The Impact of Global Warming Statistics on Child Mortality - (STAT2020 Assignment 4)

c3329700

28/10/2021

Abstract

The study examined the effects of various global warming statistics on child mortality. The purpose of the study was to identify potential factors impacting on child mortality rates and examine the degree to which the factors did so. Through studying these factors, the report intends to outline areas in which improvement will provide the greatest reduction in child mortality. The World Bank Climate Change data was visually inspected and cleaned before potential predictor variables were identified. Multiple linear regression was consequently performed using these predictor variables. Lasso Regression and Random Forest modelling were then used to examine the magnitude of impact that each predictor variable has on the response. From the study, we see that Electricity Access, Primary School Completion Rate, CO2 Emissions and Population were identified as having a statistically significant relationship with Under-5 Mortality. Furthermore, from the variety of models utilised it was consistently displayed that Electricity Access and Primary School Completion Rate have the greatest importance in predicting child mortality. From these findings, the study was able to recommend further analysis into the identified variables, with consultation from financial experts in order analysis which variables will see the greatest improvement relative to cost. These recommendations are made to ensure the best allocation of funds and resources which will in turn result in the greatest reduction of Child Mortality.

Introduction

Child mortality, also known as under-five mortality rate, is the mortality rate of children under the age of five. This rate signifies the probability of a child dying between birth and the age of 5, represented per 1,000 births.[1] The United Nations' Sustainable Development Goals outline reductions in child mortality in a multitude of the goals, [2] such as to, "by 2030, end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce under 5 mortality to at least as low as 25 per 1,000 live births." [3] By studying the data by The World Bank, we see that child mortality has been consistently decreasing since 1990, going from a global average of 93 per 1,000 to 37.7 per 1,000 in 2019. [4] However, in order to achieve these United Nations goals further improvements must be made. Hence, the objectives of this report are as follows:

1. Identify potential factors impacting on child mortality rates
2. Examine the degree of impact of potential factors
3. Outline areas for which improvement will provide the greater reduction in child mortality

In achieving these objectives, the report will aid in the efforts to further decrease child mortality and in doing so will hopefully assist in the achievement of the United Nations Sustainable Development Goals.

Data

The climate change data was obtained from The World Bank. This group provides free and open access data on global development with the aim of "working for sustainable solutions that reduce poverty and build shared prosperity in developing countries." [5] Then from inspection of the data set and the provided description of each variable, a subset of the data was produced to include the child mortality data as well as potential applicable predictor variables. These variables include:

- SH.DYN.MORT = Under-5 Mortality (per 1,000)
- EG.ELC.ACCS.ZS = Access to Electricity (%)
- EN.ATM.CO2E.KT = CO2 Emissions (kt)
- EN.ATM.GHGO.KT.CE = Other Greenhouse Emissions (kt)
- EN.CLC.DRSK.XQ = Disaster Risk Reduction Progress Score (1-5 scale)
- EN.CLC.MDAT.ZS = Droughts, Floods, Extreme Temperatures (%)
- SE.PRM.CMPT.ZS = Primary School Completion Rate (%)
- SH.MED.CMHW.P3 = Community health workers (per 1,000 people)
- SH.STA.MALN.ZS = Prevalence of underweight, weight for age (% of children under 5)
- SP.POP.TOTL = Population, total

This subset of the data was then pre-processed to remove variables for which there was too much missing data. The cut-off was chosen to be 25%, meaning columns that had over ~54 NA values were removed from the analysis. After this pre-processing the following variables comprised the data set:

- SH.DYN.MORT = Under-5 Mortality (per 1,000)
- EG.ELC.ACCS.ZS = Access to Electricity (%)
- EN.ATM.CO2E.KT = CO2 Emissions (kt)
- EN.ATM.GHGO.KT.CE = Other Greenhouse Emissions (kt)
- EN.CLC.MDAT.ZS = Droughts, Floods, Extreme Temperatures (%)
- SE.PRM.CMPT.ZS = Primary School Completion Rate (%)
- SP.POP.TOTL = Population, total

The data was then cleaned to omit NA values to aid in the exploratory analysis. A subset of the data was then produced in which the predictor variables were normalised to have mean 0 and standard deviation 1. This was done to negate the impacts that the difference between the various scales (i.e. per 1000, %, kt, etc.) would have upon the analysis.

Methods

Data Pre-Processing

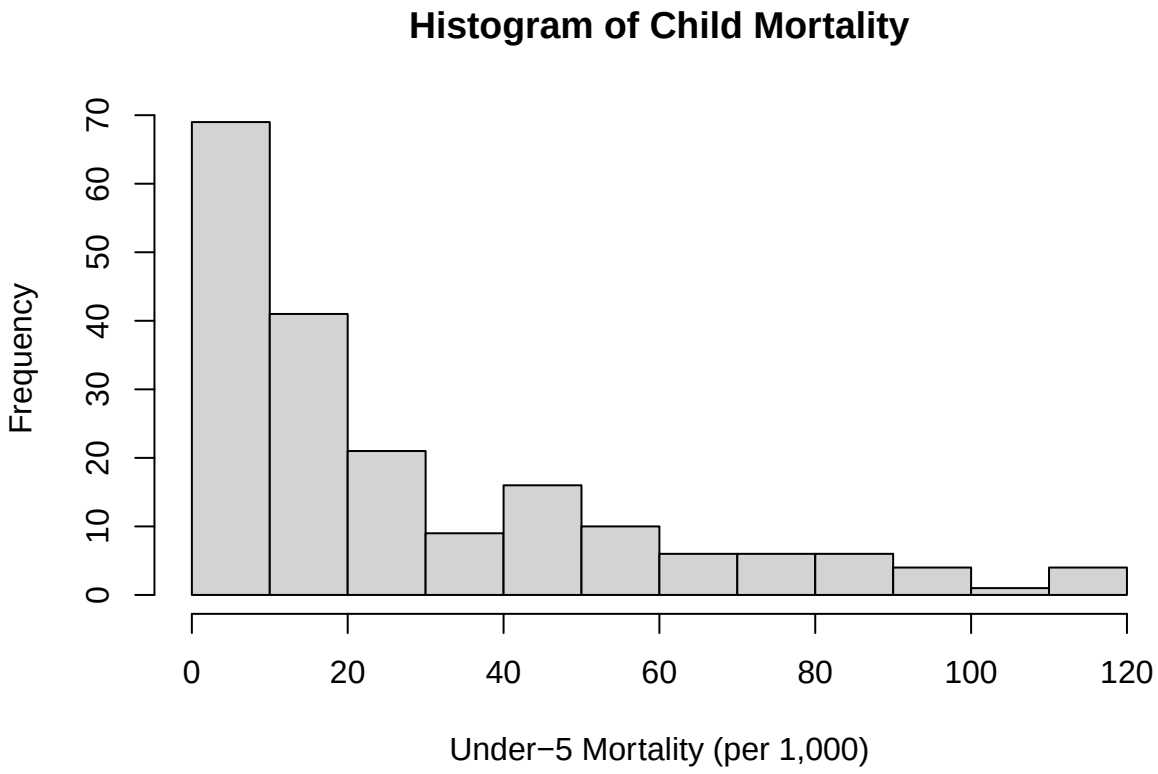
```
##  [1] "EG.ELC.ACCS.ZS : NA values = 1"     "EN.ATM.CO2E.KT : NA values = 26"
##  [3] "EN.ATM.GHGO.KT.CE : NA values = 33" "EN.CLC.DRSK.XQ : NA values = 134"
##  [5] "EN.CLC.MDAT.ZS : NA values = 49"    "SE.PRM.CMPT.ZS : NA values = 28"
##  [7] "SH.MED.CMHW.P3 : NA values = 157"   "SH.STA.MALN.ZS : NA values = 67"
##  [9] "SP.POP.TOTL : NA values = 0"        "SH.DYN.MORT : NA values = 24"
```

We use the above information to remove variables that do not have enough information for valid exploration (Variables with 25% or more entries being NA). Hence we remove the variables: EN.CLC.DRSK.XQ, SH.MED.CMHW.P3, SH.STA.MALN.ZS from the subset of selected variables.

Initial Exploration

Produce a histogram to visualise Child Mortality rates.

```
hist(wbcc$SH.DYN.MORT, main='Histogram of Child Mortality', xlab='Under-5 Mortality (per 1,000)')
```



From the histogram of Under-5 Mortality, we see that the distribution is heavily right skewed, meaning most observations have a low value for child mortality. However, one of the goal of the United Nations is to reduce child mortality so that no country has a rate greater than 25 per 1000. Clearly, a significant proportion of countries still have child mortality rates greater than this cut-off, meaning progress is still needing to be made.

Multiple Linear Regression

Normalise the predictor variables identified from pre-processing to have standard deviation 1 and mean 0 using the scale() function. Then perform multiple linear regression using these predictors against the response variable, Under-5 Mortality.

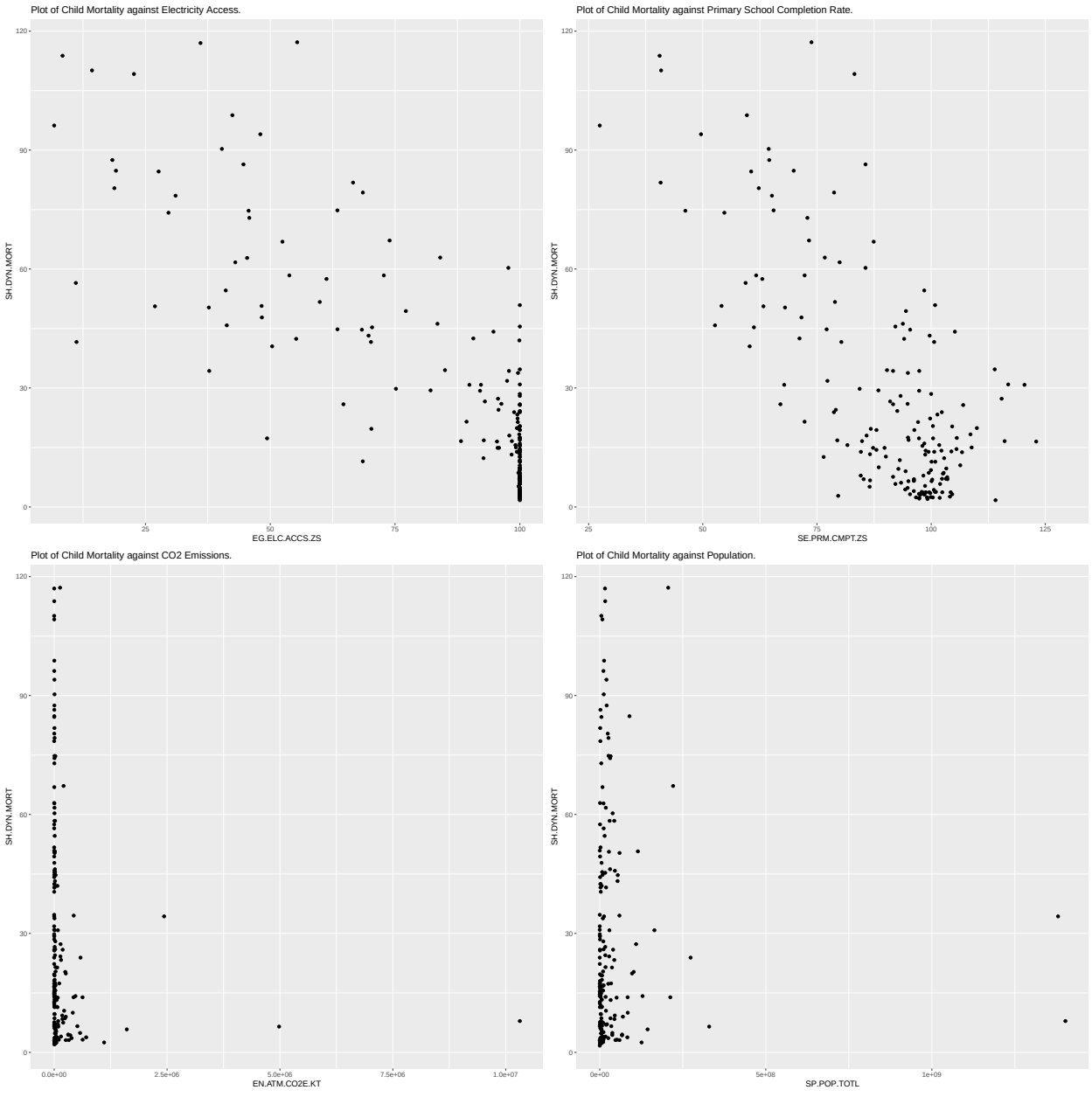
```
#normalise selected predictors
pred.new <- c('EG.ELC.ACCS.ZS','EN.ATM.CO2E.KT','EN.ATM.GHGO.KT.CE','EN.CLC.MDAT.ZS',
              'SE.PRM.CMPT.ZS','SP.POP.TOTL')
predictors <- wbcc[ ,pred.new]
pred.norm <- scale(predictors)
dat_pred.norm <- cbind(pred.norm,wbcc[,"SH.DYN.MORT"])
colnames(dat_pred.norm) <- c('EG.ELC.ACCS.ZS','EN.ATM.CO2E.KT','EN.ATM.GHGO.KT.CE',
                             'EN.CLC.MDAT.ZS','SE.PRM.CMPT.ZS','SP.POP.TOTL',
                             'SH.DYN.MORT')
dat_pred.norm <- as.data.frame(dat_pred.norm)

#perform linear regression with normalised predictors
lm.wbcc.norm <- lm(SH.DYN.MORT ~ ., data = dat_pred.norm)
```

Visualisation

Produce scatter plots to visualise the relationships between the predictors variables identified as having statistically significant relationships with the response variable.

```
#Use gridExtra package to create a 2x2 subplot grid of each variable against child mortality
grid.arrange(elec.plot,school.plot,co2.plot,pop.plot,ncol=2)
```



By plotting the identified predictors from multiple linear regression against Child Mortality, we see a clear negative relationship between both Electricity Access and Primary School Completion Rate against Child Mortality. Meaning, as Electricity Access or Primary School Completion Rate increases, Child Mortality trends to decrease. There is less visual clarity from the CO2 Emissions and Population plots, hence further investigation is required.

Lasso Regression

From the glmnet package, perform variable selection using the LASSO.

```
#perform variable selection using lasso
library(glmnet)
#remove NA entries
dat.lasso <- na.omit(dat_pred.norm)
X <- model.matrix(SH.DYN.MORT ~ ., data = dat.lasso)
#remove the first column
X <- X[,-1]
Y <- dat.lasso$SH.DYN.MORT
lasso.wbcc = glmnet(X,Y,alpha=1)
```

Perform 10 fold cross-validation with the LASSO, using the cv.glmnet() function from the glmnet package.

```
#perform cross-validation on lasso
lasso.cv <- cv.glmnet(X,Y,alpha=1)
```

Random Forest

From the randomForest package produce a random forest model with identified predictors against the response variable, Under-5 Mortality.

```
set.seed(0)
#produce random forest model
rf.wbcc <- randomForest(SH.DYN.MORT ~ ., data = dat.rf, ntree=100)
```

Results

Multiple Linear Regression

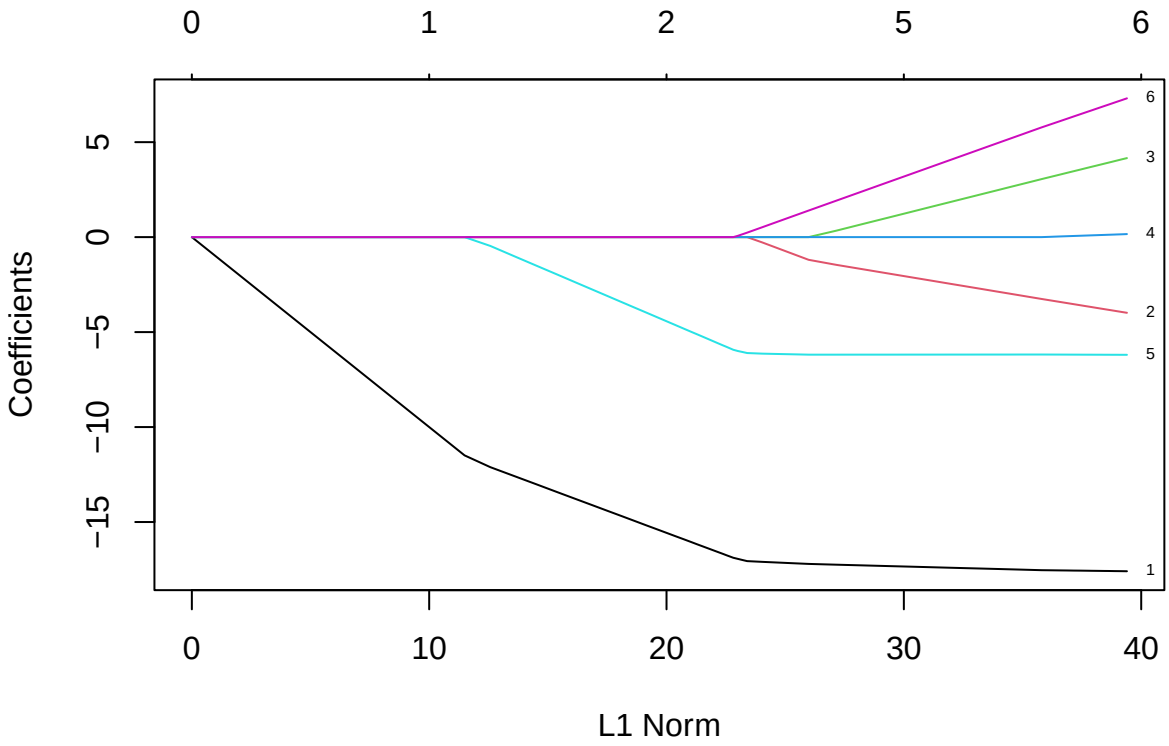
```
#produce summary of multiple linear regression model
summary(lm.wbcc.norm)
```

```
##
## Call:
## lm(formula = SH.DYN.MORT ~ ., data = dat_pred.norm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.337  -8.808  -2.446   5.810  48.678
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        25.9514     1.2639  20.532  < 2e-16 ***
## EG.ELC.ACCS.ZS    -17.6050     1.8107  -9.723  < 2e-16 ***
## EN.ATM.CO2E.KT     -4.1510     1.8810  -2.207  0.02894 *
## EN.ATM.GHGO.KT.CE   4.4208     2.2818   1.937  0.05469 .
## EN.CLC.MDAT.ZS      0.1923     1.3165   0.146  0.88405
## SE.PRM.CMPT.ZS     -6.1969     2.0002  -3.098  0.00235 **
## SP.POP.TOTL         7.6600     2.5453   3.010  0.00310 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.95 on 141 degrees of freedom
##   (69 observations deleted due to missingness)
## Multiple R-squared:  0.7335, Adjusted R-squared:  0.7222
## F-statistic: 64.69 on 6 and 141 DF,  p-value: < 2.2e-16
```

From the multiple linear regression summary, we see that EG.ELC.ACCS.ZS (Access to electricity), EN.ATM.CO2E.KT (CO2 emissions), SE.PRM.CMPT.ZS (Primary School Completion Rate) and SP.POP.TOTL (Population) all have a statistically significant relationship with SH.DYN.MORT (Under-5 Mortality) at the 5% significance level. The p-value of 0.05469 from EN.ATM.GHGO.KT.CE (Other Greenhouse Emissions) is marginally outside the 5% significance level, however we do not reject the null hypothesis due to the set 5% significance level and hence cannot conclude a statistically significant relationship.

Lasso Regression

```
#plot lasso regression
plot(lasso.wbcc,label=TRUE)
```



The Lasso plot shows the effects of the regularisation term increasing right to left from zero. As the regularisation term increases variable coefficient shrink to zero, the order in which these coefficients shrink to zero can interpreted as an inverse of the variables importance. From the Lasso regression plot, we see that the variables shrink to zero in the following order:

1. Droughts, Floods, Extreme Temperatures
2. Other Greenhouse Emissions
3. CO2 Emissions
4. Population
5. Primary School Completion Rate
6. Electricity Access

```
#examine cross-validated lasso
coef(lasso.cv)
```

```
## 7 x 1 sparse Matrix of class "dgCMatrix"
##                             s1
## (Intercept)         27.404294
## EG.ELC.ACCS.ZS     -14.121686
## EN.ATM.CO2E.KT       .
## EN.ATM.GHGO.KT.CE    .
## EN.CLC.MDAT.ZS       .
## SE.PRM.CMPT.ZS      -2.769588
## SP.POP.TOTL          .
```

From the 10 fold cross-validation for the Lasso, we see that Electricity Access and Primary School Completion

Rate are selected by the model. The co-efficients for variables are negative, indicating that an increase in either variable corresponds to a decrease in Child Mortality.

Random Forest

```
#examine importance from random forest
importance(rf.wbcc)
```

```
##                    IncNodePurity
## EG.ELC.ACCS.ZS         44800.534
## EN.ATM.CO2E.KT         14388.173
## EN.ATM.GHGO.KT.CE       6760.230
## EN.CLC.MDAT.ZS         10939.830
## SE.PRM.CMPT.ZS         31327.135
## SP.POP.TOTL             5838.771
```

The importance() function returns a measure of Gini-based importance. This calculation is made based on the reduction in sum of square errors whenever a variable is chosen to split. [6] By studying the importance from the random forest, we see that the variables are ranked in the following order:

1. Electricity Access
2. Primary School Completion Rate
3. CO2 Emissions
4. Droughts, Floods, Extreme Temperatures
5. Other Greenhouse Emissions
6. Population

Discussion

Electricity Access and Primary School Completion Rate were identified as being the most important variables in both the Random Forest model and the Lasso regression (including cross-validation). These variables along with CO2 Emissions and Population are identified as having a statistically significant relationship with Under-5 Mortality in the multiple linear regression summary. These results are interesting as they clearly confirm the initial relationship identified from the visualisations. Identifying these variables is important, as it allows for recommendations to be made in regards to areas in which changes will have greatest impact. Due to this, we disregard Population when making recommendations as it is not an area in which targeted improvement makes sense. Hence, we are left with Electricity Access, Primary School Completion Rate and CO2 Emissions as the top 3 applicable predictors in regards to Under-5 Mortality. From the results, we see that if all three areas require equal resources for change then improvements in Electricity Access followed by improvements in Primary School Completion Rate would result in the greatest reduction of Child Mortality. However, multidisciplinary collaboration, such as consultations with experts from different fields, would be required in order to develop an action plan which allocates funds to improve on these areas in the most effective way.

Conclusions

From the study, the following predictors have been identified as having a statistically significant relationship with Under-5 Mortality: Electricity Access, Primary School Completion Rate, CO2 Emissions and Population. Further Analysis utilised a range of models, such as; LASSO regression, LASSO cross-validation and Random Forest models. These models all consistently displayed that Electricity Access and Primary School have the greatest importance in predicting child mortality.

The study recommends further analysis into the identified variables (Electricity Access, Primary School Completion Rate and CO2 Emissions) in conjunction with consultation from financial experts. This consultation will allow for the identification of which variables can see the greatest improvement relative to the cost and ease of the improvement. In doing so, action plans can be formed and initiatives implemented to ensure the best allocation of funds and resources which will in turn result in the greatest reduction of Child Mortality.

The study was limited by the degree of missing data present within the data set. Data cleaning and preprocessing methods (such as: variable selection criteria and NA omits) were implemented to reduce the impact of the missing data as much as possible. In addition, the recommendations of the study were limited by a lack of information in regards to the cost that implementations or improvements would require.

References

1. Our World in Data. 2021. Child mortality. [online] Available at: https://ourworldindata.org/grapher/child-mortality [Accessed 18 October 2021].
2. Un.org. 2021. Transforming our world: the 2030 Agenda for Sustainable Development. [online] Available at: https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E [Accessed 18 October 2021].
3. United Nations Sustainable Development. 2021. Health. [online] Available at: https://www.un.org/sustainabledevelopment/health/ [Accessed 19 October 2021].
4. Data.worldbank.org. 2021. Mortality rate, under-5 (per 1,000 live births) | Data. [online] Available at: https://data.worldbank.org/indicator/SH.DYN.MORT?end=2019&start=1960&type=shaded&view=chart [Accessed 19 October 2021].
5. Data.worldbank.org. 2021. World Bank Open Data | Data. [online] Available at: https://data.worldbank.org/ [Accessed 21 October 2021].
6. Displayr. 2021. How is Variable Importance Calculated for a Random Forest?. [online] Available at: https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest/ [Accessed 23 October 2021].

Appendix: World Bank Indicators

Indicator ID	Indicator	Description
SH.DYN.MORT	Under-5 Mortality	Under-five mortality rate is the probability per 1,000 that a newborn baby will die before reaching age five, if subject to age-specific mortality rates of the specified year.
EG.ELC.ACCS.ZS	Access to Electricity	Access to electricity is the percentage of population with access to electricity. Electrification data are collected from industry, national surveys and international sources.
EN.ATM.CO2E.KT	CO2 Emissions	Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring.
EN.ATM.GHGO.KT.CE	Other Greenhouse Emissions	Other greenhouse gas emissions are by-product emissions of hydrofluorocarbons, perfluorocarbons, and sulfur hexafluoride.
EN.CLC.DRSK.XQ	Disaster Risk Reduction Progress Score	Disaster risk reduction progress score is an average of self-assessment scores, ranging from 1 to 5, submitted by countries under Priority 1 of the Hyogo Framework National Progress Reports. The Hyogo Framework is a global blueprint for disaster risk reduction efforts that was adopted by 168 countries in 2005. Assessments of "Priority 1" include four indicators that reflect the degree to which countries have prioritized disaster risk reduction and the strengthening of relevant institutions.
EN.CLC.MDAT.ZS	Droughts, Floods, Extreme Temperatures	Droughts, floods and extreme temperatures is the annual average percentage of the population that is affected by natural disasters classified as either droughts, floods, or extreme temperature events.
SE.PRM.CMPT.ZS	Primary School Completion Rate	Primary completion rate, or gross intake ratio to the last grade of primary education, is the number of new entrants (enrollments minus repeaters) in the last grade of primary education, regardless of age, divided by the population at the entrance age for the last grade of primary education. Data limitations preclude adjusting for students who drop out during the final year of primary education.
SH.MED.CMHW.P3	Community health workers	Community health workers include various types of community health aides, many with country-specific occupational titles such as community health officers, community health-education workers, family health workers, lady health visitors and health extension package workers.
SH.STA.MALN.ZS	Prevalence of underweight, weight for age	Prevalence of underweight children is the percentage of children under age 5 whose weight for age is more than two standard deviations below the median for the international reference population ages 0-59 months. The data are based on the WHO's child growth standards released in 2006.
SP.POP.TOTL	Population, total	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.