## 6. Discussion

The aim of this thesis was to estimate countries’ maternal mortality ratios between 1985 and 2018 to inform healthcare policy using interpretable machine learning techniques. A secondary aim was to identify the variables with the highest predictive power for MMR to highlight targets for healthcare policy that would result in large decreases in MMR. This chapter interprets the results presented in Section 5 and discusses them in the context of these aims and existing research in the literature. The strengths and limitations of my method were also examined in this chapter, which led naturally to a discussion of possible future directions for this research (6.4). This chapter begins with a discussion of the features with the highest predictive power for MMR, as this motivates the choice of base estimators used by the best-performing stacking ensemble to predict MMR, the primary aim of this thesis.

### 6.1 Discussion of Base Estimator Performance

#### 6.11 Feature Selection

I interpreted trends in base estimator performance, offering potential explanations for the differences in performance between the Random Forest, XGBoost, and LightGBM models trained on different feature subsets and missing data removal strategies. As described in Section 5.31 and 5.32, no single feature selection method consistently produced the highest performance.

However, the models trained for country-level prediction achieved higher performance more consistently on the feature subset hand-picked from the literature than on any other subset. This was echoed by the fact that the XGBoost base estimator model given the second highest importance score in the Random Forest Stacking Ensemble trained for country-level prediction was trained on this hand-picked feature subset (Table 13). The base estimator given the highest importance in this ensemble placed the greatest value on similar features, such as the percent of teenage mothers. This indicates that countries’ missing MMR estimates can be predicted with more consistently high accuracy using causally related variables. The more consistently high performance of models trained on this feature subset was observed much more strongly in terms of MSE than MRE. Thus, training models on features causally related to the target MMR outcome may have made the model more robust to outliers. 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, but the MMRs for these samples had a mean of 657 and standard deviation of 453 (Tables 3 and 10). This added context motivates the suggestion that using variables correlated to, but not causally related with, MMR may have led to inaccurate MMR predictions for low-income countries because the correlation did not occur for the high MMR range in addition to the lower MMR range.

While the XGBoost and LightGBM models trained to perform forecasting also had more consistently low MSE when trained on the hand-picked feature subset, the Random Forest model did not. In fact, the three model types used to perform forecasting all had their most consistently low MSE scores when trained on the ‘Correlation 0.6’ feature subset, which contained more features than the hand-picked subset (113 versus 40). Additionally, both base estimators given the highest importance scores in the Random Forest Stacking Ensemble trained to perform forecasting were fit on the ‘Correlation 0.6’ feature subset. Potentially, forecasting MMR requires use of a broader array of features than solely performing country-level prediction, as future MMR rates depend on variety of socio-economic trends that take a longer time to manifest. For example, one of the most highly weighted features in the base estimator given the second highest importance score in this RFSE was ‘cause of death, by non-communicable diseases (% of female population)’ (Table 14). The models’ lower error on the ‘Correlation 0.6’ feature subset was more noticeable when measured in terms of MSE than MRE. Thus, using this subset may have made the models more robust to outliers, potentially because at least one of the larger number of features was likely to have a relationship with MMR that the model could learn to extrapolate into the future and/or to the more extreme edges of the MMR range for low-income countries.

Despite the ‘Correlation 0.6’ and hand-picked feature subsets producing models with lower predictive error, models trained on datasets curated with no feature selection also performed well. In fact, of all the base estimators trained for country-level prediction, Random Forest and XGBoost models fit on all features produced with the lowest MRE, and LightGBM models produced the lowest MSE. Similarly, Random Forest and XGBoost models trained with all features to perform forecasting achieved the lowest MRE scores. However, the relative success of using no feature selection to perform forecasting must be qualified with the statement that there was only a small range of MRE scores achieved by these base estimators, limiting the actual improvement gained by using all features, especially when taking into account overlapping standard deviation across different cross-validation folds. The high performance of models trained without feature selection was likely due to the ability for decision-tree based models to ignore irrelevant and highly correlated features that did not improve decreases in predictive error when determining candidate splits at internal nodes, as discussed in the literature review. This finding shows the strength of the experimental method, where using a wider variety of features did not greatly increase error but did enable a comprehensive feature analysis to be conducted.

However, while features can be ignored, including too few features resulted in underfitting. More specifically, all models fit on the ‘Correlation 0.8’ feature subset had higher error. This ‘Correlation 0.8’ feature subset only contained 11 features, which primarily monitored mortality rates for broad categories of diseases and survival probabilities. Only two of these features were more explicitly related to socio-economic trends (literacy rates and use of menstrual products). Thus, models trained on this feature subset may not have had access sufficient variables monitoring socio-economic and other health-related trends to be able to accurately predict MMR. Similarly, models fit on the ‘Correlation 0.7’ feature subset generally had high predictive error. This was likely due to underfitting. While the features in this subset were highly correlated with MMR, they may not have had a sufficiently strong causal relationship with MMR to be able to predict it on out-of-sample data. Interestingly, this subset had 45 features, which was 5 more than in the hand-picked feature subset. Thus, a higher number of features correlated with the target outcome are needed to produce a similar level of accuracy as features hand-picked for their causal relationship with the target. This conclusion was shown explicitly by the higher performance of the ‘Correlation 0.6’ subset.

#### 6.12 Missing Data Removal

Missing data can 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 feature dimensions with missing data [14]. However, I found that there was no specific proportion of missing data beyond which the associated column or row should be removed that consistently produced the best performance across all feature subsets.

Therefore, the Random Forest, XGBoost, and LightGBM models could effectively handle high proportions of missing data using the surrogate splitting and default direction techniques described in the background information. As a brief recap, when the Random Forest prediction algorithm reaches an internal node whose split is governed by a feature with missing data, it uses ‘surrogate variables’ to determine the split [14]. Generally, the surrogate variable is chosen based on its correlation with the missing feature and/or its ability to produce the optimal possible split [14]. Given that over 482 pairs of features in my input dataset had an absolute pairwise correlation coefficient greater than 0.9, the surrogate splitting algorithm likely had a wide variety of similar variables to choose from when it encountered missing data. This hypothesis was lent credence by the literature, where a study by Twala (2009) found that surrogate splitting techniques have higher performance when feature variables are more highly correlated [14]. The use of default directions in the XGBoost and LightGBM base estimators worked similarly well, as the technique forces the model to explicitly learn the best path through the tree to take when feature data is missing. Additionally, features with a lot of missing data would likely be unhelpful, producing splits that only marginally diminished loss, making it less likely for these variables to be used to define internal node splits. Thus, their missing data would be less of a problem during prediction.

Nevertheless, the similar performance for different missing data thresholds was an interesting result, as researchers have hypothesised that analyses will be biased if a dataset contains more than 10% missing data, with rates of missing data between 5 and 15% necessitating sophisticated intervention [10, 14]. For example, Twala (2009) showed that, compared to a dataset with no missing data, classification error rate increase produced by 15% missing data was one and a half times smaller than the error rate increase when a dataset contains 50% missing data [14]. However, there is a debate in the literature about the exact proportion of missing data for which quality results can be attained, with other researchers arguing 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]. Nevertheless, data with a high proportion of missing data is often dropped due to researchers’ reluctance to introduce bias into their data [10]. For example, a study in Uganda predicted severe maternal morbidities by applying logistic regression to facility health data removed all samples with greater than 90% missing data, as studies have found that even principled imputation methods can introduce bias to datasets that have data missing at random [10]. In fact, there are few studies that investigate how to apply imputation to datasets with very high rates of missing data. A review of 111 papers found that only 12 investigated datasets with over 50% missing data, with the majority of studies benchmarked exploring datasets with less than 30% missing data [12].

Therefore, this thesis’s exploration of how decision-tree based models can work with datasets that have very high proportions of missing data is a strength, as the full input dataset used in this research had 78% missing data. This is a useful result, especially given the widespread appearance of missing data in facility-based healthcare datasets [10], with only 2 of 49 of the least developed countries having death registration coverage of at least 50% [11].

As well as showing that leaving samples and features with increasingly high proportions of missing data in the dataset did not harm predictive performance, I found that simply removing samples with large amounts of missing data could reduce performance. For example, Random Forest and XGBoost base estimators trained for country-level prediction experienced their highest MRE scores when trained on datasets curated with a missing data threshold of 85%. This indicates that the strict 85% threshold resulted in important information being removed from the trained dataset. This finding was validated by a large body of research that has discussed how removing samples with missing data can introduce bias into the dataset, especially when the data is missing not at random (MNAR) [14, 14]. As discussed in Section 4.21, removing missing samples when data is MNAR can obscure important trends, such as countries with more missing data, and thus less robust data collection systems, having higher MMRs. While the other base estimators trained in this analysis did not show a similar trend, this was taken as a warning sign, where utilising even stricter missing data thresholds would result in higher predictive error as more important information was removed from the dataset. Therefore, stricter missing data thresholds were not employed.

#### Model Type

As described in Section 5.33, no single model type consistently had the best performance across all feature subsets and missing data thresholds. In fact, the standard deviation in the models’ averaged performance metrics indicated large overlap in their fold-specific performance. The similarity between the model types was shown even more explicitly when the RFSE was trained with only one type of model to perform forecasting (Figure 31), where the MRE between the RFSE models trained with different base estimator subsets 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 was performed by comparing Random Forest, XGBoost, and LightGBM classifiers, it offers broadly relevant insights into regression performance due to the similar underlying architecture between the classification and regression models. The similarity of the models’ performance may be attributable to the fact that they all use ensembles of the same base decision-tree model.

Nevertheless, there were slight differences between the models. For example, the Random Forest base estimator tended to have lower average MRE than the XGBoost and LightGBM estimators across the five cross-validation folds when trained for both country-level prediction and forecasting (Figures 21 and 22). This trend may have been related boosting ensembles’ known tendency to overfit, as each subsequent base estimator in the ensemble is trained to correct the errors of its predecessor [17]. However, an analysis of the standard deviation in the Random Forest and XGBoost models’ MRE scores indicated that XGBoost performed better on specific folds than Random Forest when trained for country-level prediction. This may be due to the Random Forest models’ default feature subsampling, where it only considers a subset of samples when deciding on each internal node split. While this technique reduces the model’s potential for overfitting, by chance, it can result in important features being underutilised when deciding splits, reducing its performance [17]. However, this only occurred for models trained for country-level prediction, not the models trained to perform forecasting, reinforcing the random nature of this phenomenon.

In contrast, XGBoost models trained for country-level prediction had the lowest MSE scores (Figure 21). This may be attributable to their gradient boosting mechanism, which allows each subsequent model to explicitly correct for the mistakes of its predecessor [17]. Thus, later XGBoost base learners could focus their attention on learning trends to correctly predict outliers, as these would have resulted in the highest error for earlier learners. As a result, the XGBoost models would have higher performance on outliers, and thus lower MSE. While the LightGBM models were also based on gradient boosting, they may have had lower performance in this scenario due to their use of gradient-based one-side sampling (GOSS), as discussed in the background information. Briefly, GOSS reduces the number of samples used to split each node, where less informative samples are subsampled 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. Random Forest’s bagging algorithm would prevent it from benefiting from this iterative learning process. Interestingly, XGBoost did not achieve the highest MSE performance when trained to perform forecasting, where instead LightGBM and Random Forest models shared the best performance. Potentially, when performing forecasting, the overfitting caused by XGBoost’s ability to improve its understanding of outliers outweighed the added benefit of more accurately predicting these outliers, especially given the low incidence of high MMR estimates in the test data. There may have been less overfitting when the model was trained for country-level prediction, when models trained on the different training folds would be fit on data from different low-income countries, which by chance may or may not contain samples with high MMR, reducing overfitting to these “outliers”. In contrast, the models trained to perform forecasting would have been trained on data from all countries for a specific year.

As a final discussion point, XGBoost models tended to have higher standard deviation in their performance than the Random Forest models. This may 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 caused it to overfit to its training folds less severely, resulting in its lower standard deviation in its error across its cross-validation folds.

#### 6.14 Summary

Therefore, in general, no model type, feature subset, or missing data removal technique consistently had the highest performance.

In fact, the overlapping standard deviation of the models’ error metrics implied that the exact samples and features included in a specific cross-validation fold had more influence on the model’s performance than the feature selection and missing data removal strategies used to curate their training data. 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. This would result in some of the base estimators being trained on data that would not give them an understanding of how high MMR could be. This outcome was seen explicitly in the lower-middle income train/test split used for country-level prediction, where the highest MMR values in the test dataset were much lower than in the train/validation sets (Figure 12).

This observation motivated the use of an ensemble model to combine the predictions of the different base estimators.

### 6.2 Discussion of Random Forest Stacking Ensemble Performance

#### 6.21 Random Forest Stacking Ensemble Performance Relative to Other Models

The discussion in the previous section highlighted the importance of using a stacking or voting ensemble model to combine predictions from each of the base estimators. Rather than sifting through the 300 models trained on each cross-validation fold for both country-level prediction and forecasting, an ensemble model could learn which base estimator produced the most accurate MMR prediction in a specific scenario. Potentially, models trained on cross-validation folds solely containing samples with lower MMRs would have a clear idea of trends particular to these samples, while models trained on samples with higher MMR would have learned their associated trends. Thus, combining the two types of models would produce higher performance than using one or the other.

As a result, the Random Forest Ensemble was the best performing voting/stacking ensemble because it could most effectively learn the contexts in which different base estimators were most useful [17]. This was shown explicitly in it placing a high importance on only a small subset of base estimators (Figure 25). The RFSE placed the greatest emphasis on XGBoost models, potentially due to their higher fold-specific performance. The higher performance of XGBoost models was reinforced by the higher performance of the RFSE solely trained on XGBoost models than the RFSEs analogously trained on just LightGBM and Random Forest base estimators. However, the ability for the RFSE ensemble to ignore all base estimators that did not improve predictive performance due to its base decision-tree architecture resulted in very little difference in performance (less than 1% improvement in MRE) as a result of including all three model types in the RFSE versus a single category. This ability to ignore the predictions from less accurate models was particularly important in the context of my research, as there were many similarly performing base estimators. Using all of these base estimators may have introduced noise, where the model learned unimportant and ungeneralisable differences between the base estimators. The choice of base estimators by the RFSE trained to perform country-level prediction was shown to purposeful, as the base estimators given the highest importance scores by the ensemble did not change after retraining and permuting the order of the base estimators. Thus, the RFSE had learned which base estimators to use to have the lowest predictive error.

In contrast, the Voting Ensemble and Elastic Net Stacking Ensemble may have struggled to isolate the impact of specific base estimators from the pool of similar base estimators, as models based on linear regression are known to struggle with multicollinearity [18]. As described in the literature review, multicollinearity occurs when features of a model are linearly dependent [18]. The similarity of the base estimators used in the voting/stacking ensembles mean that they are linearly dependent, making it difficult for the Elastic Net and Voting models to isolate the effect of a specific base estimator and thus nominate certain base estimators as the most important while the others are irrelevant. This could explain why these ensembles used a much larger number of base estimators than the Random Forest Stacking Ensemble (Figures 26 and 27). It could also explain why Elastic Net’s L1 regularisation did not effectively reduce the number of base estimators used. Additionally, unlike the Elastic Net Stacking Ensemble, the Voting Ensemble did not employ either L1 regularisation for base estimator selection, preventing it from dropping any base estimators. This could explain the Voting Ensemble’s poor performance, especially in the MSE metric, as there may only have been a subset of base estimators that could effectively predict outliers while the majority of other base estimators were misled. The greater performance of the stacking ensembles versus voting ensembles was validated by the literature, as Mahajan et al.’s (2023) review found that stacking ensembles had the highest performance of all tested model architectures in 82.6% of studies where they were used, in comparison to voting ensembles’ highest accuracy in 71.4% of studies that tested them against other models [17].

The use of a wide variety of base estimators, some of which with poor performance, resulted in the Voting and Elastic Net Stacking Ensembles having worse performance than the RFSE. More specifically, the RFSE could learn which base estimators had the best performance in certain instances (e.g. if certain feature variables were especially high, a specific base estimator may better predict a high MMR). This could explain why the RFSE placed the highest importance on base estimators that had high MSE scores on the entire dataset, as these estimators may have performed extremely well for certain types of samples. In contrast, the Voting and Elastic Net Stacking Ensembles combined their base estimator predictions in a fixed way, without being able to tailor this combination to specific, local trends in the data. Therefore, they benefit less strongly from being able to combine the base estimators.

Interestingly, there was one instance where the Random Forest Stacking Ensemble (RFSE) did not have the best performance. When trained for country-level prediction, the RFSE had the lowest MRE score but the Elastic Net Stacking Ensemble (ENSE) had the lowest MSE score. This indicates that the ENSE handled outliers more effectively, potentially because the RFSE overfit on the training data, whereas the L1 and L2 regularisation employed by the ENSE prevented such overfitting. Thus, the RFSE was more at risk of outliers.

There was less variation in the ensembles’ performance when they were trained to perform forecasting, as the base estimators performed more similarly. This was seen explicitly in terms of MRE, and by how the XGBoost models did not have as significantly superior fold-specific performance as well as how differences in the MRE scores of RFSEs trained on just one model type to perform forecasting only varied by up to 2%. In contrast, the MRE scores of the RFSEs trained on different subsets of base estimators to perform country-level prediction varied by up to 8% (Figure 30). The smaller differences between base estimators could also be observed by how the importance scores that the RFSE gave to base estimator scores changed notably when the model was retrained, indicating that the importance of different base estimators could be re-weighted without large impacts to the model’s performance. The smaller differences between base estimators potentially decreased the utility of RFSE’s ability to solely use the best base estimators, reducing its edge over the other ensemble models. Additionally, when performing forecasting, overfitting in the RFSE may have been less of a problem, as there was less variation in the base estimators and therefore less noise to that the RFSE could overfit to. This could explain why the RFSE had a lower MSE score than ENSE when performing forecasting, while the opposite trend was observed when the models were trained to perform country-level prediction.

The smaller variation across base estimators trained for forecasting may be due to all the base estimators being trained on data from all countries between 1985 and 2014. This prevented some base estimators from only being trained on data from countries with lower MMR values, which was particularly relevant for low-income countries, whose MMR had a standard deviation of 453. Therefore, the base estimators would have learned similar trends in the data. In contrast, base estimators trained for country-level prediction were only trained on specific countries, as described in the method. By chance, a specific model may only have been trained on low-income countries with MMR value at the lower end of the possible range, while other models may have been trained on countries with MMR values at the higher end of the range. These models would have very different performance outcomes, accounting for the greater difference between base estimators’ performance when trained for country-level prediction versus forecasting.

As a final point, the support vector machine stacking ensemble (SVMSE) had the highest MRE and MSE when trained to perform forecasting and had the highest MRE when trained for country-level prediction. This may be due to the support vector machine’s sensitivity to noise, as its loss and predictions are based on the support vectors, or the datapoints outside of its error tolerance margin. Given the wide range of MMR values but sparse representation of high MMR values in this dataset, the model may have overfit to the few high MMR examples, which were likely used as support vectors. Additionally, unlike the RFSE, the SVMSE used all features, preventing it from dropping less informative base estimators and thus being penalised by their poor predictions. Furthermore, by using the predictions from all base estimators, the model may have identified ‘patterns’ in noisy differences between similar base estimators, again causing it to overfit to the training dataset and reducing its test performance. However, the SVMSE may have outperformed the Voting Ensemble for country-level prediction MSE because it could use its polynomial kernel (chosen through hyperparameter tuning) to learn local information and curvature in the data, allowing it to benefit a bit more strongly from being able to combine different base estimators.

#### 6.22 Variation in Random Forest Stacking Ensemble Performance Across Income Levels

The Random Forest Stacking Ensemble’s performance was expected to vary across different income-levels due to heterogeneity in the MMR estimates between countries from different income levels. Studies have found that 95% of maternal deaths occur in lower-middle and low-income counties or fragile settings [4]. This global heterogeneity in MMR estimates across different income levels was seen clearly in my input data, where the median MMR for low-income countries was 617 while the median MMR for high-income countries was 8 (Table 10). Moreover, key summary statistics about features known to be associated with maternal mortality, such as the percentage of women having access to prenatal care, also varied with income-level (Table 10) [3]. This difference across income levels was reinforced by my PCA analysis, which showed that clusters corresponding to higher-income countries were the same clusters that corresponded to low MMR estimates.

As well as different income levels having different median MMR estimates, their range of MMR values also varied. For example, the standard deviation in low-income countries’ MMRs was 453, compared to 55 for upper-middle income countries (Table 10) in my dataset. Despite having the highest variation in MMR, low-income countries also had the smallest number of available samples. More explicitly, after cleaning the input dataset and removing all samples missing an associated MMR estimate, only 78 low-income samples remained, compared to 1,405 high-income samples. As a result, I expected that the RFSE’s performance would deteriorate for lower income-levels, as it had to learn how to predict a wide range of possible MMR estimates using only a small number of samples for the lower-income countries. Additionally, it could learn clear patterns for these high-income countries, as the PCA analysis showed the strongest clustering for high-income countries, although this could solely be due to the higher availability of data for these countries.

This hypothesis was confirmed when performance was measured in terms of MSE. More explicitly, both the RFSE trained for country-level prediction and the RFSE trained to perform forecasting achieved lower MSEs when predicting on high-income samples, with differences in MSE spanning multiple orders of magnitude (Figures 33 and 34). MSE is a good metric to use to show this trend, as it squares differences between model predictions and ground truth, and thus more heavily penalises outliers. Large differences between predicted and ground truth MMR values were more likely for low-income countries, as the model had to predict a single MMR estimate from a wide range of estimates. Additionally, since the MMRs of low-income countries had higher magnitudes, the difference between the predicted and ground truth estimates could be larger. This was compounded by the lack of available training samples for low-income countries, making it more challenging for the model to accurately determine where in the wide range of MMR values a test low-income country fell. In contrast, high-income countries’ MMRs fell in a much tighter range and had lower magnitudes, with more training samples available to tell the model exactly where in the range these estimates fell. This was likely an accomplishable task given the clustering by income level observed in the PCA analysis, where the model could identify high-income specific trends. The greater uncertainty and thus higher potential for error when predicting large ground truth MMR estimates was also shown by how consensus in base estimator MMR predictions decreased substantially as MMR increased. As a result, MSE would more severely punish mispredictions for lower-income countries, explaining the trend described above.

However, trends in the RFSEs’ MRE scores were contrary to expectations. The MRE score is a better benchmark for the model’s performance on the entire dataset, with large outliers less heavily penalised than with MSE. Therefore, MRE performance was more strongly influenced by whether the distribution of the ground truth MMR values in the test distribution was similar to the distribution in the train and validation sets. Given the different train/validation/test sets used for the RFSE models trained for country-level prediction versus forecasting, they were discussed separately below.

##### 6.221 Country-Level Prediction

When the RFSE was trained for country-level prediction, test MRE generally decreased as income-level increased, as expected (Figure 33a). However, the model had the lowest MRE when predicting MMR for lower-middle income countries, which was contrary to expectations. This unexpected result was due to the test ground truth MMR distribution for lower-middle income countries being a small subset of the train/validation distribution. The Q1 to Q3 test MMR range for lower-middle income countries was between 41 and 60 while the train/validation range was 33 to 283. The Q2 of the two ranges was the same (52). Therefore, the RFSE’s training data completely covered the test data, which enabled the model to learn the necessary patterns to accurately predict the test MMR estimates, giving it very high performance on the lower-middle income dataset. Furthermore, since the test set did not contain the upper end of this distribution, it could use the patterns it had learned from the large number of higher-income samples to accurately predict the lower MMRs, while the train data may have struggled to predict higher MMR values for which there is less data. This explains why the lower-middle income’s test MRE was smaller than its 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 to Q3 train/validation MMR range was much smaller, with their Q1 and Q3 values more similar to their test sets’ Q1 and Q3 MMRs. 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 to the train and validation sets, it achieved higher performance on these sets than the test set, unlike for the lower-middle income countries.

The RFSE’s extremely high MRE performance for lower-middle income did not translate into it having the lowest MSE score because the lower-middle income test dataset contained outliers with high MMRs values of a similar magnitude 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 that for low and upper-middle income. While these outliers may have produced large predictive errors, they must have been infrequent in the lower-middle income dataset, as they did not notably reduce MRE performance.

In contrast to the lower-middle, upper-middle, and high-income test distributions, the low-income test MMR distribution was different from the 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. As discussed previously, this was responsible for the RFSE’s high MRE and MSE scores on 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 range used to train the model versus outside the model’s experience.

Despite the variation in performance between models trained on different income levels, the sensitivity analysis demonstrated that the RFSE identified patterns in the data that it could use to predict MMR regardless of income level, as there was little difference between the MRE scores of the RFSE trained on all data versus income-specific data. These common patterns were visualised on the PCA graph, where projection of samples onto their top two principal components showed a strip where samples from low, lower-middle, and upper-middle countries were overlaid. This reinforced the idea that the RFSE could find patterns in the feature data that were predictive of MMR for all income levels. For example, level of access to skilled medical practitioners during childbirth would be predictive of MMR across all income levels [4]. 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 two MSE scores was extremely small when put into context, indicating that this deviation from the general trend was likely due to random noise.

##### 6.222 Forecasting

The RFSE trained to perform forecasting had the opposite trend in MRE performance as initially expected, as its MRE scores increased as income-level increased. This was attributed to differences between the model’s train/validation and test ground truth MMR distributions.

As described in Section 5.22, there were portions of each income level’s test set outside of the associated train/validation distribution. More specifically, the lower-middle, upper-middle, and high-income test distributions all had Q2 and Q1 values lower than their associated train/validation distributions. As a result, the model’s knowledge of possible MMR values for a specific income level that it learned from the train dataset may not have generalised to the test set, reducing its performance. This would be particularly important for the high-income dataset, where the test set contained MMR values lower than the vast majority of MMR values that the model would have seen in its train/validation data, likely increasing the model’s error. More specifically, the Q1 value for the high-income test dataset was half the Q1 value of it train/validation dataset, reinforcing that the model was tested on rarer low magnitude MMR values than used to train it. The high standard deviation in the model’s MRE scores were also explained by the occurrence of test datapoints outside the range of train/validation distribution, as the model’s prediction error would vary depending on whether it was predicting for a sample that had an MMR within a familiar range versus outside this familiar range.

Figure 14 showed explicitly how each of the upper-middle and high-income datasets contained at least one year with a median MMR value significantly greater than the average MMR. For example, the high-income test median MMR was over two times greater than the train/validation median. In contrast, the test dataset for low-income countries did not have an ‘outlier year’, explaining its higher performance than expected. The lower-middle income test dataset had a year with a higher than usual, but it had years with higher median estimates. The error likely caused by this outlier year explains why the model did not perform substantially better on the lower-middle income dataset than the low-income dataset.

This difference between the train/validation and test datasets was therefore responsible for the model having worse MRE performance for higher income datasets. However, as described previously, the model had worse MSE performance for lower income levels. Therefore, the higher income models must have made many small mistakes in their predictions due to the offset train/validation and test distributions that would have added up to increase MRE. Nevertheless, these mistakes were likely low magnitude, as the actual magnitude difference between the Q1 values for the test and train/validation distributions were small, preventing them from inflating MSE.

The difference between the train/validation and test MMR distributions also explained why the RFSE trained on data from all income levels incurred a higher MSE score than the RFSEs only trained on low, lower-middle, or upper-middle income data. As discussed previously, the MSE heavily penalises outliers, or in this case, high magnitude MMR values. The outlier years in the the

Interestingly, the sensitivity analysis showed that the RFSE trained on data from all income levels incurred a higher MSE score than the RFSEs only trained on low, lower-middle, or upper-middle income data. This may be related to the fact that change in MMR over time may be due to different features for different income levels. This relates to the obstetric transition model mentioned in the background information, where 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 a country’s income level may reduce noise and allow the model to focus on relevant trends for the countries’ stage in the transition model. This may also explain why the RFSE trained on all data had higher MRE than the RFSEs trained on just low and lower-middle data.

Interestingly and contrary to this trend, the RFSE trained solely on lower-middle income data had a higher MRE than the RFSE trained on all data. This indicates that countries in this income group benefit from seeing how trends in features across countries in different stages of the transition model affect 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 in Table 10). The literature describes this stage as complex because countries in this stage benefit from both increasing access to care and increasing quality of care, bridging the early and late stages of the model [19]. This finding corroborates the hypothesis that lower-middle income countries in this intermediate transition state benefit from seeing how trends in a variety of features influence MMR, explaining why the model trained on data from all income levels performed better than the model trained on just lower-middle income data.

### 6.3 Comparison of My Results to the Literature

Building on this comprehensive understanding of my best-performing model’s results, I now discuss how and why my model’s performance varied from the BMat, CODEm, and GMatH models in the literature.

Regardless of whether my RFSE was trained to perform country-level prediction or forecasting, the majority of its MMR estimates were smaller than the associated estimates from the BMat, CODEm, and GMatH models. However, the magnitude difference between my estimates and the corresponding GMatH estimates were generally larger than the magnitude differences between my estimates and the BMat and CODEm predictions.

Interestingly, the proportion of my MMR estimates within the 95% confidence interval of the literature’s estimates was very similar to the proportion of ground truth MMR estimates used to train my model that were within those same confidence intervals. This indicates that the differences between my estimates and the literature’s estimates were due to lack of correspondence between the literature’s MMR predictions and the ground truth MMR values used to train my model rather than inherent inaccuracy in my model’s prediction.

The most likely reason why the ground truth MMR estimates used to train my model were lower than the estimates produced by the BMat, CODEm, and GMatH models was because I did not adjust these ground truth estimates for underreporting and misclassification errors. These ground truth estimates were sourced from a country’s national estimates, which are informed by its national civil registration and vital statistics (CRVS) system, national surveys and censuses, and other specialised studies [27]. However, many papers have reported on the widespread under-reporting and misclassification of maternal mortality, as described in the background information. In brief, 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 have been trained on ground truth MMR values that were underestimates of the true maternal mortality ratios for a specific country, year pair.

While I did not perform any adjustment for under-reporting and misclassification, the literature models all developed specific procedures to correct for low-quality input data, as discussed in the literature review. For example, the UN MMEIG developed the BMis model specifically to correct for errors in data from CRVS systems [22] and the Global Burden of Disease Study implemented algorithms that reassigned nonsensical deaths to more statistically probable causes, reducing misclassification [23]. The GMatH model incorporates specific parameters that capture underreporting of maternal death [24]. This analysis would explain why my MMR predictions were generally smaller than the corresponding BMat, CODEm, and GMatH MMR estimates.

However, there were also methodological differences that could account for variation between my MMR estimates and the literature’s predictions. This can be seen explicitly by how GMatH’s estimates were higher than the MMR predictions from any other model for high-income countries, sometimes by over 100 MMR points. According to the GMatH documentation, some of the parameters used in the GMatH model for high-income countries were based on the prior distribution for upper-middle income countries [26]. This substitution was done for variables like ‘number of antenatal care visits’, which were informed by Demographic and Health Surveys. These surveys only collected data from lower-income countries, thus preventing informative priors about high-income countries from being used [26]. While this allowed the model to run, it contributed to the GMatH estimates for high-income countries greatly exceeding that of lower income countries for most of the time period tested. It also resulted in very wide confidence intervals for these countries. This theory would also explain why the magnitude difference between GMatH’s MMR estimates and the other models’ estimates for upper-middle income countries was smaller than for high-income countries. It would also explain why my estimates were so much smaller than the GMatH estimates, as the majority of my test data was for higher-income countries. This highlights one of the strengths of my model, where I did not need to make assumptions about the underlying data generation process that could similarly bias my model. Nevertheless, the GMatH model estimates provide a useful counterpoint for the other models, as they encourage scientific debate about the validity of different approaches. This is particularly important given that the true MMR estimate may lie somewhere between the predictions from all the different models, as a study conducted in Germany found that the CODEm model underestimated the true mortality caused in Germany due to diabetes [25].

All three of the BMat, CODEm, and GMatH models incorporate hierarchical regression modelling in some way, where they assume a certain amount of regional heterogeneity to use regional means to predict MMR and other important features when for a country with sparse data. The consequent smoothing can be observed in Figures 40 and 41, where the literature’s estimates over time are smoother than my estimates, where no averaging over geography or time was used due to my model’s ability to handle missing data. The lack of smoothing used by my model was particularly visible in Figure 40b, where country-level predictions for Colombia were visualised. This smoothing could result in the models ignoring region specific information if the regional mean used for smoothing was not representative. This may be one of the reasons why my country-level MMR predictions for lower-middle and low-income countries (Figure 40c, 40d) were 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.

Another reason for the variation in MMR estimates between the four models was their use of different feature subsets and covariates to predict MMR. As discussed in the literature review, the BMat model only used 3 covariates while the CODEm model only used 19 [20, 21, 23]. The GMatH model used a wide variety of parameters [26]. My model used different subsets of 720 features. These differences in the covariates used to estimate MMR was one of the primary drivers for variation in the models’ MMR estimates. They could particularly explain variation between my estimates and predictions from the BMat model for countries with sparse data, as the BMat model’s MMR predictions are increasingly covariate-driven when empirical data about maternal mortality is missing [21].

By incorporating information from a wider variety of possibly meaningful variables, which represented information spanning socio-economic trends to mortality rates to quality of health care indicators, my model could estimate MMR using a more holistic approach than the BMat or CODEm models. This may have contributed to my model’s MMR estimates being higher than the corresponding BMat and CODEm estimates for low-income countries, as my model could use a wider variety of information to predict the MMR values for data-sparse areas. Additionally, the BMat and CODEm models assumed a global relationship between their covariates and the MMR. However, as described in the literature review, this relationship may change depending on local conditions, potentially introducing inaccuracy into the model. In contrast, my RFSE models have no global assumptions about the interactions between variables and MMR, as all predictions are based on more ‘local’ information, as the prediction is derived from a mapping between the specific subset of the input space to a terminal leaf node of a decision-tree. This allows more tailored predictions. As described in the literature review, the effect of skilled birth attendance (SBA) on maternal mortality is only significant when national SBA coverage is at least 40% [30]. This type of local relationship between SBA and MMR could be more effectively modelled by my RFSE, especially when missing data makes it difficult for the other models to adjust the covariate driven estimates. This could contribute toward explaining why my MMR estimates for lower-income countries exceeded the estimates from the literature models, especially given that SBA is one of the covariates used by both BMat and CODEm.

In contrast to the BMat and CODEm models, GMatH used a variety of parameters covering biological information, quality of care, and socio-economic information [26]. The GMatH model also fit the values of its were fit to local conditions [26]. However, differences in its use of features still contributed to differences between the GMatH estimates and my RFSE estimates [26]. Each of the parameters considered by the GMatH model were associated with some level of uncertainty, with the combination of all parameters’ uncertainties potentially contributing to the wide confidence intervals observed for GMatH estimates in Figures 40 and 41 and potentially impacting model estimates. In contrast, the decision-tree based models used in my thesis can ignore irrelevant features that do not contribute to reducing loss at internal nodes, preventing additive error 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 model and those generated by the BMat, CODEm, and GMatH models were driven by underestimation of MMR in my ground truth data, methodological variation, and differences in the features used to estimate MMR.

### 6.4 Discussion of Feature Importance

I first examined the features with the highest predictive power for MMR. I then used this discussion to explain differences between the performance of Random Forest, XGBoost, and LightGBM base estimators trained on the different feature subsets.

The variables with the highest predictive power for MMR were the most valued features in the base estimators given the highest importance scores in the Random Forest Stacking Ensemble.

As discussed in Section 5.53, the features with the highest predictive power for 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 women’s nutritional status (Tables 13 and 14). In addition, the variables detailing the country’s income level, national life expectancy, fertility rates, presence of skilled health staff at birth, and percentage of teenage mothers were also among the features with the highest predictive power for country-level MMR predictions. While the features with the highest predictive power were similar for the RFSEs trained for country-level prediction and forecasting, the most important features used to perform forecasting had a slightly greater emphasis on monitoring long-term non-communicable disease.

Therefore, socio-economic indicators were extremely valuable for estimating MMR for both countries with sparse data and into the future. This observation was reinforced by findings in the literature, which described how socio-economic trends like availability of contraception, women’s education level, racism in the healthcare system impacted maternal mortality [2, 5]. The literature also describes as how maternal health services are inaccessible to many due to financial constraints, validating my finding that employment is related to MMR outcomes [6]. Employment can also indicate trends in education and agency, both of which are socio-economic factors highlighted by the literature as influential for maternal mortality outcomes [2, 3]. This validates the finding that women’s national employment status has high predictive power for MMR.

These results also showed that the RFSE trained for country-level prediction placed more emphasis on current aggregate measures of health system coverage and performance, while the RFSE trained to perform forecasting placed slightly more value on longer-term health trends. This validated by the literature, as experts expect that maternal mortality rates will become increasingly determined by incidence of non-communicable diseases than direct, obstetric causes [1]. This phenomenon is referred to as the ‘obstetric transition’ [1]. For instance, indirect obstetric deaths, such as due to non-communicable diseases, were responsible for 23% of maternal deaths between 2009 and 2020, as non-communicable diseases can increase the risk of complications during pregnancy and maternal death [3,4].

The importance of socio-economic variables in MMR estimation was reinforced by the observation that base estimators given low importance scores in the RFSEs placed more emphasis on specific health-related trends, such as specific infectious disease, while those given more importance placed higher value on socio-economic trends (although the former did place some emphasis on socio-economic trends as well). This highlights policy priority areas and motivates the move for further emphasis on addressing socio-economic trends to reduce MMR in addition to solely addressing obstetric trends. These results reinforce the existing body of work that argues in favour of reducing MMR via tackling socio-economic trends. For instance, Souza et al. (2024) state that solely devoting attention to addressing the biomedical causes of maternal mortality may have caused progress in reducing MMR to stagnate [3].

Thus, this thesis’s use of a range of socio-economic and health-related features to predict MMR was a strength of in approach. This is especially true given that the two most widely used models used to predict MMR (BMat and CODEm), only use 3 and 19 covariates, respectively, with limited consideration of socio-economic variables [8, 9]. In fact, the BMat model developers expressed the interest in exploring alternative predictive variables for MMR in their 2014 paper [7]. Therefore, this thesis’s ability to explore a wider variety of possible predictor variables was a strength and enabled a wide variety of possible health policy targets to be evaluated for their utility in reducing MMR. Additionally, the fact that many of the features with the highest predictive power for MMR identified in this thesis were also established risk factors in the literature lends credence to the accuracy of my results.

However, the findings presented in this section must be qualified by saying that feature importance calculations may have been affected by the highly correlated variables used in this thesis. More specifically, feature importance was calculated based on the amount by which the feature reduced loss when used to define a split at an internal node. However, a subset of correlated features and features that contained similar information may have produced similar reductions in loss, meaning that the feature from this subset chosen first, by chance, during the model building would be attributed with generating the highest loss despite it being a random choice from a subset of features. This chosen feature would then be given an importance score, which could have been given to any one of the correlated/similar features in the subset. This explains why the importance scores of base estimators in my RFSE model changed after retraining, as the similarity and high correlation between features in my input dataset made tree structures and feature importance scores unstable.

While this could mean that other features in my dataset had similar importance scores as those discussed above, the features given high importance were validated as having meaningful relationships with MMR by the literature, as discussed above. Additionally, I calculated the Shapley values for the features in the two base estimators given the highest importance scores in my RFSE to determine whether the ordering of feature importance used in this thesis was robust. Briefly, Shapely Additive eXplanation (SHAP) values are a stable measure of feature importance based on cooperative game theory that provide information about how much each feature contributes to the final prediction [28]. These values are employed by machine learning studies to estimate the features with the highest predictive power for maternal health outcomes [29]. For example, Taye et al. (2025) used SHAP values to determine that the place of delivery and residence (e.g. rural) had high predictive power for whether a birth in Sub-Saharan Africa will be accompanied by a skilled birth attendant in a Random Forest model [29]. However, initial experimentation with SHAP values for the two base estimators with the highest importance scores in my RFSEs indicated that the features given the highest SHAP values were almost all the same as those given the highest importance scores in my thesis, such as measures of employment, skilled birth attendance, and proportion of teen mothers. While the SHAP technique placed more emphasis on mortality rates, the broad similarity between the features given high importance in my thesis and using the SHAP technique gave my feature importance findings further validity.

### 6.4 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 to inform global and national health policy about which regions are most in need of targeted health interventions to reduce MMR. This aim was accomplished via the development of my RFSE models, which can be used to advise different types of policy. The secondary aim of this thesis was to identify the features with the highest predictive power for MMR, as discussed in the previous section, to ensure that public health policies address trends to produce the highest achievable reductions in MMR. The feature importance results presented in the previous section will be further explored here to recommend policy actions.

As discussed above, socio-economic features had high predictive power for MMR, indicating the necessity of targeting socio-economic trends to reduce MMR. Measures of employment were found to be highly predictive of MMR, with employment driven by, and reflective of, a mixture of trends in women’s access to education, women’s literacy, and women’s agency, all of which are themselves have important relationships to maternal mortality [3,4].

Therefore, a recommendation for a policy focused on reducing MMR that can be made from the results of this thesis, as well as existing research, is to increase investment in encouraging girls to finish their education. Research has found a significant relationship between maternal death and low education in countries with lower human development index scores [30]. Analysis indicates that 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]. This higher risk has been attributed to the general trend of education affecting the women’s knowledge of their health condition and potential contraceptive options as well as the likelihood of a woman engaging with maternal health services [30, 31]. The importance of policy targeting women’s education outcomes is further reinforced by Ward et al. (2024)’s use of the GMatH microsimulation model to demonstrate that a global MMR between 76 and 120 could be achieved in 2030 (which would successfully meet the UN’s Sustainable Development Goal) by ensuring that all women complete secondary school [2]. The authors described how the magnitude decrease in MMR produced by this socio-economic strategy would be akin to the decrease produced by a health policy that increases the number of women delivering their child in a health facility as well as the availability of clinical services [2]. Therefore, increasing women’s education level may be an effective policy for reducing MMR.

Additionally, my research has shown that skilled birth attendance at birth is a powerful predictor for maternal mortality, as echoed by much of the literature, especially given that complications at birth like excessive bleeding are some of the primary causes of maternal mortality [3,4]. This finding and substantial evidence base highlights the importance of health policies that increase access to skilled medical practitioners at birth. For example, the government could increase incentives for skilled practitioners to work in more rural areas that may have smaller number of healthcare professionals, as well as subsidise medical training. This candidate policy could have substantial effects in the 38 countries that have been identified as having high maternal mortality burden but critical shortages in supply of medical personnel [32]. This recommendation was reinforced by existing literature that highlights the importance of both greater number of healthcare professionals and ensuring that the quality of care provided is high [32]. The importance of this recommendation was also echoed by the Ending Preventable Maternal Mortality (EPMM) strategy, which is a global, multi-partner program involving various governments and international organisations [33]. This strategy set a goal of having over 90% of births globally attended by a skilled medical practitioner [33].

Finally, knowledge of, and access to, contraceptive options was determined to have high predictive power in this thesis. This finding was also echoed throughout the literature, where many studies have recognised the strong relationship between the availability of family planning services and maternal mortality and other pregnancy-related complications [3, 34]. Such family planning services are considered to decrease both the total number of pregnancies and the number of high-risk pregnancies, such as pregnancies in young girls and much older women [34]. Therefore, increased provision of family planning services may also be able to reduce the number of teenage pregnancies, which was another feature with high predictive power identified in this thesis. This finding was reinforced by research that has 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 these high-risk pregnancies results in a lower number of women who are exposed to pregnancy-related complications and thus reducing maternal mortality [34]. As a result, estimates have placed family planning strategies as being able to prevent up to 30% of future maternal deaths [34]. As a case study, Indonesia’s national family planning program, which has been running for approximately 50 years, produced an increase in contraception use, which is estimated to have resulted in 38 to 43% fewer maternal deaths [34]. Therefore, policies that increase the availability of family planning services would likely be able to meaningfully reduce MMR.

### 6.5 Strengths of this Research

This thesis successfully produced MMR estimates for a diverse range of countries between 1985 and 2018. These MMR predictions were similar to estimates of MMR from the BMat, CODEm, and GMatH models, which are high performing models in the literature and are 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 results and these literature estimates was a major strength of this research, as it indicates that my results are valid and can thus be used to inform policy. As a result, my thesis has met its first aim. Additionally, my results provide an alternate perspective of global trends in MMR, contributing to the debate about the best methodology for estimating MMR and potentially helping to provide consensus about the true MMR values [37]. The validity of my results also suggest that my model building technique can provide a methodological framework for applying decision-tree based machine learning models to estimate other public health outcomes using sparse data.

The decision to create two different versions of my model was another strength of my thesis, as it allowed me to create models fit for different policy purposes. More specifically, the model trained for country-level prediction can be used to estimate MMR for countries with sparse and/or unreliable data. This can be used for global monitoring of MMR, aid distribution, as well as national health policy planning. In contrast, the model trained to perform forecasting can be used to simulate future MMR trends and the effect of different policy scenarios on MMR, where a subset of feature variables is changed to replicate the effects of a candidate policy. Thus, this model can be used to forecast global trends as well as determine the most effective health policy for reducing MMR.

Another strength of my research is its exploration of a wide variety of socio-economic and health-related features. As discussed in detail in the previous sections, BMat and CODEm, the most widely used models employed by governments and international organisations to monitor MMR, only use a small number of covariates to predict MMR, with a limited focus on socio-economic trends [20, 23, 24]. In contrast, the use of decision-tree based models in my thesis has allowed me to incorporate the influence of a wider variety of meaningful socio-economic and health-related trends into my estimation of MMR. My wider input feature dataset also enabled a more nuanced, robust analysis of feature importance, as I could determine which of a diverse range of features had the highest predictive power for MMR. This type of a feature analysis could be conducted due to my use of interpretable machine learning methods. This use and analysis of a wide range of features was a major strength of my research, especially given that features identified with high predictive power can then be used to inform health policy to target the socio-economic and health-related trends that would result in the highest decreases in MMR. The features identified as having the predictive power for MMR in my thesis were established risk factors, as discussed earlier, corroborating the relevance and accuracy of my results. The validity of my results was further reinforced by my SHAP analysis, with this type of robustness check being a further strength of my research.

Additionally, my model used features from a wide variety of component datasets, which were collated and combined by the WHO and World Bank [38, 39, 40, 41, 42, 43]. These datasets were initially gathered from sources like Demographic and Health Surveys (DHS), which are mainly conducted in low-income countries, as well as derived from various UN Inter-Agency Groups [39, 43]. Using these pre-processed estimates from these reputable sources increased the quality of the input datasets used by in this research.

More generally, my thesis was strengthened by its exploration of a wide variety of pre-processing techniques. A range of model types, hyperparameter settings, feature subsets, and missing data removal techniques were systematically tested and compared, providing comprehensive set of experiments justifying all methodological decisions.

Finally, a strength of my research was its lack of assumptions about the data generation process, 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 in the literature [20, 23, 24]. This prevented incorrect assumptions from reducing the validity of my MMR estimates. In large part, this strength was due to my use of decision-tree based models, which could effectively handle missing data without making assumptions about the data’s distribution.

### 6.6 Limitations and Future Extensions of this Research

Despite the strengths of my research, there were several important limitations of my thesis that must be discussed. These limitations motivated potential future extensions of this work.

The primary limitation of this thesis was its use of sparse and low-quality data, especially from low-income countries and countries without robust data collection systems [29]. Even countries with comprehensive data reporting systems suffer from underreporting and misclassification of maternal deaths, as discussed in detail in this thesis [28]. As a result, all models that estimate maternal mortality add caveats to their results about how low-quality input data likely introduced inaccuracy and bias into their results [20, 23, 37]. To address this limitation, an interesting future extension of this work could be to develop a secondary model whose sole purpose is to predict adjustment factors for maternal mortality ratio predictions given the extent of estimates underreporting and misclassification errors in a country’s data collection systems. This extension would be modelled on the UN MMEIG’s use of the BMis model to adjust for similar errors in country’s CRVS system data and the GMatH model’s use of explicit parameters to adjust for underreporting.

This low-quality and sparse input data also resulted in many country, year datapoints lacking an associated MMR estimate. The proportion of missing MMR estimates increased as income-level decreased. As a result, I only had 78 low-income samples in my input dataset, which was only 8.8% of the available datapoints 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 to accurately train my model to be able to predict low-income MMR values, as seen by my models’ higher MSE for estimates on low-income countries. Missing MMR estimates affected samples from all income levels, with even 35% of samples from high-income countries needing to be dropped.

I propose two possible extensions to address this limitation. The first is the use of semi-supervised machine learning methods. Briefly, semi-supervised machine learning techniques use a combination of unlabelled and labelled input data, making them a mixture of supervised and unsupervised machine learning [44, 45]. More specifically, in semi-supervised learning, the model is initially trained on the labelled data [45]. The initial model is then used to predict the labels of the unlabelled samples [45]. Finally, the model is retrained on all input data (both the original and newly labelled samples) [45]. Therefore, semi-supervised methods are considered particularly useful when labelled data is limited, such as in my thesis [44]. Semi-supervised models have been used for a variety of purposes, such as to help identify patients with autoimmune diseases like rheumatoid arthritis [45]. A previous study has demonstrated that semi-supervised methods had higher accuracy than a supervised model when using magnetic resonance imaging to detect Crohn’s Disease [45].

Therefore, by following a semi-supervised approach, I could use my current model to predict the MMR of samples missing their associated MMR value and then use the newly labelled samples to retrain my model. As a result, the 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 the potential for overfitting. While there is risk of my model incorrectly estimating the MMR for the unlabelled samples, my model’s similarity to the literature estimates provides confidence in its predictions.

The second possible extension to correct for having an imbalanced dataset, with 78 samples from low-income countries versus as many as 1,405 samples from high-income, was to use synthetic minority oversampling techniques (SMOTE). The SMOTE algorithm is used to generate synthetic samples of the underrepresented data by interpolating between existing samples from the underrepresented groups [46]. SMOTE is considered one of most influential pre-processing techniques in machine learning and data mining [46]. However, SMOTE is known to generate overlapping and noisy samples, which is particularly relevant for my data, as my PCA shows that the MMR estimates for the upper-middle, lower-middle, and low-income classes overlap [46]. Therefore, the neighbours of low-income samples could actually be from another income level, meaning that the sample generated from interpolating between neighbours would not be representative of low-income countries. Another limitation of SMOTE is its difficulty working with missing data and lack of 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. In particular, future work could explore potential modification to the primary SMOTE method to improve its ability to work with missing data.

In addition to missing MMR estimates, there was 80 to 90% missing feature data per year for the overwhelming majority of year recorded in my input dataset. While my earlier analysis indicated that the decision-tree based models used in my thesis could effectively work 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, which may have had a meaningful influence on the model’s results if it had been recorded. Therefore, it may be worth exploring whether imputing the missing data would improve prediction accuracy. For example, Twala (2009) found that building decision trees to impute the values of each missing feature had high performance, especially when correlation between features was high [14]. While this would likely be computationally intensive given the number of features in my dataset, and that Twala’s (2009) analysis only investigated datasets with missing data proportions of up to 50%, their observations could be used as the foundation for a future extension of this thesis. However, implementation of this potential extension must be cautioned, as imputing this large amount of missing data is likely to introduce bias into the input data, especially given the likelihood of the pattern of missing data being missing not at random.

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 (DHS), which generally did not collect information from high-income countries [46]. Additionally, high correlation between features made importance scores unstable across different instances of my model. These limitations could reduce the accuracy of my determination of the features with the highest predictive power. As a final point, due to time constraints, the features with the highest predictive were only identified in a small subset of base estimators. There may have been differences substantial feature differences between models, again 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, potentially indicating that these limitations only had a small effect on the accuracy of my results.

Another limitation of my thesis is the differences between the distribution of ground truth MMR values between its train/validation and test sets, which prevent the model being trained on data that represents the out-of-sample test set. In the future, this could be addressed by implementing additional checks after splitting the data into train/validation and test sets. Additionally, the train/test split could be increased to 75:25 to improve the probability of the test set better representing the entire train set. Many of the techniques discussed previously could also help address this limitation. Increasing the number of samples in the input data, especially of low-income countries, would reduce the chance of a very small number of datapoints representing certain MMR outcomes, which means samples representing all parts of the ground truth MMR distribution would have a greater likelihood of being placed in both the train/validation and test sets.

A further limitation of my methodology was its inability to characterise uncertainty in its estimates, unlike the BMat, CODEm, and GMatH models [20, 23, 24]. The Random Forest Stacking Ensembles could have been retrained 1,000 times while measuring the differences across their predictions. While this method could not be implemented in my thesis due to limited computation resources, it could be an interesting future extension of this work. Alternatively, a future extension of this thesis could be to conduct a thorough literature review to identify a robust method for determining uncertainty in the estimates produced by decision-tree based models. Such uncertainty measures would give policy makers more information about global and national MMR trends.

Another extension of this thesis is to determine its ability to predict sub-national MMR values. This would be a particularly important extension given that the widely used BMat model only provides country-level data at its finest granularity despite low-quality and sparse maternal mortality data in many countries as well as sub-national heterogeneity. The ability of my model 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 perform this function could be measured by comparing its performance to the GMatH microsimulation model’s simulated outcomes, as this model can also used to evaluate potential policies [2].

## 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. In contrast, comprehensive literature review found that the most widely used MMR modelling approaches used Bayesian hierarchical regression models and classic machine learning techniques with smaller feature subsets.

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 low-quality and sparse input data, the MMR estimates produced by both models were similar to those generated by the BMat, CODEm, and GMatH models, with this similarity validating the accuracy of my model. However, my model estimates were generally smaller when predicting the MMR values of high-income countries, potentially due to underestimation of MMR in my ground truth MMR dataset. Differences between these models’ estimates and the MMR estimates produced in this thesis were also attributed to models’ different choices of covariates and features, treatment of missing data, and use of smoothing.

I used my models to determine that the level and type of women’s employment, women’s knowledge of contraceptive options, women’s nutritional status, the country’s income level, the percentage of teenage mothers, and the proportion of births attended by a skilled medical practitioner within a country were socio-economic variables with high predictive power for maternal mortality. Features that benchmarked the country’s health, like fertility rates and national life expectancy, also had high predictive power. These results were existing, known risk factors for maternal mortality, reinforcing their validity. They highlighted the importance of socio-economic trends driving MMR. Consequently, I suggest that investment in women’s education, which influences their employment prospects as well as 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/