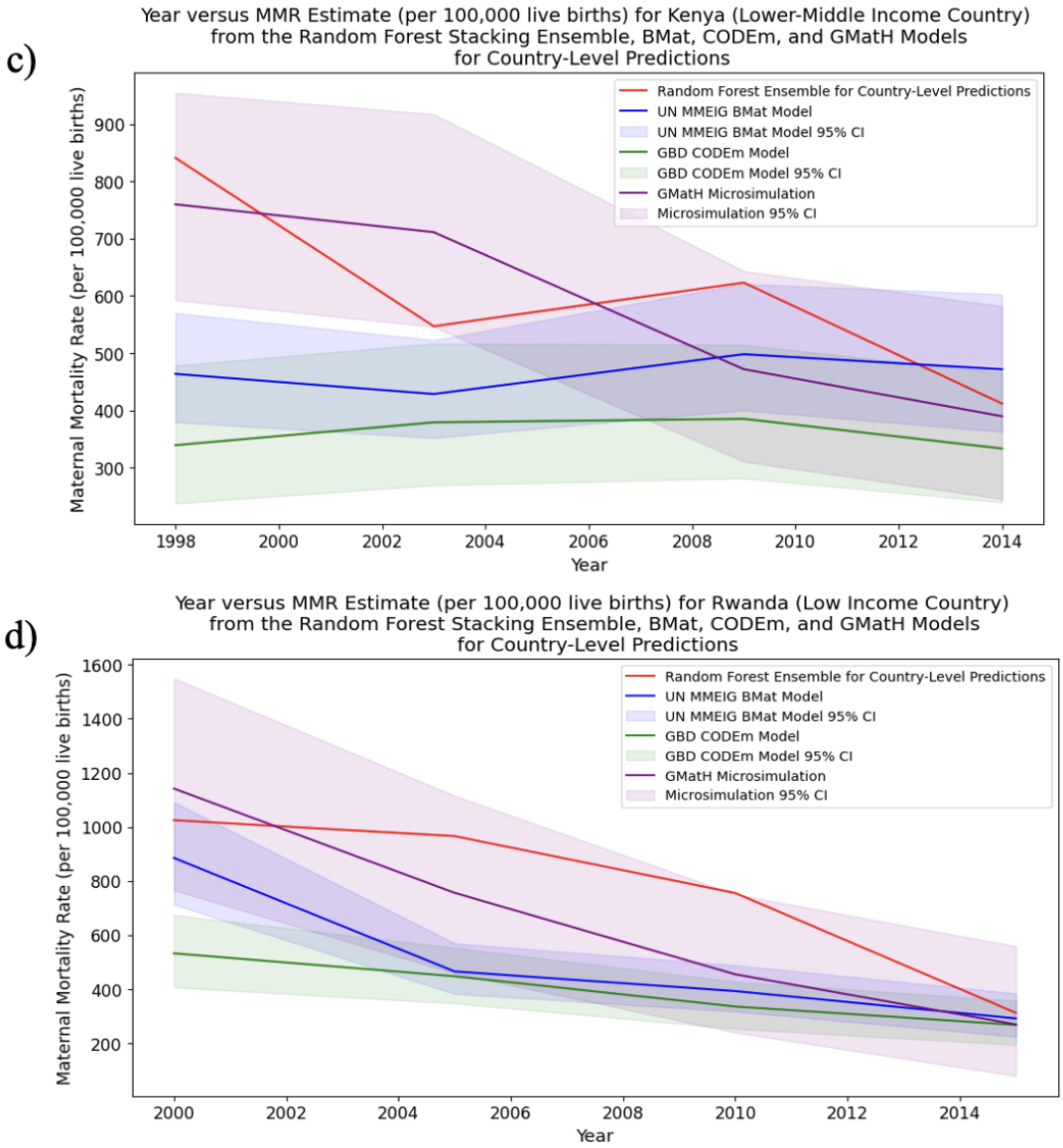


Figure 5.35: Comparison of my best-performing RFSE's country-level MMR predictions to the associated estimates from BMat, CODEm, and GMatH for a) New Zealand (high-income) (log scale), b) Colombia (upper-middle income), c) Kenya (lower-middle income), and d) Rwanda (low-income).

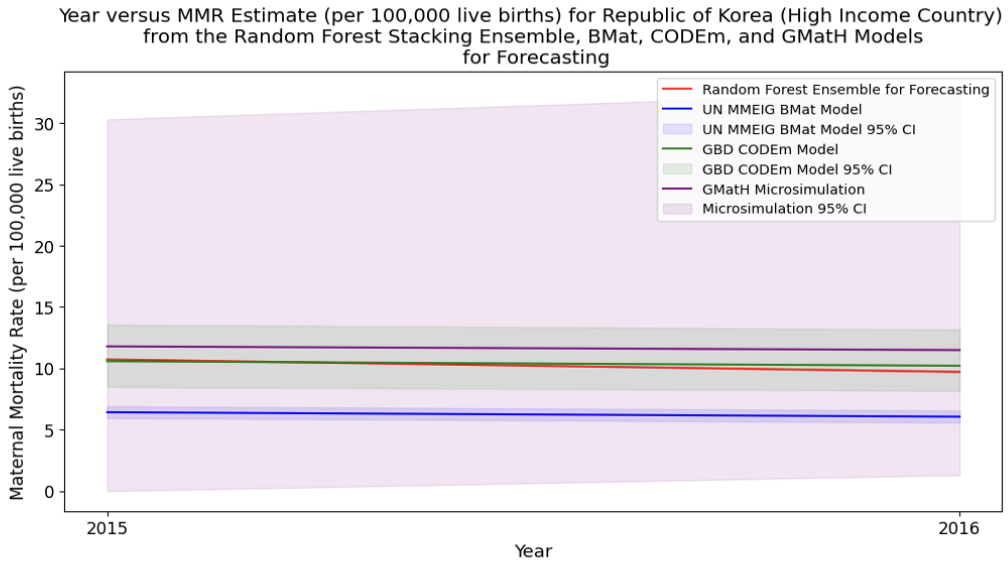
Comparisons with My RFSE Trained to Perform Forecasting

There was less room for comparison between my model's MMR forecasts and the literature's estimates because all comparisons were performed on my model's test values, which were confined between 2015 and 2018. Additionally, not all samples in my test set had non-missing MMR values, meaning some of the countries presented in this section did not have an associated MMR prediction for every year in the test set.

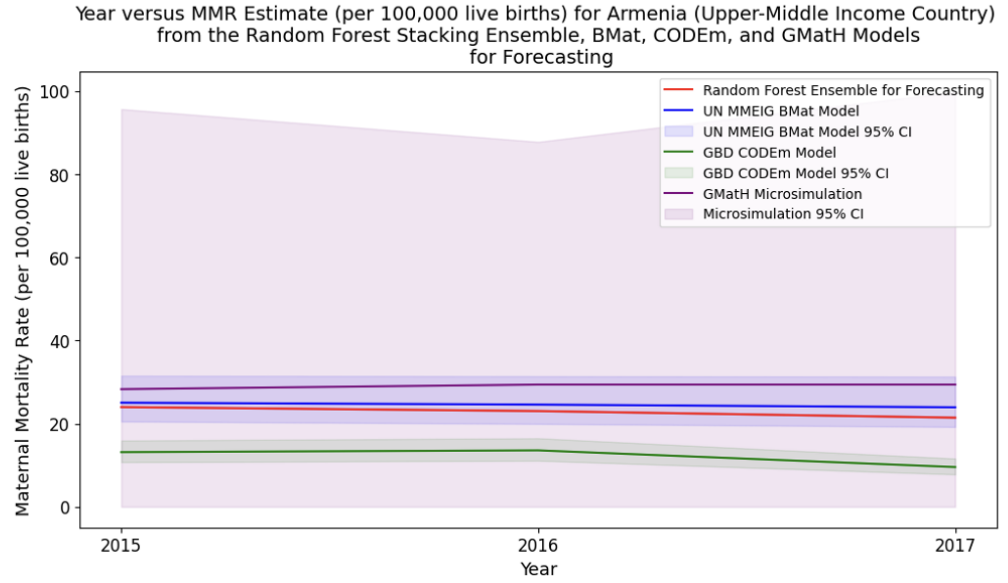
My best-performing Random Forest Stacking Ensemble's MMR forecasts were always in the 95% confidence intervals (CI) of the literature's corresponding estimates (Figure 5.36). For the high and upper-middle income countries (Republic of Korea and Armenia), my model's MMR forecasts were the second lowest, and either within the CODEm or BMat 95% CIs. The actual difference between my estimates and the CODEm estimates for the Republic of Korea's MMRs was in the single digits (Figure 5.36a). My model's MMR forecasts for Chad were also the second-lowest available, and within GMatH's 95% CI (Figure 5.36d). Unfortunately, there was only one test datapoint for Chad and every other low-income country contained in the test set.

In contrast, my model's MMR forecasts were the highest available for the first half of the testing period for Sri Lanka, a lower-middle income country (Figure 5.36c). Its estimates in the second half of this training period were very similar to the GMatH and BMat predictions. All of its estimates in the test period were within the 95% CI of the literature models.

a)


Year versus MMR Estimate (per 100,000 live births) for Republic of Korea (High Income Country) from the Random Forest Stacking Ensemble, BMat, CODEm, and GMatH Models for Forecasting
Random Forest Ensemble for Forecasting
UN MMEIG BMat Model
UN MMEIG BMat Model 95% CI
GBD CODEm Model
GBD CODEm Model 95% CI
GMatH Microsimulation
Microsimulation 95% CI
Maternal Mortality Rate (per 100,000 live births)
0
5
10
15
20
25
30
2015
2016
Year

b)


Year versus MMR Estimate (per 100,000 live births) for Armenia (Upper-Middle Income Country) from the Random Forest Stacking Ensemble, BMat, CODEm, and GMatH Models for Forecasting
Random Forest Ensemble for Forecasting
UN MMEIG BMat Model
UN MMEIG BMat Model 95% CI
GBD CODEm Model
GBD CODEm Model 95% CI
GMatH Microsimulation
Microsimulation 95% CI
Maternal Mortality Rate (per 100,000 live births)
0
20
40
60
80
100
2015
2016
2017
Year

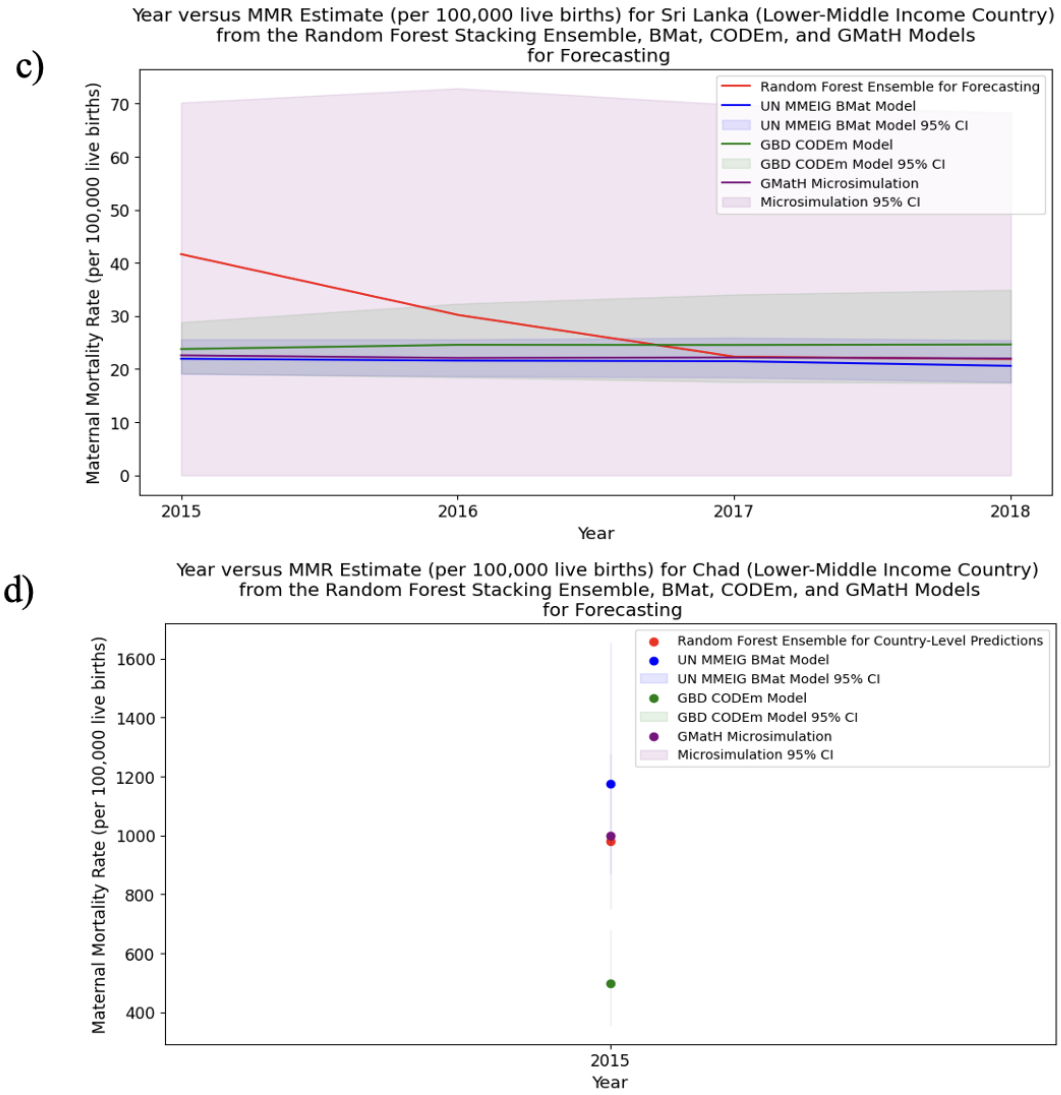


Figure 5.36: Comparison of my best-performing RFSE's MMR forecasts to the associated estimates from BMat, CODEm, and GMatH for a) Republic of Korea (high-income), b) Armenia (upper-middle income), c) Sri Lanka (lower-middle income), and d) Chad (low-income).

Chapter 6

Discussion

The primary aim of this thesis was to use decision-tree based machine learning methods to estimate countries' maternal mortality ratios between 1985 and 2018. This chapter interprets the results presented in Section 5 and discusses them in the context of this aim (6.1, 6.2) and existing research (6.3). To address the secondary aim of my research, I then discuss the socio-economic and health-related variables with the highest predictive power for MMR (6.4) and how they can be used to motivate policy recommendations (6.5). Finally, I examined the strengths, limitations, and possible extensions of my research (6.6, 6.7, 6.8).

6.1 Discussion of Base Estimator Performance

Based on the results presented in Section 5.4, there was no consistent advantage to training base estimators on a specific feature data subset or on data with less than a certain threshold proportion of missing data. Similarly, none of the Random Forest, XGBoost, or LightGBM models consistently had the highest performance. Instead, the greatest differences in base estimator performance could be attributed to the specific training fold used to fit the model. Nevertheless, there were a few context-specific performance differences across the various model types and pre-processing techniques, especially when considering whether the model was trained for country-level prediction or forecasting. In the following section, I interpret and discuss these trends and explain how they motivated the use of a stacking ensemble model to combine the predictions of the various base estimators.

6.1.1 Missing Data Removal Had No Consistent Effect on Model Performance

Missing data can potentially increase predictive error in decision-tree based models by preventing the tree building algorithm from finding the best feature-based splits to use on the tree's internal nodes (Twala, 2009). Additionally, when given a test sample, the predictive algorithm may struggle to choose the best path through the tree if splits depend on features with missing data. However, my analysis of trained decision-tree based models (Section 5.4) showed that there was no consistent difference in predictive errors between no missing data removal and removing rows and columns with more than 85%, 90%, or 95% missing data (Figures 5.10, 5.11, 5.12, 5.13, 5.14, 5.15).

Therefore, the Random Forest, XGBoost, and LightGBM models could effectively handle high proportions of missing data using the default direction technique described in Section 2.4.1 of the Background Information. This result makes intuitive sense, as the default direction algorithm forces the model to explicitly learn the best path through the tree when feature data is missing. This greedy approach also allowed the model to take advantage of the details implied by the occurrence of missing data, as the missing data instance may signal specific information about a country's circumstances, which the model can incorporate when learning the default direction. This finding was reinforced by empirical studies in the literature. For example, Dabool et al. (2024) similarly showed that using XGBoost's default direction algorithm to handle missing data had slightly higher accuracy than applying XGBoost to the same dataset with all missing data imputed.

Nevertheless, the similar performance for different missing data thresholds was an interesting result, as researchers have hypothesised that greater proportions of missing data increase inaccuracy (Mgawadere et al., 2017; Twala, 2009). For example, Twala (2009) showed that classification error due to missing data increased by a factor of 1.5 if the proportion of missing data was 50% versus 15%. However, some researchers argue that the exact proportion of missing data is less important than the amount of missing information in the dataset, where auxiliary variables can be used to compensate for missing data (Madley-Dowda et al., 2019). This argument could also be used to explain my models' high performance despite the missing data. More specifically, many of my features were highly correlated (over 482 feature pairs had an absolute pairwise correlation coefficient >0.9) and contained similar information. If a sparse feature provided very similar information to a feature with fewer missing datapoints, the model could still learn the relevant trend. Thus, the high dimensionality of my dataset could serve as a redundancy measure that prevents loss of information.

While retaining sparse samples and features did not harm predictive performance, removing them sometimes reduced performance. For example, Random Forest and XGBoost base estimators trained for country-level prediction incurred their highest MRE scores when fit on datasets for which all rows and columns with greater than 85% missing data were removed (Figures 5.10, 5.11). Therefore, the strict 85% threshold resulted in

loss of important from the training data. This finding was validated by a large body of research, which discusses how removing samples with missing data can introduce bias (Twala, 2009). The risk of introducing bias is heightened when data is missing not at random, as in this case, removing samples with missing data can obscure important trends. For example, countries with more missing data, and thus less robust data collection systems, may have higher MMRs. While the other base estimators trained in this thesis did not show a similar trend, this result was taken as a warning against using even stricter missing data thresholds, which could cause the loss of more important information and result in higher predictive error.

6.1.2 Base Estimators Did Not Have Consistently Higher Performance on a Specific Feature Subset

No single feature selection method uniformly produced the highest performance for both country-level prediction and forecasting (Section 5.4).

However, lower error for country-level prediction was observed more consistently when base estimators were fit on features discussed in the literature as having a significant relationship with MMR (Figures 5.10, 5.11, 5.12). Moreover, the Random Forest Stacking Ensemble trained for country-level prediction (RFSE-CLP) placed the second highest performance on predictions from the XGBoost base estimator fit on these literature-based features (Table 5.5). Thus, MMR predictions were more consistently accurate when based on known risk factors, as expected. Using variables significantly related to MMR also made the models more robust to outliers, as this higher performance was more noticeable when measured in terms of MSE than MRE. In the context of this thesis, outliers were likely high MMR values associated with low-income countries, as there were only 78 samples from low-income countries, which had a high mean (657 deaths per 100,000 live births) and standard deviation (453 deaths per 100,000 live births). Therefore, using variables correlated to, but not causally related with, MMR may have produced inaccurate MMR predictions for low-income countries because the correlation did not extend to the higher MMR range.

Similarly, base estimators trained to perform forecasting achieved low MSE when they were fit on the literature-based subset (Figures 5.13, 5.14, 5.15). However, they achieved low MSE scores most consistently when fit on the subset of features that had an absolute correlation coefficient with MMR of greater than 0.6 (the 'Correlation 0.6' feature subset). Additionally, the Random Forest Stacking Ensemble used for forecasting (RFSE-F) placed the highest importance on base estimators fit on the 'Correlation 0.6' subset (Table 5.6). This feature subset contained more variables than the literature-based subset (113 vs 40). Potentially, a broader array of features was needed to perform forecasting than country-level prediction, as future MMR rates depend on long-term trends as well as immediate risk factors. The models' lower error when trained on the 'Correlation 0.6' subset was more noticeable when measured in terms of MSE than MRE, indicating that models trained on this feature subset were more robust to outliers. Potentially, at least one of this subset's higher number of features had a meaningful relationship with MMR

when rates of maternal mortality were high, allowing the model to extrapolate into the future for countries with more extreme MMRs.

Interestingly, models fit on the 'Correlation 0.7' feature subset generally had higher predictive error than models fit on the 'Correlation 0.6' and literature-based subsets (Figures 5.10, 5.11, 5.12, 5.13, 5.14, 5.15). This was despite the 'Correlation 0.7' subset having 5 more features than the literature-based subset. Potentially, the features in the 'Correlation 0.7' subset had insufficiently strong causal relationships with MMR to perform out-of-sample predictions. Thus, a higher number of features correlated with the target may be needed to produce a similar level of accuracy as a smaller number of MMR risk factors, as shown by the high performance of models fit on the 'Correlation 0.6' subset.

The additional benefit of using the 'Correlation 0.6' and literature-based feature subsets was from their robustness to outliers rather than their higher predictive performance on the whole dataset. For example, models fit on all available features had similarly high MRE performance (Figures 5.10, 5.11, 5.12, 5.13, 5.14, 5.15). In fact, the Random Forest and XGBoost models with the lowest MRE scores were fit on all features, regardless of whether the models were being trained for country-level prediction or forecasting. This high performance was likely due to the ability of decision-tree based models to ignore irrelevant and highly correlated features when determining candidate splits at internal nodes, as discussed in the literature review. It would be interesting to further examine the performance of different models on the union of 'Correlation 0.6' and literature-based feature sets.

Despite these general trends, model performance appeared to be more heavily influenced by its specific training fold than its feature subset. More specifically, standard deviation in model performance, which was measured over five cross-validation folds, meant that the performance of models trained on different feature subsets overlapped (Figures 5.10, 5.11, 5.12, 5.13, 5.14, 5.15). Additionally, the range of MRE scores achieved by models trained to perform forecasting was small, limiting the actual improvement gained from using different feature subsets.

The exception to these statements was the 'Correlation 0.8' feature subset. In general, the base estimators fit on the 'Correlation 0.8' subset had the highest error, regardless of model type, missing data removal technique, training fold, and whether the models were trained for country-level prediction or forecasting (Figures 5.10, 5.11, 5.12, 5.13, 5.14, 5.15). The 'Correlation 0.8' subset contained just 11 features, only two of which were socio-economic (literacy rate and use of menstrual products). The other 9 features were survival probabilities to age 5, mortality rates due to broad categories of disease, and vitamin A deficiency. Thus, models trained on this feature subset may have lacked sufficient information about socio-economic trends to be able to accurately predict MMR, causing underfitting.

6.1.3 No Single Model Type Had the Best Performance Across Different Settings

No single model type had the best performance across all feature subsets and missing data thresholds (Figures 5.16, 5.17). Similar to the earlier discussion, the standard deviation in model performance indicated large overlap in the fold-specific performances of the three model types. This similarity was also demonstrated by how the MRE scores for RFSE-F models trained only on predictions from XGBoost, LightGBM or Random Forest base estimators differed by at most two percent. This result was corroborated by other studies in the literature, such as a review of gradient boosting methods conducted by Bentéjac et al. (2021), which found that the difference between the average performance ranks achieved by XGBoost, Random Forest and LightGBM across 28 experimental datasets was not statistically significant. While this analysis compared Random Forest, XGBoost, and LightGBM classifiers, it offers relevant insights into regression performance due to the shared underlying architecture. Some of the similarity between the models' performance may be attributed to the fact that all three models are ensembles of decision-trees.

Nevertheless, there were slight differences between the models. For example, the Random Forest base estimator tended to achieve lower average MRE than XGBoost and LightGBM across the five cross-validation folds (Figures 5.16, 5.17). This may be due to boosting ensembles' known tendency to overfit, as the boosting mechanism explicitly corrects the mistakes of base estimators in the ensemble (Mahajan et al., 2023). However, the standard deviation of the Random Forest and XGBoost models' MRE scores indicated that XGBoost models had better fold-specific performance when trained for country-level prediction ((Figures 5.10, 5.11)). This observation prevented Random Forest base estimators from being considered the best-performing model. This lower, fold-specific performance may be due to Random Forest's default feature subsampling, where it only considers a subset of samples when deciding each internal node split (Ganaie et al., 2022). While this technique reduces overfitting, it can result in important features being underutilised when deciding splits, reducing its performance (Mahajan et al., 2023). The fact that this trend was only observed for models trained for country-level prediction, not forecasting, reinforces the random nature of this training process (Figures 5.10, 5.11, 5.12, 5.13, 5.14, 5.15).

XGBoost models trained for country-level prediction had the best performance, i.e., the lowest MSE scores (Figure 5.16). This may be attributable to their boosting ensemble mechanism, where each subsequent base estimator in the ensemble is explicitly trained to correct the errors of its predecessor (Mahajan et al., 2023). Later XGBoost base learners could therefore focus on learning trends to correctly predict outliers, as these would have resulted in the highest error for earlier learners. As a result, XGBoost models would have high performance on outliers, and thus lower MSE. While LightGBM models also use boosting, they may have had lower performance due to their use of gradient-based one-side sampling (Ke et al., 2017). Briefly, this technique reduces the number of samples used to split each node by undersampling less informative datapoints from

the full input dataset. However, there is no guarantee that these samples are completely uninformative. Therefore, exclusion of these samples may have produced slightly lower predictive accuracy for LightGBM than for XGBoost. Given that the Random Forest model uses bagging, it could not benefit from this iterative learning process, potentially reducing its predictive performance for outliers. Interestingly, XGBoost models trained to perform forecasting did not achieve the lowest MSE scores. Instead, the lowest scores were shared by the LightGBM and Random Forest models. Potentially, when performing forecasting, the overfitting caused by the boosting algorithm outweighed the benefit of accurately predicting outliers (Figure 5.17). Overfitting could be particularly harmful when forecasting MMR given the differences between the train and test ground truth MMR distributions described earlier in this thesis.

As a final discussion point, XGBoost models tended to have higher standard deviation in their performance than the Random Forest models (Figures 5.16, 5.17). This may also be due to the boosting algorithm's propensity to overfit to the training data, where the variable performance was related to how well the validation fold was represented by the training data (Mahajan et al., 2023). Again, LightGBM's GOSS algorithm may have reduced overfitting, resulting in lower standard deviation in its error across the different cross-validation folds.

6.1.4 Summary

In summary, no single model type, feature subset, or missing data threshold consistently had the highest performance for both country-level prediction and forecasting (Figures 5.16, 5.17).

I observed that a model's performance was more strongly influenced by the data from its specific training fold than its model type, feature subset, or missing data threshold. This makes sense, as samples were randomly assigned to different cross-validation folds, introducing the potential for models to be trained on less representative data. For example, there was a wide range of ground truth MMR estimates for low and lower-middle income countries. By chance, samples with higher MMR values may have been randomly allocated to only some of the folds. Therefore, some of the base estimators may have been trained on data that did not allow them to learn which patterns in the feature data implied extreme MMR values. As a result, models' performance on different folds varied.

This observation motivated the use of an ensemble model to combine predictions from the different base estimators, as the ensemble could benefit from patterns learned by each individual base estimator.

6.2 Discussion of Voting and Stacking Ensemble Performance

The RFSE had the highest overall performance because it could flexibly combine predictions from most important base estimators based on the specific context of the sample whose MMR was being estimated (Figures A.9, A.10). However, the performance improvement from using a RFSE was greater when it was trained to perform country-level prediction than forecasting because the forecasting base estimators had more uniform performance (Figures 5.23, 5.24). Additionally, overfitting sometimes caused the RFSE to have lower performance than the other ensembles (Figures A.9, A.10).

As expected, the RFSEs' MSE performance decreased as income level decreased due to lower-income countries having fewer available training samples that covered a wider range of MMR values (Figures 5.28, 5.29). Trends in its MRE performance were less consistent as they were more heavily dependent on the distribution of data between the models' training and testing sets. I found that income-level specific trends were more useful for MMR estimation when the models were trained to perform forecasting than country-level prediction, as the drivers of MMR over time tended to vary across different income levels (Figures 5.31, 5.32).

I discuss and explain these trends in greater detail in the following section.

6.2.1 The Random Forest Stacking Ensemble Generally Had Higher Performance Relative to Ensemble Models

Models fit on training folds that only contained samples with low MMRs would have a nuanced understanding of the trends that imply low maternal mortality. Similarly, models trained on samples with high MMRs would have learned the patterns in the feature data that are predictive of high maternal mortality. Combining the predictions of these models would therefore produce higher performance than using one or the other. The RFSE was the best performing voting/stacking ensemble because it most effectively learned which combination of base estimators produced the highest predictive performance (Figures A.9, A.10) (Mahajan et al., 2023). This was seen explicitly by it giving high importance scores to only a small subset of base estimators, with almost the exact same subset of base estimators chosen by the RFSE-CLP between retraining instances (Figure 5.20 and Table 5.4).

The RFSE-CLP and RFSE-F models placed the greatest importance on XGBoost base estimators, likely due to their high fold-specific performance (Figure 5.20). XGBoost's low error was also observed in how the RFSE-CLP trained solely on XGBoost models had smaller error than the RFSE-CLPs trained on just LightGBM and Random Forest base models (Figure 5.25). However, the actual improvement in MRE from using a single type of base estimator to train the RFSE-CLP or RFSE-F was always less than 1% (Figures 5.25), 5.26. This was due to the RFSE's decision-tree architecture, which allowed it to ignore base estimators that did not reduce error by not using those uninformative base

estimators to define splits at internal nodes. Given many of the base models had similar performance, this ability was particularly useful.

In contrast, the Voting Ensemble and Elastic Net Stacking Ensemble (ENSE) may have found it difficult to isolate the impact of specific base estimators, as models based on linear regression struggle with multicollinearity (Chan et al., 2022). As described in the literature review, multicollinearity occurs when features are linearly dependent. The similarity of my base estimators suggests they were linearly dependent, making it difficult for the Elastic Net and Voting models to isolate the effect of a specific base estimator and thus select the most important base estimators. As a result, these ensembles placed important on a larger subset of base estimators than the RFSE, potentially causing the ensembles to learn unimportant and ungeneralisable differences between base estimators (Figures 5.21, 5.22). This may explain why the Voting Ensemble and ENSE generally lower performance than the RFSE (Figures A.9, A.10).

Additionally, after the ensemble architecture was fixed during training, the Voting Ensemble and ENSE estimated MMR by always combining predictions from their base estimators in the same, fixed proportion. Put differently, they could not tailor how they combined the base estimators' predictions based on specific, local trends in the data. In contrast, and as explained above, the RFSE could vary the combination of base estimators it used to estimate depending on the local context (Mahajan et al., 2023). Therefore, it could produce more nuanced, accurate MMR predictions than the Voting Ensemble or ENSE.

Interestingly, there was one instance where the RFSE did not have the best performance. When trained for country-level prediction, the RFSE-CLP had the lowest MRE score but the ENSE had the lowest MSE score (Figure A.9). This indicates that the ENSE handled outliers more effectively, potentially due to the RFSE's overfitting to local patterns in the training data, which the ENSE was less able to do.

ENSE also outperformed the Voting Ensemble in all evaluated cases, potentially because it could attach negative weights to base estimators, whereas the voting ensemble was only fine-tuned with positive weights (Figures A.9, A.10) (Zou and Hastie, 2005). As a result, the ENSE model could learn how to combine different over- and under-estimations of MMR from its base estimators while the Voting Ensemble was forced to use only an additive combination of base learners, reducing its ability to 'correct' inaccurate base estimator predictions. Therefore, ENSE could learn more nuanced patterns, explaining its higher accuracy.

The stronger performance of my stacking versus voting ensembles was validated by the literature. In their review, Mahajan et al. (2023) found that 82.6% of studies that used stacking ensembles found they had the highest performance of all tested model architectures. In contrast, only 71.4% of studies that used voting ensembles found they had the highest performance.

However, the support vector machine stacking ensemble (SVMSE) incurred the highest

MRE and MSE when trained to perform forecasting and the highest MRE when trained for country-level prediction (Figures A.9, A.10). This weak performance may be due to its sensitivity to noise, as its loss function and predictions are heavily influenced by datapoints outside its error tolerance margin (Smola and Schölkopf, 2004). Therefore, the model may have overfit to the small number of high MMR samples, as predictions for these outlier-like datapoints likely incurred greater error than the model's tolerance. However, the SVMSE achieved a smaller MSE score for country-level prediction than the Voting Ensemble (Figure A.9). This may be due to the SVMSE's polynomial kernel (chosen through hyperparameter tuning), which captured local information about base estimator interactions. Thus, the SVMSE benefited more strongly from being able to combine different base estimators than the Voting Ensemble, which had to use the fixed, additive combination of base estimators discussed above.

As a final discussion point, there was less variation in the voting and stacking ensembles' performance when trained to perform forecasting than country-level prediction (Figures A.9, A.10). This was likely due to the forecasting base estimators having more uniform performance (Figures 5.13, 5.14, 5.15). When trained to perform forecasting, base estimators were fit on data from every country, allowing each base estimator to 'see' extreme MMR values (Section 4.4.1). In contrast, base estimators trained for country-level prediction were fit to data for a random subset of countries, which may or may not have been associated with extreme MMR values (Section **??**). Therefore, there was less variation in the predictive error of base estimators trained to perform forecasting because there was less variation in their training data. As an example of their more uniform performance, the fold-specific performance of XGBoost models trained for forecasting was not meaningfully higher than the performance of the other model types (Figure 5.17). Furthermore, the MRE of RFSE-Fs trained on just one model type only varied by 2% (Figure 5.26). As a final example, the importance scores of base estimators in the RFSE-F changed notably after the RFSE-F was retrained, indicating the different base estimators could be re-weighted without large impacts to the RFSE-F's performance (Table 5.4). These smaller differences may have decreased the benefit produced by the RFSE-F's use of the best base estimators, reducing its edge over the other ensemble models. It also explains why the RFSE-F had the same MRE as the best-performing base estimator (Figure 5.24).

6.2.2 Variation in Random Forest Stacking Ensemble Performance Across Income Levels

Previously, I compared the Random Forest Stacking Ensemble to other stacking/voting ensembles and base estimators. In this section, I discuss how its performance varied when predicting the MMR of countries in different income levels.

95% of maternal deaths occur in lower-middle and low-income counties or fragile settings (Cresswell et al., 2025). This heterogeneity was seen clearly in my input data, where the median MMR for low-income countries was 617 (Table 5.2). In contrast, the median MMR for high-income countries was 8. Additionally, key summary statistics about

features known to be associated with maternal mortality, such as the percentage of women having access to prenatal care, varied with income-level in my input data (Souza et al., 2023).

The range of possible MMR estimates, as well as median MMR, increases as income level decreases (Cresswell et al., 2025). For example, in my dataset, the standard deviation in the MMR of low-income countries was 453, compared to 55 for upper-middle income countries (Table 5.2). As well as having the highest variation in MMR, low-income countries also had the smallest number of available samples (Table 5.1). More explicitly, only 78 low-income samples remained in my dataset after removing all samples missing an associated MMR estimate. As a result, I expected the RFSE-CLP/RFSE-F's performance to deteriorate for lower income levels, as it had to learn how to predict a wide range of possible MMR values using only a small number of samples. In contrast, 1,405 high-income samples remained after pre-processing, allowing the model to more easily learn patterns corresponding to high-income countries, the existence of which were indicated by the cluster of high-income samples visualised using PCA (Figure 5.5).

My results confirmed this hypothesis, with the MSE of both the RFSE-CLP and RFSE-F decreasing as income level increased (Figures 5.28b, 5.29b). The differences in MSE spanned multiple orders of magnitude, clearly demonstrating that large differences between predicted and ground truth MMR values were more likely for low-income countries for the reasons discussed above. The greater uncertainty and thus higher potential for error when predicting large ground truth MMR estimates was also shown by how consensus in the base estimators' predictions of MMR decreased substantially as MMR increased.

However, trends in MRE across income levels were more complex (Figures 5.28a, 5.29a). The MRE score is a better benchmark for model performance on the entire dataset, as it penalises large outliers less heavily than MSE. Therefore, MRE was more strongly influenced by whether the distribution of ground truth MMR values in the test set was similar to the distribution in the train and validation sets. Given the different train/validation/test sets used for the RFSE-CLP and RFSE-F models, their MRE scores were discussed separately.

Country-Level Prediction

When the RFSE was trained for country-level prediction, test MRE generally decreased as income-level increased, as expected (Figure 5.28a). However, the RFSE achieved its lowest MRE when predicting MMR for lower-middle income countries. This unexpected result was due to the test ground truth MMR distribution for lower-middle income countries being a small subset of the corresponding train/validation MMR distribution (Figure 5.7). More explicitly, the two distributions had the same average MMR value (52), but the test Q1 to Q3 range was between 41 and 60 while the train/validation Q1 to Q3 range was 33 to 283. Therefore, the RFSE's training data completely covered the test data, enabling the model to learn the exact patterns it needed to accurately predict

the test MMR, producing very high performance for lower-middle income countries. Furthermore, since the test set's Q3 was over 200 points lower than the train/validation set's Q3, the RFSE did not need to estimate the more difficult, high magnitude MMR values when being applied to the test set. This explains why the test MRE for lower-middle income countries was smaller than the associated validation and train MREs, which was contrary to expectations (Figure 5.28a).

While the test MMR data for upper-middle and high-income countries was also within their train/validation distributions, their Q1 and Q3 train/validation values more similar to their test Q1 and Q3 values (Figure 5.7). Therefore, the test sets were more similar to the train and validation sets. Given this similarity, and the fact that the model was fit and fine-tuned on the train and validation sets, the RFSE achieved higher performance on the train and validation sets than on the test set for these income levels (Figure 5.28a).

The RFSE's low MRE score for lower-middle income countries did not translate into its MSE score because the lower-middle income test dataset contained high, outlier MMR values of similar magnitudes as the outliers in the train/validation set (Figures 5.28a, 5.28b). As explained above, these outliers represented a wide range of possible MMR values covered by a small set of datapoints, reducing model performance and causing MSE for lower-middle income countries to be between the MSE scores of the low and upper-middle income countries. While these outliers produced large predictive errors, they were likely infrequent in the lower-middle income dataset, as they did not notably reduce MRE.

In contrast to the lower-middle, upper-middle, and high-income test distributions, the low-income test MMR distribution was higher than its train/validation distribution (Figure 5.7). The test set's Q1, Q2, and Q3 values exceeded the train/validation set's values by 126, 162, and 103, respectively. As a result, the RFSE was forced to predict on samples whose ground truth MMR values were higher than the range of MMR values used to train the model. This contributed to the RFSE's high MRE and MSE scores for low-income samples (Figure 5.28a). It was also responsible for the large standard deviation in the test MRE for low-income countries, as the model's performance would have varied greatly depending on whether it was predicting for a sample whose ground truth MMR was within its training data.

Despite this income-level specific performance, similar MRE scores were achieved by the RFSE trained on all data but tested on data from a specific income level and the RFSEs trained and tested using data from a specific income level (Figure 5.31). Therefore, the RFSE trained on all data identified patterns it could use to predict MMR regardless of income level. For example, proportion of pregnant women with non-communicable would be predictive of MMR across all income levels (Souza et al., 2023). These common patterns were visualised using PCA, where samples from low, lower-middle, and upper-middle countries were projected into the same strip of points (Figure 5.5). However, the RFSEs trained on income-specific data had higher MSE scores than the RFSE trained

on all data (Figure 5.31). Potentially, training the model on data from all income levels allowed it to better predict high, or "outlier" MMR values, as it would see a wider variety of possible MMRs. While the RFSE trained on lower-middle income data had a lower MSE than the model trained on all data, the difference between the MSE scores was negligible, suggesting this deviation from the general trend was due to noise.

Forecasting

The RFSE trained to perform forecasting had the opposite trend in MRE performance as expected, as its MRE scores increased with income-level (Figure 5.29a). This was attributed to outlier years in the model's test ground truth MMR distribution (Figure 5.8).

Each of the upper-middle and high-income datasets contained at least one year with a median MMR value meaningfully greater than any median MMR values observed in the associated train/validation sets (Figure 5.9). The RFSE may not have anticipated these relatively high MMR values for higher-income countries, increasing its predictive error (Figure 5.29a). This may have also caused high standard deviation in the RFSE's MRE scores, as its prediction error would have varied depending on whether it was predicting for a sample that had an MMR within the range seen during training. In contrast, the median MMR values in the low and lower-middle income test sets were within the range observed in their train/validation sets (Figure 5.8). This explains the model's relatively higher performance for low and lower-middle income countries (Figure 5.29a).

Therefore, differences between the train/validation and test sets were responsible for the RFSE making prediction errors that reduced its MRE performance for high-income samples. However, these mistakes were likely minor, as the actual magnitude difference between the test and train/validation MMR values were small (Figure 5.8). This explains why the RFSE still had better MSE performance for higher income countries than lower-income countries, which had outliers of higher magnitude (Figure 5.29b).

Interestingly, the sensitivity analysis showed training the RFSE on data from all income levels incurred a higher MSE score than training it on income level specific data (Figure 5.32). This may be due to change in MMR over time being driven by different features for different income levels. As described by the obstetric transition model mentioned in the background information, the main drivers of maternal mortality change from direct, pregnancy related issues in countries with high MMR values to non-communicable disease in countries with lower MMR (Souza et al., 2014). For example, countries with the highest MMR values, in the first few stages of the transition model, can significantly reduce MMR by increasing access to care. In contrast, countries with lower MMRs in the later stages of the transition model can more effectively reduce MMR by increasing quality of care and reducing overuse of medical interventions. Therefore, training the RFSE to learn trends in features specific to each income level may reduce noise and allow it to focus on relevant trends for the countries' stage in the transition model. As a result, the RFSE trained on all data generally had higher MSE and MRE scores than

when trained on income-specific data (Figure 5.32).

However, training the RFSE on all data produced a lower MSE than training it solely on lower-middle income data (Figure 5.32). Therefore, the RFSE had better predictive accuracy for lower-middle income countries when it could learn how trends in countries at different stages of the transition model affected MMR. These lower-middle income countries would likely be categorised as in Stage 3 of the transition model, where MMR is between 50 and 299 (broadly aligning with the summary statistics for lower-middle income countries) (Table 5.2). The literature describes how countries in this stage benefit from increasing both access to care and quality of care, bridging the early and late stages of the transition model (Souza et al., 2014). Thus, lower-middle income countries in the intermediate transition state benefit from learning how trends affecting countries in all transition stages influence MMR.

6.2.3 Summary

In summary, rare, high magnitude MMR values associated with lower-income countries increased absolute predictive error. Therefore, the MSE of the RFSE-CLP and RFSE-F increased as income-level decreased (Figures 5.28, 5.29). Trends in the models' MRE performance were more greatly influence by distribution of ground truth MMR values between their train/validation and test datasets. Sensitivity analysis indicated that the RFSE-CLP could learn trends generalisable to all income levels (reinforced by PCA), but the RFSE-F benefitted more greatly from learning income-level specific trends, as change in MMR was driven by different factors for countries from different income levels (Figures 5.31, 5.32).

6.3 Comparison of My Models' MMR Predictions to the Literature's Estimates

Building on the discussion of my RFSE's performance, in this section I examine the differences between its MMR estimates and those produced by the BMat, CODEm, and GMatH models (Figures 5.35, 5.36) (World Health Organization, 2025; GBD 2021 Causes of Death Collaborators, 2024; Ward et al., 2023a). Briefly, BMat used Bayesian hierarchical modelling to estimate non-HIV related MMR, with HIV-related MMR calculated separately (World Health Organization, 2025). CODEm used an ensemble of linear mixed effects regression and spatiotemporal Gaussian process regression (GBD 2021 Causes of Death Collaborators, 2024). In contrast, the GMatH microsimulation model simulated the reproductive lifecycles of thousands of women (Ward et al., 2023a). All three models used Bayesian hierarchical modelling, with BMat and CODEm using only 3 and 19 covariates while GMatH had a wide range of parameters.

The majority of my models' MMR predictions fell within the 95% confidence intervals of the predictions from at least one of the existing literature models (Table 5.7, Figures 5.35, 5.36). However, regardless of whether my RFSE was trained to perform country-level

prediction or forecasting, many of its MMR estimates were smaller than the associated estimates from the BMat, CODEm, and GMatH models (Figures 5.33, 5.33). Generally, my estimates were most different from the corresponding GMatH predictions. The proportion of my MMR estimates within the 95% confidence interval of the literature's estimates was within a few percent of the proportion of ground truth MMR estimates that were within those same confidence intervals (Table 5.7. Therefore, the differences between my predictions and the literature's estimates were primarily due to a lack of correspondence between the literature's MMR predictions and the ground truth MMR values used to train my model.

I hypothesise that the ground truth MMR values used to train my model were lower than the estimates produced by the BMat, CODEm, and GMatH models was because I did not adjust my ground truth MMR values for underreporting and misclassification errors. These ground truth estimates were sourced from national estimates. However, studies have hypothesised that maternal mortality is underestimated by at least 40%, with large differences between the quality and quantity of data collected by different countries (Ahmed et al., 2023). For example, in 2017, only 2 of 49 least developed countries had a death registration coverage of at least 50% (Mgawadere et al., 2017). Therefore, my model may have been trained on ground truth MMR values that were underestimates of the true maternal mortality.

The literature models all developed procedures to address low-quality input data, as discussed in Chapter 3. For example, the UN MMEIG developed the BMis model to correct for errors in data from CRVS systems (Peterson et al., 2024). The Global Burden of Disease Study implemented algorithms to reassign deaths attributed to nonsensical causes to more statistically probable causes, reducing misclassification (Johnson et al., 2021). The GMatH model incorporated specific parameters that captured underreporting of maternal death (Ward et al., 2023a). In contrast, I did not adjust my ground truth MMR values for under-reporting and misclassification errors, explaining why my MMR predictions were generally smaller than the corresponding BMat, CODEm, and GMatH estimates.

There were also methodological differences between the RFSE, BMat, CODEm, and GMatH models that contributed to variation across their MMR estimates. For example, GMatH estimated MMR using a variety of parameters informed by prior distributions, unlike my decision-tree based models (Ward et al., 2023a). According to GMatH's documentation, some of the parameters used for high-income countries were estimated using the prior distributions of upper-middle income countries (**?**). This substitution was used for feature supplied by Demographic and Health Surveys, which only collect data from lower-income countries, thus preventing informative priors about high-income countries from being estimated. As a result, GMatH's MMR estimates greatly exceeded the other models' predictions for high-income countries, sometimes by over 100 points (Figures 5.35, 5.36). This also explains why my model's MMR estimates were generally much smaller than GMatH's estimates, as much of my test data was for higher-income countries (Figures 5.33, 5.33). Nevertheless, the true MMR values likely lay between

the 4 models' predictions, as the other models may have underestimated MMR. For example, CODEm was found to underpredict the diabetes-induced mortality in Germany, suggesting its potential for underestimation (Rommel et al., 2018).

Another methodological difference between my model and the models in the literature was my lack of assumptions about regional homogeneity and temporal smoothness. The literature models assumed a certain amount of regional homogeneity to be able to use regional means to predict MMR and other important features for countries with sparse data (World Health Organization, 2025; GBD 2021 Causes of Death Collaborators, 2024; Ward et al., 2023a). As a result, the models may have ignored country specific information if the regional mean was not representative. Similarly, CODEm's spatiotemporal models employed smoothing over time, preventing the model from representing crisis years where MMR increased sharply(Foreman et al., 2012). These methodological differences may have contributed to my country-level MMR predictions for lower-middle and low-income countries being higher than the literature's estimates (Figure 5.33). However, these overestimations for lower-income countries must be interpreted with caution given my higher test error for low-income countries discussed in the previous section.

The previous two discussion points reinforce that one of my model's strengths is that it did not need to make assumptions about the underlying data that could bias my results.

Variation in the models' MMR estimates was also driven by their use of different features. While BMat and CODEm only had 3 and 19 covariates, respectively, my model used 720 features (World Health Organization, 2025; GBD 2021 Causes of Death Collaborators, 2024). BMat's MMR predictions were increasingly covariate-driven for countries with sparse maternal mortality data (World Health Organization, 2025). Thus, the difference between BMat's covariates and my model's features likely contributed to the discrepancies between their MMR estimates for lower-income countries, which have less data. Additionally, my model incorporated information from a more diverse pool of socio-economic, health-related, and environmental features than BMat or CODEm. As a result, it could use a wider variety of information to predict the MMRs of data-sparse areas. This may have resulted in its MMR estimates for low-income countries being higher than the corresponding BMat and CODEm estimates (Figures 5.35, 5.36).

In addition to using different features than my model, BMat and CODEm assumed a relatively global relationship between their covariates and MMR, unlike my model (World Health Organization, 2025; GBD 2021 Causes of Death Collaborators, 2024). However, this relationship may change based on local conditions, potentially reducing the validity of the literature models' predictions. For example, skilled birth attendance (SBA) was one of the covariates used by both BMat and CODEm. However, it only has a significant relationship with MMR when national SBA coverage is at least 40%, indicating its relationship with MMR was not globally applicable (McClure et al., 2007). In contrast to the literature models, my RFSE's predictions were based on 'local' information, as they were derived from mappings between specific subsets of the input space and terminal decision-tree nodes. Therefore, my model could better represent local condi-

tions, especially when missing data made it difficult for the other models to adjust their covariate-driven estimates. This may have contributed to my model's MMR estimates for lower-income countries exceeding the literature models' estimates, as my model was less likely to overestimate the effect of features like SBA on MMR based on inapplicable global trends (Figures 5.35, 5.36).

In contrast to BMat and CODEm, GMatH used a variety of parameters covering biological information, quality of health care, and socio-economic trends (Ward et al., 2023a). Each of GMatH's parameters were associated with uncertainty, the combination of which may have contributed to GMatH's wide confidence intervals and impacted model estimates. In contrast, my decision-tree based models could ignore features that did not contribute to reducing loss at internal nodes, preventing additive error and uncertainty from uninformative features from overly influencing my model. This could be particularly important in the context of low-income countries, where all estimates are more uncertain, as there is less available data to fuel predictions and parameter choices.

In summary, differences between MMR estimates produced by my RFSE, BMat, CODEm, and GMatH were driven by underestimation of MMR in my ground truth data as well as methodological variation and differences between the features used to estimate MMR.

6.4 Discussion of Feature Importance

The secondary aim of my research was to determine the socio-economic and health related features with the highest predictive power for MMR. These features were identified as the most valued variables in the base estimators given the highest importance scores by the RFSE (Tables 5.5, 5.6) in Chapter 5. I discussed the results of my investigation in this section.

The features with the highest predictive power for both country-level MMR predictions and MMR forecasts provided nationally aggregated information about the level and type of women's employment, women's knowledge of contraceptive options, and medical outcomes related to women's nutritional status (Tables 5.5, 5.6). Therefore, socio-economic trends provided valuable information for estimating MMR. This observation was corroborated by the literature, which described how maternal mortality is affected by socio-economic trends like availability of contraception and women's education (Tunçalp et al., 2014; Utomo et al., 2021). The literature also describes how maternal health services are inaccessible to many due to financial constraints, affecting maternal health outcomes(Koblinsky et al., 2016). Given that employment is influenced by education, and that financial status is influenced by employment, these findings validate my result that employment has high predictive power for MMR.

The features with the highest predictive power for MMR were similar between the RFSE-CLP and RFSE-F models (Tables 5.5, 5.6). However, the most important features in RFSE-F had slightly greater emphasis on non-communicable disease and lower empha-

sis on the national income level, life expectancy, fertility rate, presence of skilled health practitioners at birth, and percentage of teenage mothers. Therefore, the RFSE-CLP placed greater importance on current aggregate measures of health system coverage and performance, while the RFSE-F placed more value on longer-term health trends. This observation was corroborated by the literature, as experts expect maternal mortality will be increasingly determined by non-communicable disease as opposed to direct complications of pregnancy and childbirth, as described by the obstetric transition model (Souza et al., 2014).

Many of the features identified in this thesis as having the highest predictive power for MMR were established risk factors in the literature, validating my results (Cresswell et al., 2025; Koblinsky et al., 2016; Souza et al., 2023; Tunçalp et al., 2014).

However, my feature importance calculations may have been affected by the large number of highly correlated variables in my input data. A feature's importance was calculated based on the amount by which it reduced loss when used to define a split at an internal node. However, a group of correlated features may have produced similar reductions in loss. Thus, the first feature chosen from this group would be used to define the split and attributed with the decrease in loss. The chosen feature would therefore be given a higher importance score than any of the other highly correlated features. Consequently, its higher importance score would be a result of its selection order instead of an indication of it having a causal relationship with MMR. Thus, the similarity and high correlation between features in my input data made importance scores unstable.

Therefore, I conducted an additional robustness check using SHAP values to give my results further validity. Shapely Additive exPlanation (SHAP) values provide a stable measure of feature importance based on cooperative game theory (Lundberg and Lee, 2017). They represent each feature's contribution to the final prediction and are commonly used in machine learning studies (Lundberg and Lee, 2017). For example, (Taye et al., 2025) calculated the SHAP values of their Random Forest model to determine that place of delivery and place of residence (e.g. rural) had high predictive power for whether a birth in sub-Saharan Africa is attended by skilled birth personnel. I calculated the SHAP values for the two base estimators given the highest importance scores in my RFSEs. The features given the highest SHAP values were almost the same as those given the highest importance scores in my thesis, with the major difference being that SHAP values placed more emphasis on adult and infant total mortality rates. This broad similarity gave my findings further validity. As described earlier, my results were also corroborated by the literature, which describes their meaningful relationships with MMR.

6.5 Policy Implications of this Research

The primary aim of my thesis was to estimate country-level MMR values for each year between 1985 and 2018 using interpretable machine learning methods. These estimates

can be used to inform global and national health policy, giving explicit information about which regions are most in need of targeted health interventions to reduce maternal mortality. I accomplished this aim by developing my RFSE-CLP and RFSE-F models. They can be used to monitor current trends in maternal mortality for data-sparse countries and forecast MMR under various scenarios and candidate policies.

A secondary aim of this thesis was to identify the features with the highest predictive power for MMR. Upon confirmation of their causal link to MMR, they can be used to suggest targets for public health policy. While this causal inference is out of the scope of this research, the literature discusses the causal relationships between maternal mortality and some of the features I identified as having high predictive power for MMR. For example, socio-economic features had high predictive power for MMR, reinforcing the existing body of work that argues in favour of reducing MMR through targeting socio-economic trends. For instance, Souza et al. (2023) state that only addressing the biomedical causes of maternal mortality may have prevented further reductions in MMR.

The level of female employment is a socio-economic variable that is highly predictive of MMR, as discussed previously, implying that improving women's employment outcomes would reduce MMR (Tables 5.5, 5.6). Employment is driven by, and reflective of, trends in women's access to education, women's literacy, and women's agency, all of which have important relationships to maternal mortality (Souza et al., 2023; Cresswell et al., 2025; Tunçalp et al., 2014). For example, education affects women's knowledge of health conditions and contraceptive options as well as their likelihood of engaging with maternal health services (Sheikh et al., 2024; Tunçalp et al., 2014). Risk of maternal mortality for women with no education is 2.7 times higher than the risk for women with at least 12 years of education (Sheikh et al., 2024). Therefore, increased investment in encouraging girls to finish their education may decrease MMR. For example, Warda et al. (2024) used the GMatH model to demonstrate that ensuring all women complete secondary school would result in a global MMR between 76 and 120 in 2030 (which would successfully meet the UN's Sustainable Development Goal). The authors described how the magnitude decrease in MMR produced by this socio-economic strategy would be akin to the decrease produced by policy that increases availability of clinical services and the number of women delivering their child in a health facility. Therefore, increasing investment in women's education would be an effective policy for reducing MMR.

My research has also shown that the presence of skilled medical personnel at childbirth is a powerful predictor of maternal mortality (Tables 5.5, 5.6). This finding was echoed by much of the literature, as medical complications during childbirth are one of the primary causes of maternal death (Souza et al., 2023; Cresswell et al., 2025). Therefore, policies that increase the proportion of births attended by skilled medical personnel are expected to reduce MMR. As a result, the Ending Preventable Maternal Mortality strategy, a global, multi-partner program involving governmental and international organisations, set a goal of having over 90% of births attended by a skilled medical practitioner (Chou et al., 2015; Maternal Health , MAH). To encourage progress toward this goal, the government could increase incentives for skilled medical practitioners to

work in rural areas, which likely have fewer healthcare professionals (Koblinsky et al., 2016). This candidate policy could have substantial effects in the 38 countries that have high maternal mortality burden but critical shortages in medical personnel (Chou et al., 2015).

Finally, my research demonstrated that knowledge of, and access to, contraceptive options is highly predictive of MMR (Tables 5.5, 5.6). Many studies have recognised the strong relationship between the availability of family planning services and maternal mortality (Utomo et al., 2021; Souza et al., 2023). Such family planning services decrease both the total number of pregnancies and the number of high-risk pregnancies, including pregnancies in young girls and older women (Utomo et al., 2021). Therefore, increased provision of family planning services may reduce the proportion of teenage pregnancies and mothers, which was also found to be highly predictive of MMR in my research. This finding was reinforced by research that identified complications due to pregnancy and childbirth as being one of the primary causes of death in women between 15 and 19 years old (McQueston et al., 2013). Reducing the incidence of these high-risk pregnancies would result in a lower number of women who are exposed to pregnancy-related complications and thus reduce MMR (Utomo et al., 2021). Consequently, models have predicted that family planning strategies can prevent up to 30% of future maternal deaths. As a case study, estimates suggest that Indonesia's national family planning program prevented 38 to 40% of the maternal deaths that would have otherwise occurred over the last 50 years in the absence of the program (Utomo et al., 2021). Therefore, policies that increase the availability of family planning services could meaningfully reduce MMR (Koblinsky et al., 2016).

6.6 Strengths of this Research

My proposed RFSE models successfully estimated MMR for a diverse range of countries between 1985 and 2018. My MMR predictions were similar to the MMR estimates produced by the BMat, CODEm, and GMatH models, which are high performing models in the literature, with BMat and CODEm routinely used by governments and international organisations to plan resource allocation and aid, as well as monitor global trends (World Health Organization, 2025; Murray, 2022). The similarity between my model's predictions and these literature estimates was a major strength of my research, as it indicates that my results are valid and can thus be used to inform policy. As a result, my thesis met its first aim. This finding also shows that my models can achieve comparable predictive accuracy as the literature models without needing to have a similarly heavy reliance on domain knowledge. Thus, my models may have wider applicability in low-resource settings, where domain knowledge is still developing. My results' validity also suggests that my method can provide a framework for applying decision-tree based machine learning models to sparse data to estimate other public health outcomes.

Another strength of my research was its exploration of a wide variety of socio-economic and health-related features. As discussed previously, BMat and CODEm, the most

widely used models in the literature, only use a small number of covariates to predict MMR, with a limited focus on socio-economic trends (World Health Organization, 2025; GBD 2021 Causes of Death Collaborators, 2024). The BMat model developers expressed interest in exploring alternative predictive variables for MMR in their 2010 paper (Wilmoth et al., 2012). This highlights the utility of my decision-tree based methods, which could consider the influence of a wide range of socio-economic and health-related trends when estimating MMR. My larger feature dataset also enabled a more nuanced analysis of feature importance, as I could determine which of a diverse range of features had the highest predictive power for MMR. This feature analysis could be conducted due to my use of interpretable machine learning methods, which was another strength of my methodology. The features identified as having high predictive power for MMR were established risk factors in the literature, with their robustness confirmed by SHAP analysis. This external validity was a further strength of my research.

My feature data was sourced from a variety of datasets. These datasets were collated and combined by the WHO and World Bank from sources like the Demographic and Health Surveys and various UN Inter-Agency Groups. Using pre-processed estimates from reputable sources increased the quality of my feature data.

My comprehensive investigation of a variety of pre-processing techniques was a further strength of my methodology. A range of model architectures, hyperparameter settings, feature subsets, and missing data thresholds were systematically tested and compared, providing an extensive set of experiments justifying all methodological choices.

In particular, a strength of my thesis was its exploration of how decision-tree based models can effectively handle high proportions of missing data. Missing samples are often dropped due to researchers' reluctance to introduce bias into their data with incorrect imputation (Memon et al., 2022). For example, a study in Uganda that predicted severe maternal morbidities by applying logistic regression to facility health data removed all samples with greater than 90% missing data (Memon et al., 2022). Therefore, demonstrating my model's ability to work effectively with high levels of missing data was a useful result for future maternal mortality studies, especially given that only 2 of 49 of the least developed countries having death registration coverage of at least 50% (Ahmed et al., 2023).

Another strength of my research was its lack of assumptions about the underlying data distribution, as discussed in detail in previous sections. For example, I did not need to make assumptions about prior parameter distributions, regional homogeneity, or the order of events in a micro-simulation, unlike other models used to estimate MMR (World Health Organization, 2025; GBD 2021 Causes of Death Collaborators, 2024; Ward et al., 2023a). This prevented incorrect assumptions from reducing my model's accuracy. The ability to handle sparse data without making assumptions about the data's distribution was largely due to my use of decision-tree based models, a strength of my methodology.

Finally, using an alternate model to estimate MMR encourages debate about the validity of different modelling approaches and may help provide consensus about which of the

models' various estimates is the true MMR value (Ward et al., 2025).

6.7 Limitations of this Research

Despite the strengths of my research, it had several important limitations.

My research was most limited by its use of sparse and low-quality data, as countries' official maternal mortality estimates are substantially affected by underreporting and misclassification errors (Ahmed et al., 2023). The quality of data reported by countries without robust data collection systems was particularly low. Consequently, the authors of the BMat, CODEm, and GMatH models qualify their results by cautioning readers that their MMR estimates may be biased by their low-quality input data, as it is more representative of countries with more robust data collection systems, which tend to have lower MMRs (World Health Organization, 2025; Naghavi, 2024; Ward et al., 2023a). My results were similarly limited by the low-quality data.

More specifically, due to the low-quality and sparse input data, many of my country, year samples lacked an associated MMR estimate. While there were instances of samples from all income levels being reported without an associated MMR estimate, the proportion of missing MMR estimates increased as income-level decreased (Table 5.1). As a result, there were only 78 samples from low-income countries in my input dataset, representing just 8.8% of the available low-income samples from the original, un-cleaned data. Given the wide range of possible MMR values for low-income countries, this small number of samples may have been insufficient for my model to learn how to accurately predict low-income MMR values, as seen by its higher MSE for estimates for low-income countries (Figures 5.28, 5.28).

In addition to missing MMR estimates, there was 80 to 90% missing feature data per year for the majority of years in my input data (Figure 5.3). While I have demonstrated that my decision-tree based models can work effectively with this high quantity of missing data, having 80 to 90% missing feature data per year resulted in the loss of a significant amount of useful information. If it had been recorded, this data may have had a meaningful influence on the model's results, enabling the model to learn important relationships between MMR and the features with missing data, allowing it to more accurately predict MMR.

The sparse and low-quality feature data used in this thesis also adversely affected my feature importance analysis. The high proportion of missing feature data may have masked important relationships between features and MMR. This could have been particularly relevant for the Demographic and Health Survey, which did not collect information about high-income countries (dhsprogram.com, 2025). Additionally, high correlation between features may have made importance scores unstable across different instances of my model, reducing the accuracy with which I could determine the features with the highest predictive power for MMR. As a final point, due to time constraints, the features with the highest predictive power were only identified in a small subset of base estima-

tors. There may have been meaningful differences in the most predictive features across base models, limiting my results. However, as discussed above, the features with high predictive power identified in this study were corroborated by the literature and SHAP analysis, indicating that these limitations only had a small effect on my results' accuracy.

A further limitation of my methodology was its lack of characterisation of uncertainty in its estimates, unlike the BMat, CODEm, and GMatH models (World Health Organization, 2025; GBD 2021 Causes of Death Collaborators, 2024; Ward et al., 2023a). This uncertainty could have been calculated by retraining the Random Forest Stacking Ensembles 1,000 times and measuring differences across their predictions. However, this method was not implemented in my thesis due to limited computation resources.

6.8 Future Extensions of this Research

The limitations discussed above motivated a range of possible extensions for this research. Additionally, this research can be extended through applying my models to model MMR trends under specific policy scenarios and in certain demographic groups.

An interesting extension of this work would be to develop a secondary model that calculates adjustment factors for MMR values using estimates of the extent of underreporting and misclassification errors in a country's data collection systems. This extension could be modelled on the UN MMEIG's BMis model, which adjusts for similar errors in data from country's CRVS systems (Peterson et al., 2024). This future work could investigate whether the adjustment factors should be applied to the ground truth MMR estimates used to train the model, as done for the BMat model, or to the model's predictions.

I propose two extensions to address the lack of samples from low-income countries in my dataset. The first is the use of semi-supervised machine learning methods. Briefly, semi-supervised learning is a mixture of supervised and unsupervised methods, as it uses a combination of labelled and unlabelled input data (Dabool et al., 2024; Sadeghi et al., 2024). In semi-supervised learning, the model is initially trained on the labelled data (Sadeghi et al., 2024). This model is then used to predict the labels of unlabelled samples. Finally, this initial model is retrained on all input data (both the original and newly labelled samples). Consequently, semi-supervised methods are particularly useful when labelled data is limited, such as in my thesis (Dabool et al., 2024). Semi-supervised models have been applied to health contexts, such as to identify patients with autoimmune disease (Sadeghi et al., 2024). For example, a previous study demonstrated that semi-supervised methods could more accurately detect Crohn's Disease from magnetic resonance imaging than supervised learning.

Therefore, by following a semi-supervised approach, I could use my current RFSE models to predict the MMR of samples missing their ground truth MMR values. These newly labelled samples could then be used to retrain my model. The resulting model would be trained on a much greater number of samples from low-income countries, increasing its ability to learn nuanced patterns in low-income country data. Additionally, having

access to these previously unlabelled samples would increase the size of my input data by over 3,000 samples, reducing my model's risk of overfitting (Table 5.1). While my model may incorrectly estimate the unlabelled samples' ground truth MMR, my model's similarity to the literature estimates provides confidence in its predictions.

An alternative method for increasing the number of samples from low-income countries was to use synthetic minority oversampling techniques (SMOTE). The SMOTE algorithm generates synthetic samples of underrepresented data by interpolating between the neighbours of samples from the underrepresented group (Fernández et al., 2018). SMOTE is one of most influential pre-processing techniques in machine learning. However, it is known to generate overlapping and noisy samples, which may limit its utility on my data, as PCA shows that my ground truth MMR values for upper-middle, lower-middle, and low-income countries overlap. Therefore, the neighbours of low-income samples could actually be from another income level, meaning the sample generated from interpolation between low-income neighbours may not be representative of low-income countries. SMOTE is also limited by its difficulty working with missing data. Additionally, it has been shown to produce insubstantial performance improvement when applied to high-dimensional data. Given these limitations, the technique was not applied in my thesis. However, it would be an interesting future avenue to explore to increase the availability of samples representing low-income countries. Future work could investigate modifications to SMOTE that improve its performance on sparse, high-dimensional data. The modified version of SMOTE could then be used to generate additional low-income samples, which would be added to my model's training set, thus improving its knowledge of trends in low-income data and ability to predict the MMR of low-income countries.

Additionally, it may be worth exploring whether imputing the missing feature data improves prediction accuracy by allowing the model to learn a more comprehensive set of relationships between features and MMR. For example, Twala (2009) found that training a decision tree to predict the missing data of a specific feature using the rest of the input dataset had high performance, especially when correlation between features was high. While it would be computationally intensive to impute all 720 features in my dataset using this method, it may be a worthwhile future extension of this thesis. However, implementation of this potential extension must be done carefully, as imputing this large amount of missing data is likely to introduce bias, especially given the likelihood of the pattern of missing data being missing not at random.

Another extension of this thesis is to use its models to predict sub-national MMR values. The widely used BMat model only provides country-level data at its finest granularity, preventing monitoring of sub-national heterogeneity in MMR. The ability of my models to estimate sub-national MMR values could be tested by altering feature data to represent specific sub-national geographic areas or demographic subgroups.

Similarly, specific values for my model's feature variables could be modified to simulate the effects of different candidate health policies. My model's ability to estimate the

impact of different policies on MMR could be measured by comparing its predictions to GMatH's simulated outcomes, as GMatH has also been used to evaluate potential policies (Warda et al., 2024). This would be a particularly useful extension for policymakers.

Finally, this work motivates further research into the causal relationships between maternal mortality and various socio-economic and health-related features. Specifically, I found that base estimators tended to be more robust to outliers when fit on established risk factors for MMR (Section **??**). Given that high MMR estimates tended to manifest as outliers, this result suggests the importance of causal research into the MMR risk factors specific to low and lower-middle income countries to ensure that MMR estimates and thus maternal mortality management is specific to local trends. This is particularly important given the observation that the relationship between MMR and established risk factors like skilled birth attendance are affected by local conditions (McClure et al., 2007). Investigation into causal relationships can start with the features identified as having high predictive power for MMR. It can also explore the features in the 'Correlation 0.6' subset, as base estimators trained on these features tended to have high performance, indicating that some of the features in this subset may have a causal relationship with MMR.

Chapter 7

Concluding Remarks

In this thesis, I have proposed and developed interpretable machine learning models to predict the maternal mortality ratio of 172 countries between 1985 and 2018. I used a wide variety of socio-economic and health-related indicators sourced from the World Health Organisation and World Bank. A comprehensive literature review found that, in contrast, the most widely used MMR modelling approaches used Bayesian hierarchical regression models and classical machine learning techniques with substantially smaller feature subsets (3 and 19 features versus 720 in my research). While their estimates are widely accepted, they are based on assumptions about the underlying data distribution. Therefore, my research proposes a new, alternative method for estimating global maternity mortality ratios that does not make similar assumptions about the underlying data distribution and is informed by a wider range of features.

The best-performing model architecture evaluated in this research was the Random Forest Stacking Ensemble (RFSE), which used the Random Forest bagging algorithm to combine 300 predictions from component Random Forest, XGBoost, and LightGBM base estimators. The highest performing RFSE trained for country-level prediction achieved a test mean relative error of 0.07. It can be used to monitor global and national trends in MMR, particularly in data sparse areas. In contrast, the highest performing RFSE trained to perform forecasting incurred a test mean relative error of 0.37. This model can be used to predict future MMR values and simulate the effects of candidate policies. Despite model development being limited by low-quality and sparse input data, the MMR estimates produced by both models were similar to those generated by the regression or simulation models in the literature (BMat, CODEm, and GMatH models), with this similarity validating the accuracy of my models. However, my model predictions were generally smaller than the literature's estimates when predicting the MMR values of high-income countries, potentially due to underestimation of MMR in my ground truth dataset. Differences between the literature models' predictions and my MMR estimates were also attributed to variation in the models' choice of covariates

and features, treatment of missing data, and assumptions about the underlying data distribution. My models' alterative set of MMR estimates could be used to resolve disagreement in the literature about the true MMR value when my predicted values are closer to the estimates from one model than another.

I used my models to determine that the level and type of women's employment, women's knowledge of contraceptive options, and a country's income level were socio-economic variables with high predictive power for maternal mortality. Features that benchmarked the country's fertility rates and national life expectancy, as well as and medical outcomes related to women's nutritional status and the proportion of births attended by a skilled medical practitioner also had high predictive power. These were existing, known risk factors for maternal mortality with identified causal relationships to MMR in the literature, emphasising the accuracy of this feature analysis. My results highlighted the importance of addressing the socio-economic trends driving MMR. Consequently, I suggest that investment in women's education, which influences their employment prospects, incentives for skilled medical personnel to practice in more remote areas and provision of family planning services would reduce MMR by targeting important drivers of maternal mortality.

Appendix A

Appendix

This appendix provides further information about base estimator and ensemble performance by providing their predictive accuracy in terms of Mean Average Error (MAE), Root Mean-Squared Error (RMSE), and R^2.

More specifically, I provide additional performance measures for base estimators trained on datasets curated with various pre-processing, feature selection, and missing data removal techniques (A.1, A.2). I also give extra performance metrics to compare voting and stacking ensembles (A.3) and to compare best-performing ensemble with base estimators (A.4). Then, I present additional performance metrics for the best-performing ensemble trained on various subsets of base estimators and for the highest performing ensemble's sensitivity analysis (A.5, A.6).

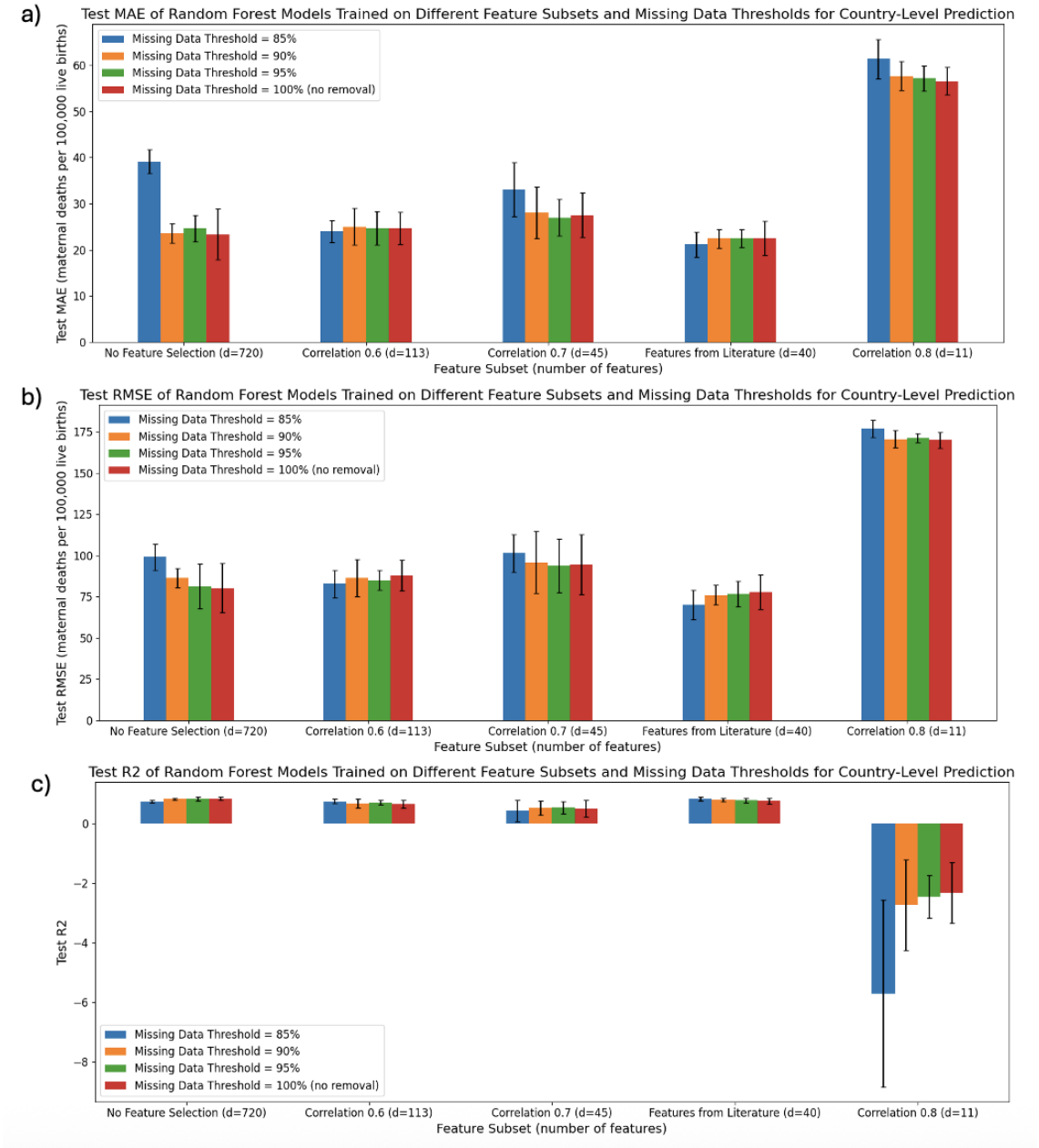


Figure A.1: a) Mean average error, b) root mean-squared error, and c) R^2 for Random Forest base estimators fit on different feature subsets and missing data thresholds for country-level prediction.

A.1 Additional Metrics to Quantify Base Estimator Performance on Different Feature Subsets and Missing Data Thresholds

A.1.1 Country-Level Prediction

Random Forest

XGBoost

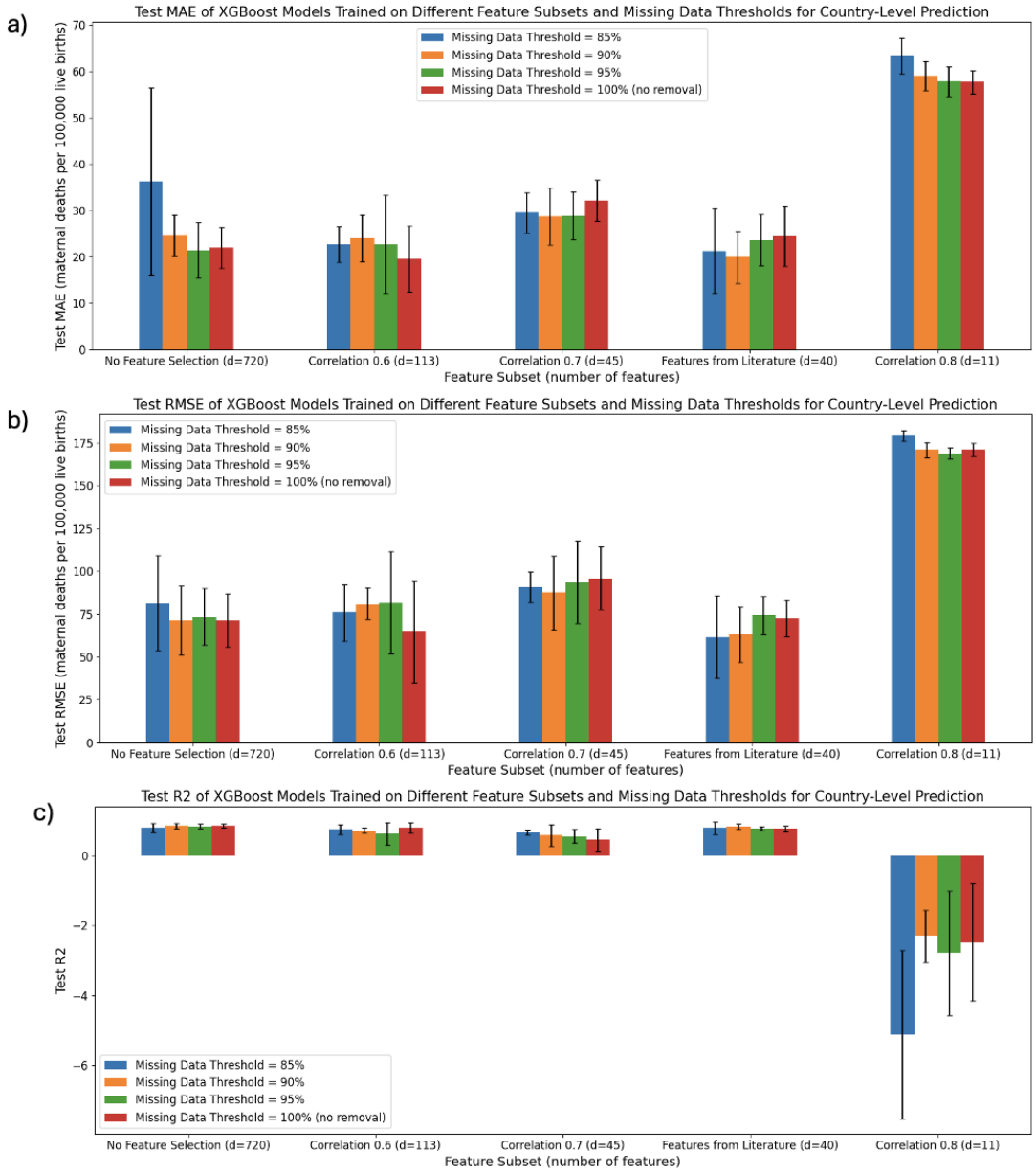


Figure A.2: a) Mean average error, b) root mean-squared error, and c) R^2 for XGBoost base estimators fit on different feature subsets and missing data thresholds for country-level prediction.

LightGBM

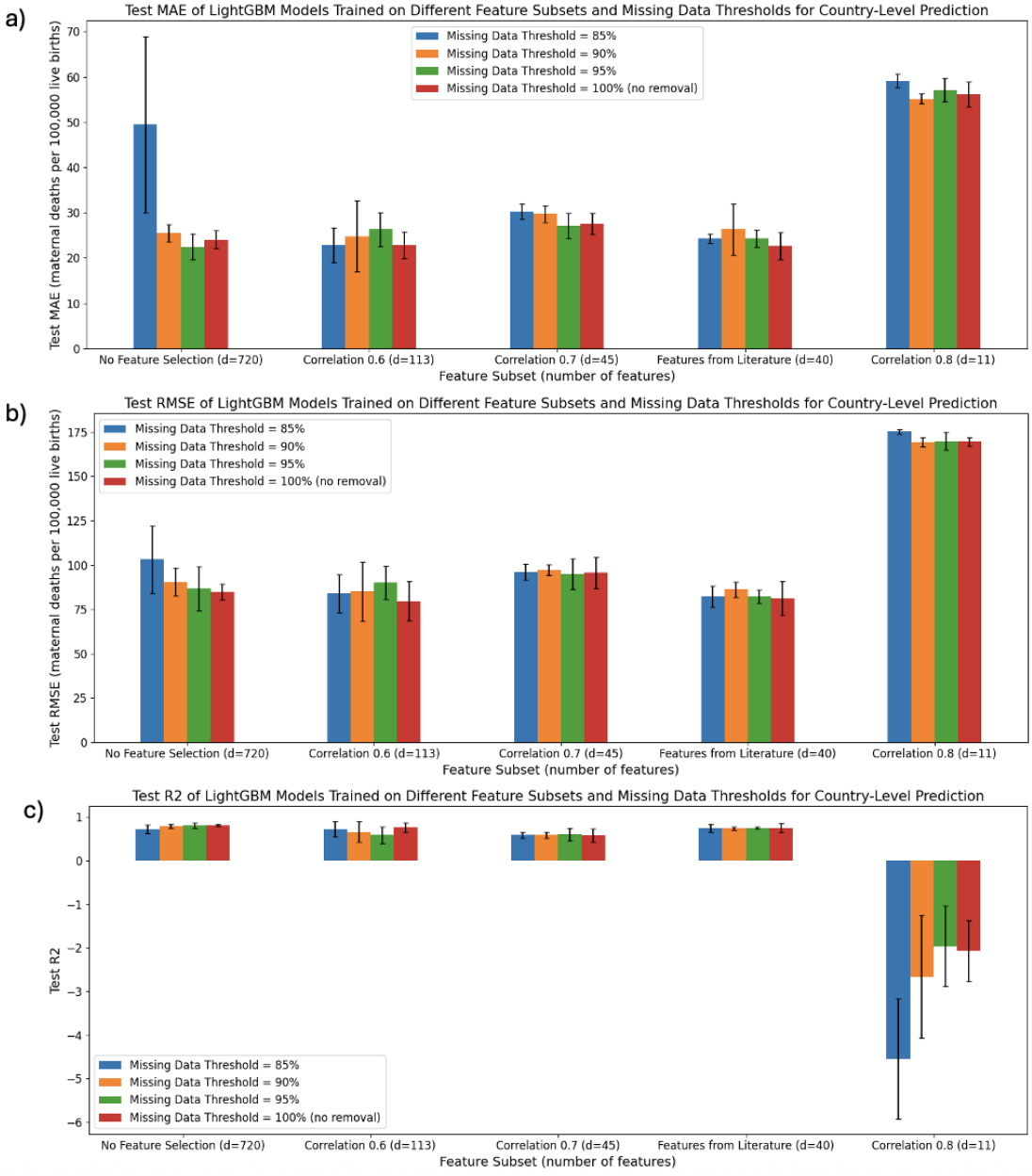


Figure A.3: a) Mean average error, b) root mean-squared error, and c) R^2 for LightGBM base estimators fit on different feature subsets and missing data thresholds for country-level prediction.

A.1.2 Forecasting

Random Forest

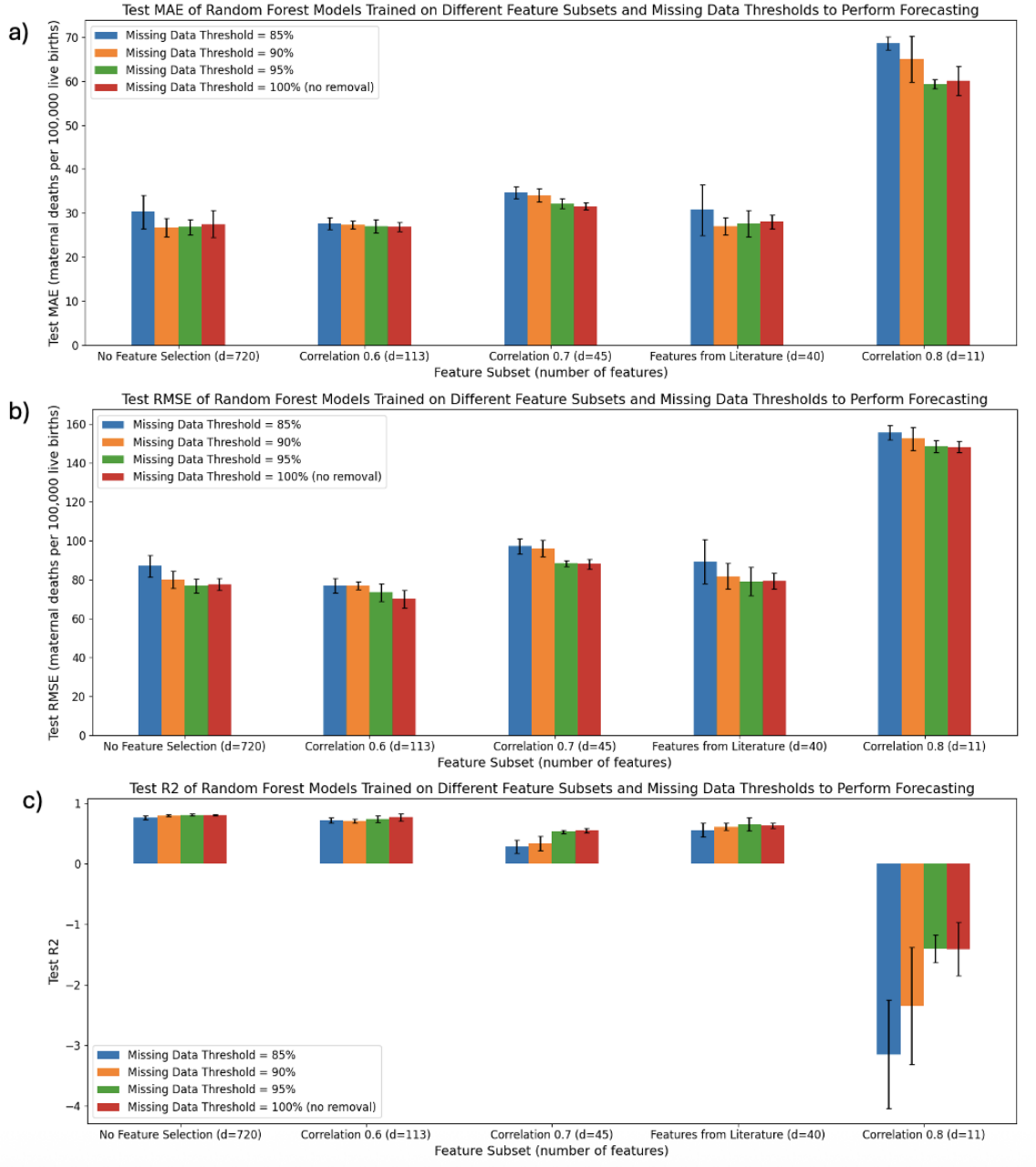


Figure A.4: a) Mean average error, b) root mean-squared error, and c) R^2 for Random Forest base estimators fit on different feature subsets and missing data thresholds to perform forecasting.

XGBoost

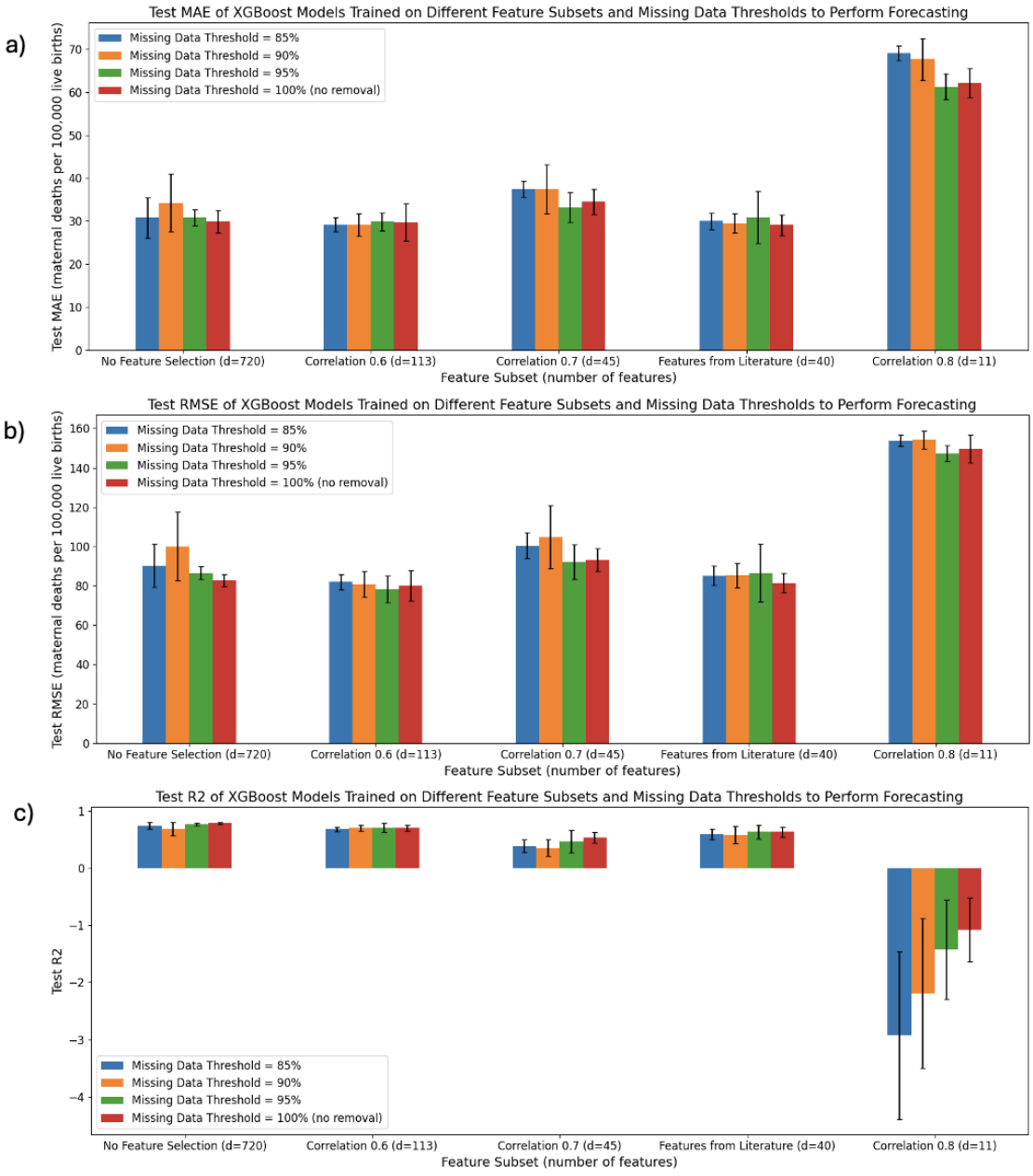


Figure A.5: a) Mean average error, b) root mean-squared error, and c) R^2 for XGBoost base estimators fit on different feature subsets and missing data thresholds to perform forecasting.

LightGBM

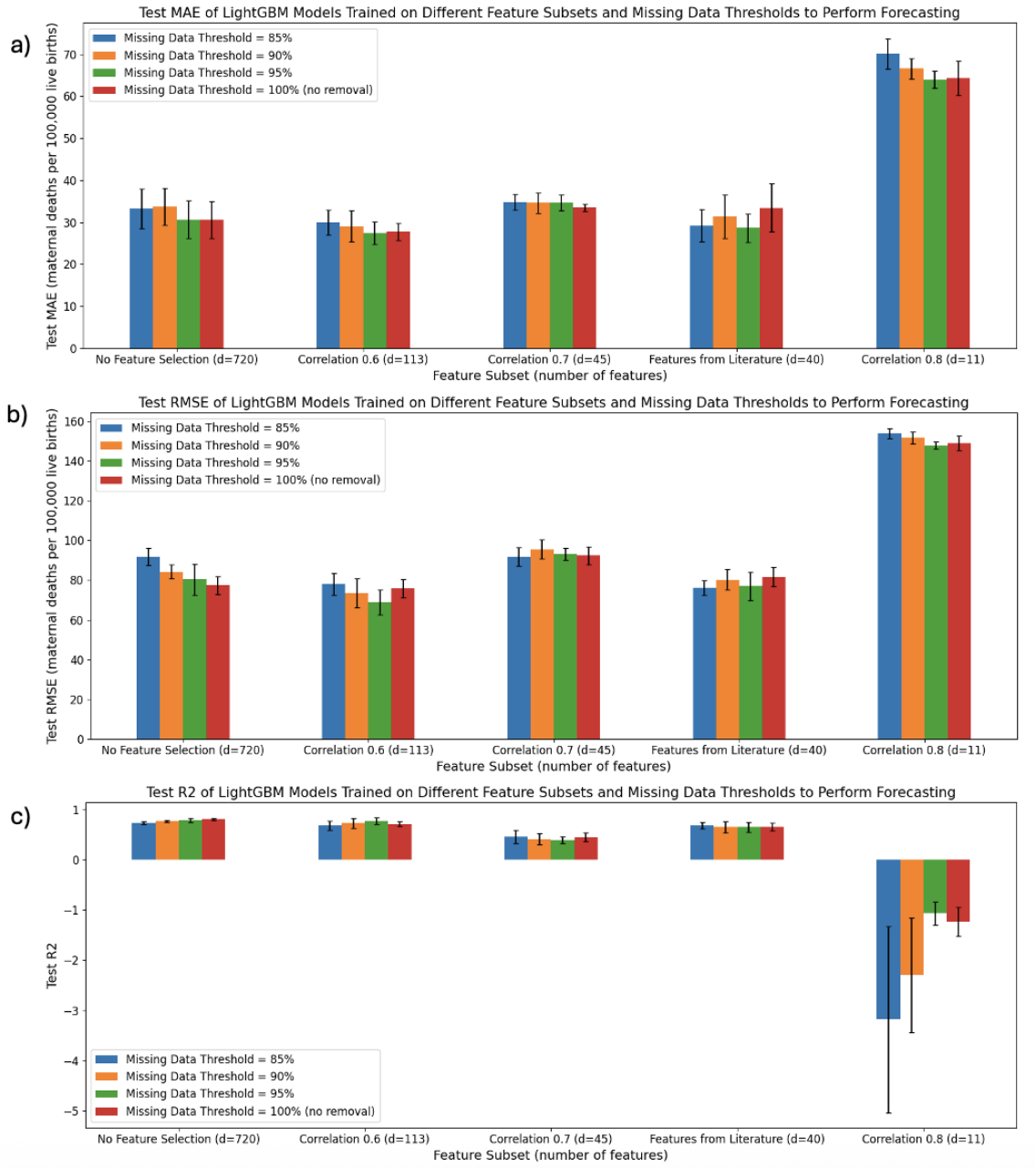
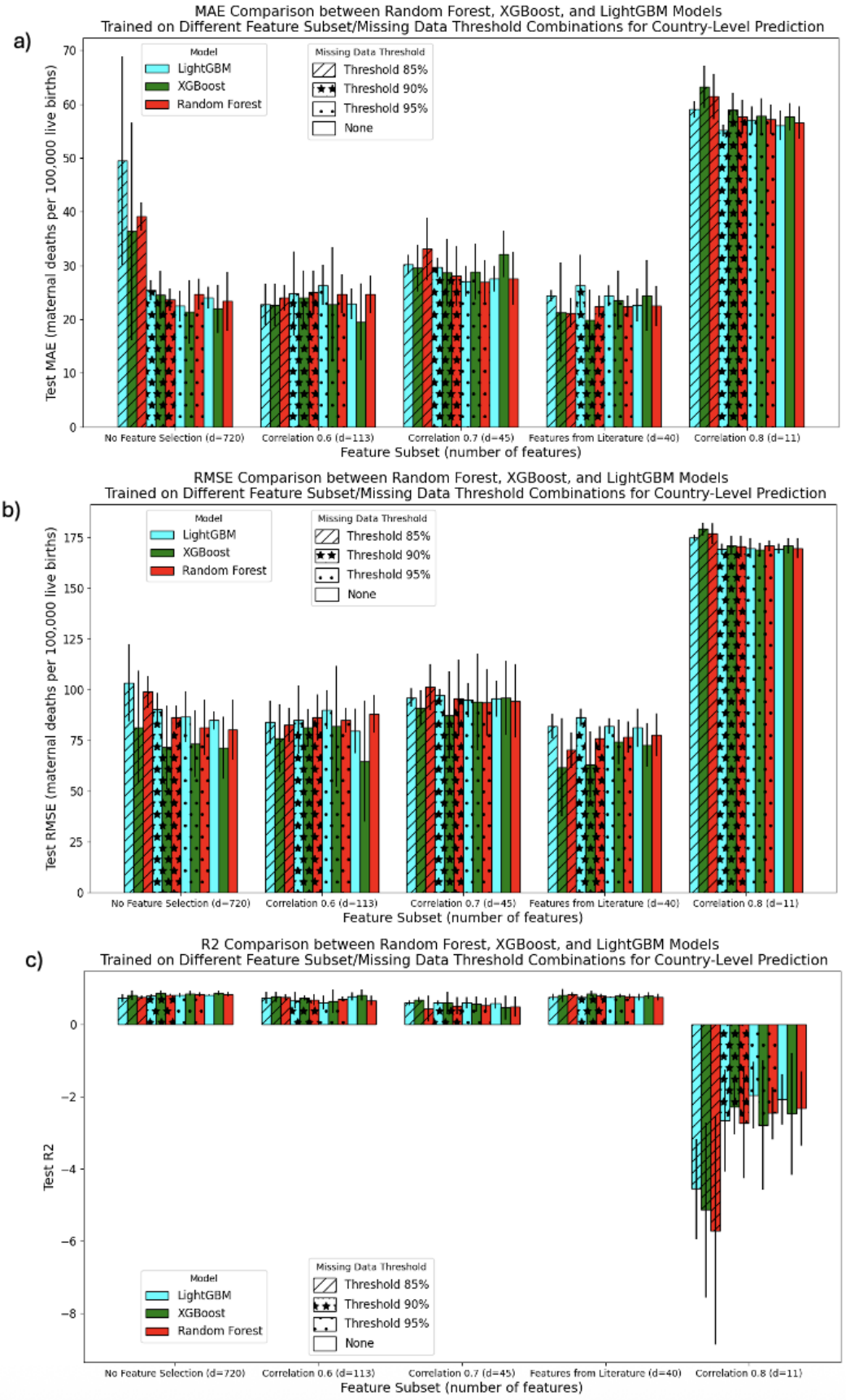


Figure A.6: a) Mean average error, b) root mean-squared error, and c) R^2 for LightGBM base estimators fit on different feature subsets and missing data thresholds to perform forecasting.

A.2 Additional Performance Metrics to Compare Predictive Performance of Base Estimator Model Types

A.2.1 Country-Level Prediction



A.2.2 Forecasting

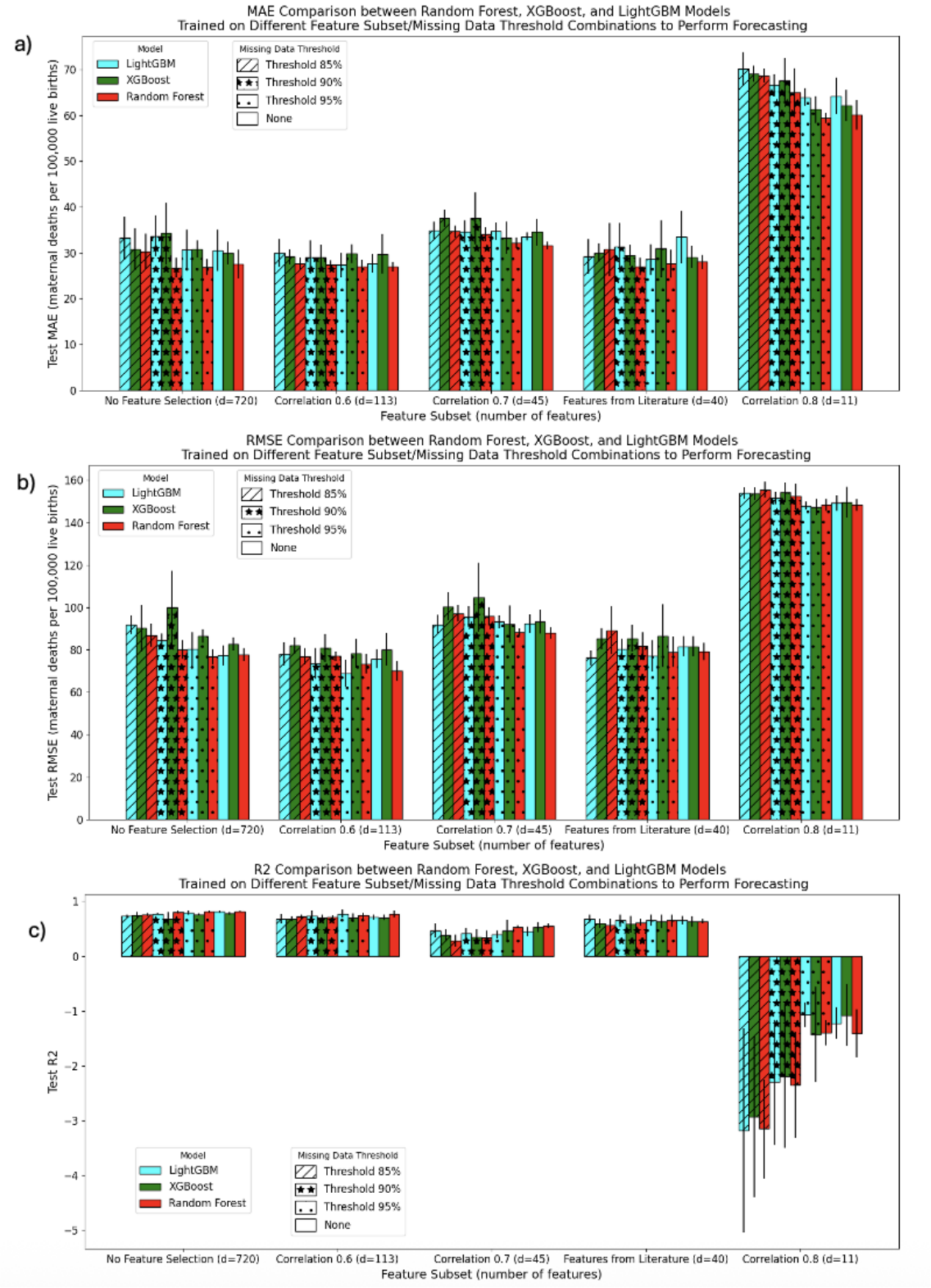


Figure A.8: a) Mean average error, b) root mean-squared error, and c) R^2 for Random Forest (red), XGBoost (green), and LightGBM (blue) models fit on different feature subsets and missing data thresholds to perform forecasting.

A.3 Additional Performance Metrics to Compare Stacking versus Voting Ensemble Models

A.3.1 Country-Level Prediction

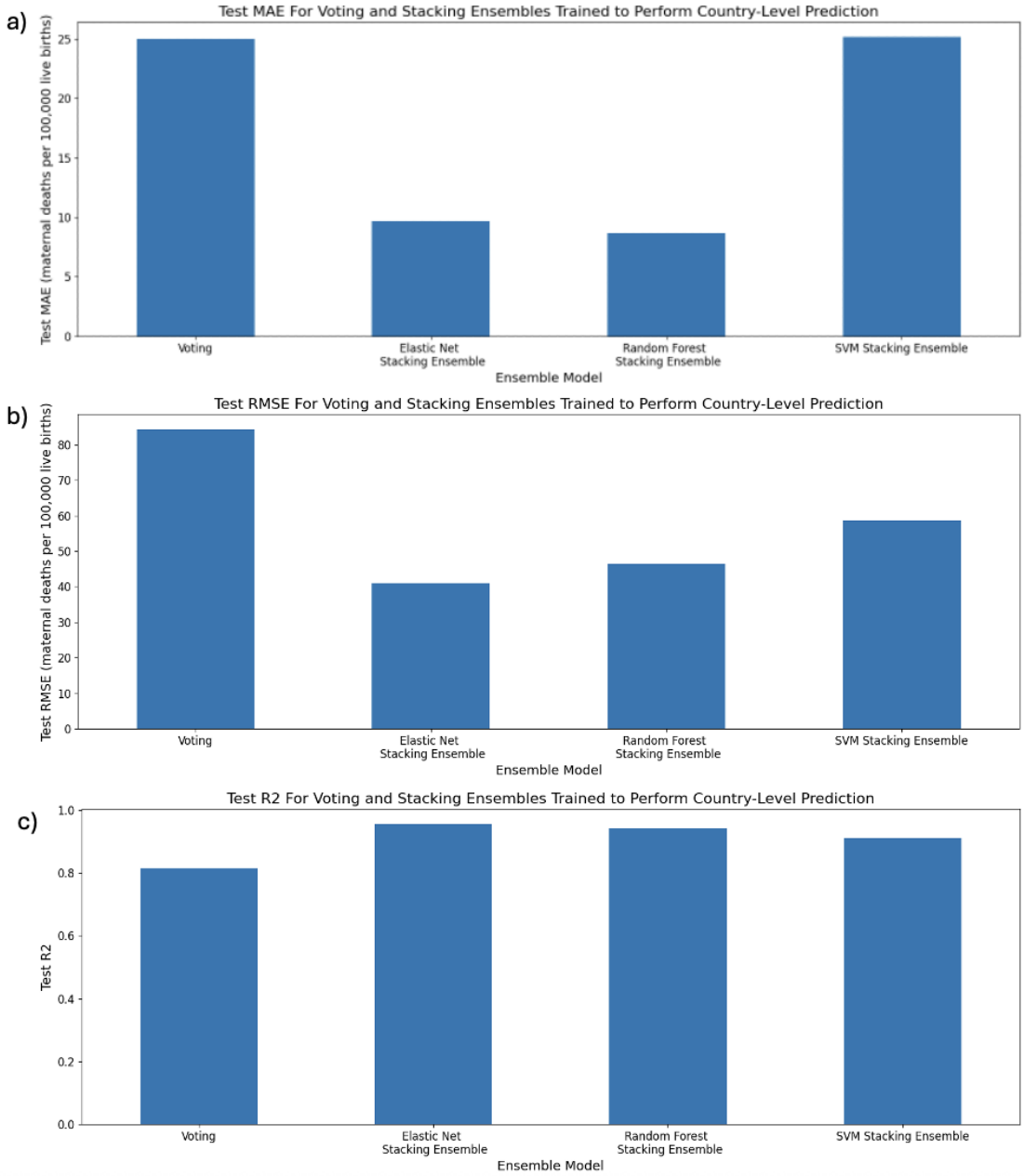


Figure A.9: a) Mean average error, b) root mean-squared error, and c) R^2 for voting and stacking ensembles trained on all base models to perform country-level prediction.

A.3.2 Forecasting

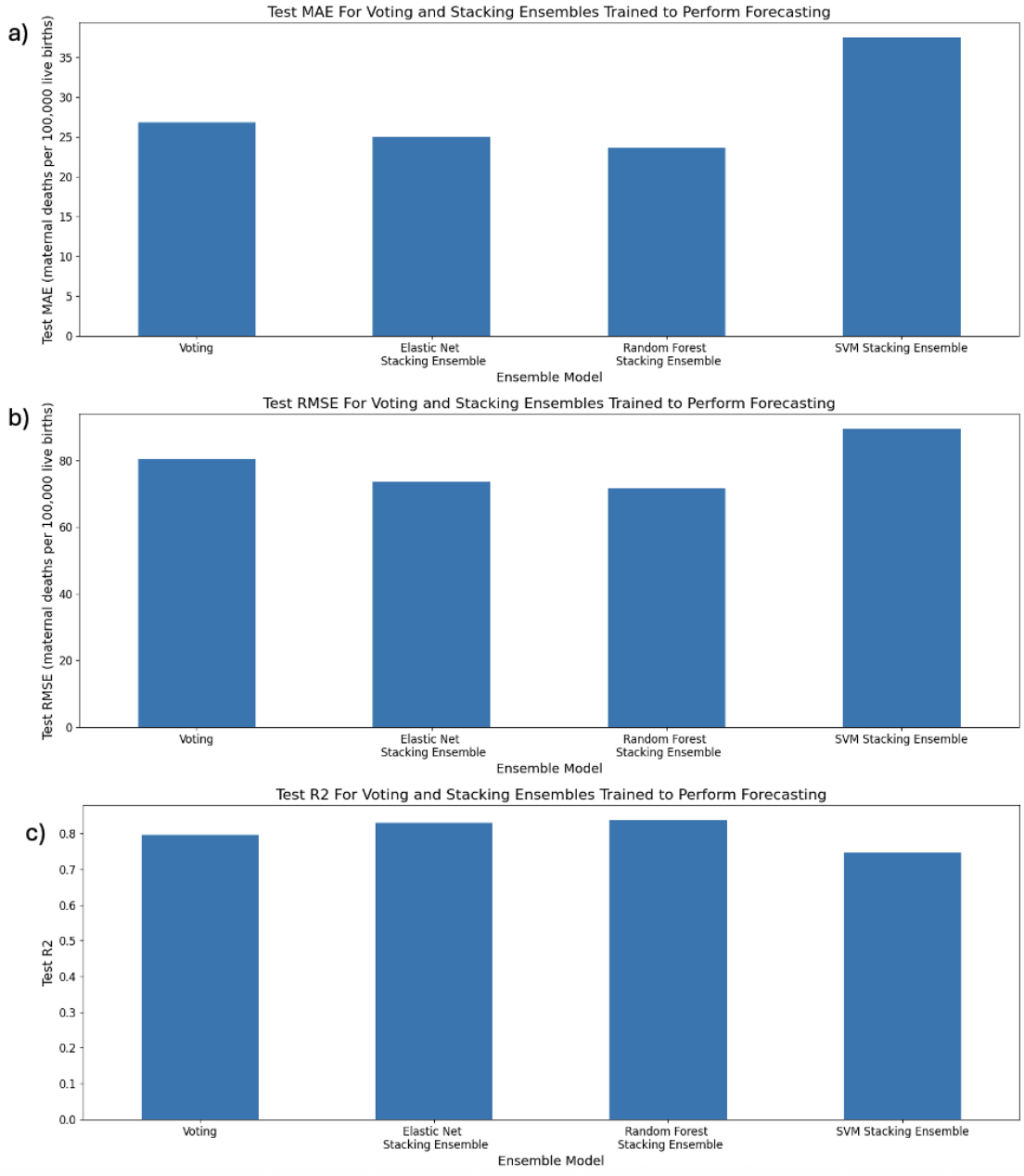


Figure A.10: a) Mean average error, b) root mean-squared error, and c) R^2 for voting and stacking ensembles trained on all base models to perform forecasting.

A.4 Additional Performance Metrics to Compare RFSE with Base Estimators

A.4.1 Country-Level Prediction

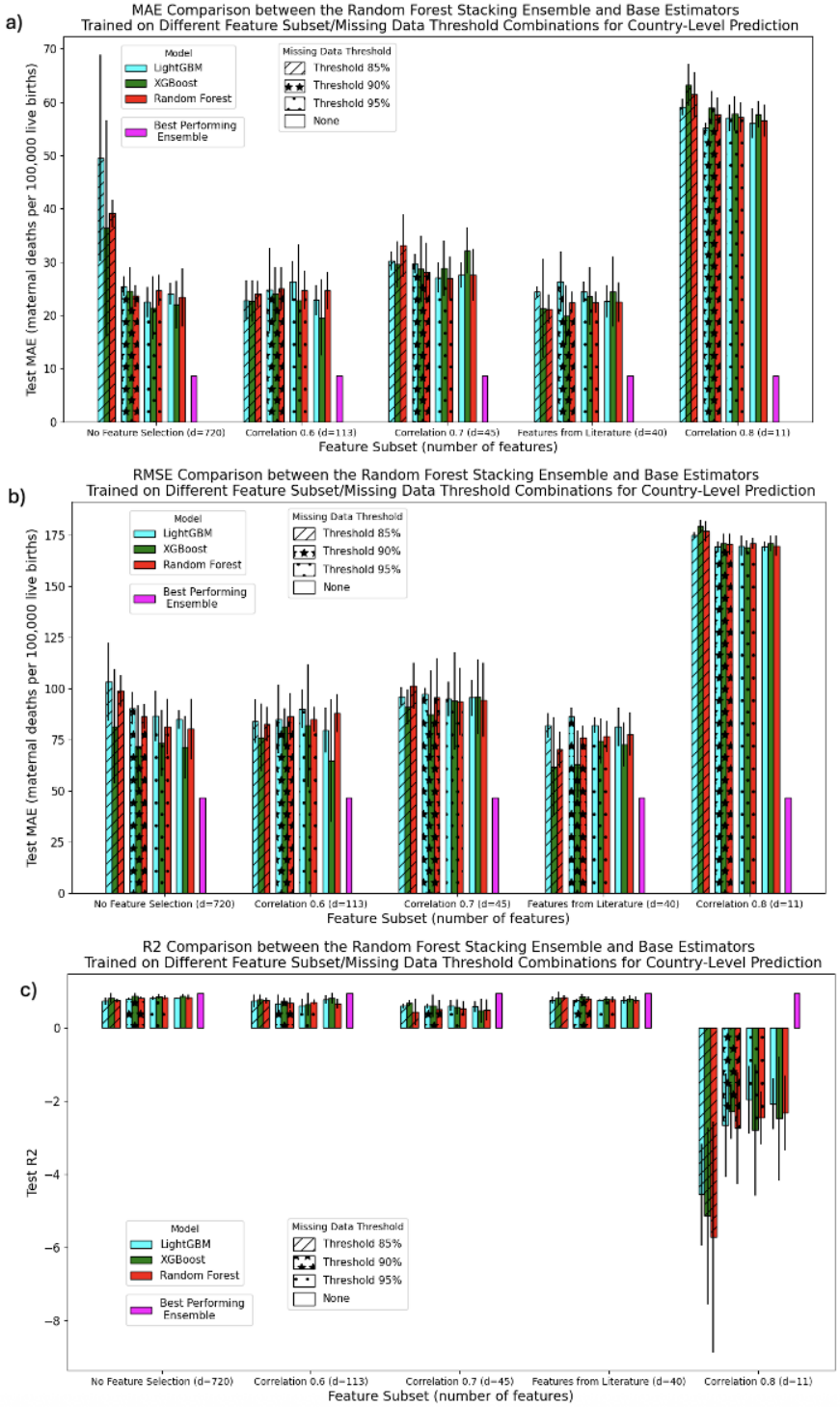


Figure A.11: a) Mean average error, b) root mean-squared error, and c) R^2 for the Random Forest Stacking Ensemble (purple) and the Random Forest (red), XGBoost (green), and LightGBM (blue) base estimators trained for country-level prediction.

A.4.2 Forecasting

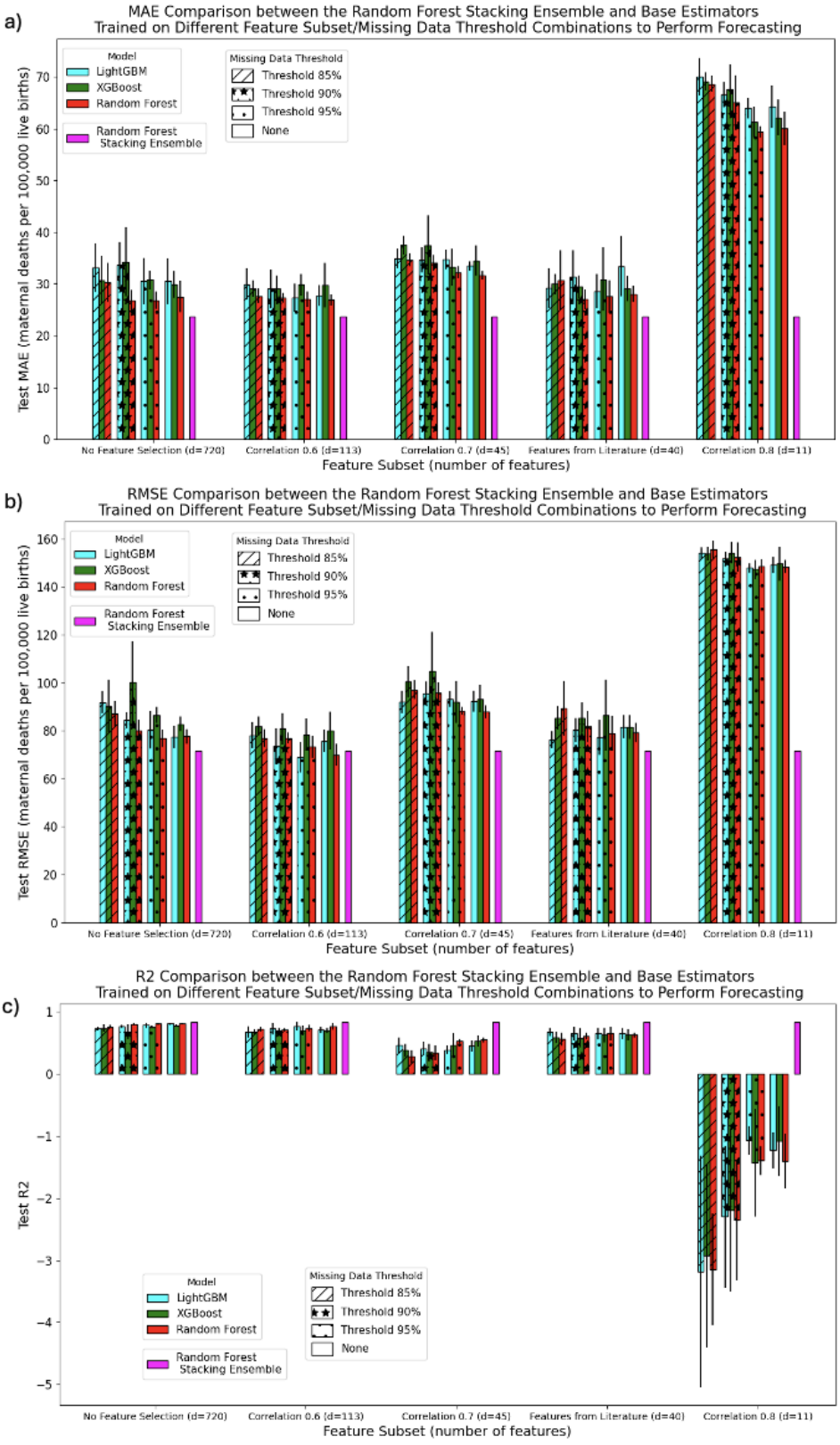


Figure A.12: a) Mean average error, b) root mean-squared error, and c) R^2 for for the Random Forest Stacking Ensemble (purple) and the Random Forest (red), XGBoost (green), and LightGBM (blue) base estimators trained for forecasting.

A.5 Additional Performance Metrics to Compare Random Forest Stacking Ensemble Performance When Trained with Different Subsets of Base Estimators

A.5.1 Country-Level Prediction

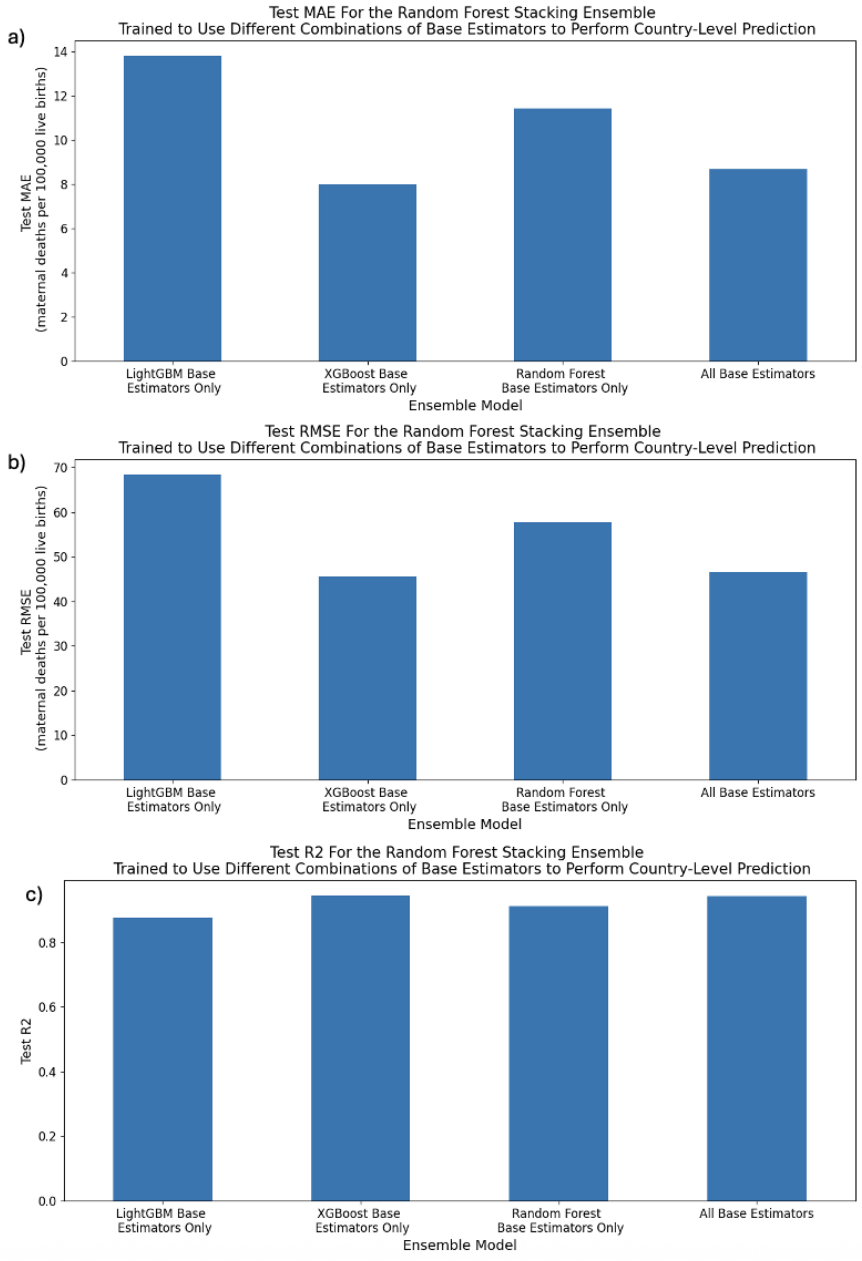


Figure A.13: a) Mean average error, b) root mean-squared error, and c) R^2 for the Random Forest Stacking Ensemble trained to perform country-level prediction using different combinations of base estimators.

A.5.2 Forecasting

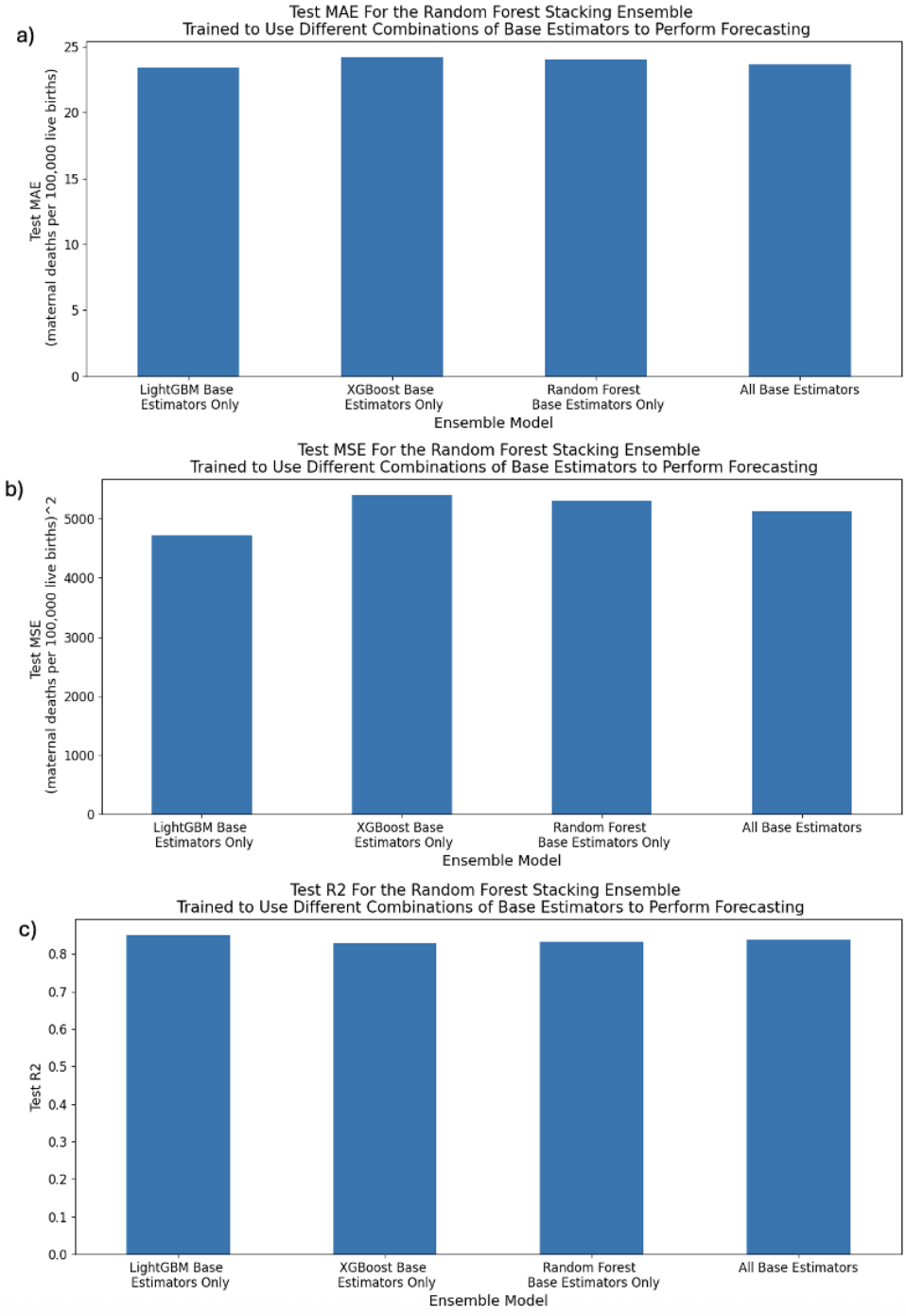


Figure A.14: a) Mean average error, b) root mean-squared error, and c) R^2 for the Random Forest Stacking Ensemble trained to perform forecasting using different combinations of base estimators.

A.6 Additional Performance Metrics for Sensitivity Analysis

A.6.1 Country-Level Prediction

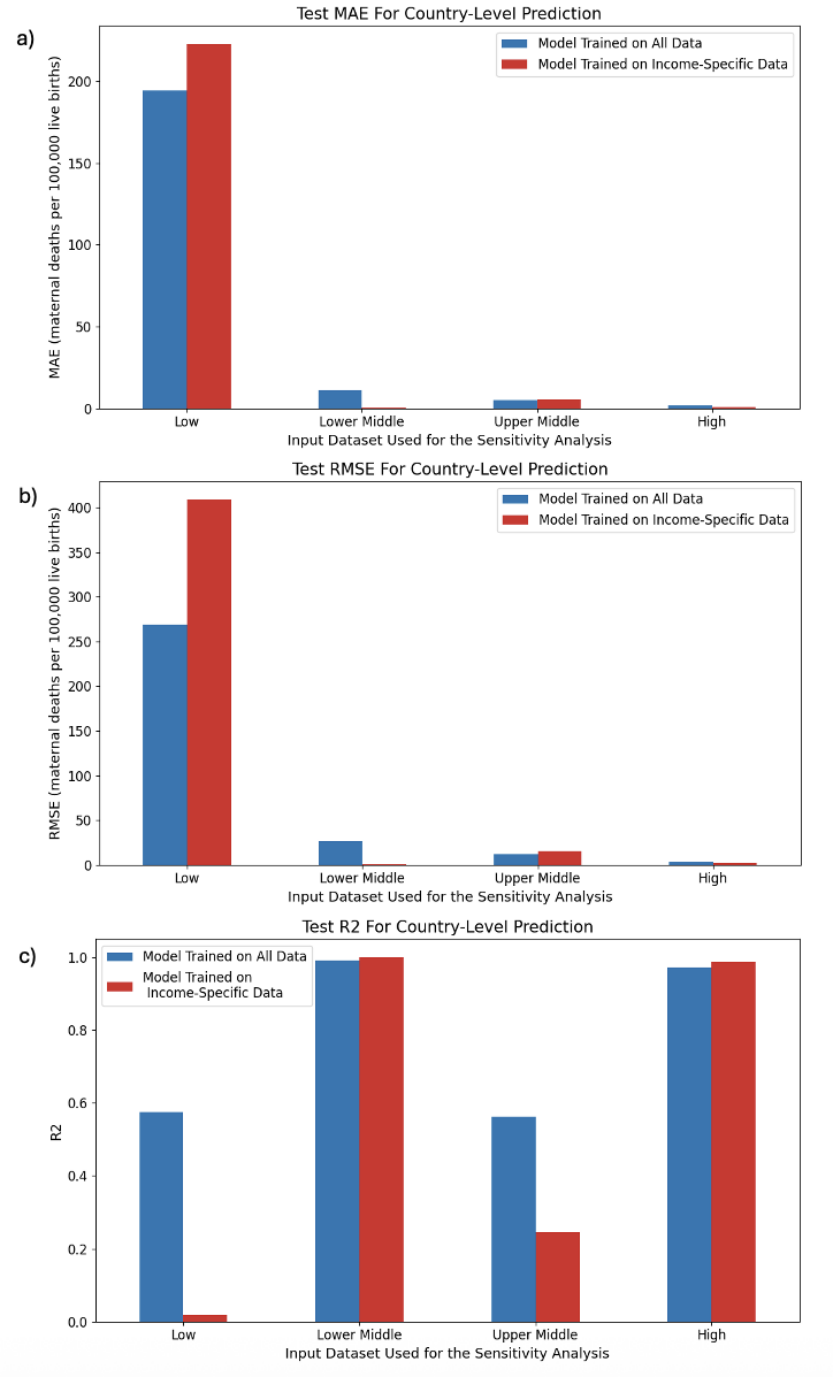


Figure A.15: a) Mean average error, b) root mean-squared error, and c) R^2 for the Random Forest Stacking Ensemble trained on data from all income levels (blue) and RFSEs trained on data from a specific income level (red) to perform country-level prediction. The models being compared were tested on data from the same income level.

A.6.2 Forecasting

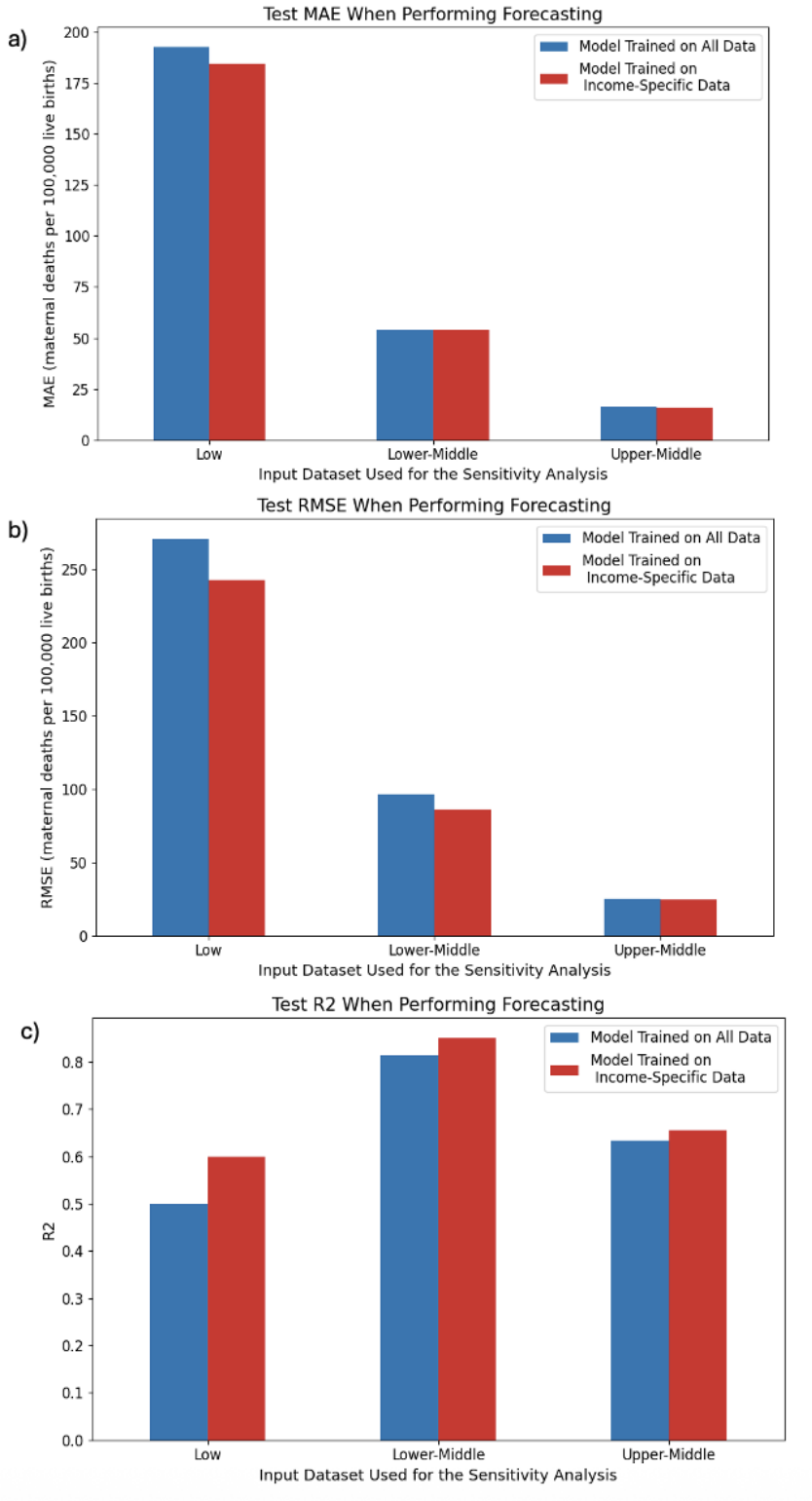


Figure A.16: a) Mean average error, b) root mean-squared error, and c) R^2 for the Random Forest Stacking Ensemble trained on data from all income levels (blue) and RFSEs trained on data from a specific income level (red) to perform forecasting. The models being compared were tested on data from the same income level.

Bibliography

AbouZahr, C., 2011. New estimates of maternal mortality and how to interpret them: choice or confusion? *Reproductive Health Matters*, 19, 37 (2011), 117–128. doi:10.1016/S0968-8080(11)37550-7. [Cited on pages 2 and 25.]

Ahmed, S. M. A.; Cresswell, J. A.; and Say, L., 2023. Incompleteness and misclassification of maternal death recording: a systematic review and meta-analysis. *BMC Pregnancy and Childbirth*, 23, 794 (2023). doi:10.1186/s12884-023-06077-4. [Cited on pages 1, 2, 6, 7, 130, 136, and 137.]

Akazawa, M.; Hashimoto, K.; Katsuhiko, N.; and Kaname, Y., 2021. Machine learning approach for the prediction of postpartum hemorrhage in vaginal birth. *Scientific Reports*, 11 (2021). doi:10.1038/s41598-021-02198-y. [Cited on page 34.]

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19 (Anchorage, AK, USA, 2019), 2623–2631. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3292500.3330701. https://doi.org/10.1145/3292500.3330701. [Cited on page 51.]

Akselrod, S.; Banerjee, A.; Collins, T. E.; Acharya, S.; Artykova, N.; Askew, I.; Berdzuli, N.; Diorditsa, S.; Eggers, R.; Farrington, J.; Jakab, Z.; Ferreira-Borges, C.; Mikkelsen, B.; Azzopardi-Muscat, N.; Olsavszky, V.; Park, K.; Sobel, H.; Tran, H.; Vujnovic, M.; Weber, M.; Were, W.; Yaqub, N.; Berlina, D.; Dunlop, C. L.; and Allen, L. N., 2023. Integrating maternal, newborn, child health and non-communicable disease care in the sustainable development goal era. *Frontiers in Publich Health*, 11 (2023). doi:10.3389/fpubh.2023.1183712. [Cited on pages 27, 28, and 48.]

Alkema, L.; Zhang, S.; Chou, D.; Gemmel, A.; Moller, A.-B.; Fat, D. M.; Say, L.; Mathers, C.; and Hogan, D., 2017. A bayesian approach to the global estimation of maternal mortality. *The Annals of Applied Statistics*, 11, 3 (2017), 1245–1274. doi:10.1214/16-AOAS1014. [Cited on pages 2, 24, 25, 26, 27, and 28.]

Baldi, P., 1995. Gradient descent learning algorithm overview: a general dynamical systems perspective. *IEEE Transactions on Neural Networks*, 6, 1 (1995), 182–195. doi:10.1109/72.363438. [Cited on page 10.]

Bentéjac, C.; Csörgő, A.; and Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54 (2021), 1937–1967. doi:10.1007/s10462-020-09896-5. [Cited on page 121.]

Breiman, L., 1996. Bagging predictors. *Machine Learning*, 24 (1996), 123–140. doi:10.1007/BF00058655. [Cited on pages 17 and 18.]

Breiman, L., 2001. Statistical modeling: The two cultures. *Statistical Science*, 16 (2001), 199–231. doi:10.1214/ss/1009213726. [Cited on page 32.]

Chan, J. Y.-L.; Leow, S. M. H.; Bea, K. T.; Cheng, W. K.; Phoong, S. W.; Hong, Z.-W.; and Chen, Y.-L., 2022. Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics*, 10, 8 (2022). doi:10.3390/math10081283. https://www.mdpi.com/2227-7390/10/8/1283. [Cited on pages 32 and 124.]

Chen, T. and Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (San Francisco, California, USA, 2016), 785–794. Association for Computing Machinery, New York, NY, USA. doi:10.1145/2939672.2939785. https://doi.org/10.1145/2939672.2939785. [Cited on pages 18, 19, 20, 32, 51, 52, and 54.]

Chou, D.; Daelmans, B.; Jolivet; Rima, R.; Kinney, M.; and Say, L., 2015. Ending preventable maternal and newborn mortality and stillbirths. *BMJ*, 351 (2015). doi:10.1136/bmj.h4255. https://www.bmj.com/content/351/bmj.h4255. [Cited on pages 134 and 135.]

Conway, F.; Portela, A.; Filippi, V.; Chou, D.; and Kovats, S., 2024. Climate change, air pollution and maternal and newborn health: An overview of reviews of health outcomes. *Journal of Global Health*, 14 (2024). doi:10.7189/jogh.14.04128. [Cited on pages 28 and 48.]

Costa, V. G. and Pedreira, C. E., 2023. Recent advances in decision trees: an updated survey. *Artificial Intelligence Review*, 56 (2023), 4765–4800. doi:10.1007/s10462-022-10275-5. [Cited on pages 3, 8, 14, 15, and 32.]

Cresswell, J. A.; Alexander, M.; Chong, M. Y. C.; Link, H. M.; Pejchinovska, M.; Gazeley, U.; Ahmed, S. M. A.; Chou, D.; Moller, A.-B.; Simpson, D.; Alkema, L.; Villanueva, G.; Sguassero, Y.; Özge Tunçalp; Xiao, Q. L. A. S.; and Say, L., 2025. Global and regional causes of maternal deaths 2009–20: a who systematic analysis. *The Lancet Global Health*, 13, 4 (april 2025),

e626–e634. doi:10.1016/S2214-109X(24)00560-6. [Cited on pages 1, 2, 6, 7, 27, 125, 126, 133, and 134.]

Dabool, H.; Alashwal, H.; and Moustafa, A., 2024. Enhancing diagnostic accuracy by bypassing traditional imputation and leveraging missing data in alzheimer's disease detection models. In *Proceedings of the 2024 8th International Conference on Information System and Data Mining*, ICISDM '24, 33–38. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3686397.3686403. https://doi.org/10.1145/3686397.3686403. [Cited on pages 118 and 138.]

dhsprogram.com, 2025. The dhs program countries. dhsprogram.com. https://dhsprogram.com/countries/. Accessed: 2025-10-02. [Cited on page 137.]

Emmanuel, T.; Maupong, T.; Mpoeleng, D.; Semong, T.; Mphago, B.; and Tabona, O., 2021. A survey on missing data in machine learning. *Journal of Big Data*, 8 (2021). doi:10.1186/s40537-021-00516-9. [Cited on pages 43 and 44.]

Fernández, A.; García, S.; Herrera, F.; and Chawla, N. V., 2018. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *J. Artif. Int. Res.*, 61, 1 (Jan. 2018), 863–905. [Cited on page 139.]

Foreman, K. J.; Lozano, R.; Lopez, A. D.; and Murray, C. J., 2012. Modeling causes of death: an integrated approach using codem. *Population Health Metrics*, 10, 1 (2012). doi:10.1186/1478-7954-10-1. [Cited on pages 26, 27, and 131.]

Fredriksson, A.; Fulcher, I. R.; Russell, A. L.; Li, T.; Tsai, Y.-T.; Seif, S. S.; Mpembeni, R. N.; and Hedt-Gauthier, B., 2022. Machine learning for maternal health: Predicting delivery location in a community health worker program in zanzibar. *Frontiers in Digital Health*, 4 (2022). doi:10.3389/fdgth.2022.855236. [Cited on page 35.]

Ganaie, M.; Hu, M.; Malik, A.; Tanveer, M.; and Suganthan, P., 2022. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115 (2022), 105151. doi:10.1016/j.engappai.2022.105151. https://www.sciencedirect.com/science/article/pii/S095219762200269X. [Cited on pages 17, 18, 21, and 121.]

Ganie, A. G. and Dadvandipour, S., 2023. From big data to smart data: a sample gradient descent approach for machine learning. *Journal of Big Data*, 10, 162 (2023). doi:10.1186/s40537-023-00839-9. [Cited on page 10.]

GBD 2021 Causes of Death Collaborators, 2024. Supplement to: Global burden of 288 causes of death and life expectancy decomposition in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the global burden of disease study 2021. *The Lancet*, (2024), Supplementary Appendix. doi:10.1016/S0140-6736(24)00367-2. https://doi.org/10.1016/S0140-6736(24)00367-2. Published online April 3. [Cited on pages 2, 25, 26, 28, 42, 60, 129, 131, 136, and 138.]

Greener, J. G.; Kandathil, S. M.; Moffat, L.; and Jones, D. T., 2022. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol*, 13 (2022), 40–55. doi:10.1038/s41580-021-00407-0. [Cited on pages 8, 9, 10, 11, and 13.]

Hu, Q.; Liao, H.; and Yu, H., 2025. Global, regional, and national burden of maternal hypertensive disorder: 1990–2021 analysis and future projections. *BMC Public Health*, 25 (2025). doi:10.1186/s12889-025-23528-z. [Cited on page 33.]

IBAN.com, 2025. List of country codes by alpha-2, alpha-3 (iso.3166). IBAN.com. https://www.iban.com/country-codes. Accessed: 2025-07-13. [Cited on page 42.]

Institute for Health Metrics and Evaluation, 2023. Institute for health metrics and evaluation (ihme) global burden of disease (gbd) study estimates. WHO Health Inequality Data Repository. https://www.who.int/data/sets/health-inequality-monitor-dataset#ihme-gbd. Dataset. [Cited on pages 40 and 41.]

Johnson, S. C.; Cunningham, M.; Dippenaar, I. N.; Sharara, F.; Wool, E. E.; Agesa, K. M.; Han, C.; Miller-Petrie, M. K.; Wilson, S.; Fuller, J. E.; Balassyano, S.; Bertolacci, G. J.; Weaver, N. D.; of Death Collaborators, G. C.; Lopez, A. D.; Murray, C. J. L.; and Naghavi, M., 2021. Public health utility of cause of death data: applying empirical algorithms to improve data quality. *BMC Medical Informatics and Decision Making*, 21, 175 (2021). doi:10.1186/s12911-021-01501-1. [Cited on pages 25 and 130.]

Jordan, M. I. and Mitchell, T. M., 2015. Machine learning: Trends, perspectives, and prospects. *Science*, 349, 6245 (2015), 255–260. doi:10.1126/science.aaa8415. https://www.science.org/doi/abs/10.1126/science.aaa8415. [Cited on pages 8 and 9.]

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf. [Cited on pages 18, 19, 20, 51, 53, 54, and 121.]

Khadidos, A. O.; Saleem, F.; Selvarajan, S.; Ullah, Z.; and Khadidos, A. O., 2024. Ensemble machine learning framework for predicting maternal health risk during pregnancy. *Scientific Reports*, 14 (2024). doi:10.1038/s41598-024-71934-x. [Cited on pages 34 and 56.]

Kia, S. M.; Rad, N. M.; van Opstal, D.; van Schie, B.; Marquand, A. F.; Pluim, J.; Cahn, W.; and Schnack, H. G., 2022. Promissing: Pruning missing values in neural networks. https://arxiv.org/abs/2206.01640. [Cited on pages 3, 17, and 51.]

Koblinsky, M.; Moyer, C. A.; Calvert, C.; Campbell, J.; Campbell, O. M. R.; Feigl, A. B.; Graham, W. J.; Hatt, L.; Hodgins, S.; Matthews, Z.; McDougall, L.; Moran, A. C.; Nandakumar, A. K.; and Langer, A., 2016. Quality maternity care for every woman, everywhere: a call to action. *The Lancet*, 388, 10057 (2016), 2307–2320. doi:https://doi.org/10.1016/S0140-6736(16)31333-2. https://www.sciencedirect.com/science/article/pii/S0140673616313332. [Cited on pages 1, 28, 48, 132, 133, and 135.]

Krzywinski, N. A. . M., 2018. The curse(s) of dimensionality. *Nature Methods*, 15 (2018), 399–400. doi:10.1038/s41592-018-0019-x. A Point of Significance. [Cited on page 33.]

LeCun, Y.; Bengio, Y.; and Hinton, G., 2015. Deep learning. *Nature*, 521 (2015), 436–444. doi:10.1038/nature14539. https://www.nature.com/articles/nature14539. [Cited on pages 3, 10, 15, 16, and 17.]

Li, J. and O'Donoghue, C., 2013. A survey of dynamic microsimulation models: uses, model structure and methodology. *INTERNATIONAL JOURNAL OF MICROSIMULATION*, 6, 2 (2013), 3–55. doi:10.34196/ijm.00082. [Cited on pages 30 and 31.]

Loh, W.-Y., 2014. Fifty years of classification and regression trees. *International Statistical Review*, 82, 3 (2014), 329–348. doi:https://doi.org/10.1111/insr.12016. https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12016. [Cited on pages 14 and 32.]

Lundberg, S. M. and Lee, S.-I., 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (Long Beach, California, USA, 2017), 4768–4777. Curran Associates Inc., Red Hook, NY, USA. doi:10.48550/arXiv.1705.07874. [Cited on page 133.]

Madley-Dowda, P.; Hughesa, R.; Tilling, K.; and Herona, J., 2019. The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110 (2019), 63–73. doi:10.1016/j.jclinepi.2019.02.016. [Cited on page 118.]

Mahajan, P.; Uddin, S.; Hajati, F.; and Moni, M. A., 2023. Ensemble learning for disease prediction: A review. *Healthcare (Basel)*, 11, 12 (2023), 1808. doi:10.3390/healthcare11121808. [Cited on pages 17, 18, 20, 21, 26, 33, 56, 121, 122, 123, and 124.]

Marić, I.; Tsur, A.; Aghaeepour, N.; Montanari, A.; Stevenson, D. K.; Shaw, G. M.; and Winn, V. D., 2020. Early prediction of preeclampsia via machine learning. *American Journal of Obstetrics Gynecology MFM*, 2, 2 (2020), 100100. doi:10.1016/j.ajogmf.2020.100100. https://www.sciencedirect.com/science/article/pii/S2589933320300306. [Cited on page 34.]

Maternal Health (MAH), N. C. . A. H. . A. M., Maternal, 2021. Ending preventable maternal mortality (epmm): a renewed focus for improving maternal and newborn health and well-being. Technical report, World Health Organization. [Cited on page 134.]

Maternal Mortality Estimation Inter-Agency Group, 2025. Maternal mortality ratio (modeled estimate, per 100,000 live births). World Bank Group Gender Data Portal. https://genderdata.worldbank.org/en/indicator/sh-sta-mmrt?estimate=National. Dataset. [Cited on pages 40 and 41.]

Mathers, C. D., 2020. History of global burden of disease assessment at the world health organization. *Archives of Public Health*, 78, 77 (2020). doi:10.1186/s13690-020-00458-3. [Cited on page 25.]

McClure, E. M.; Goldenberg, R. L.; and Bann, C. M., 2007. Maternal mortality, stillbirth and measures of obstetric care in developing and developed countries. *International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics*, 92 (2007), 139–146. doi:10.1016/j.ijgo.2006.10.010. [Cited on pages 27, 131, and 140.]

McQueston, K.; Silverman, R.; and Glassman, A., 2013. The efficacy of interventions to reduce adolescent childbearing in low- and middle-income countries: A systematic review. *Studies in Family Planning*, 44 (2013), 369–481. doi:10.1111/j.1728-4465.2013.00365.x. [Cited on page 135.]

Memon, S. M. Z.; Wamala, R.; and Kabano, I. H., 2022. Missing data analysis using statistical and machine learning methods in facility-based maternal health records. *SN Computer Science*, 3 (2022). doi:10.1007/s42979-022-01249-z. [Cited on pages 49 and 136.]

Mgawadere, F.; Kana, T.; and van den Broek, N., 2017. Measuring maternal mortality: a systematic review of methods used to obtain estimates of the maternal mortality ratio (mmr) in low- and middle-income countries. *British Medical Bulletin*, 121, 1 (2017), 121–134. doi:10.1093/bmb/ldw056. [Cited on pages 6, 7, 118, and 130.]

Miranda, J.; Scholz, I.; Agard, J.; Al-Ghanim, K.; Bobylev, S. N.; Dube, O. P.; Hathie, I.; Kanie, N.; Madise, N. J.; Malekpour, S.; Montoya, J. C.; Pan, J.; Åsa Persson; Sagar, A.; Shackell, N.; and who are the Independent Group of Scientists appointed by the Secretary-General, 2023. Global sustainable development report 2023: Times of crisis, times of change: Science for accelerating transformations to sustainable development. Technical report, United Nations, New York. United Nations Digital Library recommends this citation format. [Cited on page 5.]

Molla, M.; Hossain, S.; Ali, A.; Islam, R.; Sultana, P.; and Roy, D. C., 2025. Exploring the achievements and forecasting of sdg 3 using machine learning algorithms:

Bangladesh perspective. *PLOS One*, 20, 3 (2025). doi:10.1371/journal.pone.0314466. [Cited on page 35.]

Mukherjee, K.; Gunsoy, N. B.; Kristy, R. M.; Cappelleri, J. C.; Roydhouse, J.; Stephenson, J. J.; Vanness, D. J.; Ramachandran, S.; Onwudiwe, N. C.; Pentakota, S. R.; Karcher, H.; and Tanna, G. L. D., 2023. Handling missing data in health economics and outcomes research (heor): A systematic review and practical recommendations. *PharmacoEconomics*, 41 (2023), 1589–1601. doi:10.1007/s40273-023-01297-0. [Cited on page 43.]

Murray, C. J. L., 2022. The global burden of disease study at 30 years. *Nature Medicine*, 28 (2022), 2019–2026. doi:10.1038/s41591-022-01990-1. [Cited on pages 25, 26, and 135.]

Naghavi, M. e. a., 2024. Global burden of 288 causes of death and life expectancy decomposition in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the global burden of disease study 2021. *The Lancet*, 403 (2024), 2100 – 2132. doi:10.1016/S0140-6736(24)00367-2. [Cited on pages 25, 32, and 137.]

Neha Margret, I.; Rajakumar, K.; Arulalan, K. V.; Manikandan, S.; and Valentina, 2024. Statistical insights into machine learning-based box models for pregnancy care and maternal mortality reduction: A literature survey. *IEEE Access*, 12 (2024), 68184–68207. doi:10.1109/ACCESS.2024.3399827. [Cited on page 33.]

Onambele, L.; Guillen-Aguinaga, S.; Guillen-Aguinaga, L.; Ortega-Leon, W.; Montejo, R.; Alas-Brun, R.; Aguinaga-Ontoso, E.; Aguinaga-Ontoso, I.; and Guillen-Grima, F., 2023. Trends, projections, and regional disparities of maternal mortality in africa (1990–2030): An arima forecasting approach. *Epidemiologia*, 4 (2023), 322–351. doi:10.3390/epidemiologia4030032. [Cited on page 40.]

Patel, S. S., 2023. Explainable machine learning models to analyse maternal health. *Data Knowledge Engineering*, 146 (2023), 102198. doi:10.1016/j.datak.2023.102198. https://www.sciencedirect.com/science/article/pii/S0169023X23000587. [Cited on page 48.]

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12 (2011), 2825–2830. [Cited on pages 19, 44, 46, 51, 52, 54, 56, 57, and 58.]

Peterson, E.; Chou, D.; Moller, A.-B.; Gemmill, A.; Say, L.; and Alkema, L., 2022. Estimating misclassification errors in the reporting of maternal mortality in national civil registration vital statistics systems: A bayesian hierarchical bivariate

random walk model to estimate sensitivity and specificity for multiple countries and years with missing data. *Statistics in Medicine*, 41, 14 (2022), 2483–2496. doi:https://doi.org/10.1002/sim.9335. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9335. [Cited on pages 1 and 7.]

Peterson, E. N.; Guranich, G.; Cresswell, J. A.; and Alkema, L., 2024. A bayesian approach to estimate maternal mortality globally using national civil registration vital statistics data accounting for reporting errors, statistics and public policy. *Statistics and Public Policy*, 11, 1 (2024). doi:10.1080/2330443X.2023.2286313. [Cited on pages 1, 6, 7, 23, 24, 26, 130, and 138.]

Ramson, J. A.; Williams, M. J.; Afolabi, B. B.; Colagiuri, S.; Finlayson, K. W.; Hemmingsen, B.; Venkatesh, K. K.; and Chou, D., 2024. Pregnancy, childbirth and the postpartum period: opportunities to improve lifetime outcomes for women with non-communicable diseases. *Medical Journal of Australia*, 221, 7 (2024), 350–353. doi:https://doi.org/10.5694/mja2.52452. https://onlinelibrary.wiley.com/doi/abs/10.5694/mja2.52452. [Cited on pages 27 and 48.]

Rizkallah, L., 2025. Enhancing the performance of gradient boosting trees on regression problems. *Journal of Big Data*, 12, 35 (2025). doi:10.1186/s40537-025-01071-3. [Cited on pages 18 and 19.]

Rommel, A.; Elena von der Lippe, D. P.; Wengler, A.; Aline Anton, C. S.; Schüssel, K.; Brückner, G.; Schröder, H.; Porst, M.; Leddin, J.; Tobollik, M.; Baumert, J.; Scheidt-Nave, C.; and Ziese, T., 2018. Burden 2020—burden of disease in germany at the national and regional level. *Bundesgesundheitsblatt*, 61 (2018), 1159–1166. doi:10.1007/s00103-018-2793-0. [Cited on pages 28 and 131.]

Sadeghi, P.; Karimi, H.; Lavafian, A.; Rashedi, R.; Samieefar, N.; Shafiekhani, S.; and Rezaei, N., 2024. Machine learning and artificial intelligence within pediatric autoimmune diseases: applications, challenges, future perspective. *Expert Review of Clinical Immunology*, 20, 10 (2024), 1219–1236. doi:10.1080/1744666X.2024.2359019. https://doi.org/10.1080/1744666X.2024.2359019. PMID: 38771915. [Cited on page 138.]

Sadr, H.; Nazari, M.; Khodaverdian, Z.; Farzan, R.; Yousefzadeh-Chabok, S.; Ashoobi, M. T.; Hemmati, H.; Hendi, A.; Ashraf, A.; Pedram, M. M.; Hasannejad-Bibalan, M.; and Yamaghani, M. R., 2025. Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches. *European Journal of Medical Research*, 30 (2025). doi:10.1186/s40001-025-02680-7. [Cited on page 33.]

Scornet, E.; Biau, G.; and Vert, J.-P., 2015. Consistency of random forests. *The Annals of Statistics*, 43, 4 (2015), 1716 – 1741. doi:10.1214/15-AOS1321. https://doi.org/10.1214/15-AOS1321. [Cited on page 32.]

Serghiou, S. and Rough, K., 2023. Deep learning for epidemiologists: An introduction to neural networks. *American Journal of Epidemiology*, 192, 11 (05 2023), 1904–1916. doi:10.1093/aje/kwad107. https://doi.org/10.1093/aje/kwad107. [Cited on pages 16 and 17.]

Sheikh, J.; Allotey, J.; Kew, T.; Khalil, H.; Galadanci, H.; Hofmeyr, G. J.; Abalos, E.; Vogel, J. P.; Lavin, T.; Souza, J. P.; Kaur, I.; Ram, U.; Betran, A. P.; Bohren, M. A.; Oladapo, O. T.; and Thangaratinam, S., 2024. Vulnerabilities and reparative strategies during pregnancy, childbirth, and the postpartum period: moving from rhetoric to action. *eClinicalMedicine*, 67 (2024), 102264. doi:10.1016/j.eclinm.2023.102264. https://www.sciencedirect.com/science/article/pii/S2589537023004418. [Cited on page 134.]

Smola, A. J. and Schölkopf, B., 2004. A tutorial on support vector regression. *Statistics and Computing*, 14 (2004), 199–222. doi:10.1023/B:STCO.0000035301.49549.88. [Cited on pages 14, 56, and 125.]

Souza, J.; Tunçalp, ; Vogel, J.; Bohren, M.; Widmer, M.; Oladapo, O.; Say, L.; Gülmezoglu, A.; and Temmerman, M., 2014. Obstetric transition: the pathway towards ending preventable maternal deaths. *BJOG: An International Journal of Obstetrics & Gynaecology*, 121, s1 (2014), 1–4. doi:https://doi.org/10.1111/1471-0528.12735. https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1111/1471-0528.12735. [Cited on pages 6, 128, 129, and 133.]

Souza, J. P.; Day, L. T.; Rezende-Gomes, A. C.; Zhang, J.; Mori, R.; Baguiya, A.; Jayaratne, K.; Osoti, A.; Vogel, J. P.; Campbell, O.; Mugerwa, K. Y.; Lumbiganon, P.; Özge Tunçalp; Cresswell, J.; Say, L.; Moran, A. C.; and Oladapo, O. T., 2023. A global analysis of the determinants of maternal health and transitions in maternal mortality. *The Lancet Global Health*, 12 (2023), e306 – e316. doi:10.1016/S2214-109X(23)00468-0. [Cited on pages 1, 126, 127, 133, 134, and 135.]

Spielauer, M., 2011. What is social science microsimulation? *Social Science Computer Review*, 29, 1 (2011), 9–20. doi:10.1177/0894439310370085. [Cited on page 30.]

Sylvain, M. H.; Nyabyenda, E. C.; Uwase, M.; Komezusenge, I.; Ndikumana, F.; and Ngaruye, I., 2025. Prediction of adverse pregnancy outcomes using machine learning techniques: evidence from analysis of electronic medical records data in rwanda. *BMC Medical Informatics and Decision Making*, 25, 76 (2025). doi:10.1186/s12911-025-02921-z. [Cited on page 34.]

Taye, E. A.; Woubet, E. Y.; Hailie, G. Y.; Arage, F. G.; Zerihun, T. E.; Zegeye, A. T.; Zeleke, T. C.; and Kassaw, A. T., 2025. Application of the random forest algorithm to predict skilled birth attendance and identify determinants among reproductive-age women in 27 sub-saharan african countries; machine learning analysis. *BMC Public Health*, 25 (2025). doi:10.1186/s12889-025-22007-9. [Cited on pages 34 and 133.]

Terven, J.; Cordova-Esparza, D.-M.; Romero-González, J.-A.; Ramírez-Pedraza, A.; and Chávez-Urbiola, E. A., 2025. A comprehensive survey of loss functions and metrics in deep learning. *Artifial Intelligence Review*, 58, 195 (2025). doi:10.1007/s10462-025-11198-7. [Cited on pages 9, 10, 11, and 12.]

The World Bank Data Catalog, 2024. Health determinants (the world bank data catalog). WHO Health Inequality Data Repository. https://www.who.int/data/sets/health-inequality-monitor-dataset#wb. Dataset. [Cited on pages 40 and 41.]

Tunçalp, ; Souza, J.; Hindin, M.; Santos, C.; Oliveira, T.; Vogel, J.; Togoobaatar, G.; Ha, D.; Say, L.; Gülmezoglu, A.; and on behalf of the WHO Multicountry Survey on Maternal AND Newborn Health Research Network , 2014. Education and severe maternal outcomes in developing countries: a multicountry cross-sectional survey. *BJOG: An International Journal of Obstetrics & Gynaecology*, 121, s1 (2014), 57–65. doi:10.1111/1471-0528.12634. https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1111/1471-0528.12634. [Cited on pages 28, 132, 133, and 134.]

Twala, B., 2009. An empirical comparison of techniques for handlingincomplete data using decision trees. *Applied Artificial Intelligence*, 23 (2009), 373–405. doi:10.1080/08839510902872223. [Cited on pages 118, 119, and 139.]

Utomo, B.; Sucahya, P. K.; Romadlona, N. A.; Robertson, A. S.; and Magnani, R. I. A. A. R. J., 2021. The impact of family planning on maternal mortality in indonesia: what future contribution can be expected? *Population Health Metrics*, 19 (2021). doi:10.1186/s12963-020-00245-w. [Cited on pages 132 and 135.]

van Imhoff, E. and Post, W., 1998. Microsimulation methods for population projection. *Population: An English Selection*, 10, 1 (1998), 97–138. https://www.jstor.org/stable/2998681. [Cited on pages 30 and 31.]

Veenman, M.; Stefan, A. M.; and Haaf, J. M., 2024. Bayesian hierarchical modeling: an introduction and reassessment. *Behavior Research Methods*, 56 (2024), 4600–4631. doi:10.3758/s13428-023-02204-3. [Cited on pages 24 and 27.]

Ward, Z. J.; Atun, R.; King, G.; Dmello, B. S.; and Goldie, S. J., 2023a. Simulation-based estimates and projections of global, regional and country-level maternal mortality by cause, 1990–2050. *Nature Medicine*, 29 (2023), 1253–1261. doi:10.1038/s41591-023-02310-x. [Cited on pages 2, 28, 29, 30, 31, 42, 60, 129, 130, 131, 132, 136, 137, and 138.]

Ward, Z. J.; Atun, R.; King, G.; Dmello, B. S.; and Goldie, S. J., 2025. Assessing differences in country-level estimates of maternal mortality: a comparison of gmath, un, and gbd model results for 2020. *eClinicalMedicine*, 88 (2025), 103505. doi:10.1016/j.eclinm.2025.103505. https://www.sciencedirect.com/science/article/pii/S2589537025004389. [Cited on pages 2, 31, and 137.]

Warda, Z. J.; Atunb, R.; Kinge, G.; Dmellof, B. S.; and Goldiea, S. J., 2024. Global maternal mortality projections by urban/rural location and education level: a simulation-based analysis. *eClinicalMedicine*, 72 (2024). doi:10.1016/j.eclinm.2024.102653. [Cited on pages 30, 48, 134, and 140.]

WHO Collaborating Center for Health Equity Monitoring, 2024. Women's empowerment index (swper) (dhs re-analyzed by iceh). WHO Health Inequality Data Repository. https://www.who.int/data/sets/health-inequality-monitor-dataset#swper. Dataset. [Cited on pages 40 and 41.]

Wilmoth, J. R.; Mizoguchi, N.; Oestergaard, M. Z.; Say, L.; Mathers, C. D.; Zureick-Brown, S.; Inoue, M.; and Chou, D., 2012. A new method for deriving global estimates of maternal mortality. *Statistics, politics, and policy*, 3 (2012), 2151–7509. doi:10.1515/2151-7509.1038. [Cited on pages 25 and 136.]

World Bank Group Gender Data Portal, 2025. Health. World Bank Group Gender Data Portal. https://genderdata.worldbank.org/en/topics/health#idAllIndicators. Dataset. [Cited on pages 40 and 41.]

World Health Organization, 2022. *International Classification of Diseases Eleventh Revision (ICD-11)*. https://icdcdn.who.int/icd11referenceguide/en/html/index.html. Online edition. [Cited on page 5.]

World Health Organization, 2025. Trends in maternal mortality estimates 2000 to 2023: estimates by who, unicef, unfpa, world bank group and undesa/population division. Technical report, World Health Organization. [Cited on pages 1, 2, 5, 6, 7, 23, 24, 25, 26, 27, 42, 43, 48, 60, 129, 131, 135, 136, 137, and 138.]

World Health Organization's (WHO) Global Health Observatory (GHO), 2024. Who global health observatory (gho). WHO Health Inequality Data Repository. https://www.who.int/data/sets/health-inequality-monitor-dataset#gho. Dataset. [Cited on pages 40 and 41.]

Zaba, B.; Calvert, C.; Marston, M.; Isingo, R.; Nakiyingi-Miiro, J.; Lutalo, T.; Crampin, A.; Robertson, L.; Herbst, K.; Newell, M.-L.; Todd, J.; Byass, P.; Boerma, T.; and Ronsmans, C., 2013. Effect of hiv infection on pregnancy-related mortality in sub-saharan africa: secondary analyses of pooled community-based data from the network for analysing longitudinal population-based hiv/aids data on africa (alpha). *The Lancet*, 381, 9879 (2013), 1763–1771. doi:https://doi.org/10.1016/S0140-6736(13)60803-X. https://www.sciencedirect.com/science/article/pii/S014067361360803X. [Cited on page 25.]

Zaman, B.; Sharma, A.; Garg, J.; Ram, C.; Kushwah, R.; and Muradia, R., 2024. Analysis of factors influencing maternal mortality and newborn health—a machine learning approach. *Journal of Medical Artificial Intelligence*, 7, 1 (2024). doi:10.21037/jmai-23-107. https://jmai.amegroups.org/article/view/8590. [Cited on page 48.]

Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67, 2 (march 2005), 301–320. doi:10.1111/j.1467-9868.2005.00503.x. https://doi.org/10.1111/j.1467-9868.2005.00503.x. [Cited on pages 13, 56, and 124.]

Zuhair, V.; Babar, A.; Ali, R.; Oduoye, M. O.; Noor, Z.; Chris, K.; Okon, I. I.; and Rehman, L. U., 2024. Exploring the impact of artificial intelligence on global health and enhancing healthcare in developing nations. *Journal of primary care community health*, 15 (2024). doi:10.1177/21501319241245847. [Cited on page 8.]