Cluster Analysis Assignment

Nyasha Mashanda

2020-10-15

Contents

List of Figures

List of Tables

Abstract

The goal of this assignment is to predict if an individual will develop Coronary Heart Disease (CHD) over a 10 year period using various risk factors. This will be achieved by using support vector machines (SVM) and neural networks (NN).

There are three kinds of SVM models that will be built: Logistic Regression without regularization, Lasso Logistic Regression and a model using all variables. The three model types will also be used in building a neural network. In order to make sure that the models are optimised, the support vector machine will be tuned using tune.svm from the e1071 package and the neural networks will be tuned using h2o.grid from the h2o package. The best model will be selected based on classification accuracy.

```r
# Display this on the first page to see what and from where each variable was collected.
owid_covid_codebook <- read_csv("owid-covid-codebook.csv")
```

```
## Parsed with column specification:
## cols(
##   column = col_character(),
##   description = col_character(),
##   source = col_character()
## )
```

```r
print.data.frame(owid_covid_codebook)
```

```
##                              column
## 1                          iso_code
## 2                         continent
## 3                          location
## 4                              date
## 5                       total_cases
## 6                         new_cases
## 7                new_cases_smoothed
## 8                      total_deaths
## 9                        new_deaths
## 10              new_deaths_smoothed
## 11          total_cases_per_million
## 12            new_cases_per_million
## 13   new_cases_smoothed_per_million
## 14         total_deaths_per_million
## 15           new_deaths_per_million
## 16  new_deaths_smoothed_per_million
## 17               population_density
## 18                       median_age
## 19                    aged_65_older
## 20                    aged_70_older
## 21                   gdp_per_capita
## 22                  extreme_poverty
## 23            cardiovasc_death_rate
## 24              diabetes_prevalence
## 25                   female_smokers
## 26                     male_smokers
## 27           handwashing_facilities
## 28       hospital_beds_per_thousand
## 29                  life_expectancy
## 30          human_development_index
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13                                                                                            New confirmed
## 14
## 15
## 16                                                                                          New deaths attr
## 17                                                                 Number of people divided by land area, r
## 18
## 19                                                                                   Share of the populatio
## 20                                                                                                       Sh
## 21                                              Gross domestic product at purchasing power parity (constant
## 22                                                                         Share of the population living i
## 23                                                                      Death rate from cardiovascular dise
## 24                                                                                                      Dia
## 25
## 26
## 27                                                                Share of the population with basic handwa
## 28                                                                                            Hospital beds
## 29
## 30 Summary measure of average achievement in key dimensions of human development: a long and healthy life,
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17                                          World Bank - World Development Indicators, sourced from Food an
## 18                                                                                           UN Population
## 19 World Bank - World Development Indicators, based on age/sex distributions of United Nations Population D
## 20                        United Nations, Department of Economic and Social Affairs, Population Division (2
## 21                                                World Bank - World Development Indicators, source from Wo
## 22                                                               World Bank - World Development Indicators,
## 23                                                                 Global Burden of Disease Collaborative N
## 24                                                   World Bank - World Development Indicators, sourced fro
## 25                                World Bank - World Development Indicators, sourced from World Health Orga
## 26                                World Bank - World Development Indicators, sourced from World Health Orga
## 27
## 28                                                                                   OECD, Eurostat, World
## 29                                                                                              James C. Ri
## 30
```

```
#View(owid_covid_codebook)
```

1 What about standardising the data????

2 Exploratory Data Analysis

2.1 Distribution

2.2 Missing values

3 Cluster Analysis

3.1 K-means

3.2 K-mediods

3.3 Clara

3.4 DBSCAN

4 Dimension Reduction

4.1 Results

The data shows that handwashing_facilties, extreme poverty, male smokers and female smokers are among the columns with the highest percentages of missing data.

Over 78% of rows have missing data therefore simply omitting the rows with missing will lead to a loss a huge size of important data.

Population density missing values will be imputed using the data from wikipedia on the following site https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population_density

It is important to indicate that the data was given in 2019 but given that population densities change very slowly, this is better than replacing with an average/median.

Median age missing values will be imputed using the data found from the following wikipedia site: https://en.wikipedia.org/wiki/List_of_countries_by_median_age

Median ages data was added using data from the 2019 data https://www.cia.gov/library/publications/the-world-factbook/fields/343rank.html

Only data for Syria was found.

Median age above 75 years old was ignored

Searching the source of the data maps to a site: https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-codebook.csv which says the gdp_per_capita Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available. Given that the most recent data from the world bank (https://databank.worldbank.org/source/jobs/Series/NY.GDP.PCAP.PP.KD#) which matches this data is for the year 2016, the data was used for imputing the data. Note the data is not available

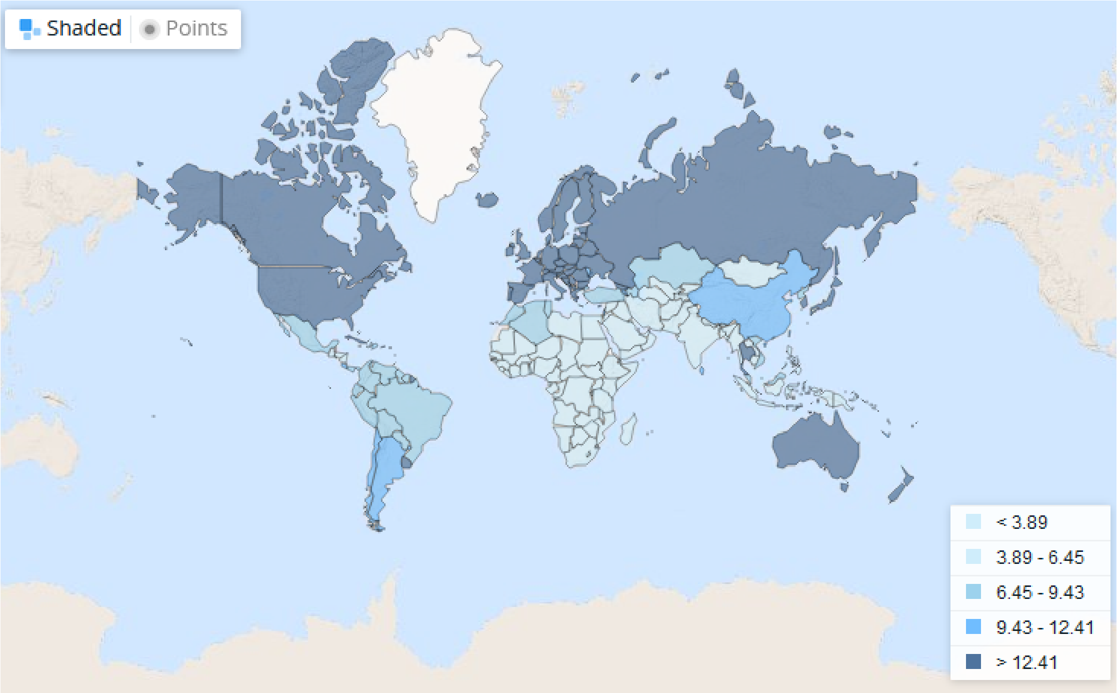
Extreme poverty is a population with an income of less than $1.90 a day. From the data the Share of the population living in extreme poverty, most recent year available since 2010. The data from the following site https://en.wikipedia.org/wiki/List_of_countries_by_percentage_of_population_living_in_poverty the data is very similar to that available without missing values, therefore it will be better to use the data for imputing.

Most countries with 0 percent were having missing values

This was skipped due to insufficient data to fill in the missing values.

female and male smokers info not enough handwashing facilities skipped because data is not enough. hospital bed per thousand informatiion was found to be too old for example Chad which had data for 2005. Given that there are many years between the time the data was collected and this year the data was ignored.

From the world bank data on life expectancy (https://data.worldbank.org/indicator/SP.DYN.LE00.IN), the life expectancy of Guernsey, Jersey and Kosovo are 82.6, 80.6, 71.95 years respectively.



The results show that handwashing_facilites is missing 56.25% of the values followed by extreme_poverty missing 42.3% of the values. In total, 78.4% of the samples have missing values.

The histograms also suggest that there is a negative value in new_cases_smoothed_per_million.

Upon further investigation, Luxembourg seems to have negative values for new_cases_smoothed and new_cases_smoothed_per_million. This is probably because the new_cases_smoothed_per_million is derived from new_cases_smoothed.

The mice package allows multivariate imputation by chained equations

MICE assumes that the missing data are Missing at Random (MAR), which means that the propensity for a data point to be missing is not related to the missing data but is related to some of the observed data. https://www.theanalysisfactor.com/mar-and-mcar-missing-data/

Therefore it is important to determine which columns have data points that may be missing at random. The imputation process is not a one size fits all method.

There is also the Amelia package which also makes an assumption that the missing data is random in nature (MSR)

Hmisc is a multiple purpose package useful for data analysis, high – level graphics, imputing missing values, advanced table making, model fitting & diagnostics (linear regression, logistic regression & cox regression) etc. Hmisc assumes linearity in the variables being predicted.

mi (Multiple imputation with diagnostics) package provides several features for dealing with missing values. Like other packages, it also builds multiple imputation models to approximate missing values. And, uses predictive mean matching method like the mice package.

The following variables need need impiutation using some of the methods.

4.2 female__smokers and male__smokers, cardiovasc__death__rate and diabetes prevalence are MACR variables therefore the MICE package will be used on this data.

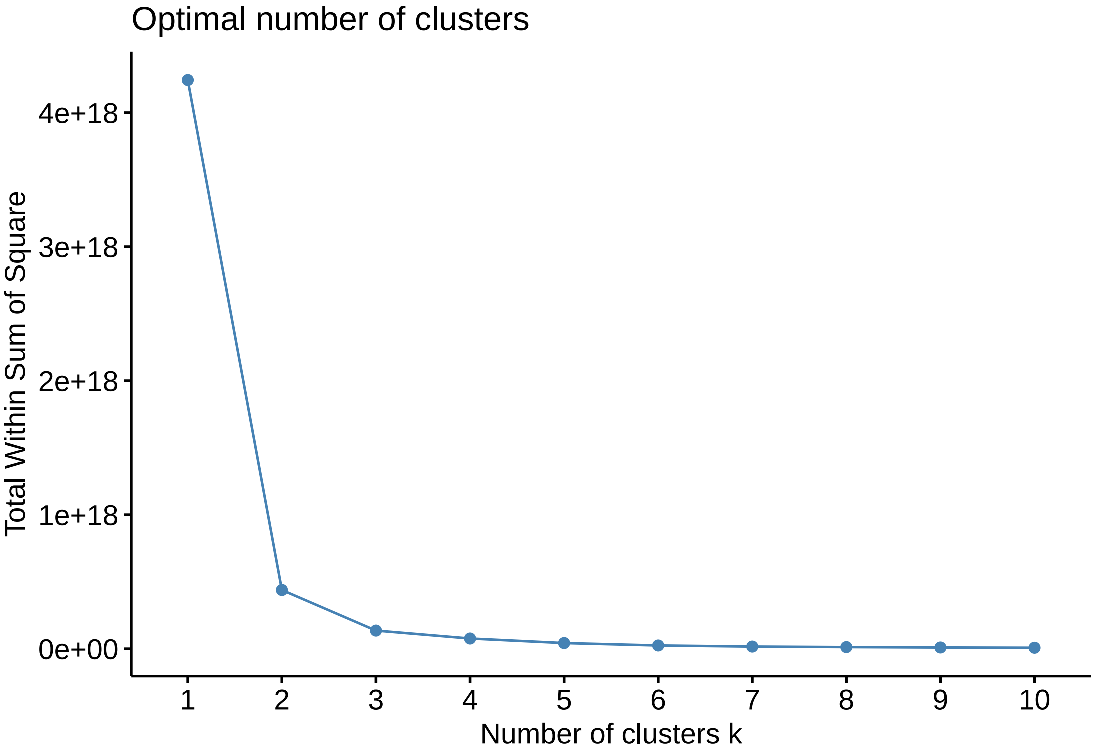
various packages will be used for imputation and the one that yields the best result will be chosen.

This will be support for dimension reduction.

It is not enough to try out one package for imputing data therefore try out more! The question is if any one value of the observation is missing, will it affect the missingness of a specific variable.

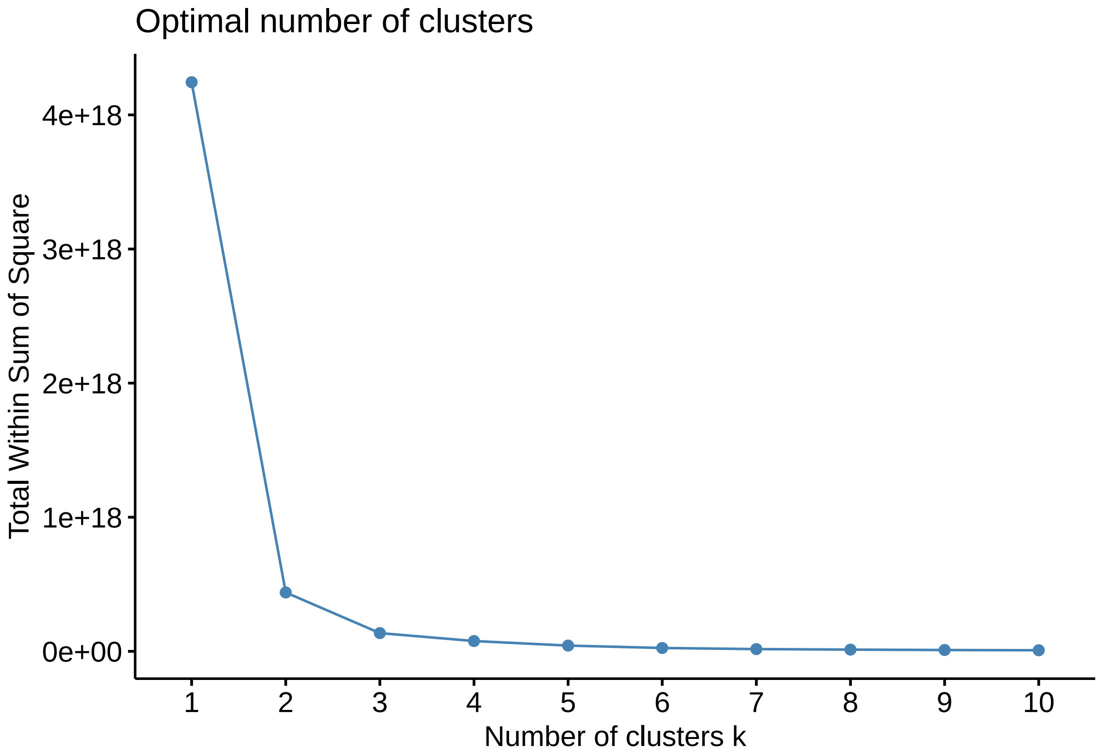
The percentage of population aged__65__older looks to have regional patterns therefore it is justified to use regional medians/averages for imputation. Since aged__70__older can is closely related to aged__65__older it is decided to use the same method.

The randomizing, re-running, averaging and final run is a very good advice.



The dataset is very small and this results in Clara performing worse than PAM.

The randomizing, re-running, averaging and final run is a very good advice.



The dataset is very small and this results in Clara performing worse than PAM.
This code was optimised for reading