Predicting Childhood Mortality Based on Health and Socio-Economic Indicators

Aicha Sidiya, Hanin Alzaher, Razan Almahdi

2023-06-01

#loading libraries  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.1 ✔ purrr 1.0.1  
## ✔ tibble 3.2.1 ✔ dplyr 1.1.0  
## ✔ tidyr 1.3.0 ✔ stringr 1.5.0  
## ✔ readr 2.1.3 ✔ forcats 1.0.0

## Warning: package 'tibble' was built under R version 4.2.3

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(dplyr)  
library(readr)  
library(caret)

## Warning: package 'caret' was built under R version 4.2.3

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(RANN)

## Warning: package 'RANN' was built under R version 4.2.3

library(skimr)

## Warning: package 'skimr' was built under R version 4.2.3

library(ggplot2)  
library(stringr)  
library(gbm)

## Warning: package 'gbm' was built under R version 4.2.3

## Loaded gbm 2.1.8.1

#loading the data set  
mortality\_rate <- read.csv('data/Mortality rate, under-5 (per 1,000 live births).csv')  
health\_expenditure <- read.csv('data/Current health expenditure per capita (current US$).csv')  
health\_expenditure\_per <- read.csv('data/Current health expenditure (% of GDP).csv')  
education\_expenditure <- read.csv('data/Current education expenditure, total (%).csv')  
literacy\_rate <- read.csv('data/literacy\_rate.csv')  
domestic\_health\_expenditure <- read.csv('data/Domestic private health expenditure (% of current health expenditure).csv')  
economic\_inequality <- read.csv('data/economic-inequality-gini-index.csv')  
water\_invest <- read.csv('data/Investment in water and sanitation (current US$).csv')  
vacinnation <- read.csv('data/vaccination-coverage-by-income-in.csv')  
water\_productivity <- read.csv('data/Water productivity\_per cubic meter of total freshwater withdrawal.csv')  
healthcare\_access <- read.csv('data/healthcare-access-and-quality-index.csv')

#selecting from year 2000 till 2020  
mortality\_rate <- select(mortality\_rate, country, 'X2000':'X2020')  
health\_expenditure <- select(health\_expenditure, country, 'X2000':'X2020')  
health\_expenditure\_per<- select(health\_expenditure\_per, country, 'X2000':'X2020')  
literacy\_rate <- select(literacy\_rate, country, 'X2000':'X2020')  
education\_expenditure <- select(education\_expenditure, country, 'X2000':'X2020')  
water\_invest <- select(water\_invest, country, 'X2000':'X2020')  
water\_productivity <- select(water\_productivity, country, 'X2000':'X2020')  
domestic\_health\_expenditure <- select(domestic\_health\_expenditure, country, 'X2000':'X2020')  
economic\_inequality <- filter(economic\_inequality, year >= 2000)  
vacinnation <- filter(vacinnation, year >= 2000)  
healthcare\_access <- filter(healthcare\_access, year >= 2000)

#renaming columns  
mortality\_rate\_years <- select (mortality\_rate, 'X2000':'X2020')  
names(mortality\_rate\_years) <- str\_sub(names(mortality\_rate\_years),2)  
mortality\_rate <- select(mortality\_rate, country)  
mortality\_rate <- bind\_cols(mortality\_rate,mortality\_rate\_years)  
  
health\_expenditure\_years <- select (health\_expenditure, 'X2000':'X2020')  
names(health\_expenditure\_years) <- str\_sub(names(health\_expenditure\_years),2)  
health\_expenditure <- select(health\_expenditure, country)  
health\_expenditure <- bind\_cols(health\_expenditure, health\_expenditure\_years)  
  
health\_expenditure\_per\_years <- select (health\_expenditure\_per, 'X2000':'X2020')  
names(health\_expenditure\_per\_years) <- str\_sub(names(health\_expenditure\_per\_years),2)  
health\_expenditure\_per <- select(health\_expenditure\_per, country)  
health\_expenditure\_per <- bind\_cols(health\_expenditure\_per, health\_expenditure\_per\_years)  
  
education\_expenditure\_years <- select (education\_expenditure, 'X2000':'X2020')  
names(education\_expenditure\_years) <- str\_sub(names(education\_expenditure\_years),2)  
education\_expenditure <- select(education\_expenditure, country)  
education\_expenditure <- bind\_cols(education\_expenditure, education\_expenditure\_years)  
  
domestic\_health\_expenditure\_years <- select (domestic\_health\_expenditure, 'X2000':'X2020')  
names(domestic\_health\_expenditure\_years) <- str\_sub(names(domestic\_health\_expenditure\_years),2)  
domestic\_health\_expenditure <- select(domestic\_health\_expenditure, country)  
domestic\_health\_expenditure <- bind\_cols(domestic\_health\_expenditure, domestic\_health\_expenditure\_years)  
  
literacy\_rate\_years <- select (literacy\_rate, 'X2000':'X2020')  
names(literacy\_rate\_years) <- str\_sub(names(literacy\_rate\_years),2)  
literacy\_rate <- select(literacy\_rate, country)  
literacy\_rate <- bind\_cols(literacy\_rate, literacy\_rate\_years)  
  
water\_invest\_years <- select (water\_invest, 'X2000':'X2020')  
names(water\_invest\_years) <- str\_sub(names(water\_invest\_years),2)  
water\_invest <- select(water\_invest, country)  
water\_invest <- bind\_cols(water\_invest, water\_invest\_years)  
  
water\_productivity\_years <- select (water\_productivity, 'X2000':'X2020')  
names(water\_productivity\_years) <- str\_sub(names(water\_productivity\_years),2)  
water\_productivity <- select(water\_productivity, country)  
water\_productivity <- bind\_cols(water\_productivity, water\_productivity\_years)

#pivoting tables  
mortality\_rate1 <- pivot\_longer(mortality\_rate, cols="2000":"2020",  
 names\_to = "year",  
 values\_to = "mortality\_rate")  
health\_expenditure1 <- pivot\_longer(health\_expenditure, cols="2000":"2020",  
 names\_to = "year",  
 values\_to = "health\_expenditure")  
health\_expenditure\_per1 <- pivot\_longer(health\_expenditure\_per, cols="2000":"2020",  
 names\_to = "year",  
 values\_to = "health\_expenditure\_per")  
education\_expenditure1 <- pivot\_longer(education\_expenditure, cols="2000":"2020",  
 names\_to = "year",  
 values\_to = "education\_expenditure")  
domestic\_health\_expenditure1 <- pivot\_longer(domestic\_health\_expenditure, cols="2000":"2020",  
 names\_to = "year",  
 values\_to = "domestic\_health\_expenditure")  
literacy\_rate1 <- pivot\_longer(literacy\_rate, cols="2000":"2020",  
 names\_to = "year",  
 values\_to = "literacy\_rate")  
water\_invest1 <- pivot\_longer(water\_invest, cols="2000":"2020",  
 names\_to = "year",  
 values\_to = "water\_invest")  
water\_productivity1 <- pivot\_longer(water\_productivity, cols="2000":"2020",  
 names\_to = "year",  
 values\_to = "water\_productivity")

#merging data  
merge\_data <- merge(mortality\_rate1, health\_expenditure1, by = c("country", "year"), all = TRUE)  
merge\_data <- merge(merge\_data, health\_expenditure\_per1, by = c("country", "year"), all = TRUE)  
merge\_data <- merge(merge\_data, education\_expenditure1, by = c("country", "year"), all = TRUE)  
merge\_data <- merge(merge\_data, domestic\_health\_expenditure1, by = c("country", "year"), all = TRUE)  
merge\_data <- merge(merge\_data, literacy\_rate1, by = c("country", "year"), all = TRUE)  
merge\_data <- merge(merge\_data, water\_invest1, by = c("country", "year"), all = TRUE)  
merge\_data <- merge(merge\_data, water\_productivity1, by = c("country", "year"), all = TRUE)  
merge\_data <- merge(merge\_data, vacinnation, by = c("country", "year"), all = TRUE)

glimpse(merge\_data)

## Rows: 7,403  
## Columns: 12  
## $ country <chr> "Abkhazia", "Afghanistan", "Afghanistan", …  
## $ year <chr> "2015", "2000", "2001", "2002", "2003", "2…  
## $ mortality\_rate <dbl> NA, 129.3, 125.3, 121.2, 117.0, 112.8, 108…  
## $ health\_expenditure <dbl> NA, NA, NA, 17.00759, 17.81492, 21.42946, …  
## $ health\_expenditure\_per <dbl> NA, NA, NA, 9.443391, 8.941258, 9.808474, …  
## $ education\_expenditure <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ domestic\_health\_expenditure <dbl> NA, NA, NA, 85.37560, 86.06919, 84.52759, …  
## $ literacy\_rate <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ water\_invest <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ water\_productivity <dbl> NA, NA, NA, 0.3725069, 0.4054078, 0.411140…  
## $ immunazation <int> NA, 24, 33, 36, 41, 50, 58, 58, 63, 64, 63…  
## $ GDP\_per\_capita <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…

skimmed <- skim\_to\_wide(merge\_data)

## Warning: 'skim\_to\_wide' is deprecated.  
## Use 'skim()' instead.  
## See help("Deprecated")

skimmed

Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 7403 |
| Number of columns | 12 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 2 |
| numeric | 10 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| country | 0 | 1 | 4 | 52 | 0 | 385 | 0 |
| year | 0 | 1 | 4 | 4 | 0 | 22 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| mortality\_rate | 2279 | 0.69 | 40.90 | 4.085000e+01 | 1.80 | 9.80 | 2.42000e+01 | 6.252000e+01 | 2.285000e+02 | ▇▂▁▁▁ |
| health\_expenditure | 2477 | 0.67 | 876.37 | 1.567730e+03 | 4.45 | 62.73 | 2.40640e+02 | 7.461300e+02 | 1.170241e+04 | ▇▁▁▁▁ |
| health\_expenditure\_per | 2477 | 0.67 | 6.14 | 2.740000e+00 | 1.26 | 4.24 | 5.40000e+00 | 7.720000e+00 | 2.423000e+01 | ▇▅▁▁▁ |
| education\_expenditure | 5669 | 0.23 | 90.86 | 7.220000e+00 | 32.81 | 88.92 | 9.22500e+01 | 9.509000e+01 | 1.000000e+02 | ▁▁▁▁▇ |
| domestic\_health\_expenditure | 2477 | 0.67 | 42.30 | 1.852000e+01 | 0.52 | 28.05 | 4.22700e+01 | 5.585000e+01 | 8.794000e+01 | ▃▆▇▆▂ |
| literacy\_rate | 5742 | 0.22 | 80.21 | 1.719000e+01 | 14.38 | 66.54 | 8.54000e+01 | 9.450000e+01 | 1.000000e+02 | ▁▁▃▃▇ |
| water\_invest | 7117 | 0.04 | 953377952.62 | 1.280216e+09 | 0.00 | 96250000.00 | 3.67710e+08 | 1.338034e+09 | 6.272480e+09 | ▇▂▁▁▁ |
| water\_productivity | 2827 | 0.62 | 56.87 | 1.424400e+02 | 0.22 | 6.65 | 1.56500e+01 | 4.772000e+01 | 3.072790e+03 | ▇▁▁▁▁ |
| immunazation | 2978 | 0.60 | 86.79 | 1.470000e+01 | 19.00 | 82.00 | 9.30000e+01 | 9.700000e+01 | 9.900000e+01 | ▁▁▁▂▇ |
| GDP\_per\_capita | 3753 | 0.49 | 19489.65 | 2.203415e+04 | 251.09 | 4219.56 | 1.15329e+04 | 2.762972e+04 | 3.032066e+05 | ▇▁▁▁▁ |

#including counties of the world  
all\_countries <- c("Afghanistan", "Albania", "Algeria", "Andorra", "Angola", "Antigua and Barbuda",  
 "Argentina", "Armenia", "Australia", "Austria", "Azerbaijan", "Bahamas", "Bahrain",  
 "Bangladesh", "Barbados", "Belarus", "Belgium", "Belize", "Benin", "Bhutan",  
 "Bolivia", "Bosnia and Herzegovina", "Botswana", "Brazil", "Brunei", "Bulgaria",  
 "Burkina Faso", "Burundi", "Cabo Verde", "Cambodia", "Cameroon", "Canada",  
 "Central African Republic", "Chad", "Chile", "China", "Colombia", "Comoros",  
 "Congo", "Costa Rica", "Croatia", "Cuba", "Cyprus", "Czech Republic", "Denmark",  
 "Djibouti", "Dominica", "Dominican Republic", "East Timor", "Ecuador", "Egypt",  
 "El Salvador", "Equatorial Guinea", "Eritrea", "Estonia", "Eswatini", "Ethiopia",  
 "Fiji", "Finland", "France", "Gabon", "Gambia", "Georgia", "Germany", "Ghana",  
 "Greece", "Grenada", "Guatemala", "Guinea", "Guinea-Bissau", "Guyana", "Haiti",  
 "Honduras", "Hungary", "Iceland", "India", "Indonesia", "Iran", "Iraq", "Ireland",  
 "Israel", "Italy", "Jamaica", "Japan", "Jordan", "Kazakhstan", "Kenya", "Kiribati",  
 "Korea, North", "Korea, South", "Kosovo", "Kuwait", "Kyrgyzstan", "Laos", "Latvia",  
 "Lebanon", "Lesotho", "Liberia", "Libya", "Liechtenstein", "Lithuania", "Luxembourg",  
 "Madagascar", "Malawi", "Malaysia", "Maldives", "Mali", "Malta", "Marshall Islands",  
 "Mauritania", "Mauritius", "Mexico", "Micronesia", "Moldova", "Monaco", "Mongolia",  
 "Montenegro", "Morocco", "Mozambique", "Myanmar", "Namibia", "Nauru", "Nepal",  
 "Netherlands", "New Zealand", "Nicaragua", "Niger", "Nigeria", "North Macedonia",  
 "Norway", "Oman", "Pakistan", "Palau", "Panama", "Papua New Guinea", "Paraguay",  
 "Peru", "Philippines", "Poland", "Portugal", "Qatar", "Romania", "Russia", "Rwanda",  
 "Saint Kitts and Nevis", "Saint Lucia", "Saint Vincent and the Grenadines", "Samoa",  
 "San Marino", "Sao Tome and Principe", "Saudi Arabia", "Senegal", "Serbia", "Seychelles",  
 "Sierra Leone", "Singapore", "Slovakia", "Slovenia", "Solomon Islands", "Somalia",  
 "South Africa", "South Sudan", "Spain", "Sri Lanka", "Sudan", "Suriname", "Sweden",  
 "Switzerland", "Syria", "Taiwan", "Tajikistan", "Tanzania", "Thailand", "Togo",  
 "Tonga", "Trinidad and Tobago", "Tunisia", "Turkey", "Turkmenistan", "Tuvalu",  
 "Uganda", "Ukraine", "United Arab Emirates", "United Kingdom", "United States",  
 "Uruguay", "Uzbekistan", "Vanuatu", "Vatican City", "Venezuela", "Vietnam",  
 "Yemen", "Zambia", "Zimbabwe")  
  
merge\_data <- subset(merge\_data, country %in% all\_countries)

#saving the final data  
write.csv(merge\_data, "data/my\_data.csv")

#remove rows with all na  
filtered\_data <- merge\_data %>%  
 select(-country, -year) %>%  
 filter(rowSums(is.na(.)) != ncol(.))

# Create the knn imputation model on the training data  
preProcess\_missingdata\_model <- preProcess(filtered\_data, method='knnImpute')  
preProcess\_missingdata\_model

## Created from 39 samples and 10 variables  
##   
## Pre-processing:  
## - centered (10)  
## - ignored (0)  
## - 5 nearest neighbor imputation (10)  
## - scaled (10)

# Use the imputation model to predict the values of missing data points  
my\_data\_imputed <- predict(preProcess\_missingdata\_model, newdata = filtered\_data)  
anyNA(my\_data\_imputed)

## [1] FALSE

#saving the imputed data  
write.csv(my\_data\_imputed, "data/my\_data\_imputed.csv")