IS517 Final Project: World Health

Abhinav Choudhry
Kerstin Wolf

> “The greatest wealth is health.”
>
> – Virgil

Research Questions

1. How do countries compare when evaluating healthcare factors versus economic factors (GDP)?
2. Are we able to successfully predict life expectancy at birth and healthy life expectancy at birth by a) health infrastructure and b) health behaviors?
3. Is there a significant difference between factors that predict healthy life expectancy at birth versus life expectancy, and what factors cause this?

Problem Setting

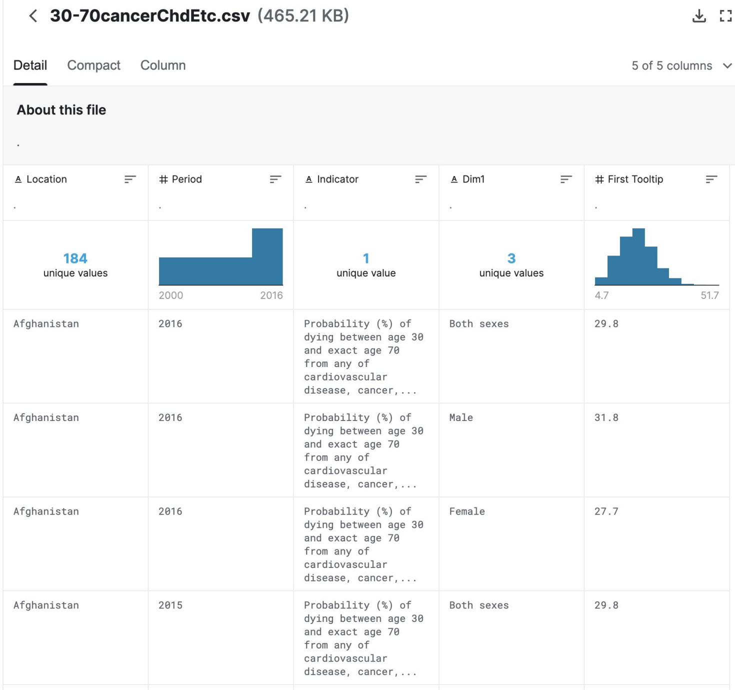
- Industry
  - Healthcare
- Audience
  - Healthcare workers
  - Politicians and government workers
  - Health Organizations
  - General population

Importance of Study

- Help governments raise the life expectancy and healthy life expectancy in their country by learning key areas to target and dedicate funds toward
- Help health organizations in better focusing their efforts and funding
- Help hospitals and healthcare workers identify factors that are decreasing healthy life expectancy the most

Datasets

- Kaggle
  - World Health Statistics 2020
- The World Bank
  - Gross Domestic Product 2019 Data



About this file

Location	Period	Indicator	Dim1	First Tooltip
184 unique values	2000 – 2016	1 unique value	3 unique values	4.7 – 51.7
Afghanistan	2016	Probability (%) of dying between age 30 and exact age 70 from any of cardiovascular disease, cancer,...	Both sexes	29.8
Afghanistan	2016	Probability (%) of dying between age 30 and exact age 70 from any of cardiovascular disease, cancer,...	Male	31.8
Afghanistan	2016	Probability (%) of dying between age 30 and exact age 70 from any of cardiovascular disease, cancer,...	Female	27.7
Afghanistan	2015	Probability (%) of dying between age 30 and exact age 70 from any of cardiovascular disease, cancer,...	Both sexes	29.8

World Health Statistics 2020 Data

Consists of 39 CSV files.

Most files contain information from around 2000 to 2019.

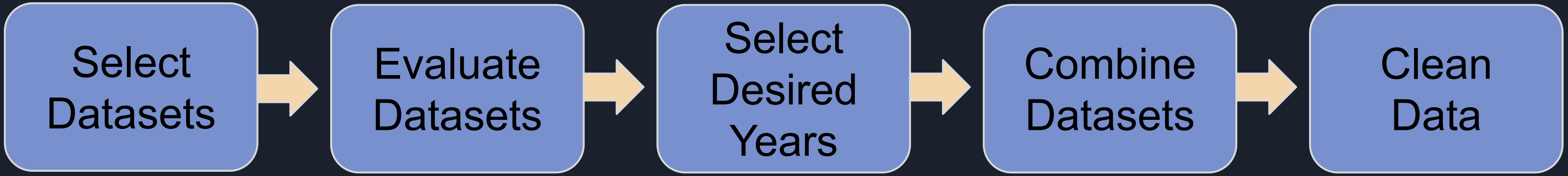
- Air Pollution Death Rate
- Number of Dentists
- Infant Mortality Rate
- Life Expectancy at Birth
- Number of Pharmacists
- Road Traffic Death Rate
- Alcohol Consumption
- Violence Against Women

Data Characteristics

- Numeric data
- No Null Values
- Some zero values are incorrect
- Each CSV file has only one numeric attribute
- Total of 39 attributes in dataset
- Number of rows in each file = Countries * number of years * sex (only in some of the files)

Attribute Selection

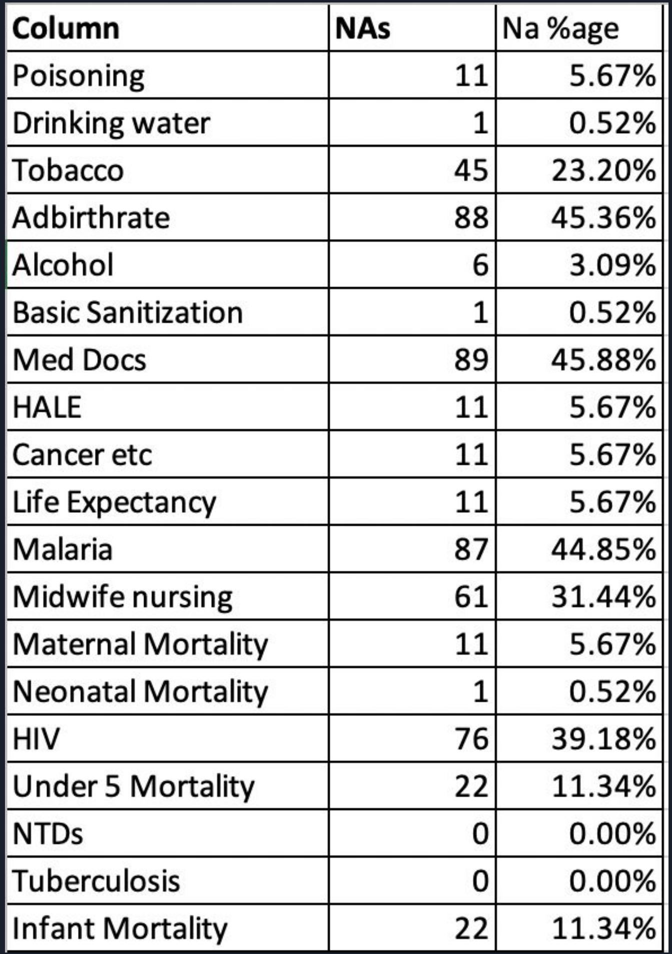
- Of the 39, only 22 attributes made it through the first thorough inspection.
- 19 attributes met the row and year requirements.
- Attributes were further trimmed to limit the null values.



Tackling Null Values

Merging datasets led to issues with null values.

Solution was to eliminate a couple attributes.



Column	NAs	Na %age
Poisoning	11	5.67%
Drinking water	1	0.52%
Tobacco	45	23.20%
Adbirthrate	88	45.36%
Alcohol	6	3.09%
Basic Sanitization	1	0.52%
Med Docs	89	45.88%
HALE	11	5.67%
Cancer etc	11	5.67%
Life Expectancy	11	5.67%
Malaria	87	44.85%
Midwife nursing	61	31.44%
Maternal Mortality	11	5.67%
Neonatal Mortality	1	0.52%
HIV	76	39.18%
Under 5 Mortality	22	11.34%
NTDs	0	0.00%
Tuberculosis	0	0.00%
Infant Mortality	22	11.34%

Final Attributes Going Into Modeling

- Location
- Poisoning
- Tuberculosis
- Under 5 Mortality
- Drinking Water
- Alcohol
- Basic Sanitization
- Cancer Etc.
- Healthy Life Expectancy
- Infant Mortality
- Life Expectancy
- Maternal Mortality
- NTDs
- Neonatal Mortality
- Tobacco

World Bank GDP 2019 Data

Single CSV file containing about 205 rows and 3 columns not including the index.

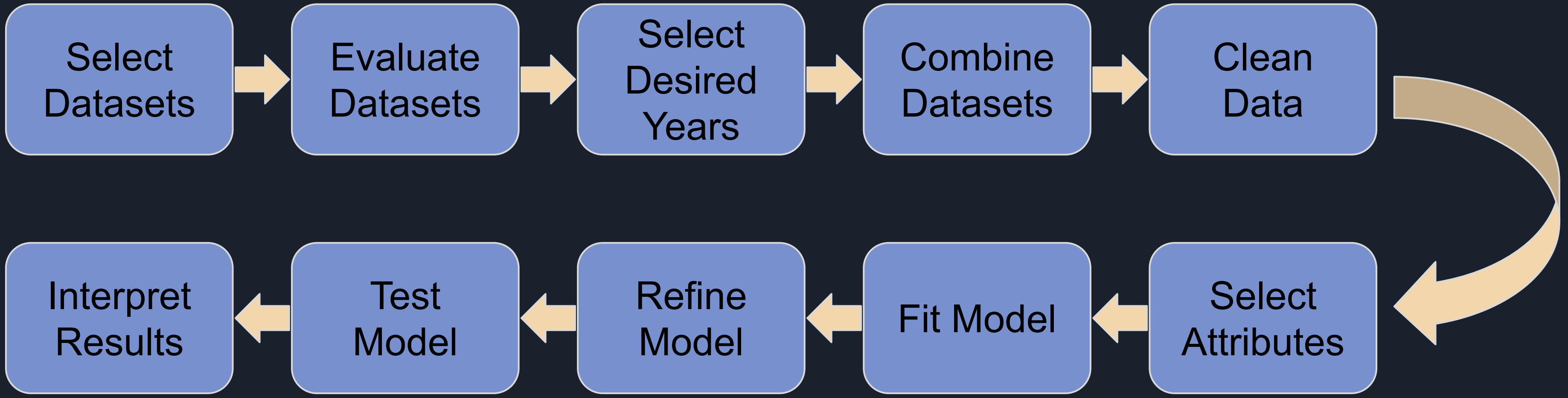
File contains country and region GDP data.

Ranking

Country Name

GDP (in Millions of US Dollars)

Analysis


Select Datasets
Evaluate Datasets
Select Desired Years
Combine Datasets
Clean Data
Select Attributes
Fit Model
Refine Model
Test Model
Interpret Results

Methods and Models

- Linear Regression
- Random Forests
- Bagging and Boosting
- Principal Components Analysis
- Support Vector machines
- Logistic Regression

Also added a binary class for Healthy life expectancy

Model Performance

- Linear Regression using 6 parameters -> R square of 80.4% and 3.7% error rate
- Linear regression with Health Behaviour -> R square of 17.8% & 7.6% error rate
- Linear regression with mortalities -> R square of 84% . Multicollinearity might exist
- Principal Components Analysis -> 13 PCAs and 67% for PCA1
- Logistic Regression -> 16.25% on classification

Model Performance (Continued)

- Random Forests -> 8.13% for bernoulli & 2.3% for gaussian
- Bagging -> 2.3% error rate
- Boosting -> 2.25% error rate
- Support Vector machines -> 15% error rate on classification using 91 support vectors at min cost

Findings: Model Fit

Random Forest performs best with lowest error rate in both regression and classification.

Support Vector machines run better than Logistic Regression.

Linear Regression performs well.

Linear Regression on GDP data has poorer prediction than health infrastructure and behaviours

Findings: Factors

Tuberculosis, Poisoning, Basic Sanitization, and Alcohol are most significant by general linear regression.

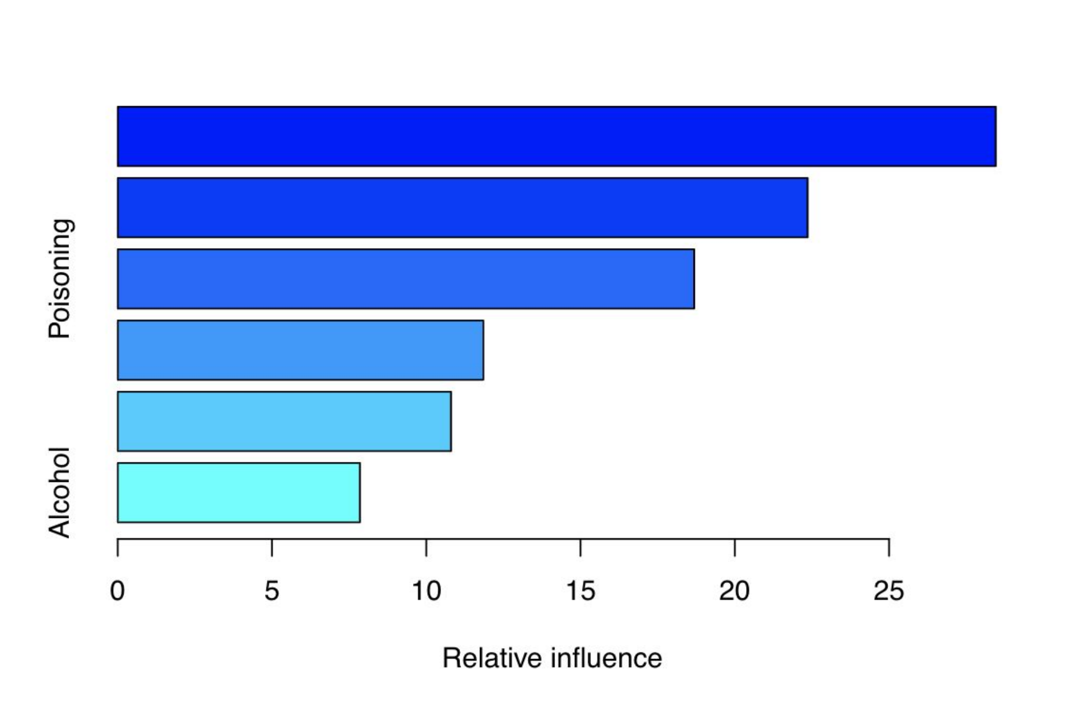
Regression Model on health behaviours lists alcohol most significant but tobacco too.

Basic Sanitization, Drinking water, and Poisoning predicted as most significant by Random Forests, Boosting, and Bagging.

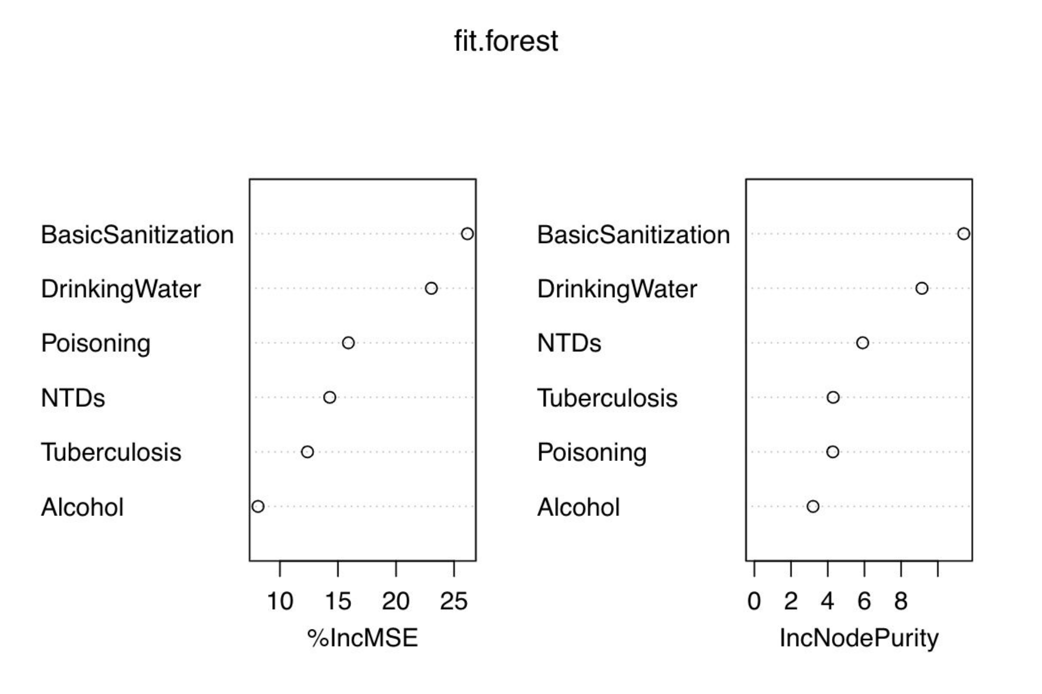
PCA1 explains the most variance at 66.9% followed by PCA 2 at 7.7%, PCA3 at 7.28% and PCA4 at 6.5%. The other components explain less than 5%.

Looking at PCA1 decomposition, we see that it is made of different types of mortalities and intuitively we can understand that they would be close to each other for the same country.

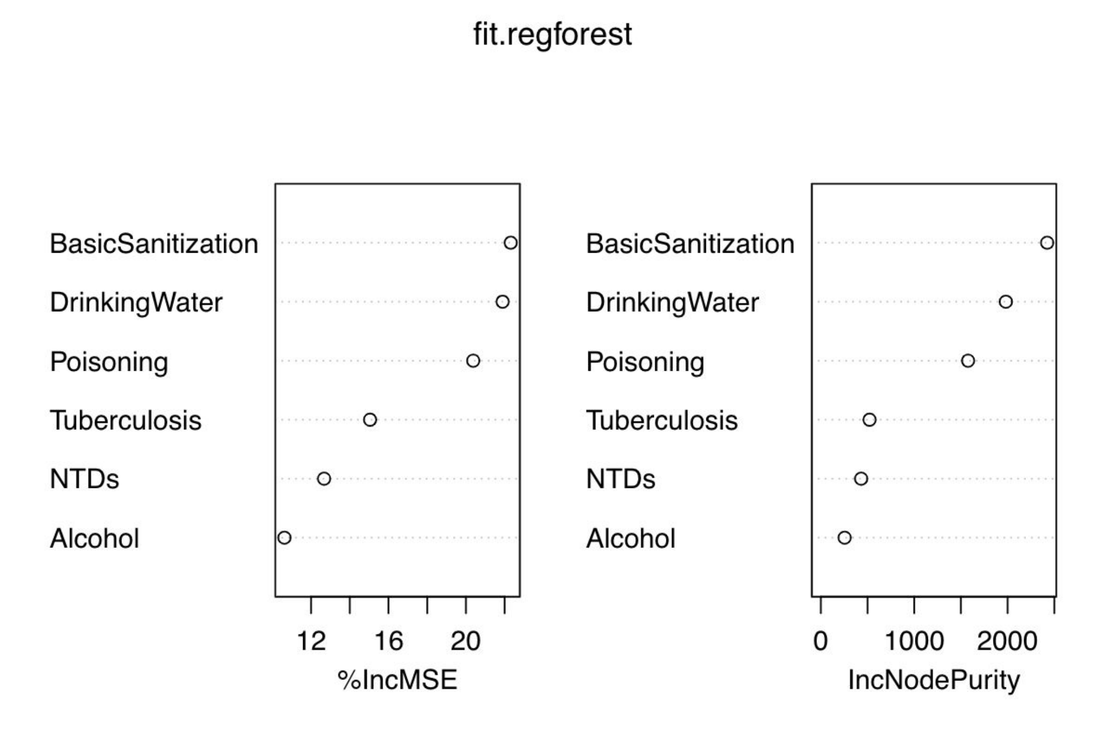
Boosting



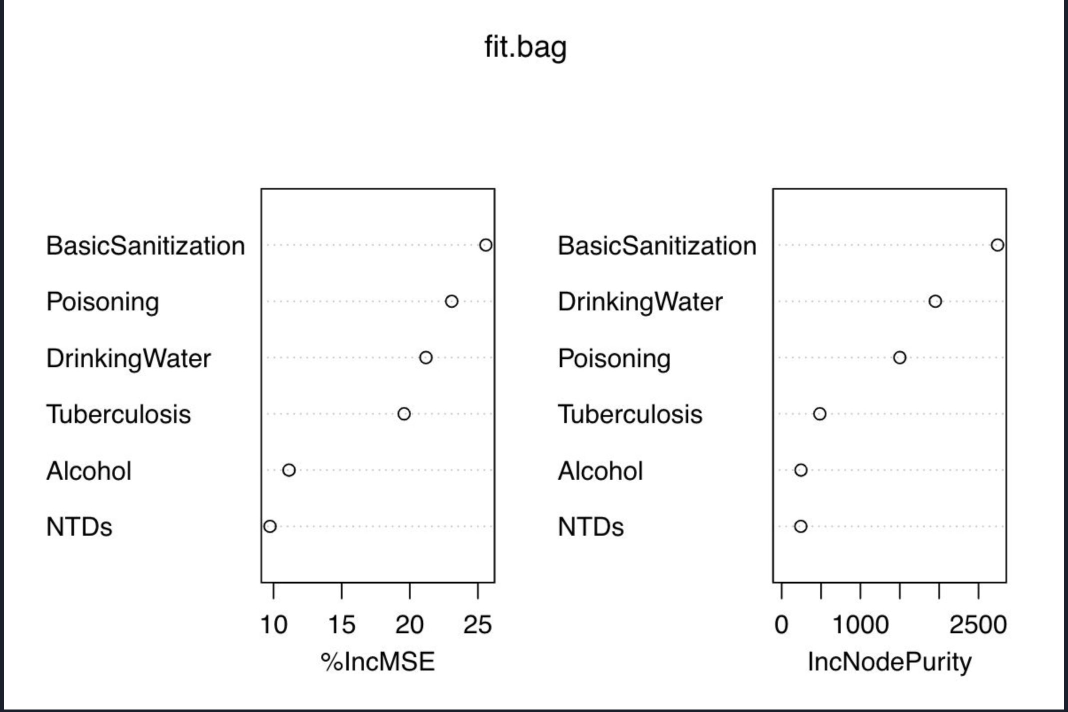
Random Forest



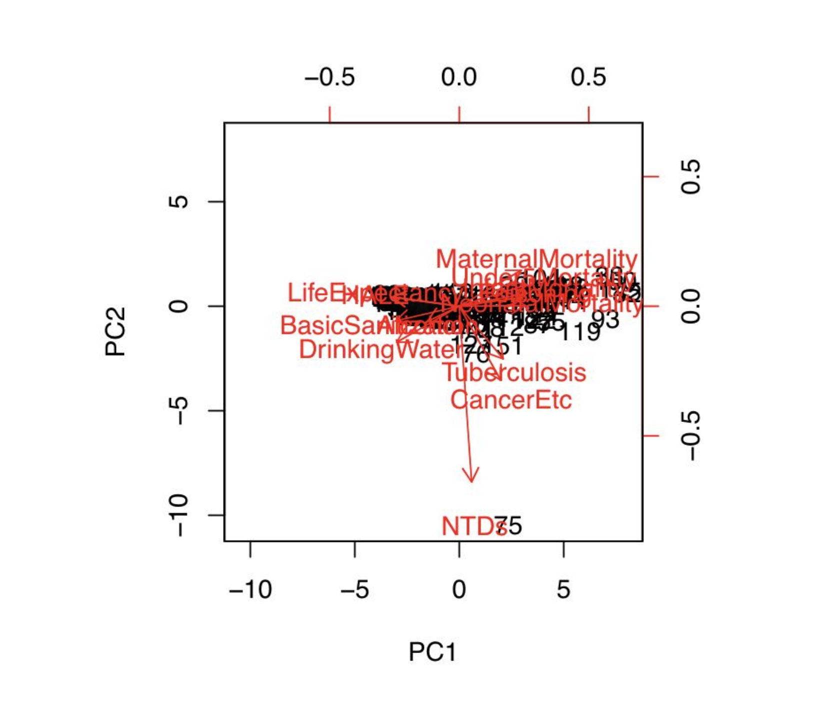
Random Forest



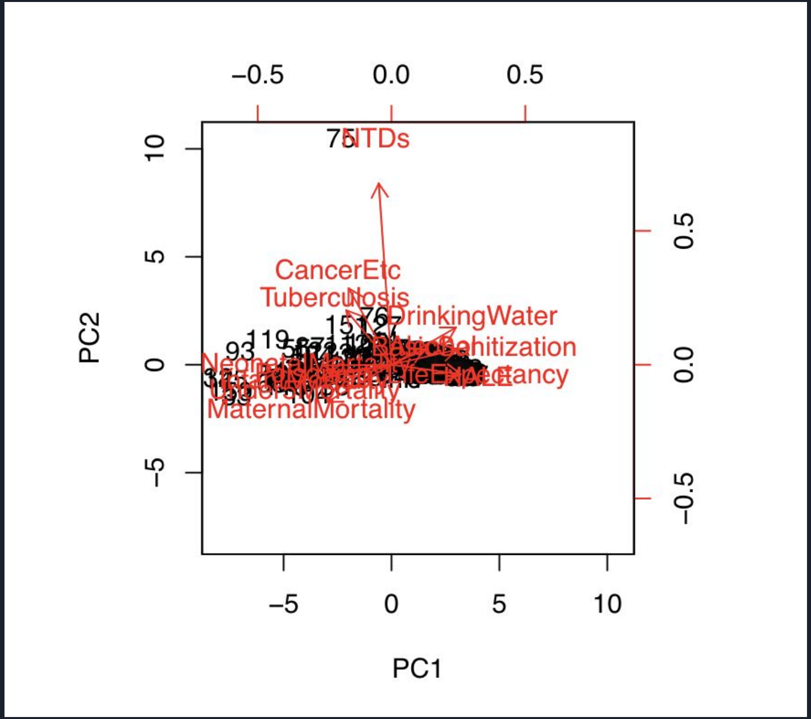
Bagging


fit.bag
BasicSanitization
Poisoning
DrinkingWater
Tuberculosis
Alcohol
NTDs
10
15
20
25
%IncMSE
BasicSanitization
DrinkingWater
Poisoning
Tuberculosis
Alcohol
NTDs
0
1000
2500
IncNodePurity

PCA


−0.5
0.0
0.5
0.5
0.0
−0.5
5
0
−5
−10
−10
−5
0
5
PC1
PC2
MaternalMortality
Under5Mortality
LifeExpectancy
NeonatalMortality
BasicSanitation
Alcohol
DrinkingWater
Tuberculosis
CancerEtc
NTDs
75
93
119
51
76
36

PCA



−0.5
0.0
0.5
10
5
0
−5
PC2
0.5
0.0
−0.5
−5
0
5
10
PC1
75
NTDs
CancerEtc
Tuberculosis
DrinkingWater
93
119
151
Sanitization
LifeExpectancy
MaternalMortality

Next Steps and Recommendations

- Add overweight and sedentary behaviour attributes missing from Kaggle set.
- Add economic factors to our data set and run models including that and checking results.
- Also check for clusters.
- Also run on other dependent variables to answer all RQs.
- Both Health Behaviours and Health Infrastructure are significant for healthy life expectancy.
- Could be an alternative to economic prediction models

Reflection

- Selecting which attributes to utilize while keeping in mind the year and dealing a small row count is very time consuming and labor intensive.
- Merging many datasets with different row counts into one can be disastrous in regard to null values.
- Handling null values in a small dataset is significantly more challenging than with a large dataset.
- It is essential to run models again after removing parameters and keep tweaking it

Questions?