## 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). I then discuss the socio-economic and health-related variables with the highest predictive power for MMR (6.3) and how they can be used to motivate policy recommendations (6.4, 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 (Figures 16-21), 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.

#### 6.11 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 [14]. 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 [14]. However, my analysis of trained decision-tree based models (section 5.5) 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 16-21).

Therefore, the Random Forest, XGBoost, and LightGBM models could effectively handle high proportions of missing data using the default direction technique described in the background information (Chapter XX). 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 [48].

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 [10, 14]. 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% [14]. 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 [13]. 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. 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 [14]. 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 [14]. 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.

#### 6.12 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.

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. 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 (Figure XXX). 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 (Figure XXX). 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. 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 (Figure XXX). 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. 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. 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. 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.

#### 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. 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 Bentéjac et al.’s (2020) review of gradient boosting methods, 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 [16]. 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 (Figure XXX). This may be due to boosting ensembles’ known tendency to overfit, as the boosting mechanism explicit corrects the mistakes of base estimators in the ensemble [17]. 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. 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. While this technique reduces overfitting, it can result in important features being underutilised when deciding splits, reducing its performance [17]. 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.

XGBoost models trained for country-level prediction had the best performance, i.e., the lowest MSE scores (Figure XXX). 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 [17]. 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. 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. 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. 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 [17]. Again, LightGBM’s GOSS algorithm may have reduced overfitting, resulting in lower standard deviation in its error across the different cross-validation folds.

#### 6.14 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.

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. 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. Additionally, overfitting sometimes caused the RFSE to have lower performance than the other ensembles.

As expected, the RFSE’s 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. 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.

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

#### 6.21 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 24, 25) [17]. 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.

The RFSE-CLP and RFSE-F models placed the greatest importance on XGBoost base estimators, likely due to their high fold-specific performance. 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. 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%. This was due to the RFSE’s decision-tree architecture, which allowed it to ignore base estimators that did not reduce error. 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 [18]. As described in the literature review, multicollinearity occurs when features are linearly dependent [18]. 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. This may explain why the Voting Ensemble and ENSE generally lower performance than the RFSE.

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 [13]. 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. 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 [49]. 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. Mahajan et al.’s (2023) review found that 82.6% of studies that used stacking ensembles found they had the highest performance of all tested model architectures [17]. In contrast, only 71.4% of studies that used voting ensembles found they had the highest performance [17].

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. 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. 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. Additionally, unlike the RFSE, the SVMSE used all base estimators, increasing the probability it learned ungeneralisable, noisy patterns in the base models’ predictions, reducing its test performance. However, the SVMSE achieved a smaller MSE score for country-level prediction than the Voting Ensemble. 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. This was likely due to the forecasting base estimators having more uniform performance. When trained to perform forecasting, base estimators were fit on data from every country, allowing each base estimator to ‘see’ extreme MMR values. 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. Therefore, there was less variation in the performance 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. Furthermore, the MRE of RFSE-Fs trained on just one model type only varied by 2%. 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. 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.

#### 6.22 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 [4]. This heterogeneity was seen clearly in my input data, where the median MMR for low-income countries was 617. 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 [3].

The range of possible MMR estimates, as well as median MMR, increased as income level decreased. 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. As well as having the highest variation in MMR, low-income countries also had the smallest number of available samples. 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.

My results confirmed this hypothesis, with the MSE of both the RFSE-CLP and RFSE-F decreasing as income level increased. 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. 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.

##### 6.221 Country-Level Prediction

When the RFSE was trained for country-level prediction, test MRE generally decreased as income-level increased, as expected. 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. 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.

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. 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.

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. 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. 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. 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. 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 [3]. 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. However, the RFSEs trained on income-specific data had higher MSE scores than the RFSE trained on all data. 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.

##### 6.222 Forecasting

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

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. The RFSE may not have anticipated these relatively high MMR values for higher-income countries, increasing its predictive error. 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. This explains the model’s relatively higher performance for low and lower-middle income countries.

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. This explains why the RFSE still had better MSE performance for higher income countries than lower-income countries, which had outliers of higher magnitude.

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. 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 [19]. 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 [19]. 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 [19]. 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.

However, training the RFSE on all data produced a lower MSE than training it solely on lower-middle income data. 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). 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 [19]. Thus, lower-middle income countries in the intermediate transition state benefit from learning how trends affecting countries in all transition stages influence MMR.

#### 6.23 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. 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.

### 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 41 and 42 in Section 5) [8, 9, 24]. Briefly, BMat used Bayesian hierarchical modelling to estimate non-HIV related MMR, with HIV-related MMR calculated separately [8]. CODEm used an ensemble of linear mixed effects regression and spatiotemporal Gaussian process regression [9]. In contrast, the GMatH microsimulation model simulated the reproductive lifecycles of thousands of women [24]. 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 [8, 9, 24].

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. 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. 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. 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 [28]. For example, in 2017, only 2 of 49 least developed countries had a death registration coverage of at least 50% [29]. 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 Section 3. For example, the UN MMEIG developed the BMis model to correct for errors in data from CRVS systems [22]. The Global Burden of Disease Study implemented algorithms to reassign deaths attributed to nonsensical causes to more statistically probable causes, reducing misclassification [23]. The GMatH model incorporated specific parameters that captured underreporting of maternal death [24]. 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. 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 [26]. 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 [26]. As a result, GMatH’s MMR estimates greatly exceeded the other models’ predictions for high-income countries, sometimes by over 100 points. 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. 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 [25].

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 [20, 21, 23]. 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 [21]. 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. 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 [20, 21, 23]. BMat’s MMR predictions were increasingly covariate-driven for countries with sparse maternal mortality data. 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 [21]. 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.

In addition to using different features than my model, BMat and CODEm assumed a global relationship between their covariates and MMR, unlike my model. 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 [30]. 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 conditions, 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.

In contrast to BMat and CODEm, GMatH used a variety of parameters covering biological information, quality of health care, and socio-economic trends [26]. 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 [26]. 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 12, 13) in the Analysis chapter. 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. 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 [2, 5]. The literature also describes how maternal health services are inaccessible to many due to financial constraints, affecting maternal health outcomes [6]. 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. However, the most important features in RFSE-F had slightly greater emphasis on non-communicable disease and lower emphasis 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 [1].

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.

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 [28]. They represent each feature’s contribution to the final prediction and are commonly used in machine learning studies [28, 29]. 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 [29]. 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. (2024) state that only addressing the biomedical causes of maternal mortality may have prevented further reductions in MMR [3].

The level of female employment is a socio-economic variable that is highly predictive of MMR, as shown previously, implying that improving women’s employment outcomes would reduce MMR. 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 [3,4]. For example, education affects women’s knowledge of health conditions and contraceptive options as well as their likelihood of engaging with maternal health services [30, 31]. 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 [30]. Therefore, increased investment in encouraging girls to finish their education may decrease MMR. For example, Ward 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) [2]. 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 [2]. 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. This finding was echoed by much of the literature, as medical complications during childbirth are one of the primary causes of maternal death [3,4]. 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 [33, 32]. 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 [50]. This candidate policy could have substantial effects in the 38 countries that have high maternal mortality burden but critical shortages in medical personnel [32].

Finally, my research demonstrated that knowledge of, and access to, contraceptive options is highly predictive of MMR. Many studies have recognised the strong relationship between the availability of family planning services and maternal mortality [3, 34]. 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 [34]. 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 [35]. 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 [34]. Consequently, models have predicted that family planning strategies can prevent up to 30% of future maternal deaths [34]. 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 [34]. Therefore, policies that increase the availability of family planning services could meaningfully reduce MMR [50].

### 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 [20, 36]. 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 [20, 23, 24]. The BMat model developers expressed interest in exploring alternative predictive variables for MMR in their 2012 paper [7]. 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 [38, 39, 40, 41, 42, 43]. 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 [10]. 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 [10]. 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% [11].

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 [20, 23, 24]. 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 [37].

### 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 [28]. 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 [20, 23, 37]. 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. 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.

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. 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 [46]. Additionally, high correlation between features 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 estimators. 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 [20, 23, 24]. 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. 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 [44, 45]. In semi-supervised learning, the model is initially trained on the labelled data [45]. This model is then used to predict the labels of unlabelled samples [45]. Finally, this initial model is retrained on all input data (both the original and newly labelled samples) [45]. Therefore, semi-supervised methods are particularly useful when labelled data is limited, such as in my thesis [44]. Semi-supervised models have been applied to health contexts, such as to identify patients with autoimmune disease [45]. For example, a previous study demonstrated that semi-supervised methods could more accurately detect Crohn’s Disease from magnetic resonance imaging than supervised learning [45].

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. 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 [46]. SMOTE is one of most influential pre-processing techniques in machine learning [46]. 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 [46]. 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 [46]. Additionally, it has been shown to produce insubstantial performance improvement when applied to high-dimensional data [46]. 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 [14]. 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 [2]. 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. 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 [30]. 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.

## 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.

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.

Citations:

1. <https://obgyn.onlinelibrary.wiley.com/doi/epdf/10.1111/1471-0528.12735>
   1. Obstetrics – maternal and perinatal morbidity
2. [**https://www.sciencedirect.com/science/article/pii/S2589537024002323**](https://www.sciencedirect.com/science/article/pii/S2589537024002323)
   1. Microsimulation to explore effects of education and geographic location on MMR
3. [**https://www.thelancet.com/action/showPdf?pii=S2214-109X%2823%2900468-0**](https://www.thelancet.com/action/showPdf?pii=S2214-109X%2823%2900468-0)
   1. Global determinants of maternal mortality
4. [**https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(24)00560-6/fulltext**](https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(24)00560-6/fulltext)
   1. Global and regional maternal mortality causes
5. <http://dx.doi.org/10.15585/mmwr.mm6835a3>
   1. Health care disparities by race
6. [**https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)31333-2/fulltext**](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)31333-2/fulltext)
   1. Financial
7. <https://pmc.ncbi.nlm.nih.gov/articles/PMC3886639/>
   1. Bmat 2014 paper
8. <https://www.thelancet.com/cms/10.1016/S0140-6736(24)00367-2/attachment/b7b1d025-a598-408a-b62e-025ebc1f08d1/mmc1.pdf>
   1. Supplement to: GBD 2021 Causes of Death Collaborators. 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. Lancet 2024; published online April 3. https://doi.org/10.1016/ S0140-6736(24)00367-2.
9. Trends in maternal mortality estimates 2000 to 2023: estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division. Geneva: World Health Organization; 2025. Licence: CC BY-NC-SA 3.0 IGO
10. <https://link.springer.com/article/10.1007/s42979-022-01249-z>
11. <https://academic.oup.com/bmb/article/121/1/121/2926155>
    1. Recording MMR methods
12. <https://link.springer.com/article/10.1007/s10462-019-09709-4>
    1. Missing value imputation review
13. <https://www.jclinepi.com/article/S0895-4356(18)30871-0/fulltext>
    1. Auxiliary variables and missing info
14. <https://www.tandfonline.com/doi/epdf/10.1080/08839510902872223?needAccess=true>
    1. Imputation analysis with surrogate variable splitting
15. <https://link.springer.com/article/10.1186/s40537-021-00516-9>
    1. Survey of missing data in ML
16. <https://link.springer.com/article/10.1007/s10462-020-09896-5>?
    1. Gradient boosting algorithm review
17. <https://pmc.ncbi.nlm.nih.gov/articles/PMC10298658/>
    1. Ensemble review
18. <https://www.mdpi.com/2227-7390/10/8/1283>
    1. Multicollinearity
19. <https://obgyn.onlinelibrary.wiley.com/doi/epdf/10.1111/1471-0528.12735>
    1. Obstetrics – maternal and perinatal morbidity
20. Trends in maternal mortality estimates 2000 to 2023: estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division. Geneva: World Health Organization; 2025. Licence: CC BY-NC-SA 3.0 IGO
21. The Annals of Applied Statistics 2017, Vol. 11, No. 3, 1245–1274 DOI: 10.1214/16-AOAS1014 © Institute of Mathematical Statistics, 2017
    1. 2017 un mmeig estimates
22. <https://www.tandfonline.com/doi/full/10.1080/2330443X.2023.2286313#abstract>
    1. Bmis model information
23. <https://www.thelancet.com/cms/10.1016/S0140-6736(24)00367-2/attachment/b7b1d025-a598-408a-b62e-025ebc1f08d1/mmc1.pdf>
    1. Supplement to: GBD 2021 Causes of Death Collaborators. 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. Lancet 2024; published online April 3. https://doi.org/10.1016/ S0140-6736(24)00367-2.
24. <https://www.nature.com/articles/s41591-023-02310-x>
    1. Ward, Z.J., Atun, R., King, G. *et al.* Simulation-based estimates and projections of global, regional and country-level maternal mortality by cause, 1990–2050. *Nat Med* **29**, 1253–1261 (2023). <https://doi.org/10.1038/s41591-023-02310-x>
25. <https://link.springer.com/article/10.1007/s00103-018-2793-0>
    1. Germany gbd
26. <https://static-content.springer.com/esm/art%3A10.1038%2Fs41591-023-02310-x/MediaObjects/41591_2023_2310_MOESM1_ESM.pdf>
    1. Gmath methods
27. https://genderdata.worldbank.org/en/indicator/sh-sta-mmrt?estimate=National
28. https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-023-06077-4
29. https://academic.oup.com/bmb/article/121/1/121/2926155?login=true
30. <https://journals-sagepub-com.virtual.anu.edu.au/doi/epdf/10.1177/0894439310370085>
31. https://dl.acm.org/doi/pdf/10.5555/3295222.3295230
32. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-025-22007-9
33. https://obgyn.onlinelibrary.wiley.com/doi/10.1111/1471-0528.12634
34. https://www.thelancet.com/action/showPdf?pii=S2589-5370%2823%2900441-8
35. https://www.bmj.com/content/351/bmj.h4255#:~:text=Doris%20Chou%2C%20medical%20officer1,EPMM)%20working%20groups%20Author%20affiliations
36. <https://cdn.who.int/media/docs/default-source/mca-documents/maternal-nb/ending-preventable-maternal-mortality_epmm_brief-230921.pdf?sfvrsn=f5dcf35e_5>
37. https://pophealthmetrics.biomedcentral.com/articles/10.1186/s12963-020-00245-w
38. https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1728-4465.2013.00365.
39. <https://www.nature.com/articles/s41591-022-01990-1>
    1. Gbd at 30 years
40. https://www.sciencedirect.com/science/article/pii/S2589537025004389
41. <https://genderdata.worldbank.org/en/indicator/sh-sta-mmrt?estimate=National>
    1. Mmr data
42. <https://genderdata.worldbank.org/en/topics/health#idAllIndicators>
    1. World bank data
43. <https://www.who.int/data/sets/health-inequality-monitor-dataset#swper>
    1. Empowerment data
44. <https://www.who.int/data/sets/health-inequality-monitor-dataset#wb>
    1. Health determinants data (mostly economic)
45. <https://www.who.int/data/sets/health-inequality-monitor-dataset#ihme-gbd>
    1. Illness prevalence data
46. <https://www.who.int/data/sets/health-inequality-monitor-dataset#gho>
    1. Income level data
47. <https://alz-journals.onlinelibrary.wiley.com/doi/epdf/10.1002/alz.13441?getft_integrator=scopus&src=getftr&utm_source=scopus>
    1. Semi-supervised methods 1
48. <https://www.tandfonline.com/doi/epdf/10.1080/1744666X.2024.2359019?src=getftr&utm_source=scopus&getft_integrator=scopus>
    1. Semi-supervised methods 2
49. https://dhsprogram.com/Countries/
50. <https://onlinelibrary.wiley.com/doi/10.5694/mja2.52452>
    1. Non-communicable disease and pregnancy
    2. Ramson et al., 2024
51. https://dl.acm.org/doi/10.1145/3686397.3686403
    1. xgboost missing data
52. <https://academic.oup.com/jrsssb/article/67/2/301/7109482>
    1. Elastic net
53. [**https://pubmed.ncbi.nlm.nih.gov/27642018/**](https://pubmed.ncbi.nlm.nih.gov/27642018/)