NANYANG TECHNOLOGICAL UNIVERSITY
NANYANG BUSINESS SCHOOL

NANYANG TECHNOLOGICAL UNIVERSITY

BC2406 Analytics I

Machine Learning Proof of Concept (POC) for EIU

Health Information and Resource Allocation Analyser (H.I.R.A.A)

Seminar Group: 9
Team: 7

Name	Matriculation Number
Benedict Leong Wei Xin	U1923641A
Ho Wei Ling Charmaine	U1911802F
Tan Kiang Hwee Jeremy	U2021940D
Lim Shi Min	U2010855H

Table of Contents

1 Executive Summary

In the forecasting and advisory market, EIU has severe competition. In this sector, there are countless other significant industry players, such as Oxford Economics, Forrester Research, CEIC Data, Stat Diagnostics and more, that can take over EIU's target audience. In order for EIU to maintain its strategic position, EIU must be able to deliver the best forecasting services available and those that address gaps in the market that their clients need. EIU should thus constantly improve its forecasting technique by embracing new technologies such as Machine Learning. We believe there is merit in adopting machine learning specifically in opening a new line of product and creating new sources of revenue, such as with our proposed model.

We propose that the EIU should make use of machine learning models to come up with a 'healthcare' score that can be used to rank countries' healthcare statuses. This is especially relevant for non-profit healthcare organisations, such as Red Cross. EIU can tap into this market and provide useful information for non-profits. By doing so, EIU can open up a new source of revenue by providing a ranking of countries based on their health environments. With this new product segment, EIU will be able to fill the current market gap and the non-profit organisations can rely on EIU's data to make more informed and reliable decisions with regards to their humanitarian plans.

This is because in determining which countries to offer assistance to, non-profit organisations face a lot of difficulties as there are an overwhelming number of factors to consider. The concept of demand and supply comes into play as many countries require the assistance of non-profit organisations, however, these organisations have scarce resources and they need to ponder about what would be the best and most efficient allocation of their limited resources.

Currently, these non-profit organisations rely on traditional demographic census - which is typically not very reliable given that some countries may not provide accurate data (e.g. Nigeria), and data is updated every few years or even in decades.

With the advent of Covid-19, this problem is more pressing than ever - there is a significant need for this data as there is a need to allocate medical supplies such as vaccines, testing kits and medication to countries with limited access to these resources, in a timely manner.

This report serves to describe our process of creating a Health Information Resource Allocation (H.I.R.A) score that will accurately predict which country is in urgent need of the most help in terms of medical resources. In order to predict the H.I.R.A score, we identify important indicators and use both Linear Regression and Classification and Regression Trees (CART) models to generate the score.

In this report, we elaborate in detail about the following points:

1. Identify the opportunities in machine learning
2. Highlight the specific business opportunity which EIU can tap upon
3. Identify the machine learning approaches EIU can consider adopting
4. Recommendation to EIU

2 Introduction

In the forecasting and advisory market, EIU is up against severe competition. In this sector, there are various significant industry players such as Oxford Economics, Forrester Research, CEIC Data and Stat Diagnostics that can erode EIU's market share. In order for EIU to maintain its position, EIU must deliver the best forecasting services available on the market. EIU should thus constantly improve its forecasting technique by embracing new technologies such as Machine Learning. We believe there is merit in adopting machine-learning specifically in opening a new line of product and creating new sources of revenue.

3 Opportunities in Machine Learning

Machine learning is the application of artificial intelligence (AI) and computer science which provides systems the ability to learn automatically and improve from experience through the use of data and algorithms, instead of programming it explicitly. This allows systems to imitate the way that humans learn and gradually improve its accuracy. (IBM Cloud Education,2021)

3.1 Machine Learning vs Non-Machine Learning

Currently, EIU does not use machine learning forecasting methodology, but instead a robust and complex traditional statistical forecasting method to make forecasts. Their accuracy is backed by a unique comprehension of each country and information and data that has been gathered is verified for accuracy and consistency. However, this method largely depends on individual knowledge of the country each analyst is working on, and their knowledge changes the way they interpret the data and hence the accuracy of the resulting forecast. Thus, to reduce human bias, EIU should consider implementing machine learning methodologies to improve on the accuracy of their forecast.

There are many advantages to using machine learning to make forecasts. To list a few, machine learning forecasting is able to identify patterns in data using past and historical data which would otherwise be hard to identify using traditional methods. Machine learning forecast is also especially helpful in cases where there are massive amounts of data. By training and testing the machine learning model, it is able to apply what it has learnt to a new set of data, hence being able to produce forecasts with higher accuracy and performance rates.

3.2 Machine Learning Used by Other Companies

Vodafone Group is a British multinational telecommunications company that predominantly operates services in Asia, Africa, Europe, and Oceania. 5G technology is a game-changer for the telecommunications industry. As a step of moving towards an AI-powered 5G network, Vodafone worked with Datatonic to build traffic forecasting machine learning models. This allows Vodafone to forecast the internet traffic on a network of 10000 location points. The forecast of the model is used to suggest when the network resources should be upscaled or downscaled, in order to ensure that their customers have access to their services without lags or delays, and at the same time, avoid wasting operational resources by providing more network capacity than what is

needed. Hence, this machine learning forecasting methodology helps Vodafone to use AI to enable optimal provisioning of resources on the network at a large scale and thus minimise the operation costs and energy consumption for a complex network. This example serves to show that machine learning is a tool which EIU could likewise leverage on to increase their forecasting accuracy. (Datatonic, 2021)

4 Business Opportunity

4.1 Business Problem

In determining which countries to offer assistance to, non-profit organisations face a lot of difficulty as there is an overwhelming number of factors to consider. The concept of demand and supply comes into play as many countries require the assistance of non-profit organisations. However, these organisations have limited resources as they rely on donations, and as a consequence, they need to make decisions about how to allocate their limited resources.

Currently, these non-profit organisations rely on traditional demographic census - which is typically not very reliable given that 1. some countries may not provide accurate data and 2. data is updated every few years or even in decades.

With the advent of Covid-19, this problem is more pressing than ever - there is a huge emphasis and importance placed on these data as there is a need to allocate medical supplies such as vaccines, testing kits and medication to countries with limited access to these resources, in a timely manner.

4.2 Opportunity for EIU

We believe that EIU can tap into this new market and provide useful information for non-profit organisations - specifically healthcare charity organisations such as Red Cross.

Granted, the EIU currently does provide general rankings on countries' business and economic environments, however, these are not particularly indicative of the countries' need for humanitarian help. A country could have a good business outlook, but may not be sufficiently advanced in their healthcare systems.

EIU can open up a new source of revenue by providing a ranking of countries based on their health environments. By providing this data, EIU will be able to fill the current market gap and non-profit organisations can rely on EIU's data to make more informed and reliable decisions with regards to their humanitarian plans.

By expanding the scope of the reports they provide from their current business-centric reports to one that includes healthcare considerations, this would open up a new target market for EIU to tap into and expand their clientele. Additionally, EIU would be able to profit from a new revenue source and this would give EIU a competitive edge over their competitors.

5 Proposed Approach

5.1 The Health Information and Resource Allocation (H.I.R.A) Score

We propose to use machine learning to derive a score that is reflective of a country's healthcare environment.

The first step was to come up with a score that encompassed the socioeconomic and health care aspects of a country. Since we were not very well versed with what factors would have an effect, we had to do research on it and eventually came up with a subset of indicators that are often used in healthcare studies, and those we thought would have an effect on our indicator.

This is because when performing machine learning, we have to train the model first, showing the program which inputs results in which outputs. Hence, we needed to have existing data for the outputs (ie. the score), and we could not pull the data for the outputs out of thin air. Thus, we needed to take a subset of existing data from reputable sources and combine them to come up with a score that would meet our needs and be suitable for use.

We will use the infant mortality, or neonatal mortality rate, for each country as the scoring and ranking indicator. Various hygiene, political, socio-economical, technological, and cultural factors will be taken into consideration to predict this score, which will be achieved through analytics and machine learning techniques. This is because the rate of infant mortality is a general indicator that is used in the healthcare industry, and can be pretty reflective of a country's current demographic situation.

We will further detail why we chose the other indicators in the below section.

5.2 The Indicators Affecting the Score

Our methodology for deciding the indicators was to source for a large number of indicators that we thought could be relevant, and were sufficiently indicative of a country's healthcare situation. Subsequently through either linear regression analysis or CART analysis, we would identify the indicators that were statistically significant in determining the score, and remove the ones that were not significant, to end up with our final model.

The table below shows the reason for using the factors we selected.

Detailed descriptions of the indicators, their data type, values and references can be found in the data dictionary.

General factors selected	
Factor	**Reason For Including Factor(s)**
Annual GDP Growth	As a broad measure of overall domestic production, it functions as a comprehensive scorecard of a given country's economic health. This will tell us how a country is performing economically, a factor that could affect spending on healthcare.
Annual PM2.5 Exposure	Fine particulate matter (PM2.5) is the air pollutant that poses the greatest risk to health globally, affecting more people than any other pollutant. Since chronic exposure to PM2.5 considerably increases the risk of respiratory and cardiovascular diseases, it is a factor that can show a country's extent of air pollution, and hence, possible healthcare issues.
DPT Immunisation and Measles Immunisation	DPT (Diphtheria, pertussis, tetanus) and Measles vaccines can prevent the spread of these diseases. - Diphtheria (D) can lead to difficulty breathing, heart failure, paralysis, or death. - Tetanus (T) causes painful stiffening of the muscles. Tetanus can lead to serious health problems, including being unable to open the mouth, having trouble swallowing and breathing, or death. - Pertussis (P), also known as "whooping cough," can cause uncontrollable, violent coughing that makes it hard to breathe, eat, or drink. Pertussis can be extremely serious especially in babies and young children, causing pneumonia, convulsions, brain damage, or death. In teens and adults, it can cause weight loss, loss of bladder control, passing out, and rib fractures from severe coughing. - Measles is a serious respiratory disease in the lungs and breathing tubes, which can be especially threatening for babies and young children. As can be seen, having a high immunisation coverage will prevent a high prevalence of these serious illnesses in a country, and a high immunisation coverage also implies a somewhat robust healthcare administration.
Gini Coefficient	Gini coefficient (Gini index or Gini ratio) is a statistical measure of economic inequality in a population. The ratio measures the dispersion of income or distribution of wealth among the members of a population. It can take any value between 0 to 1(or 0% to 100%). A value of 0 indicates a perfectly equal distribution of income within a population. A coefficient of 1 represents a perfect inequality, when one person receives all the

	income, while other people earn nothing. Hence, we can account for income disparities using this factor. However, the Gini index is prone to systematic and random data errors, and inaccurate data can distort the validity of the coefficient, thus we cannot only use the Gini index.
Neonatal Mortality Rate	The score that we are trying to predict, and will use for our rankings. This is because the infant mortality rate is an indicator that is regularly used in the healthcare industry, studies and research, and can shed light on a country's current demographic situation.
Population Living in Slums	A high population living in slums would indicate a large income gap. The living conditions of slums are also not very hygienic, indirectly resulting in poor health of the residents.
Prevalence of Undernourishment	A high prevalence of undernourishment indicates that a large proportion of the population are unable to obtain sufficient food and nutrients. This would lead to severe health problems further down in the future.
Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)	A high % of causes of death from diseases, nutrition conditions etc can suggest that the country does not have a particularly robust healthcare system, and hence requires assistance in that sector.
Political Stability	Index that captures the likelihood of government destabilisation through unconstitutional or violent means. Political instability might cause a country's social and healthcare situation to deteriorate significantly.
Gender Gap	Gender gap is the difference between women and men as reflected in social, political, intellectual, cultural, or economic attainments or attitudes. Firstly, male babies have a lower chance of surviving infancy due to biological factors. In addition, certain cultures have a strong preference for male babies, so they may provide less help to the other 'unwanted' babies.

6 Machine Learning Methodology

We identify 2 potential machine learning models we could adopt, namely:

1. Linear Regression
2. CART

Before we can start working on the machine learning models, we need to perform data cleaning.

6.1 Data cleaning

The original datasets were structured in the same format, with rows indicating country and columns indicating year. The entries of the dataset were the values of the indicators.

For each dataset, NA values found in each row were replaced by the mean of available values of the row. After which, only values from 2010 to 2020 were selected to be part of the final dataset.

Data for the years 2010 to 2020 were chosen for two reasons. First, the most recent and complete data sources available are up until 2020. Second, 2010 was a significant year in which the UN's Millennium Development Goals were reviewed. Thus, looking at statistical data after certain milestones acts as a review of progress made in global humanitarian efforts. Coincidentally, this range also provided sufficient data points to work with.

The data gathered from the individual datasets were then combined to form a CSV file with 14 columns - 2 for Country Name and Year, and the rest for the indicators.

6.2 Linear Regression Model

Linear regression is a linear approach for modelling the relationship between independent and dependent variables. We can apply the linear regression model to predict the Infant Mortality Rates based on the indicators identified previously.

6.2.1 Building the model

We can build this model, dt.lm, using the lm() function.

```
set.seed(2021) # set seed for replicable results
dt <- fread("dataset.csv")
dt.lm <- lm(Infant.Mortality.Rate ~ . - Country.Name - Year, data=dt)
```

Figure 1: Building the linear regression model

After building the model, it is important to look at a summary of the model and identify insignificant variables by examining the p-value for each term. The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.1) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to the model because changes in the predictor's value are related to changes in the dependent variable, infant mortality rate.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     30.03862    4.70939   6.378 4.54e-10 ***
GDP.Growth                      -0.36223    0.06645  -5.451 8.37e-08 ***
DPT.Immunisation                -0.35140    0.05251  -6.692 6.75e-11 ***
Gini.Coefficient                -0.07452    0.04694  -1.588 0.113093
Measles.Immunisation             0.17560    0.04795   3.662 0.000281 ***
Annual.PM.2.5.Exposure           0.03869    0.02241   1.726 0.085037 .
Population.Living.In.Slums       0.02079    0.02248   0.925 0.355432
Prevalence.Of.Undernourishment   0.04822    0.03462   1.393 0.164348
Sever.wasting..weight.for.height 0.66178    0.10552   6.272 8.55e-10 ***
Death.From.Communicable.Diseases 0.33237    0.02458  13.522  < 2e-16 ***
Political.Stability             -0.91550    0.31931  -2.867 0.004341 **
Overall.Global.Gender.Gap.Index -10.03318    4.86193  -2.064 0.039641 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2: Summary of the linear regression model

According to Figure 2, the following variables are insignificant: Gini Coefficient, Population Living In Slums and Prevalence of Undernourishment. These insignificant variables should be removed and a new linear regression model should be built.

```
# remove insignificant variables
dt2 <- dt[,c(5,9,10):=NULL]
dt.lm2 <- lm(Infant.Mortality.Rate ~ . -Country.Name -Year, data=dt2)
summary(dt.lm2)
```

Figure 3: Removal of insignificant variables

Next, it is important to check if there are any multicollinearity issues in the linear regression. Multicollinearity occurs when two or more independent variables are highly correlated with one another. This means that an independent variable can be predicted from another independent variable in a regression model. If there are independent variables which are multicollinear, it is important to remove those variables. By running vif() on the model, we can detect multicollinearity.

```
> vif(dt.lm2)
                  GDP.Growth                DPT.Immunisation            Measles.Immunisation
                    1.102047                       4.778334                        4.647800
       Annual.PM.2.5.Exposure Sever.wasting..weight.for.height Death.From.Communicable.Diseases
                    1.888218                       2.290888                        1.686483
          Political.Stability  Overall.Global.Gender.Gap.Index
                    1.166201                       1.078278
```

Figure 4: Variance inflation factor

Using the benchmark that multicollinearity is present when VIF>5, according to Figure 4, there are no multicollinearity issues in this linear regression and thus there is no need to remove any variables from the linear regression.

Next, we need to split the dataset randomly into train and test sets. We will train the model using the test set and test the model by comparing the model predicted infant mortality rates with the actual infant mortality rates in the test set. We have chosen to use a split ratio of 0.7 which is a common practice.

6.2.2 Evaluation of the model

Evaluation 1

We can use the summary function to identify the multiple r-squared and adjusted r-squared value. These two values represent the explanation power of the model - the higher the values, the stronger the model.

```
Residual standard error: 6.278 on 442 degrees of freedom
Multiple R-squared:  0.7451,    Adjusted R-squared:  0.7405
F-statistic: 161.5 on 8 and 442 DF,  p-value: < 2.2e-16
```

Figure 5: Summary of the linear regression model

According to Figure 5, about 74% of the data can be explained by the linear regression equation, which we derived from the R-squared values. Given that the multiple R-squared and adjusted R-squared values are relatively high, it is a strong model.

Evaluation 2

The linear regression model has 3 assumptions:

1. Linear Association between Y and X variables
2. Errors has a normal distribution with mean 0
3. Errors are independent of X and has constant standard deviation

In order to measure if our linear regression model is a good fit, we need to examine the model diagnostic plots and determine if the 3 assumptions are satisfied. To measure Assumption 1, we can use a graph that shows the Residuals vs Fitted Values, as shown in the graph below:

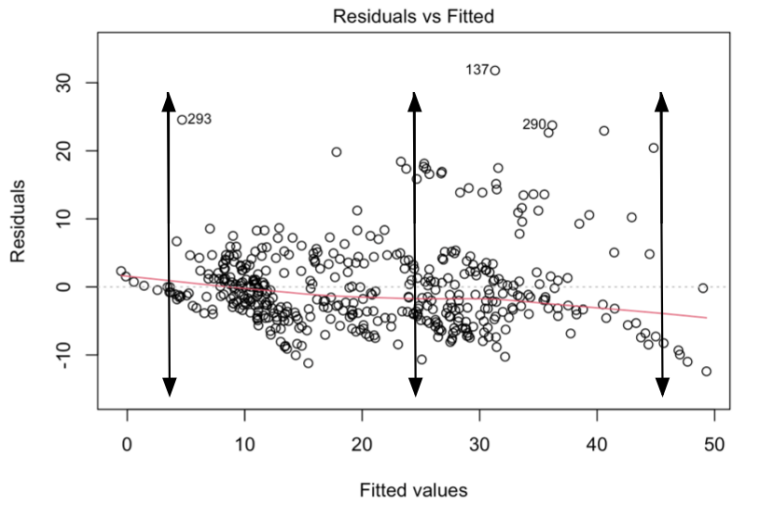


Figure 6: Residuals vs Fiited Plot

In layman terms, residuals are “leftovers” of the outcome variable after fitting a model (predictors) to data and they could reveal unexplained patterns in the data by the fitted model. A residual vs. fitted plot can show if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and our outcome variable, and the pattern would show up in this plot if the model doesn’t capture the non-linear relationship.

If there are equally spread residuals around a horizontal line without distinct patterns, it is a good indication that there are no non-linear relationships.

According to Figure 6, there are relatively equal spread residuals about a horizontal line and thus it is indicative that there are no non-linear relationships. The 1st assumption of a linear association between Y and X variables is thus satisfied.

Next, we will look at the Normal Q-Q graph. This plot shows if the residuals are normally distributed. If residuals follow a straight line well and do not deviate severely, residuals are normally distributed.

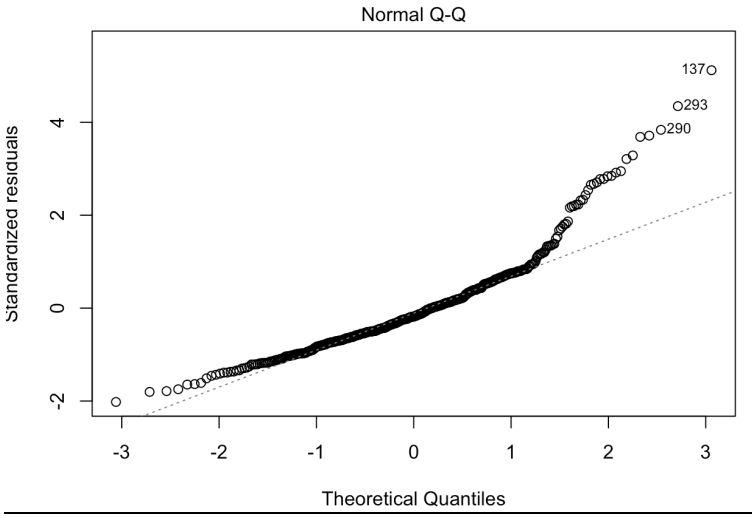


Figure 7: Normal Q-Q plot

In our case, analysing Figure 7 above, residuals deviate from the straight line. This suggests that residuals are not normally distributed and the 2nd assumption of errors having a normal distribution with mean 0 is not satisfied.

Finally, let's take a look at the Scale-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). A horizontal line with equally, and randomly, spread points is indicative that residuals have equal variance.

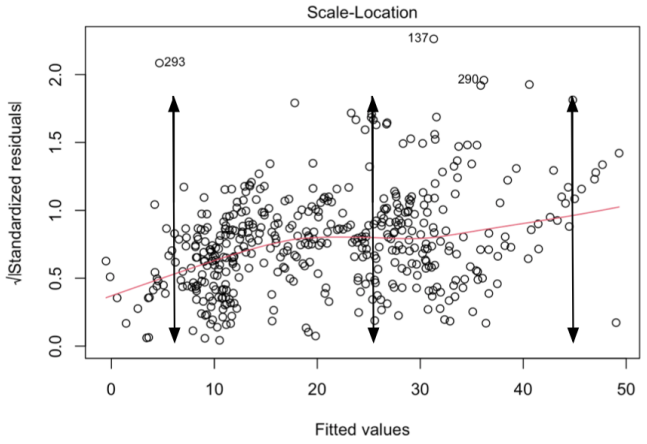


Figure 8: Scale-Location Plot

According to Figure 8, residuals appear equally and randomly spread. This suggests that residuals have equal variance and the 3rd assumption that errors are independent of X and have constant standard deviation is thus satisfied.

Evaluation 3

Finally, we will take a look at the Residuals vs Leverage graph. This plot helps identify if there are any influential outliers.

Contrary to popular belief, not all outliers may be influential in linear regression analysis. Even though data have extreme values, they might not be influential to determine a regression line. That means, the results wouldn't be much different whether we included or excluded them from the analysis. They follow the trend in the majority of cases and they don't really matter; they are not influential. On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. Another way to put it is that they don't get along with the trend in the majority of the cases.

We are looking for cases outside of a dashed line, also known as Cook's distance. When cases are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential.

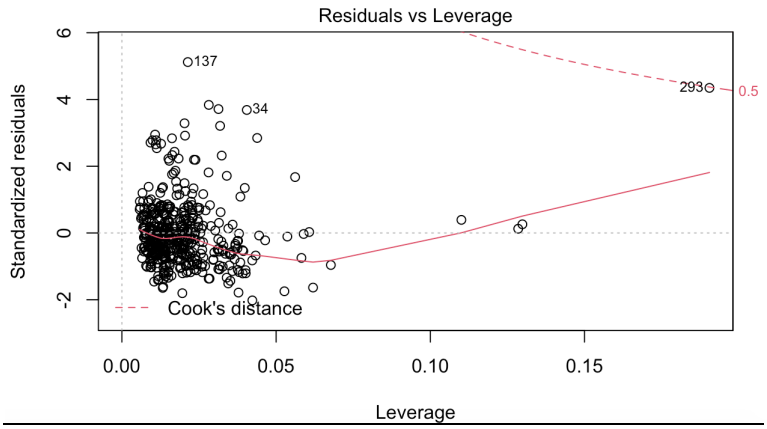


Figure 9: Residuals vs Leverage Plot

According to Figure 9, the Cook's distance lines are barely visible for our model, therefore, all the cases are well inside of the Cook's distance lines. Thus, it suggests that there is no influential case, and our model would be a good model to predict the data.

Evaluation 4

We can use root mean square error (RMSE) to determine the model accuracy. RMSE is the average difference between the observed known values of the dependent variable and the predicted value. The lower the RMSE, the better the accuracy of the model.

```
> RMSE.lm3.train
[1] 6.207103
> RMSE.lm3.test
[1] 6.267764
```

Figure 10: RMSE of train and test set

In this case, since the root mean square error of both the train and test set is low, the model is relatively accurate.

Evaluation 5

We can also plot a graph of actual infant mortality rates against the model predicted infant mortality rates to determine the accuracy of the model

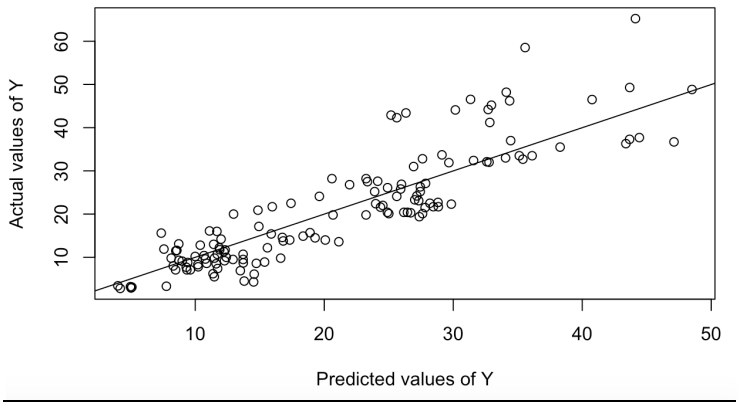


Figure 11: Graph of actual against predicted value of Infant Mortality Rate

Given that there is a positive linear relationship between the two, the model is quite accurate.

6.3 CART model

6.3.1 Brief Overview of CART

A Classification and Regression Tree (CART) is a predictive algorithm used in machine learning, which explains how a dependent variable can be predicted based on other independent variables. The CART output is a decision tree where each fork is split in a predictor variable and each node at the end contains a prediction for the outcome variable. (Digital Vidya, 2021)

In our case, our dependent variable Y is Infant Mortality Rate and our CART model aims to predict Infant Mortality Rate (Y) from the significant indicators identified previously from the Linear Regression. As such, we would split our dataset into a train set and a test set, such that we can train the CART model using the train set, where patterns in the data and independent variables

that are more important can be identified, before applying the CART model to the test set to predict the values of the dependent variable Y, and evaluate the accuracy of the model.

```
> summary(trainset$Infant.Mortality.Rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.30   10.55   19.80   20.61   28.50   63.53
> summary(testset$Infant.Mortality.Rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.80    9.80   19.90   20.89   27.75   65.23
```

Figure 12: Distribution of dependent variable Y for train set and test set

6.3.2 Regression Tree

The CART methodology refers to two types of decision trees - classification tree or regression tree. In our case, our CART output would be a regression tree because our dependent variable Y is continuous. In a regression tree, the model is fit to the dependent variable using each of the independent variables. After this, the data is split at several points for each independent variable. (Digital Vidya, 2021)

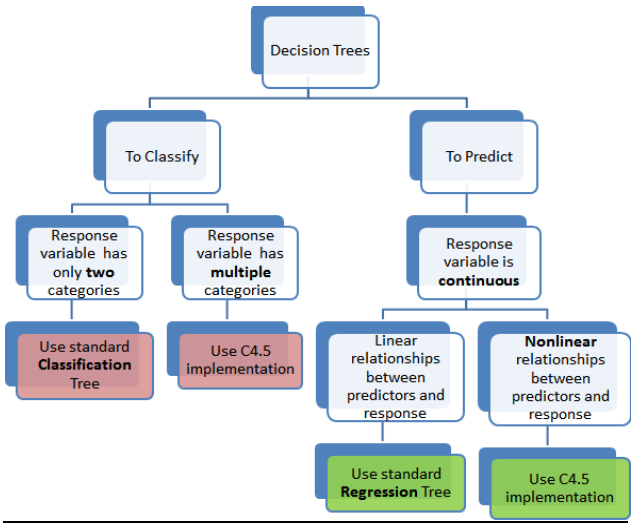


Figure 13: CART methodology (Digital Vidya, 2021)

6.3.3 Complexity Parameter

The complexity parameter (cp) imposes a penalty to the tree for having too many splits and is thus used to control the size of the decision tree and to select the optimal tree size. The lower the cp, the bigger the tree. (STHDA, 2018)

With reference to Figure 14, we can create a fully grown tree showing all the independent variables in our dataset by setting the minimum number of splits as 2 and setting the cp as 0. However, a fully grown tree will overfit the training data and lead to poor test set performance, hence it is not useful and undesirable.

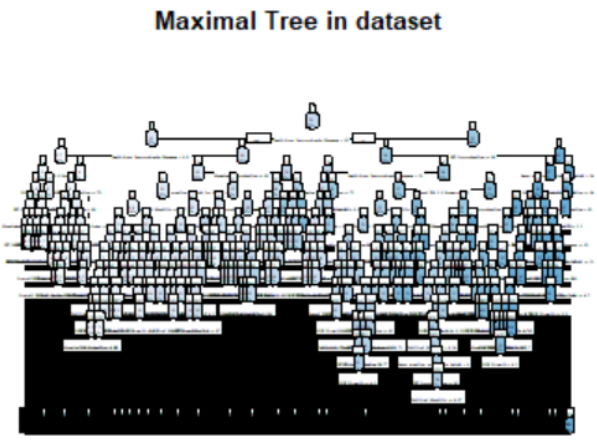


Figure 14: Fully Grown Tree

A strategy to limit this overfitting is to prune the tree in order to get a simpler tree with fewer splits. (Refer to Appendix 1 to get a guide on how we create the CART model and how we prune the tree) A too small value of cp leads to overfitting and a too large cp value results in a too small tree. Either case would decrease the predictive performance of the model. Hence, we can code to find the optimal cp, and thereafter use this cp value to prune the tree. Thus, we managed to get our optimal regression tree as shown in Figure 15.

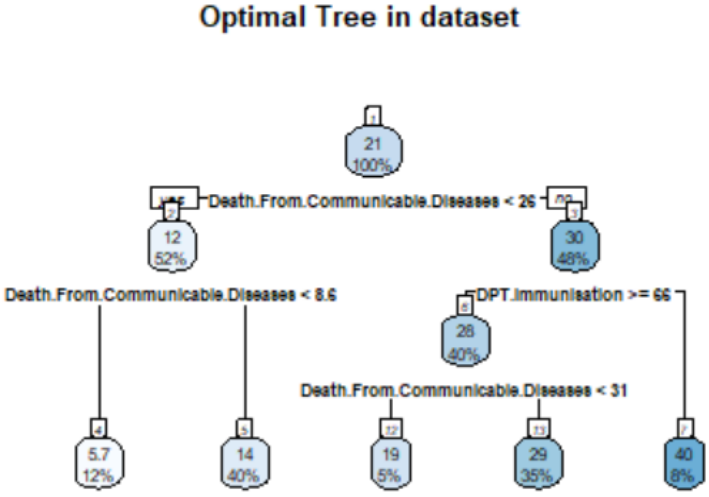


Figure 15: Optimal Regression Tree

6.3.4 Variable Importance

After getting our optimal regression tree, we can evaluate the relative importance of the independent variables in the regression tree and find out which variables are more important. An important variable is a variable that is used as a primary or surrogate splitter in the regression tree. The variable with higher number means there's higher importance to the tree. (Mini Tab, 2021)

As seen from Figure 15, there are 7 variables which have been identified as more important. Listing in order of importance starting with the highest importance - Death From Communicable

Diseases, Sever Wasting Weight for Height, Diphtheria, Pertussis and Tetanus (DPT) Immunisation, Measles Immunisation, GDP growth, Annual PM 2.5 Exposure and Overall Global Gender Gap Index.

```
Variable importance
Death.From.Communicable.Diseases Sever.wasting..weight.for.height                DPT.Immunisation
                              35                              20                              15
            Measles.Immunisation                      GDP.Growth          Annual.PM.2.5.Exposure
                              15                               8                               6
 Overall.Global.Gender.Gap.Index
                               1
```

Figure 15: Variable Importance of Optimal Regression Tree (based on scaling)

6.3.5 Surrogates

One key advantage of CART is that it automatically handles missing values through surrogates. When there is a missing value for the current best splitting variable, surrogate splits are used and CART calculates which alternative split resembles the best split. Any observation with a missing value for the best split is then classified using the first (most resembling) surrogate split, or if that value is missing too, the second surrogate split, and so on. (Webspace, 2021)

In our case, as seen from Figure 16, we did not have any missing values in the primary splits, hence surrogates were not activated.

```
Node number 1: 315 observations,    complexity param=0.5535719
  mean=20.60564, MSE=142.58
  left son=2 (165 obs) right son=3 (150 obs)
  Primary splits:
      Death.From.Communicable.Diseases < 25.86255  to the left,  improve=0.5535719, (0 missing)
      Sever.wasting..weight.for.height < 5.266667  to the left,  improve=0.4163257, (0 missing)
      DPT.Immunisation                 < 78.5      to the right, improve=0.2853342, (0 missing)
      Measles.Immunisation             < 92.5      to the right, improve=0.2081031, (0 missing)
      Annual.PM.2.5.Exposure           < 42.79498  to the left,  improve=0.1094304, (0 missing)
  Surrogate splits:
      Sever.wasting..weight.for.height < 5.825     to the left,  agree=0.835, adj=0.653, (0 split)
      Measles.Immunisation             < 93.5      to the right, agree=0.730, adj=0.433, (0 split)
      DPT.Immunisation                 < 91.275    to the right, agree=0.711, adj=0.393, (0 split)
      GDP.Growth                       < 5.359727  to the left,  agree=0.641, adj=0.247, (0 split)
      Annual.PM.2.5.Exposure           < 38.98936  to the left,  agree=0.606, adj=0.173, (0 split)
```

Figure 16: Summary of Optimal Regression Tree

6.3.6 Evaluating Model Accuracy

Previously, we have split our dataset into a train set and a test set. Now that we have built our CART model using the train set, we can apply the CART model to the test set to predict the values of the dependent variable, and subsequently evaluate the accuracy of the model.

To evaluate the accuracy of the CART model, we can calculate the Root Mean Square Error (RMSE) of both the train set and test set. The prediction error which is measured by the RMSE, is the average difference between the observed known values of the dependent variable and the predicted value by the CART model. The lower the RMSE, the more accurate the model. As seen from Figure 17, since the RMSE of both the train and test set is low, the model is relatively accurate.

```
> RMSE.cart2.train
[1] 6.531862
> RMSE.cart2.test
[1] 7.224243
```

Figure 17: RMSE of train set and test set

Besides calculating the RMSE, we also plotted the observed known values of the dependent variable against the predicted value by the CART model. As seen from Figure 18, the plot shows a positive linear relationship, hence the model is quite accurate.

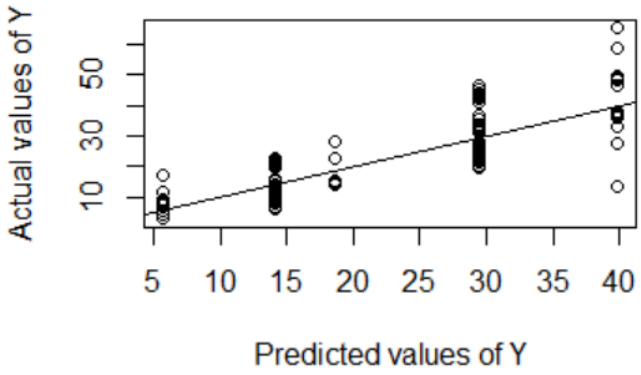


Figure 18: Plot of actual against predicted values of Infant Mortality Rate

6.4 Evaluation of our Proposed Solution

Currently, the EIU uses several forecasting methodologies such as surveys, inputs from experts, and in-house EIU expert ratings. While these are certainly effective and they produce valuable insights, the use of machine learning can definitely help to add value to their products and services.

Firstly, using machine learning is faster, automatic and requires less human effort. In addition, they also tend to be more accurate at making predictions, especially in areas such as pattern recognition, where it has an edge over humans, as we are prone to make errors. Machine learning has the ability to identify patterns and make accurate predictions at a scale and speed that humans are unable to. It is therefore in the EIU's interest to start incorporating and embracing machine learning models into their research and forecasting process.

This is not to say that our model and our application of machine learning is perfect, but we believe that it will add much value to the EIU's research processes. The opinions and qualitative insights from experts are still very much needed, but with machine learning, these can now be supported with statistical data that is produced in a quicker and more accurate scale.

However, one of the limitations that we faced in the process of completing this project was the limited amount of data. We only used data from 2010 to 2020 in our predictive models. Although it was sufficient for the objectives of such a project, it goes without saying that the model would have been much more accurate if trained with more data.

A limitation of the model itself is that it is only a truly accurate predictor when it works with countries that have data collected for all the indicators we used for our model. As a result, this score may not be very inclusive for all countries, since there are definitely many countries that have some lack of data collection for the indicators. We could only work with publicly available data that we could find online. In contrast, the EIU has many researchers worldwide and extensive resources that allow them to conduct large scale research, and so this limitation could be easily overcome as they are able to do the necessary research and obtain the data for these indicators for all the 200 plus countries they have produced country reports for.

7 Model Summary

In order to determine the H.I.R.A scores , the variable importance values from the models can be multiplied with their individual values, and then summed up together in a new column.

```
dataset <- dt %>% rowwise() %>%
  mutate(Overall.Score = sum(35*Death.From.Communicable.Diseases+15*Measles.Immunisation
                             +1*Overall.Global.Gender.Gap.Index+20*Sever.wasting..weight.for.height
                             +15*DPT.Immunisation+6*Annual.PM.2.5.Exposure+8*GDP.Growth))
```

Figure 19: Computing the H.I.R.A score

In order to provide better readability for our target audience, it is important to filter out unnecessary data, compile a timeline of the scores and calculate an average score. After sorting the data by descending order, we will be able to obtain the following results for the first 20 countries:

Country.Name	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Average.Score
Burkina Faso	5626.998681	5373.737237	5342.565271	5147.265163	5282.735107	5222.442993	5321.112954	5353.469653	5339.639855	5030.489347	4618.034203	5241.680951
Kenya	5197.785548	5191.437783	5246.300718	5164.191049	5205.102638	5078.85765	5222.080424	5003.465484	5165.464866	4749.11081	4866.596947	5099.12672
Malawi	5284.950018	5276.548593	5151.239904	5036.441272	5013.866905	4725.982896	4812.49655	4887.467305	4998.590158	4752.333183	4798.820853	4976.248876
Mozambique	4846.297466	4868.907188	4958.193805	5029.450501	5092.885571	4945.912428	5057.430152	5055.720398	5060.767599	4820.682081	4385.6348	4920.17109
Mauritania	4643.680461	4749.13633	4916.318369	5001.349982	4959.711917	4763.566419	4806.959127	5024.371677	4968.542144	4806.554687	4282.124362	4811.119589
Mali	5033.171012	4798.573926	4709.865104	4630.56541	4711.202455	4841.992362	4963.725155	4972.623275	4932.556756	4764.290157	4373.156695	4793.792937
Liberia	4476.391694	4769.66305	4948.970784	4862.556316	4297.524989	4796.545027	4919.125445	5079.931297	5121.89725	4580.457685	4471.342917	4756.764223
Lesotho	4863.368249	4922.702496	4905.618863	4829.87327	4712.812399	4477.62434	4731.382148	4675.685553	4617.500979	4447.859289	4415.291128	4690.88352
Nepal	4662.878486	4843.86624	4755.402269	4798.209011	4850.690412	4423.689352	4621.679444	4868.340949	4862.312744	4510.90786	3993.417167	4653.763085
Bangladesh	4645.017549	4800.912294	4676.606257	4787.81911	4790.713907	4679.058489	4812.417008	4818.121937	4729.237104	4372.873974	3936.620466	4640.854372
India	4658.930215	4667.384525	4597.537487	4641.777334	4623.528517	4615.620898	4793.860739	4805.04863	4882.744592	4579.855166	3843.616039	4609.991286
Namibia	4523.880112	4481.719357	4540.233231	4681.898779	4708.779008	4634.601414	4492.500333	4602.879711	4674.586512	4414.411749	4324.628789	4552.73809
Benin	4533.278716	4526.006899	4672.190888	4552.99856	4392.348365	4410.021276	4549.268508	4601.233295	4561.509387	4432.043862	4327.718993	4505.328977
Nigeria	4649.805661	4407.54958	4237.537633	4248.725288	4205.521795	4206.640918	4591.412505	4679.035656	4573.85596	4544.559638	4098.549643	4403.926752
Madagascar	4152.407807	4163.719737	4062.171009	4202.79331	4208.741761	3922.279329	4195.255562	4162.66037	4132.54208	4183.461222	3796.9352	4107.54249
Myanmar	4377.027514	4168.65057	4116.005028	4021.911515	4227.872626	3980.369613	4203.672912	4126.805726	4285.848592	3811.216063	3575.633354	4081.364865
Pakistan	3715.606486	3930.923673	3963.579386	3975.816443	4150.591081	4146.313304	4302.093711	4315.211973	4149.563334	4076.649819	3558.272642	4025.874714
Guyana	3936.598374	3983.025348	4052.262027	4041.15825	3971.851212	3890.885788	4012.50071	4011.943555	3990.258186	3934.19958	3909.327615	3975.81915
Lao PDR	3900.589968	3872.953875	4028.910702	4298.683744	4374.227344	3988.331225	3708.992804	3808.534069	3792.385344	3376.475984	3106.665503	3841.522778
Guatemala	4046.275647	3966.481793	4125.914773	3847.726049	3382.542504	3296.988369	3742.050744	3776.870456	3881.095293	3654.854817	3337.630984	3732.584675

Figure 20: H.I.R.A scores

From the table as shown, we can see that Burkina Faso has the highest H.I.R.A score. This implies that Burkina Faso has the greatest need for medical assistance.

8 Recommendations for EIU

As mentioned earlier, humanitarian non-profit companies have scarce resources, and thus a ranking of countries in terms of their need for healthcare assistance will help these companies greatly, to be able to better allocate their resources to helping the countries that need it the most.

Without machine learning, it would be a gargantuan challenge for non-profit healthcare organisations to plan their resources, because there is no objective and clear way to decide on which country would require the most, and how much, of their assistance. With any excessive spending of resources in a country, comes with it a huge loss in capital and opportunity costs.

As such, there is a gap in the market currently. We saw an opportunity for EIU to provide useful information for this non-profit healthcare industry by offering recommendations tailored to the needs of these organisations planning to distribute resources to other countries. Our product would be able to solve the needs that these organisations have and allow them to make well-informed decisions. This would give EIU a competitive edge over their competitors.

Furthermore, in the future, to further help EIU make more profit, similar models could be built by using the model with other subsets of data, such as for social welfare groups and social enterprises. The model can provide an objective and statistical way of deciding which country requires assistance.

9 Conclusion

In conclusion, EIU faces stiff competition in the forecasting market, and constantly needs to improve their game to stay afloat in this cutthroat environment. There is significant value in the EIU adopting machine learning for their strategies, and hence we believe that the EIU should definitely consider adopting machine learning models. Statistical data with higher accuracy coupled with opinions and qualitative insights from experts would give EIU the competitive edge.

Finally, the team definitely recommends EIU to adopt this machine learning proof of concept (POC). Pilot studies are small-scale, preliminary studies which aim to investigate whether crucial components of a main study – usually a randomized controlled trial (RCT) – will be feasible. This POC serves to act as a pilot study for EIU to evaluate machine learning as our POC allows readers to interpret the results and implications correctly, and evaluate the feasibility of some crucial components, such as the collection of data. Based on our POC, it can be seen that the proposed machine learning methodology works and is relatively accurate. However, we have evaluated that the current publicly available dataset is insufficient. There is a need to gather more data and include more countries into our analysis. With EIU's extensive capabilities, EIU could easily provide the data points needed for the model to be more accurate and effective. With its many

researchers worldwide and extensive resources, EIU can collect all the data required to evaluate and look more deeply into a country's economy to find new relationships in the healthcare administration that have not been researched on before.

10 Appendices

10.1 Appendix I: CART Model

A Classification and Regression Tree (CART) is a predictive algorithm used in machine learning, which explains how a dependent variable can be predicted based on other independent variables. The CART output is a decision tree where each fork is split in a predictor variable and each node at the end contains a prediction for the outcome variable. The CART model aims to predict Infant Mortality Rate (Y) from the significant variables identified in the Linear Regression.

Firstly, we need to load the packages needed and read the csv file. Thereafter, we need to set.seed( ) to generate a random number sequence that can be reproduced to verify results.

```
library(caTools)
library(rpart)
library(rpart.plot)

dt <- fread("dataset.csv")

# Generate a random number sequence that can be reproduced to verify results.
# for randomization in 10-fold CV
set.seed(2021)
```

Before building the CART model, we need to split the dataset into training and testing sets with a split ratio of 0.7. We would first use the train set to train the CART model, before applying the CART model to the test set to predict the values of the dependent variable Y, and evaluate the accuracy of the model by using RMSE.

```
# split dataset into train-test set
# 70% trainset. Stratify on Y = mortalityRate
train <- sample.split(Y = dt$Infant.Mortality.Rate, SplitRatio = 0.7)
trainset <- subset(dt, train == T)
testset <- subset(dt, train == F)
```

In order to ensure that the distribution of Y is similar in both the train set and the test set, we can use summary( ) and the results show that they are indeed of similar distribution.

```
# Checking the distribution of Y is similar in trainset vs testset.
summary(trainset$Infant.Mortality.Rate)
summary(testset$Infant.Mortality.Rate)
```

```
> summary(trainset$Infant.Mortality.Rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.30   10.55   19.80   20.61   28.50   63.53
> summary(testset$Infant.Mortality.Rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.80    9.80   19.90   20.89   27.75   65.23
```

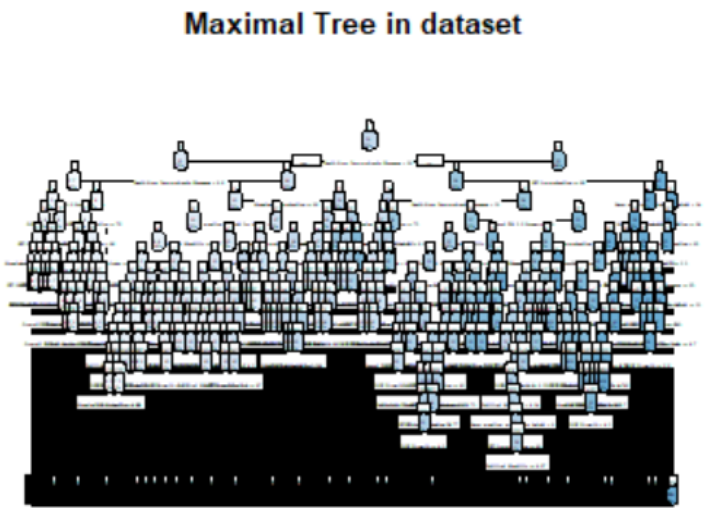
Next, we build the CART model using the rpart( ) function. Since we are predicting a continuous Y variable, the method should be set to “anova” so that it will create a regression tree rather than

a classification tree. The cp is used to control the size of the decision tree. The higher the cp, the smaller the tree. With a small sample size, we set the minsplit value as 2 and set the cp value as 0 to ensure the tree grows to the maximum.

```
# build CART model
# Continuous Y: Set method = 'anova'
# minsplit value changed to 2 due to the small sample size
# cp value changed to 0 to ensure grow tree to the max
cart1 <- rpart(Infant.Mortality.Rate ~ . - Country.Name - Year - Gini.Coefficient
               - Population.Living.In.Slums - Prevalence.Of.Undernourishment,
               data = trainset, method = 'anova', control = rpart.control(minsplit = 2, cp = 0))
```

cart1 is hence the maximal tree that we have grown. Next, we plotted and printed out the maximal tree structure.

```
# plot the maximal tree and results
rpart.plot(cart1, nn = T, main = "Maximal Tree in dataset")

# print the maximal tree (Cart1) onto the console
print(cart1)
```




```
> print(cart1)
n= 315

node), split, n, deviance, yval
      * denotes terminal node

     1) root 315 4.491271e+04 20.605640
       2) Death.From.Communicable.Diseases< 25.86255 165 6.760816e+03 12.134920
         4) Death.From.Communicable.Diseases< 8.567268 39 3.976201e+02  5.689093
           8) Annual.PM.2.5.Exposure< 28.38488 16 6.500009e+01  3.276605
            16) GDP.Growth>=-0.4590135 14 1.072929e+01  2.607143
              32) DPT.Immunisation< 92.5 6 8.933333e-01  1.766667
                64) Measles.Immunisation< 61 4 1.475000e-01  1.525000
                 128) DPT.Immunisation< 86.5 2 5.000000e-03  1.350000
                   256) GDP.Growth< 4.57041 1 0.000000e+00  1.300000 *
                   257) GDP.Growth>=4.57041 1 0.000000e+00  1.400000 *
                 129) DPT.Immunisation>=86.5 2 2.000000e-02  1.700000
                   258) DPT.Immunisation< 88 1 0.000000e+00  1.600000 *
```

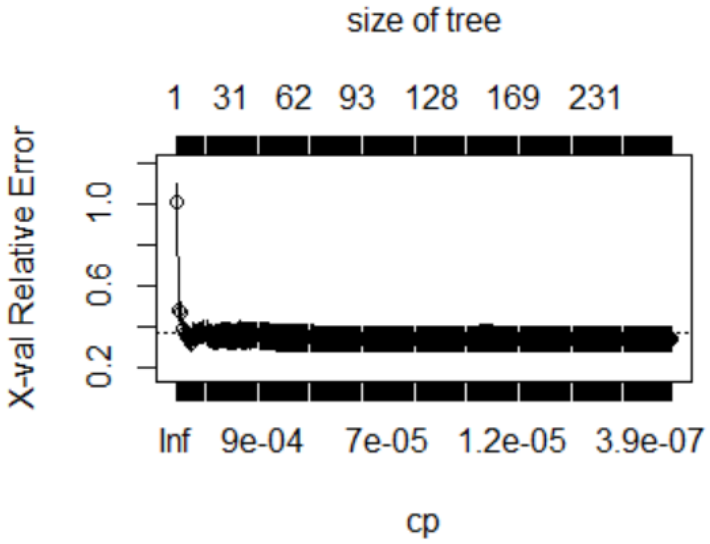
Since a fully grown tree will overfit the training data and lead to poor test set performance, we need to prune the tree. Hence, we printed out the pruning sequence and 10-fold CV errors both

as a table using printcp( ) and as a chart using plotcp( ). We would be able to identify the optimal tree from here.

```
# prints out the pruning sequence and 10-fold CV errors, as a table
# can be used to identify optimal tree
printcp(cart1)
```

```
> printcp(cart1)

Regression tree:
rpart(formula = Infant.Mortality.Rate ~ . - Country.Name - Year -
    Gini.Coefficient - Population.Living.In.Slums - Prevalence.Of.Undernourishment,
    data = trainset, method = "anova", control = rpart.control(minsplit = 2,
        cp = 0))

Variables actually used in tree construction:
[1] Annual.PM.2.5.Exposure          Death.From.Communicable.Diseases DPT.Immunisation
[4] GDP.Growth                      Measles.Immunisation             Overall.Global.Gender.Gap.Index
[7] Political.Stability             Sever.wasting..weight.for.height

Root node error: 44913/315 = 142.58

n= 315

           CP nsplit  rel error  xerror     xstd
1   5.5357e-01      0 1.0000e+00 1.00531 0.087171
2   6.3429e-02      1 4.4643e-01 0.48215 0.058486
3   4.7246e-02      2 3.8300e-01 0.46417 0.058932
4   3.6516e-02      3 3.3575e-01 0.39148 0.047597
5   2.9994e-02      4 2.9924e-01 0.36500 0.045824
6   2.0993e-02      5 2.6924e-01 0.34534 0.044006
7   1.9331e-02      6 2.4825e-01 0.34552 0.043784
```

```
# display the pruning sequence and 10-fold CV errors, as a chart
# can be used to identify optimal tree
## unable to find the optimal tree as the plot is too compact
plotcp(cart1)
```



Since the plot is too compact, we are unable to identify the optimal tree from the plot. Hence, we can compute the minimum CV error + 1SE in the maximal tree as CVerror.cap and then find the optimal CP region whose CV error is just below the CVerror.cap in the maximal tree. Thereafter, we will be able to get the geometric mean of the two identified CP values in the optimal region as the cp value.

```
# Compute min CVerror + 1SE in maximal tree cart1.
CVerror.cap <- cart1$cptable[which.min(cart1$cptable[,"xerror"]), "xerror"] +
  cart1$cptable[which.min(cart1$cptable[,"xerror"]), "xstd"]

# Find the optimal CP region whose CV error is just below CVerror.cap
# in maximal tree cart1.
i <- 1; j<- 4
while (cart1$cptable[i,j] > CVerror.cap) {
  i <- i + 1
}

# Get geometric mean of the two identified CP values in the optimal region
#if optimal tree has at least one split.
cp.opt = ifelse(i > 1, sqrt(cart1$cptable[i,1] * cart1$cptable[i-1,1]), 1)
```

Thereafter, we can prune the tree to get our optimal tree using the cp value calculated earlier.

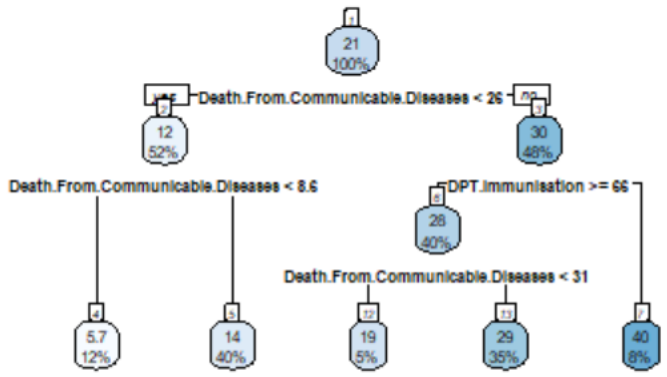
```
# get a specific subtree by pruning the maximal tree (cart1) with a specific value of cp
cart2 <- prune(cart1, cp = cp.opt)
```

Next, we plotted and printed out the optimal regression tree structure. From the plot, we are able to identify which independent variables are more important, since the importance of the variables are ordered according to how high up they are in the tree structure.

```
# print the optimal tree (Cart2) onto the console
print(cart2)

# plot the optimal tree and results
# The number inside each node represent the mean value of Y.
rpart.plot(cart2, nn = T, main = "Optimal Tree in dataset")
```

```
> print(cart2)
n= 315

node), split, n, deviance, yval
      * denotes terminal node

 1) root 315 44912.7100 20.605640
   2) Death.From.Communicable.Diseases< 25.86255 165  6760.8160 12.134920
     4) Death.From.Communicable.Diseases< 8.567268 39   397.6201  5.689093 *
     5) Death.From.Communicable.Diseases>=8.567268 126  4241.2430 14.130060 *
   3) Death.From.Communicable.Diseases>=25.86255 150 13289.4800 29.923420
     6) DPT.Immunisation>=65.5 126  7921.2530 28.021450
      12) Death.From.Communicable.Diseases< 30.50391 16   523.6284 18.561790 *
      13) Death.From.Communicable.Diseases>=30.50391 110  5757.6060 29.397400 *
     7) DPT.Immunisation< 65.5 24  2519.4480 39.908780 *
```

Optimal Tree in dataset



Next, we printed out the variable importance and summary of the CART model. The variable with higher number means there's higher importance to the tree. From the summary of the CART model, the variable importance has been scaled (adds up to 100%), and there are 6 variables which have been identified as more important. Listing in order of importance starting with the highest importance - Death From Communicable Diseases, Sever Wasting Weight For Height, Population Living In Slums, Prevalence of Undernourishment, Measles Immunisation, and DPT Immunisation.

```
cart2$variable.importance

summary(cart2)
```

```
> cart2$variable.importance
Death.From.Communicable.Diseases Sever.wasting..weight.for.height                DPT.Immunisation
                       29099.1829                       16755.9499                    12627.9967
            Measles.Immunisation                       GDP.Growth          Annual.PM.2.5.Exposure
                       12297.0402                        6187.1380                     4665.5827
 Overall.Global.Gender.Gap.Index                Political.Stability
                         474.7967                         217.6362
```

```
> summary(cart2)
Call:
rpart(formula = Infant.Mortality.Rate ~ . - Country.Name - Year -
    Gini.Coefficient - Population.Living.In.Slums - Prevalence.Of.Undernourishment,
    data = trainset, method = "anova", control = rpart.control(minsplit = 2,
        cp = 0))
  n= 315

          CP nsplit rel error    xerror       xstd
1 0.55357189      0 1.0000000 1.0053055 0.08717085
2 0.06342927      1 0.4464281 0.4821450 0.05848631
3 0.04724616      2 0.3829988 0.4641713 0.05893217
4 0.03651568      3 0.3357527 0.3914751 0.04759654
5 0.03309480      4 0.2992370 0.3649971 0.04582382

Variable importance
Death.From.Communicable.Diseases Sever.wasting..weight.for.height                DPT.Immunisation
                              35                                20                              15
            Measles.Immunisation                       GDP.Growth          Annual.PM.2.5.Exposure
                              15                                 8                               6
 Overall.Global.Gender.Gap.Index
                               1

Node number 1: 315 observations,    complexity param=0.5535719
  mean=20.60564, MSE=142.58
```

Next, we display the residuals of the optimal tree. We calculated the Root Mean Square Error (RMSE) of the train set based on our CART model and check the minimum and maximum abs error to measure the accuracy of the CART model.

```
residuals(cart2)

# Trainset Errors
# Residuals = Error = Actual Infant.Mortality.Rate - Model Predicted Infant.Mortality.Rate
# RMSE on trainset based on cart2 model.
RMSE.cart2.train <- sqrt(mean(residuals(cart2)^2))

# Check Min Abs Error and Max Abs Error.
summary(abs(residuals(cart2)))
```

```
> residuals(cart2)
          1           2           3           4           5           6           7
-1.90877838 -2.30877838 -2.50877838 -2.80877838 -3.00877838 -4.00877838 -0.53006165
          8           9          10          11          12          13          14
-0.83006165 -1.33006165 -1.73006165 -3.13006165 -3.53006165  6.67646009 -2.09740141
         15          16          17          18          19          20          21
-3.19740141 -8.39740141  5.10259859  4.70259859  3.90259859  9.99122162  9.19122162
         22          23          24          25          26          27          28
 7.29122162  3.40259859 13.80259859 12.80259859 23.62622162 -2.63006165 -3.03006165
         29          30          31          32          33          34          35
-3.33006165 -3.63006165  2.90259859 -4.03006165 16.66993835 -5.03006165 -5.23006165
         36          37          38          39          40          41          42
-5.53006165  6.34827168 -5.03006165 -5.43006165  9.83820685 -5.83006165 -6.23006165
         43          44          45          46          47          48          49
-6.53006165 -6.93006165 -7.53006165 -7.73006165 13.75827168 -3.56179315  2.10259859
         50          51          52          53          54          55          56
-3.76179315 -3.96179315 -0.26993835 -4.26179315 -0.83006165  1.77154018 -12.00877838
```

```
> summary(abs(residuals(cart2)))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.01091  2.35780  3.93821  5.02357  6.53006 33.71260
```

Since we have already gotten our CART model, we can now apply the model to the test set to make predictions on the variable Y. Similarly, we calculated the RMSE of the test set based on our CART model and check the minimum and maximum Abs error to measure the accuracy of the CART model.

```
# Apply model from trainset to predict on testset.
predict.cart2.test <- predict(cart2, newdata = testset)

# Testset Errors
# Residuals = Error = Actual Infant.Mortality.Rate - Model Predicted Infant.Mortality.Rate
# RMSE on testset based on cart2 model.
testset.error <- testset$Infant.Mortality.Rate - predict.cart2.test
RMSE.cart2.test <- sqrt(mean(testset.error^2))

# Check Min Abs Error and Max Abs Error.
summary(abs(testset.error))
```

```
> summary(abs(testset.error))
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.06994  2.57434  4.31633  5.63762  7.57810 26.30878
```

Next, we display the RMSE for both the train set and test set that we calculated earlier. The prediction error which is measured by the RMSE, is the average difference between the observed

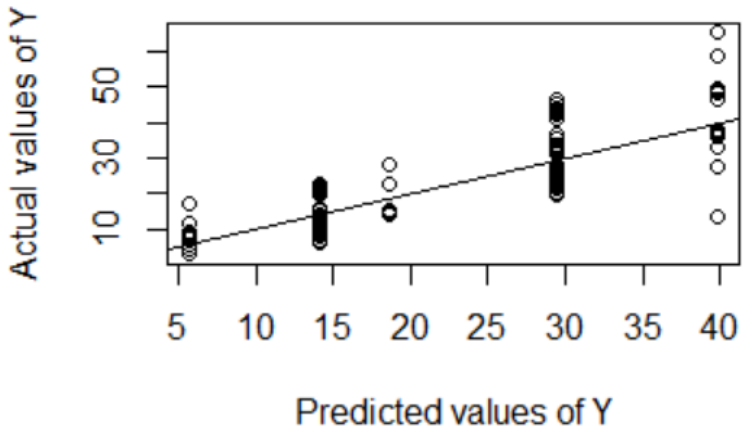
known values of the dependent variable and the predicted value by the CART model. The lower the RMSE, the more accurate the model. Since the RMSE of both the train and test set is low, the model is relatively accurate.

```
# RMSE for both trainset and testset
RMSE.cart2.train
RMSE.cart2.test
```

```
> RMSE.cart2.train
[1] 6.531862
> RMSE.cart2.test
[1] 7.224243
```

Lastly, we plotted the actual value of Infant Mortality Rate against the predicted values of Infant Mortality Rate from the CART model. The plot shows a positive linear relationship between the two, hence the model is quite accurate.

```
# plot of actual against predicted value
plot(predict.cart2.test,testset$Infant.Mortality.Rate,
     xlab="Predicted values of Y",
     ylab="Actual values of Y")
abline(a=0,b=1)
```



10.2 References

Bommae, W. by. (n.d.). University of Virginia Library Research Data Services + Sciences. Research Data Services + Sciences. Retrieved November 6, 2021, from https://data.library.virginia.edu/diagnostic-plots/.

By: IBM Cloud Education. (n.d.). What is machine learning? IBM. Retrieved November 6, 2021, from https://www.ibm.com/cloud/learn/machine-learning.

Datatonic & Vodafone: AI-Powered 5G. Datatonic. (2021, July 23). Retrieved November 6, 2021, from https://datatonic.com/insights/vodafone-5g-traffic-forecasting/.

Editor, M. B. (n.d.). How to interpret regression analysis results: P-values and coefficients. Minitab Blog. Retrieved November 6, 2021, from https://blog.minitab.com/en/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients.

The evolution of forecasting techniques. Genpact. (n.d.). Retrieved November 6, 2021, from https://www.genpact.com/insight/technical-paper/the-evolution-of-forecasting-techniques-traditional-versus-machine-learning-methods.

Kassambara. (2018, March 10). CART model: Decision tree essentials. STHDA. Retrieved November 6, 2021, from http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/141-cart-model-decision-tree-essentials/.

Multicollinearity: Detecting multicollinearity with VIF. Analytics Vidhya. (2020, April 16). Retrieved November 6, 2021, from https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/.

PKDD99 - science.uu.nl Project CSG. (n.d.). Retrieved November 6, 2021, from https://webspace.science.uu.nl/~feeld101/pkdd99.pdf.

Relative variable importance chart for CART® Classification. Minitab. (n.d.). Retrieved November 6, 2021, from https://support.minitab.com/en-us/minitab/19/help-and-how-to/statistical-modeling/predictive-analytics/how-to/cart-classification/interpret-the-results/all-statistics-and-graphs/relative-variable-importance-chart/.

writer, A. M. A. creative. (2021, April 21). A beginner's guide to classification and Regression Trees. Digital Vidya. Retrieved November 6, 2021, from https://www.digitalvidya.com/blog/classification-and-regression-trees/.

What is a pilot study? Students 4 Best Evidence. (2018, June 11). Retrieved November 6, 2021, from https://s4be.cochrane.org/blog/2017/07/31/pilot-studies/.