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
```

```r
#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)
```

```r
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_rat	227	0.69	40.90	4.085	1.8	9.80	2.420	6.252	2.285	▙

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
e	9			000e+01	0		00e+01	000e+01	000e+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")
```