Phân tích các yếu tố ảnh hưởng đến tuổi thọ

- Author: Nguyen Minh Khoa
- Gmail: nmk14062001@gmail.com

1. Tóm tắt

Có thể thấy rằng trong 15 năm qua, lĩnh vực y tế đã có sự phát triển vượt bậc, điều đó giúp cải thiện tỷ lệ tử vong ở người, đặc biệt là ở các nước đang phát triển. Vì thế tôi thực hiện nghiên cứu với các phương pháp thống kê cơ bản và machine learning với mục tiêu tìm ra được các yếu tố chính làm ảnh hưởng tới tuổi thọ trung bình.

2. Giới thiệu

Với câu hỏi nghiên cứu chung là tìm ra các yếu tố ảnh hưởng tới tuổi thọ. Tôi sẽ sử dụng phương pháp thống kê suy diễn để thực hiện 2 câu hỏi nghiên cứu.

1. Tuổi thọ trung bình giữa các nước phát triển và đang phát triển có thực sự khác biệt hay không?
- **Output**: Tuổi thọ trung bình (Life.expectancy)
- **Input**: Trạng thái của quốc gia (Status)
2. Tác động của việc đi học đối với tuổi thọ trung bình là như thế nào?
- **Output**: Tuổi thọ trung bình (Life.expectancy)
- **Input**: Số năm đi học (Schooling)

Áp dụng thêm phương pháp machine learning để dự đoán tuổi thọ trung bình của dân số ở các quốc gia.

- **Output**: Life.expectancy
- **Input**: Adult.Mortality, infant.deaths, Alcohol, percentage.expenditure, Hepatitis.B, Measles, BMI, under.five.deaths, Polio, Total.expenditure, Diphtheria, HIV.AIDS, GDP, Population, thinness..1.19.years, thinness.5.9.years, Income.composition.of.resources, Schooling
- Các thuật toán được sử dụng: Linear regression, Ridge regression, Lasso regression.

3. Dữ liệu

Dữ liệu được thu thập bởi WHO và trang web của liên hợp quốc. Gồm 2938 dòng và 22 cột.

1. **Country**: quốc gia
2. **Year**: năm
3. **Status**: trạng thái đã phát triển hay đang phát triển
4. **Life.expectancy**: tuổi thọ trung bình
5. **Adult.Mortality**: số ca tử vong ở người trưởng thành (từ 15-60 tuổi) trên 1000 người

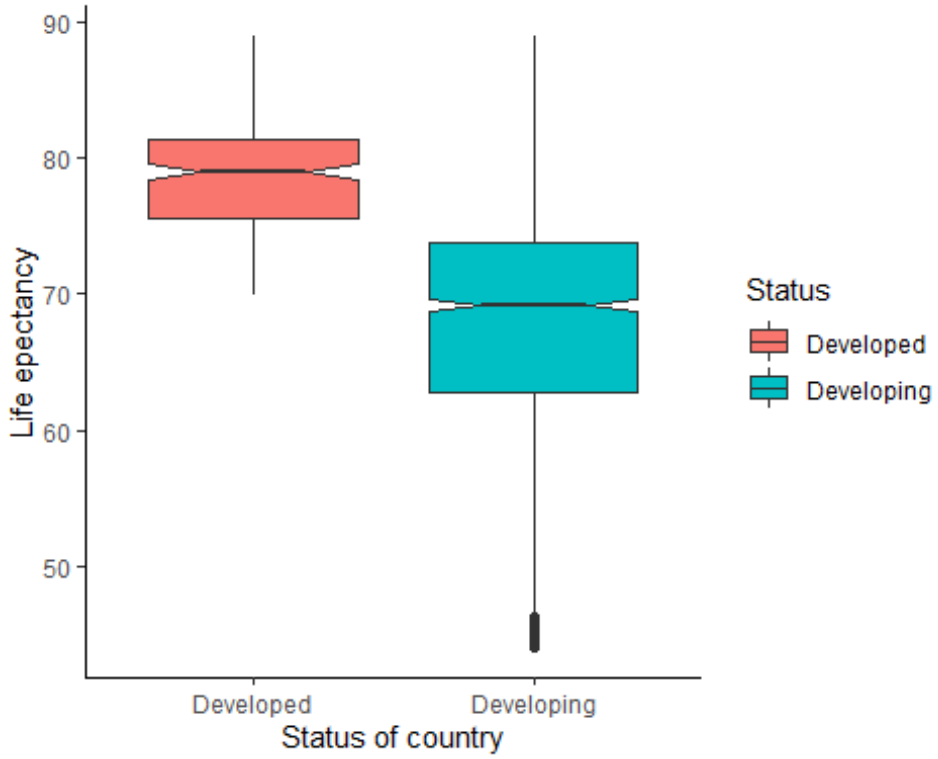
6. **infant.deaths**: số ca tử vong ở trẻ em trên 1000 người
7. **Alcohol**: mức tiêu thụ rượu tính theo bình quân đầu người (lit)
8. **percentage.expenditure**: chi tiêu cho y tế tính theo phần trăm tổng sản phẩm quốc nội bình quân đầu người
9. **Hepatitis.B**: tỷ lệ tiêm chủng viêm gan B ở trẻ 1 tuổi (%)
10. **Measles**: số cả sởi ghi nhận
11. **BMI**: chỉ số khối cơ thể
12. **under.five.deaths**: sô ca trẻ dưới 5 tuổi tử vong trên 1000 người
13. **Polio**: tỷ lệ tiêm chủng viêm bại liệt ở trẻ 1 tuổi (%)
14. **Total.expenditure**: chi tiêu chung của chính phủ cho y tế tính theo tỷ lệ phần trăm trong tổng chi tiêu của chính phủ
15. **Diphtheria**: tỷ lệ tiêm chủng viêm bạch hầu, ho gà, uống ván ở trẻ 1 tuổi (%)
16. **HIV.AIDS**: số ca tử vong ở trẻ em (từ 0-4 tuổi) do HIV/AIDS trên 1000 người
17. **GDP**: tổng sản phẩm quốc nội ($)
18. **Population**: dân số của quốc gia
19. **thinness..1.19.years**: tỷ lệ gầy ốm của thanh niên (từ 10-19 tuổi)
20. **thinness.5.9.years**: tỷ lệ gầy ốm của thanh niên (từ 5-9 tuổi)
21. **Income.composition.of.resources**: chỉ số phát triển con người dựa trên thành phần thu nhập của các nguồn lực (từ 0-1)
22. **Schooling**: số năm đi học

Thực hiện loại bỏ missing values.

Encode biến phân loại.

4. Trực quan hóa dữ liệu và thống kê suy diễn

Câu 1: Tuổi thọ trung bình giữa các nước phát triển và đang phát triển có thực sự khác biệt hay không?



Nhìn vào biểu đồ ta có thể thấy:

- Median của các nước phát triển trong khoảng 78 trong khi những nước đang phát triển chỉ ở mức 70.
- Tuổi thọ trung bình của các nước phát triển chủ yếu trong khoảng 76-82 còn với những nước đang phát triển thì đa số trong khoảng 63-74.
- Có thể thấy rằng sự khác biệt về tuổi thọ giữa 2 nhóm nước là khá rõ rệt.
- Những chấm đen ở đuôi, ứng với những nước đang phát triển cho ta thấy đó có thể là giá trị outlier, những quốc gia này có mức tuổi thọ trung bình là khá thấp, rơi vào khoảng dưới 50 tuổi.

Để kiểm chứng cho câu hỏi được đặt ra ban đầu thì tôi sử dụng thêm phương pháp t test.

- H_0: Không có sự khác biệt về tuổi thọ trung bình giữa các nước phát triển và đang phát triển.
- H_a: Có sự khác biệt về tuổi thọ trung bình giữa các nước phát triển và đang phát triển.

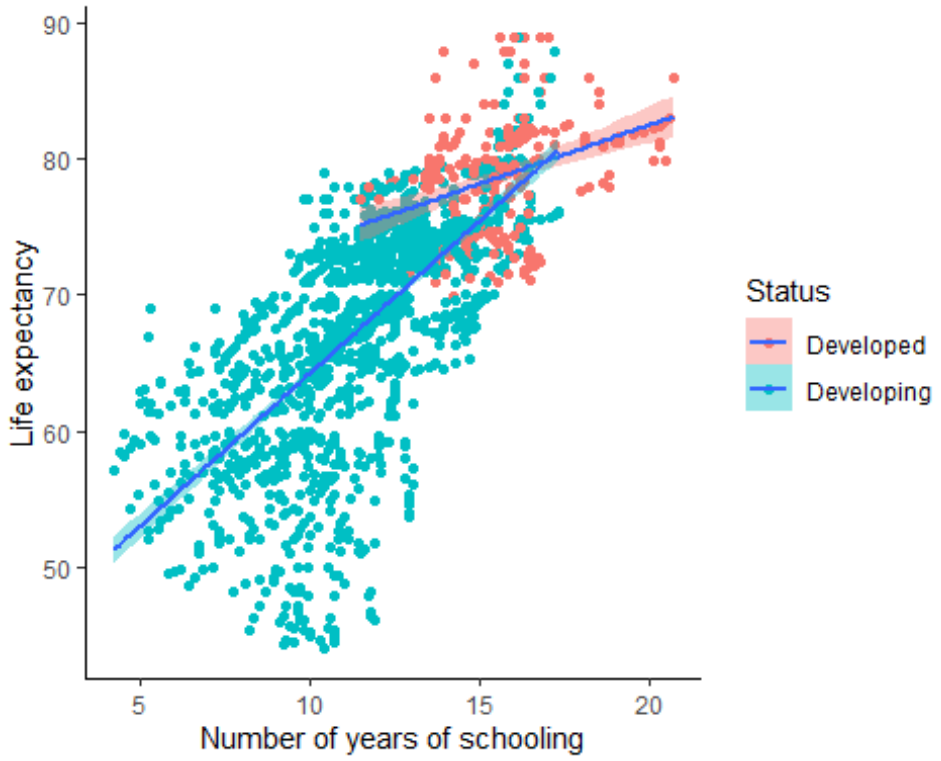
```
##
##  Welch Two Sample t-test
##
```

```
## data:  Life.expectancy by Status
## t = 47.868, df = 1807, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Developed
and group Developing is not equal to 0
## 95 percent confidence interval:
##  11.59118 12.58159
## sample estimates:
##  mean in group Developed mean in group Developing
##                 79.19785                 67.11147
```

Kết luận

- Chỉ số p rất nhỏ, từ đó có thể loại bỏ giả thuyết H_0 và có thể kết luận rằng sự khác biệt về tuổi thọ trung bình giữa các nước phát triển và đang phát triển là do thực tế.
- t = 47.868 cho ta thấy sự khác biệt về tuổi thọ trung bình giữa 2 nhóm nước phát triển và đang phát triển lớn gấp 47.868 lần so với độ lệch chuẩn của sự khác biệt này.
- Mức ý nghĩa 95% có thể hiểu rằng nếu ta lặp lại nghiên cứu 100 lần thì sẽ có 95 nghiên cứu cho thấy sự khác biệt về tuổi thọ trung bình giữa 2 nhóm nước phát triển và đang phát triển sẽ trong khoảng 11.59 - 12.58.

Câu 2: Tác động của việc đi học đối với tuổi thọ trung bình là như thế nào?



Nhìn vào biểu đồ ta có thể thấy:

- Các quốc gia phát triển thì có số năm đi học trong khoảng từ 12-20 năm, tương ứng với số năm đi học đó thì tuổi thọ của người dân ở các quốc gia này cũng khá cao, luôn nằm trong khoảng 70-90 tuổi.
- Trong khi đó những quốc gia đang phát triển có số năm đi học thấp hơn, và trong khoảng từ 5-16 năm thì tuổi thọ trung bình của họ đa phần sẽ trải dài từ 50-80 tuổi.
- Có một số khá ít quốc gia đang phát triển có tuổi thọ từ 80-90 tuổi, đây có thể là do những quốc gia này có tốc độ phát triển nhanh nên đời sống cũng tăng theo.
- Từ biểu đồ ta có thể thấy việc đi học có tác động rất lớn đối với tuổi thọ trung bình của người dân.
- 2 đường chéo thể hiện việc dự đoán tuổi thọ của dân số của các quốc gia.

Để kiểm chứng cho câu hỏi được đặt ra ban đầu thì tôi sử dụng thêm phương pháp cor-test.

- H_0: Không có mối tương quan giữa việc đi học và tuổi thọ trung bình.
- H_a: Có mối tương quan giữa việc đi học và tuổi thọ trung bình.

```
##
##  Pearson's product-moment correlation
##
## data:  Life.expectancy and Schooling
## t = 59.995, df = 2766, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7353198 0.7677234
## sample estimates:
##       cor
## 0.7519755
```

Kết luận

- Chỉ số p rất nhỏ, từ đó có thể loại bỏ giả thuyết H_0 và có thể kết luận rằng số năm đến trường ảnh hưởng tới tuổi thọ trung bình.
- Hệ số tương quan bằng 0.75 cho thấy mối tương quan giữa số năm đến trường và tuổi thọ trung bình là khá cao.
- Từ hệ số tương quan này có thể suy ra hệ số xác định là 0.75^2 = 0.56
- Có thể hiểu rằng 56% sự khác biệt về tuổi thọ có thể giải thích bằng sự khác biệt về số năm đến trường.
- Mức ý nghĩa 95% có thể hiểu rằng nếu ta lặp lại nghiên cứu 100 lần thì sẽ có 95 nghiên cứu cho ra hệ số tương quan trong khoảng 0.74 - 0.77.

5. Mô hình hóa dữ liệu

Tôi sẽ sử dụng 3 mô hình Linear, Ridge và Lasso regression để thực hiện train model.

- Mô hình Linear regression chỉ tập trung vào việc thực hiện quá trình cực tiểu hóa lỗi trên tập huấn luyện, điều đó sẽ dễ dẫn đến tình trạng overfitting.

- Chia dữ liệu thành 2 phần: tập train chiếm 80%, tập test chiếm 20%.
- Áp dụng phương pháp cross validation lên tập train, chia tập train thành 10 phần, 9 phần dùng để train và 1 phần dùng validation.
- Sử dụng độ đo chính để so sánh các mô hình là RMSE và R^2.

```
## Linear Regression
##
## 1320 samples
##   19 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1188, 1188, 1188, 1188, 1186, 1189, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   3.576265  0.8359181  2.750881
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

- Sau khi train ta có thể thấy giá trị RMSE là khoảng 3.576 và R^2 là khoảng 0.836.
- Mô hình Ridge và Lasso sẽ thêm phần chỉnh hóa vào sau mô hình Linear thông thường

- Đối với mô hình Ridge regression thì mô hình này sẽ tránh được trường hợp ma trận suy biến, tăng được khả năng tổng quát hóa của mô hình nhưng lỗi trong tập huấn luyện có thể cao hơn phương pháp Linear thông thường.

```
## glmnet
##
## 1320 samples
##   19 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1188, 1188, 1188, 1188, 1186, 1189, ...
## Resampling results across tuning parameters:
##
##   lambda  RMSE      Rsquared   MAE
##   0.01    3.668390  0.8273919  2.832568
##   0.02    3.668390  0.8273919  2.832568
##   0.03    3.668390  0.8273919  2.832568
##   0.04    3.668390  0.8273919  2.832568
##   0.05    3.668390  0.8273919  2.832568
##   0.06    3.668390  0.8273919  2.832568
##   0.07    3.668390  0.8273919  2.832568
```

```
##   0.08    3.668390  0.8273919  2.832568
##   0.09    3.668390  0.8273919  2.832568
##   0.10    3.668390  0.8273919  2.832568
##   0.11    3.668390  0.8273919  2.832568
##   0.12    3.668390  0.8273919  2.832568
##   0.13    3.668390  0.8273919  2.832568
##   0.14    3.668390  0.8273919  2.832568
##   0.15    3.668390  0.8273919  2.832568
##   0.16    3.668390  0.8273919  2.832568
##   0.17    3.668390  0.8273919  2.832568
##   0.18    3.668390  0.8273919  2.832568
##   0.19    3.668390  0.8273919  2.832568
##   0.20    3.668390  0.8273919  2.832568
##   0.21    3.668390  0.8273919  2.832568
##   0.22    3.668390  0.8273919  2.832568
##   0.23    3.668390  0.8273919  2.832568
##   0.24    3.668390  0.8273919  2.832568
##   0.25    3.668390  0.8273919  2.832568
##   0.26    3.668390  0.8273919  2.832568
##   0.27    3.668390  0.8273919  2.832568
##   0.28    3.668390  0.8273919  2.832568
##   0.29    3.668390  0.8273919  2.832568
##   0.30    3.668390  0.8273919  2.832568
##   0.31    3.668390  0.8273919  2.832568
##   0.32    3.668390  0.8273919  2.832568
##   0.33    3.668390  0.8273919  2.832568
##   0.34    3.668390  0.8273919  2.832568
##   0.35    3.668390  0.8273919  2.832568
##   0.36    3.668390  0.8273919  2.832568
##   0.37    3.668390  0.8273919  2.832568
##   0.38    3.668390  0.8273919  2.832568
##   0.39    3.668390  0.8273919  2.832568
##   0.40    3.668390  0.8273919  2.832568
##   0.41    3.668390  0.8273919  2.832568
##   0.42    3.668390  0.8273919  2.832568
##   0.43    3.668390  0.8273919  2.832568
##   0.44    3.668390  0.8273919  2.832568
##   0.45    3.668390  0.8273919  2.832568
##   0.46    3.668390  0.8273919  2.832568
##   0.47    3.668390  0.8273919  2.832568
##   0.48    3.668390  0.8273919  2.832568
##   0.49    3.668390  0.8273919  2.832568
##   0.50    3.668390  0.8273919  2.832568
##   0.51    3.668390  0.8273919  2.832568
##   0.52    3.668390  0.8273919  2.832568
##   0.53    3.668390  0.8273919  2.832568
##   0.54    3.668390  0.8273919  2.832568
##   0.55    3.668390  0.8273919  2.832568
##   0.56    3.668390  0.8273919  2.832568
##   0.57    3.668390  0.8273919  2.832568
```

```
##   0.58    3.668390  0.8273919  2.832568
##   0.59    3.668390  0.8273919  2.832568
##   0.60    3.668390  0.8273919  2.832568
##   0.61    3.668390  0.8273919  2.832568
##   0.62    3.668390  0.8273919  2.832568
##   0.63    3.668390  0.8273919  2.832568
##   0.64    3.668465  0.8273878  2.832646
##   0.65    3.668704  0.8273737  2.832925
##   0.66    3.668972  0.8273567  2.833249
##   0.67    3.669244  0.8273395  2.833574
##   0.68    3.669521  0.8273221  2.833899
##   0.69    3.669803  0.8273045  2.834226
##   0.70    3.670088  0.8272868  2.834576
##   0.71    3.670363  0.8272700  2.834926
##   0.72    3.670639  0.8272535  2.835294
##   0.73    3.670918  0.8272368  2.835685
##   0.74    3.671202  0.8272198  2.836077
##   0.75    3.671491  0.8272026  2.836473
##   0.76    3.671783  0.8271852  2.836882
##   0.77    3.672078  0.8271678  2.837289
##   0.78    3.672372  0.8271507  2.837696
##   0.79    3.672668  0.8271334  2.838113
##   0.80    3.672969  0.8271159  2.838530
##   0.81    3.673274  0.8270983  2.838948
##   0.82    3.673583  0.8270804  2.839365
##   0.83    3.673897  0.8270623  2.839782
##   0.84    3.674214  0.8270440  2.840199
##   0.85    3.674527  0.8270262  2.840610
##   0.86    3.674837  0.8270089  2.841017
##   0.87    3.675149  0.8269915  2.841425
##   0.88    3.675465  0.8269739  2.841837
##   0.89    3.675785  0.8269561  2.842249
##   0.90    3.676109  0.8269382  2.842661
##   0.91    3.676437  0.8269200  2.843081
##   0.92    3.676769  0.8269016  2.843500
##   0.93    3.677101  0.8268835  2.843917
##   0.94    3.677431  0.8268655  2.844330
##   0.95    3.677765  0.8268475  2.844744
##   0.96    3.678103  0.8268292  2.845168
##   0.97    3.678445  0.8268108  2.845602
##   0.98    3.678790  0.8267922  2.846038
##   0.99    3.679140  0.8267733  2.846474
##   1.00    3.679493  0.8267543  2.846910
##
## Tuning parameter 'alpha' was held constant at a value of 0
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were alpha = 0 and lambda = 0.63.
```

- alpha = 0 thể hiện rằng mô hình đang được train theo phương pháp Ridge regression, lambda tối ưu nhất được chọn là 0.63 ứng với giá trị RMSE là khoảng 3.668 và R^2 là khoảng 0.827.

```
## 20 x 1 sparse Matrix of class "dgCMatrix"
##                                        s1
## (Intercept)                  5.564864e+01
## Status                      -1.300289e+00
## Adult.Mortality             -1.742653e-02
## infant.deaths                1.438294e-03
## Alcohol                     -1.124047e-01
## percentage.expenditure       3.010404e-04
## Hepatitis.B                 -2.778906e-03
## Measles                      1.428290e-05
## BMI                          4.306321e-02
## under.five.deaths           -3.803610e-03
## Polio                        8.704240e-03
## Total.expenditure            5.745040e-02
## Diphtheria                   2.134113e-02
## HIV.AIDS                    -4.215459e-01
## GDP                          2.895973e-05
## Population                   2.760560e-09
## thinness..1.19.years        -4.183057e-02
## thinness.5.9.years           1.049064e-02
## Income.composition.of.resources  1.029614e+01
## Schooling                    7.703399e-01
```

- Đây là bảng hệ số của mô hình Ridge regression.

- Đối với mô hình Lasso regression, khi tăng siêu tham số lambda lên cao thì nó sẽ làm cho 1 số hệ số của mô hình dần về 0, mô hình này khá phù hợp trong việc chọn tham số cho mô hình.

```
## glmnet
##
## 1320 samples
##   19 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1188, 1188, 1188, 1188, 1186, 1189, ...
## Resampling results across tuning parameters:
##
##   lambda  RMSE      Rsquared   MAE
##   0.0001  3.577061  0.8358460  2.748744
##   0.0002  3.577061  0.8358460  2.748744
##   0.0003  3.577061  0.8358460  2.748744
##   0.0004  3.577061  0.8358460  2.748744
##   0.0005  3.577061  0.8358460  2.748744
##   0.0006  3.577061  0.8358460  2.748744
```

```
##   0.0007  3.577061  0.8358460  2.748744
##   0.0008  3.577061  0.8358460  2.748744
##   0.0009  3.577061  0.8358460  2.748744
##   0.0010  3.577061  0.8358460  2.748744
##   0.0011  3.577061  0.8358460  2.748744
##   0.0012  3.577061  0.8358460  2.748744
##   0.0013  3.577061  0.8358460  2.748744
##   0.0014  3.577059  0.8358459  2.748718
##   0.0015  3.577053  0.8358462  2.748673
##   0.0016  3.577039  0.8358475  2.748633
##   0.0017  3.577028  0.8358485  2.748599
##   0.0018  3.577041  0.8358470  2.748586
##   0.0019  3.577156  0.8358346  2.748665
##   0.0020  3.577274  0.8358220  2.748745
##   0.0021  3.577395  0.8358092  2.748827
##   0.0022  3.577518  0.8357963  2.748911
##   0.0023  3.577642  0.8357836  2.748997
##   0.0024  3.577775  0.8357702  2.749088
##   0.0025  3.577915  0.8357563  2.749175
##   0.0026  3.578057  0.8357424  2.749261
##   0.0027  3.578199  0.8357283  2.749349
##   0.0028  3.578344  0.8357140  2.749437
##   0.0029  3.578436  0.8357048  2.749453
##   0.0030  3.578520  0.8356964  2.749454
##   0.0031  3.578605  0.8356879  2.749455
##   0.0032  3.578683  0.8356795  2.749454
##   0.0033  3.578760  0.8356711  2.749452
##   0.0034  3.578840  0.8356625  2.749450
##   0.0035  3.578955  0.8356512  2.749490
##   0.0036  3.579084  0.8356389  2.749548
##   0.0037  3.579216  0.8356264  2.749606
##   0.0038  3.579352  0.8356134  2.749668
##   0.0039  3.579495  0.8356000  2.749734
##   0.0040  3.579641  0.8355864  2.749800
##   0.0041  3.579789  0.8355724  2.749866
##   0.0042  3.579939  0.8355582  2.749920
##   0.0043  3.580092  0.8355437  2.749974
##   0.0044  3.580247  0.8355290  2.750027
##   0.0045  3.580406  0.8355140  2.750093
##   0.0046  3.580567  0.8354987  2.750163
##   0.0047  3.580731  0.8354831  2.750243
##   0.0048  3.580898  0.8354673  2.750340
##   0.0049  3.581067  0.8354512  2.750438
##   0.0050  3.581240  0.8354349  2.750541
##   0.0051  3.581415  0.8354183  2.750657
##   0.0052  3.581593  0.8354015  2.750773
##   0.0053  3.581773  0.8353844  2.750889
##   0.0054  3.581956  0.8353670  2.751005
##   0.0055  3.582141  0.8353496  2.751118
##   0.0056  3.582328  0.8353319  2.751229
```

```
##   0.0057  3.582519  0.8353140  2.751340
##   0.0058  3.582713  0.8352958  2.751450
##   0.0059  3.582909  0.8352773  2.751561
##   0.0060  3.583109  0.8352585  2.751676
##   0.0061  3.583312  0.8352394  2.751803
##   0.0062  3.583518  0.8352201  2.751931
##   0.0063  3.583726  0.8352005  2.752058
##   0.0064  3.583938  0.8351807  2.752186
##   0.0065  3.584152  0.8351606  2.752313
##   0.0066  3.584365  0.8351406  2.752445
##   0.0067  3.584579  0.8351204  2.752583
##   0.0068  3.584796  0.8351000  2.752744
##   0.0069  3.585016  0.8350793  2.752905
##   0.0070  3.585238  0.8350584  2.753066
##   0.0071  3.585463  0.8350371  2.753226
##   0.0072  3.585686  0.8350161  2.753378
##   0.0073  3.585913  0.8349946  2.753520
##   0.0074  3.586143  0.8349728  2.753659
##   0.0075  3.586376  0.8349508  2.753798
##   0.0076  3.586612  0.8349285  2.753937
##   0.0077  3.586850  0.8349060  2.754076
##   0.0078  3.587091  0.8348832  2.754216
##   0.0079  3.587333  0.8348603  2.754355
##   0.0080  3.587561  0.8348387  2.754481
##   0.0081  3.587794  0.8348166  2.754614
##   0.0082  3.588029  0.8347943  2.754747
##   0.0083  3.588267  0.8347718  2.754881
##   0.0084  3.588507  0.8347490  2.755014
##   0.0085  3.588750  0.8347260  2.755158
##   0.0086  3.588996  0.8347027  2.755328
##   0.0087  3.589244  0.8346791  2.755499
##   0.0088  3.589491  0.8346556  2.755665
##   0.0089  3.589740  0.8346321  2.755827
##   0.0090  3.589992  0.8346082  2.755990
##   0.0091  3.590246  0.8345842  2.756152
##   0.0092  3.590502  0.8345599  2.756314
##   0.0093  3.590761  0.8345353  2.756477
##   0.0094  3.591022  0.8345105  2.756639
##   0.0095  3.591287  0.8344853  2.756800
##   0.0096  3.591550  0.8344603  2.756949
##   0.0097  3.591814  0.8344351  2.757098
##   0.0098  3.592081  0.8344096  2.757246
##   0.0099  3.592349  0.8343840  2.757409
##   0.0100  3.592621  0.8343581  2.757606
##
## Tuning parameter 'alpha' was held constant at a value of 1
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were alpha = 1 and lambda = 0.0017.
```

- alpha = 1 thể hiện rằng mô hình đang được train theo phương pháp Lasso regression, lambda tối ưu nhất được chọn là 0.0017 ứng với giá trị RMSE là khoảng 3.577 và R^2 là khoảng 0.836.

```
## 20 x 1 sparse Matrix of class "dgCMatrix"
##                                           s1
## (Intercept)                     5.572909e+01
## Status                         -1.283155e+00
## Adult.Mortality                -1.689903e-02
## infant.deaths                   8.709639e-02
## Alcohol                        -1.029681e-01
## percentage.expenditure          4.499953e-04
## Hepatitis.B                    -3.011896e-03
## Measles                        -5.037436e-06
## BMI                             3.882932e-02
## under.five.deaths              -6.502955e-02
## Polio                           4.910252e-03
## Total.expenditure               6.632954e-02
## Diphtheria                      1.582135e-02
## HIV.AIDS                       -4.407891e-01
## GDP                             3.700476e-06
## Population                     -1.496331e-09
## thinness..1.19.years           -3.313654e-02
## thinness.5.9.years              1.640762e-04
## Income.composition.of.resources 1.018556e+01
## Schooling                       8.417884e-01
```

- Đây là bảng hệ số của mô hình Lasso regression.

6. Thực nghiệm, kết quả, và thảo luận

- Khi điều chỉnh tham số lambda trong mô hình Ridge và Lasso ta sẽ đánh đổi giữa việc. Nếu giảm lambda dần về 0 sẽ trở thành Linear thông thường, nó sẽ dễ dẫn đến overfitting, nếu tăng lambda lên thì mô hình sẽ tăng được khả năng tổng quát hóa nhưng đồng nghĩa với việc lỗi trên tập huấn luyện sẽ cao hơn so với việc dùng mô hình Linear. Nếu lambda đạt đến vô cùng thì mô hình sẽ bị underfitting.

- Sử dụng RMSE và R^2 để so sánh giữa các mô hình.

```
##
## Call:
## summary.resamples(object = ., metric = "RMSE")
##
## Models: linear, ridge, lasso
## Number of resamples: 10
##
## RMSE
##            Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## linear 3.221138 3.371918 3.556250 3.576265 3.775411 4.035166    0
## ridge  3.290496 3.426751 3.619122 3.668390 3.800870 4.215461    0
## lasso  3.227227 3.370740 3.557169 3.577028 3.766098 4.037857    0
```

```
##
## Call:
## summary.resamples(object = ., metric = "Rsquared")
##
## Models: linear, ridge, lasso
## Number of resamples: 10
##
## Rsquared
##             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## linear 0.7959470 0.8156388 0.8397577 0.8359181 0.8559542 0.8662799    0
## ridge  0.7705930 0.8081095 0.8376301 0.8273919 0.8486677 0.8614420    0
## lasso  0.7989824 0.8155832 0.8399274 0.8358485 0.8555931 0.8664658    0
```

Sau khi so sánh ta có thể thấy:

- Mô hình Linear sẽ cho ra giá trị RMSE tối ưu nhất, sau đó đến mô hình Lasso.
- Mô hình Lasso sẽ cho ra kết quả R2 cao nhất.

Chọn mô hình Lasso làm mô hình cuối để train lại trên toàn bộ tập train với siêu tham số lambda tối ưu nhất là 0.0017.

```
## glmnet
##
## 1320 samples
##   19 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 1320, 1320, 1320, 1320, 1320, 1320, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   3.589354  0.8347196  2.757824
##
## Tuning parameter 'alpha' was held constant at a value of 1
## Tuning
##  parameter 'lambda' was held constant at a value of 0.0017
```

- Sau khi train thì mô hình cho ra kết quả RMSE là 3.589 và R^2 là 0.835 là khá tốt

Dự đoán trên tập test

- Tạo ra biến predictions chưa giá trị dự đoán

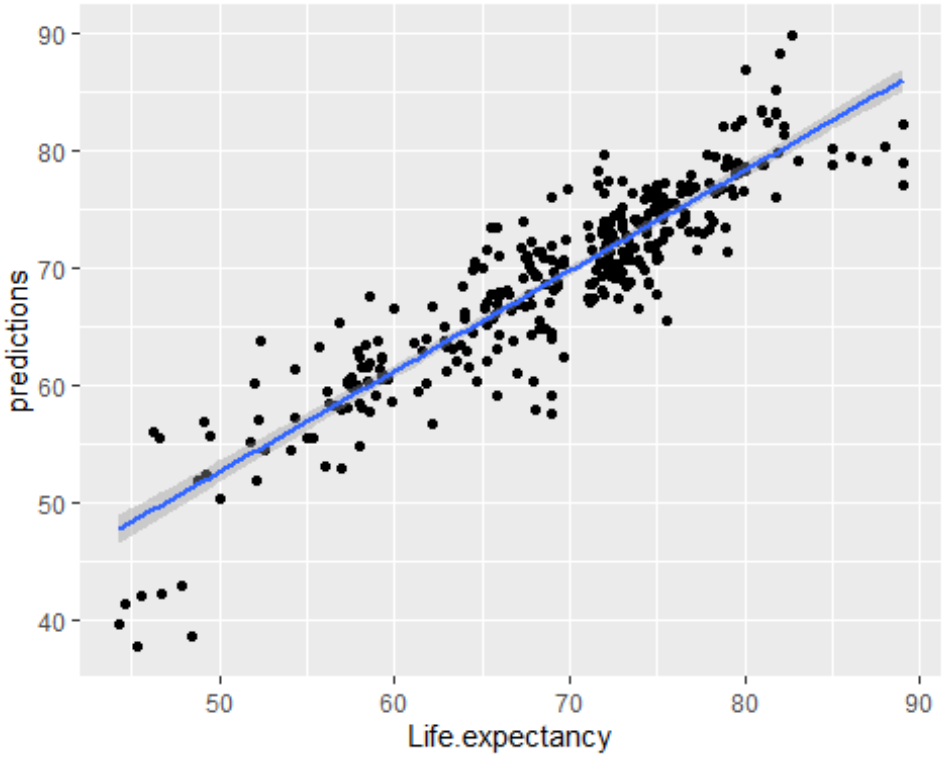
```
##    Life.expectancy predictions
## 1             58.6    61.86407
## 2             58.1    61.49118
## 3             57.5    60.71021
## 4             57.3    60.28178
## 5             57.3    60.10190
## 6             57.0    57.89099
## 7             76.6    74.16520
```

```
## 8              74.2    69.87700
## 9              75.3    75.45814
## 10             75.1    75.27591
```

- Ta so sánh giá trị dự đoán và giá trị thực tế có thể thấy sự chênh lệch không quá lớn

Tính lại giá trị RMSE và R^2 để có thể đánh giá mô hình có bị overfiting hay underfitting không

```
##       RMSE   Rsquare
## 1 3.809214 0.8185347
```



- Kết quả cho thấy mô hình dự đoán khá tốt, không lệch nhiều so với dự đoán trên tập train

7. Kết luận

- Sử dụng độ đo là R^2 và RMSE. Nên khi so sánh 3 mô hình thì mô hình Lasso cho ra kết quả tốt nhất trên tập train. Mặc dù R^2 trên mô hình Linear cho kết quả tốt hơn nhưng khả năng tổng quát hóa của mô hình này không tốt bằng mô hình Lasso. Vì thế tôi quyết định train lại mô hình Lasso trên toàn bộ tập train sau đó thực hiện dự đoán trên tập test. Kết quả R^2 dự đoán được chưa thực sự cao, nếu có điều kiện thì tôi sẽ tiếp tục nghiên cứu 1 số mô hình khác để có thể mang lại kết quả cao hơn.