

Merged data

Sourced from the World Bank’s Gender Data Portal and the WHO’s Health Inequality Data Repository.

721 features

2,789 samples

1985-2018

172 countries

Split data into train, validation, and test sets

Data split into train: test sets (90:10) for predictive or missing data analysis.

Training set split using 5-fold cross validation (80:20).





**Predictive analysis:**

Train/validation 1985 -> 2014 Test: 2015 -> 2018.

**Missing data analysis:**

All data from the same country is either in the train, validation or test set.

Feature selection

5 versions of each fold to test feature selection techniques.

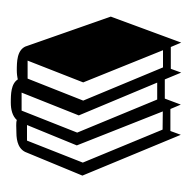


Feature selected if pairwise correlation with MMR is:

Selection via literature review

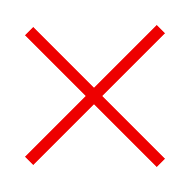
No feature selection

>= 0.6



>= 0.7

>= 0.8



n=11

n=45

n=113

n=40

n=720

Missing data removal

Row and columns removed if they have a proportion of missing data >= threshold, producing 4 versions of each dataset version.

95% threshold

90% threshold

85% threshold

No removal

100 versions of the dataset for each of predictive and missing data analysis:

5 x cross-validation folds x 5 feature selection methods x 4 missing data thresholds

Training base estimators

Random Forest, LightGBM, and XGBoost trained on each fold.

Hyperparameter tuning 1,000 Optuna trials.

Random Forest Stacking

Elastic Net Stacking

Support Vector Machine Stacking

Voting

Combine the 300 predictions (100 per model) using:

Training ensemble models

Comparison to literature

Sensitivity analysis

2,789 samples

721 features

172 countries

1985-2018

Sourced from the World Bank’s Gender Data Portal and the WHO’s Health Inequality Data Repository.

Merged data

Data split into train: test sets (90:10) for predictive and missing data analysis.



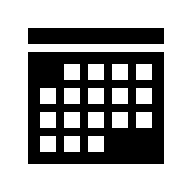
**Predictive analysis:**

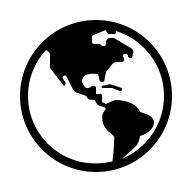
Train/validation 1985 -> 2014

Test: 2015 -> 2018.

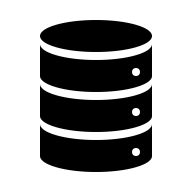
**Missing data analysis:**

All data from the same country is either in the train, validation or test set.





Training sets split using 5-fold cross validation (80:20).

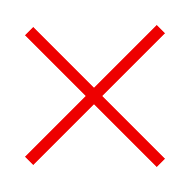
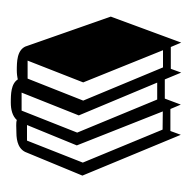


5 versions of each fold to test feature selection techniques.

Selection via literature review

Feature selected if pairwise correlation with MMR is:

No feature selection



>= 0.6

>= 0.7

>= 0.8



k

n=720

n=40

n=11

n=45

n=113

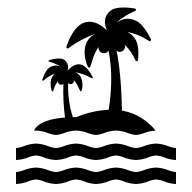
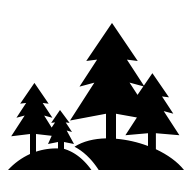
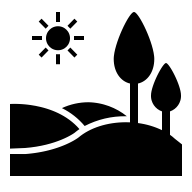
Row and columns removed if they have a proportion of missing data >= threshold, producing 4 versions of each dataset version.

85% missing data threshold

No missing data removal

95% missing data threshold

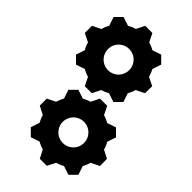
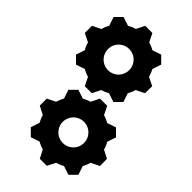
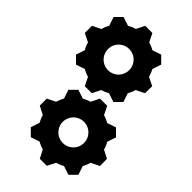
90% missing data threshold



Random Forest

XGBoost

LightGBM



Voting Ensemble

SVM Stacking Ensemble

Elastic Net Stacking Ensemble

Random Forest Stacking Ensemble

Ensemble Model

|  |  |  |  |
| --- | --- | --- | --- |
| **Data from Low Income Countries** | **Data from Lower-Middle Income Countries** | **Data from Upper-Middle Income Countries** | **Data from High Income Countries** |
| Training data | Training data | Training data | Training data |
| Testing data | Testing data | Testing data | Testing data |

**Testing data**

**Training data**

1985

2018

2014

2015

|  |
| --- |
| **Data from High Income Countries** |
| Training data |
| Testing data |

Ensemble model trains on these predictions and is thus fit solely on high-income data.

Pre-processing techniques create 100 versions of this high-income dataset, which are used to train the 300 base predictors.

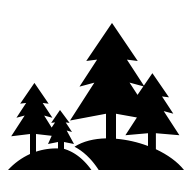
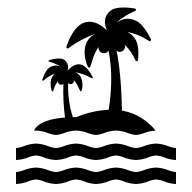
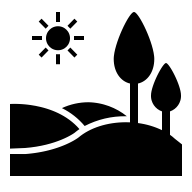
Ensemble Model:

Combines 300 predictions from base estimators

100 Predictions

100 Predictions

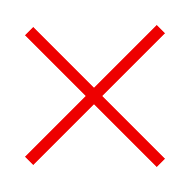
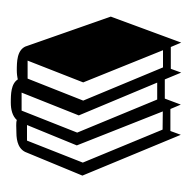
100 Predictions



Random Forest

XGBoost

LightGBM



300 MMR Predictions

1

1

1

1

1

Fold 0

Fold 4

Fold 3

Fold 2

Fold 1

5

5

5

5

No Missing Data Threshold

Missing Data Threshold = 95%

Missing Data Threshold = 90%

Missing Data Threshold = 85%

20

20

20

20

20

Correlation >0.8

Correlation >0.7

Correlation >0.6

Literature

No Feature Selection



Final MMR Prediction

No Missing Data Threshold

Missing Data Threshold = 95%

Missing Data Threshold = 90%

Missing Data Threshold = 85%

1

1

1

1

5

5

5

5

5

Correlation >= 0.8

Correlation >= 0.7

Correlation >= 0.6

Literature

No Feature Selection

25

25

25

25

25

Fold 4

Fold 3

Fold 2

Fold 1

Fold 0

XGBoost

LightGBM

Random Forest



**Final MMR value**

**Ensemble Model:**

Combines the predictions from 300 base estimators

100

100

100

**Ensemble Model:**

Combines the predictions from 300 base estimators

100

100

100

Random Forest

LightGBM

XGBoost

20

20

20

20

20

Fold 0

Fold 1

Fold 2

Fold 3

Fold 4

4

4

4

4

4

Correlation >= 0.8

Correlation >= 0.7

Correlation >= 0.6

Literature

No Feature Selection

1

1

1

1

No Missing Data Threshold

Missing Data Threshold = 95%

Missing Data Threshold = 90%

Missing Data Threshold = 85%