Xây Dựng Mô Hình Máy Học Dự Đoán Tuổi Thọ Của Người

Nguyễn Hoàng Minh
Khoa Khoa học và Kỹ thuật Thông tin
Trường Đại học Công nghệ Thông Tin
Hồ Chí Minh, Việt Nam
20521609@gm.uit.edu.vn

Tạ Nhật Minh
Khoa Khoa học và Kỹ thuật Thông tin
Trường Đại học Công nghệ Thông Tin
Hồ Chí Minh, Việt Nam
20521614@gm.uit.edu.vn

Nguyễn Minh Tiến
Khoa Khoa học và Kỹ thuật Thông tin
Trường Đại học Công nghệ Thông Tin
Hồ Chí Minh, Việt Nam
20522010@gm.uit.edu.vn

Nguyễn Thiện Thuật
Khoa Khoa học và Kỹ thuật Thông tin
Trường Đại học Công nghệ Thông Tin
Hồ Chí Minh, Việt Nam
20521998@gm.uit.edu.vn

***Tóm tắt* - Tuổi thọ là một trong những chỉ số quan trọng để con người có thể đánh giá chất lượng sống ở từng nơi hay từng khu vực cụ thể nào đó, việc dự đoán tuổi thọ còn có thể giúp đánh giá được tình hình sức khỏe của các quốc gia. Bài báo này trình bày các phân tích về yếu tố ảnh hưởng đến tuổi thọ và xây dựng các mô hình máy học dự đoán dựa trên các yếu tố đó. Các mô hình dự đoán được chúng tôi sử dụng để xây dựng là Ridge và Lasso Regression, Random Forest Regression, Support Vector Regression. Việc đánh giá mô hình được dựa vào các độ đo Mean Absolute Error (MAE), Mean Squared Error (MSE), Root - Mean Square Deviation (R-MSE) & R2 score. Phân tích so sánh cho kết quả các yếu tố Adult Mortality, BMI, Under-five deaths, Diphtheria, Thinness 1-19 years, Income composition of resources là những yếu tố quan trọng ảnh hưởng đến tuổi thọ. Kết quả thực nghiệm cho thấy mô hình Random Forest Regressor cho kết quả tốt nhất cho mô hình dự đoán tuổi thọ trên bộ dữ liệu, với các điểm R2 score trên tập test là 0.96, MAE: 1.28, MSE: 3.39, R-MSE: 1.98. Nghiên cứu được thực hiện cho thấy rằng các mô hình được lựa chọn phù hợp trên bộ dữ liệu và có thể sử dụng trong thực tế. Những kết quả về báo cáo này có thể sử dụng được cho các chính phủ và y tế để cải thiện chất lượng xã hội.**

***Keywords* - Life expectancy, Ridge regression, Lasso regression, Support vector regression (SVR), Random forest regression (RFR)**

I. GIỚI THIỆU

A. Bối cảnh

Tuổi thọ phản ánh một phần chất lượng sống, các điều kiện kinh tế xã hội của một quốc gia. Sức khỏe, hạnh phúc, tuổi thọ của con người có thể được giải thích thông qua bản thống kê tuổi thọ. Các bảng dữ liệu thống kê này có thể được sử dụng ở bất kỳ quốc gia nào. Nó có thể được dùng để thống kê tuổi thọ của cả thực vật, động vật và con người.

Học máy bao gồm các yếu tố của toán học, thống kê và khoa học máy tính. Trong sự phát triển của trí tuệ nhân tạo (AI), Máy học (ML) đã phát triển và đóng một vai trò quan trọng trong việc giải quyết các vấn đề xã hội.

Việc dự đoán tuổi thọ cũng là một yêu cầu giúp cho con người có những quyết định đúng đắn để nâng cao chất lượng xã hội.

B. Mục tiêu

Bài báo cáo này thực hiện phân tích và xây dựng mô hình dự đoán tuổi thọ trung bình của người dựa trên các đặt điểm của họ đến từ 193 quốc gia trên thế giới như: Mỹ, Anh, Pháp, Đức, Úc, Nhật Bản, Hàn Quốc, Ý,... Tầm quan trọng về tuổi thọ trung bình của một quốc gia còn phụ thuộc vào một số yếu tố như hoàn cảnh kinh tế, bệnh nền, bệnh thể chất, uống rượu, GDP, chi tiêu cho hệ thống chăm sóc sức khỏe và các yếu tố khác. Tuổi thọ của con người đã được cải thiện đáng kể từ 66.8 năm 2000 lên 73.4 vào năm 2019 [1].

Mục tiêu của bài toán máy học trong việc dự đoán được tuổi thọ là phải chọn được một thuật toán máy học phù hợp để từ đó xây dựng được một mô hình máy học với bộ tham số tối ưu giúp cho mô hình có thể dự đoán được chính xác nhất tuổi thọ của con người thông qua các yếu tố cụ thể.

C. Ứng dụng

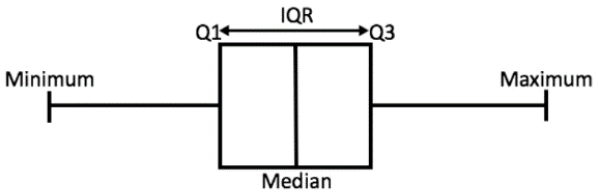
Máy học (ML) được ứng dụng nhiều trong việc nâng cao tuổi thọ bằng cách theo dõi sức khỏe và giảm tỷ lệ tử vong [2].

II. DỮ LIỆU

A. Tiền xử lý dữ liệu và phân tích các thuộc tính

Dữ liệu được chúng tôi thu thập từ WHO, trang web Liên hợp quốc lấy từ Kaggle [3]. Được tham chiếu vào trang web của WHO [4] để xác minh độ chính xác. Dữ liệu được thực hiện cho nghiên cứu cung cấp khoảng thời gian từ năm 2000 đến năm 2015. Có 22 cột trong đó 21 cột là các đặc trưng & 1 là đầu ra mong muốn. Trong số 21 đặc trưng: Country, year & status của quốc gia là các đặc trưng thừa, vì vậy những đặc trưng đó không được xem xét.

Các dòng dữ liệu có giá trị trống hoặc không có giá trị được điền bằng giá trị trung bình của đặc trưng đó. Với các giá trị bị nhiễu (outlier) chúng tôi dựa trên đồ thị hộp và phương pháp IQR để phát hiện chúng:



IQR có công thức:

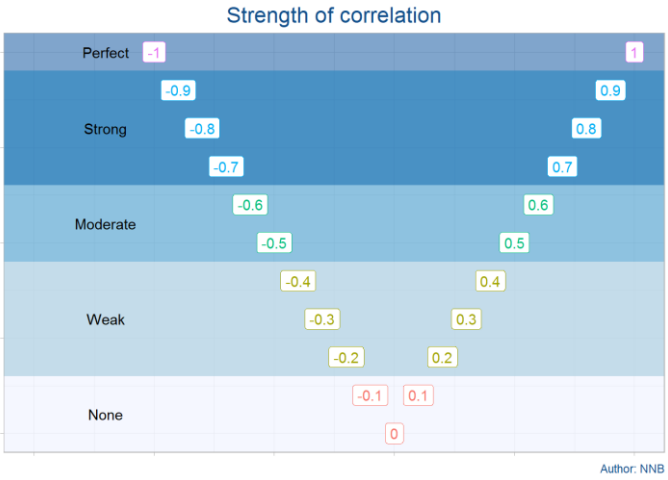
$$IQR = Q_3 - Q\_1$$
$$Lower\ Bound = Q_1 - 1,5 * IQR$$
$$Higher\ Bound\ = Q_3 + 1,5 * IQR$$

Với các giá trị thấp hơn **Lower Bound** và cao hơn **Higher Bound** là các giá trị nhiễu (outlier), được điền bằng cách gán lại bằng giá trị Min và Max của phân phối ứng với từng đặc trưng. Chúng tôi đã thực hiện trực quan hóa dữ liệu bằng các hình ảnh để kiểm tra mối tương quan giữa các đặc trưng và đầu ra mong muốn.

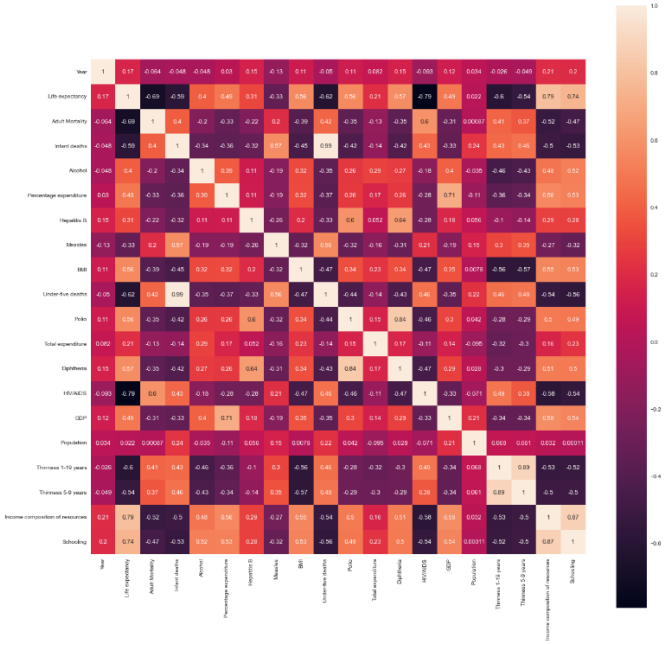
Ma trận tương quan tương quan các thuộc tính được vẽ biểu đồ nhiệt *(Hình 2)*, cho chúng ta nhìn một cách trực quan về mối tương quan giữa tuổi thọ (life expectancy) và các thuộc tính khác. Chú thích bên phải của hình ảnh cho biết rằng các giá trị càng tiến về ở hai đầu của cột màu sắc (nhạt nhất và đậm nhất) có mối tương quan cao với nhau. Ngược lại, các giá trị càng tiến về 0 (nằm ở giữa cột màu sắc) sẽ có mối tương quan yếu giữa hai thuộc tính. Mối quan hệ tuyến tính giữa thuộc tính đích và thuộc tính khác có thể được ước tính thông qua ma trận tương quan. Trong phân tích đa biến, nó đóng một vai trò quan trọng, vì nó làm rõ mối quan hệ giữa các thành phần khác nhau. Ma trận này được tính từ công thức tính hệ số tương quan (Correlation Coefficient), nó là chỉ số thống kê đo lường mức độ mạnh yếu của mối quan hệ giữa hai biến số, được biểu diễn bằng công thức toán học sau:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Các giá trị thể hiện mức độ tương quan nằm trong khoảng [-1,1], mức độ tương quan giữa hai thuộc tính mạnh hay yếu được quy ước như hình:

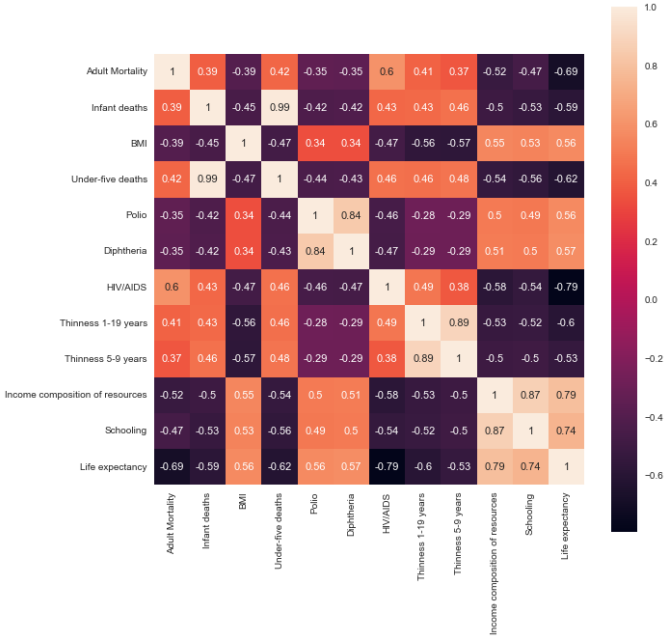


Chúng tôi lấy các giá trị bé hơn -0.5 và lớn hơn 0.5 là các giá trị thể hiện có sự tương quan giữa hai yếu tố, theo Hình 2.1 bên dưới, ta có thể suy ra rằng các thuộc tính: 'Adult Mortality', 'Infant deaths', 'Under-five deaths', 'HIV/AIDS', 'Thinness 1-19 years', 'Thinness 5-9 years' có tương quan tỉ lệ nghịch với thuộc tính 'Life expectancy' và các thuộc tính: , 'BMI', 'Polio', 'Diphtheria', 'Income composition of resources', 'Schooling' có tương quan tỉ lệ thuận với thuộc tính 'Life expectancy'.



Hình 2.1 Biểu đồ ma trận tương quan giữa các thuộc tính

Chúng tôi thực hiện kiểm tra mối tương quan giữa các thuộc tính đã chọn ở phía trên với nhau cho việc lọc ra các thuộc tính có mức tương quan tương đối cao để loại bỏ. Nhằm đảm bảo cho các yếu tố không tương tác lẫn nhau trong mô hình dự đoán thuộc tính đích (Life expectancy). Nhìn vào hình 3, ta thấy thuộc tính 'Under-five deaths' có tương quan tuyến tính mạnh với thuộc tính 'Infant deaths', trường hợp này ta tiến hành so sánh mối tương quan của 2 thuộc tính này với thuộc tính 'Life expectancy', cả 2 thuộc tính có giá trị lần lượt là 0.62 và 0.59. Như vậy ta loại bỏ thuộc tính có tương quan thấp hơn là 'Infant deaths'. Với cách trên ta cũng loại bỏ thêm cả ba thuộc tính nữa là 'Polio', ' Thinness 5-9 years ', Schooling'.



Hình 2.2 Biểu đồ ma trận tương quan giữa các thuộc tính ảnh hưởng đến tuổi thọ

B. Trích xuất đặc trưng và phân chia tập dữ liệu huấn luyện

Để tránh tình trạng những thuộc có giá trị lớn sẽ thiên vị các thuộc tính có giá trị nhỏ làm ảnh hưởng đến mô hình máy học, nên chúng tôi đã thực hiện phương pháp trích xuất đặc trưng để cải thiện hiệu xuất mô hình máy học. Ở bài báo cáo này chúng tôi đã sử dụng phương pháp chuẩn hóa Standardizing scores (Z-score), có công thức như sau:

$$\boldsymbol{X_{new} = \frac{X - \mu}{\sigma}}$$

Với X_{new} là giá trị mới, X là giá trị cần chuẩn hóa, μ là trung bình và σ là độ lệch chuẩn của phân phối thuộc tính đó.

Dữ liệu được chúng tôi chia thành 2 thập train và test với tỉ lệ 75:25. Mục đích của việc phân tách dữ liệu là để tránh overfitting. Nếu bị overfitting, các thuật toán học máy có thể hoạt động tốt trong các tập dữ liệu train nhưng hoạt động kém trong tập dữ liệu test. Từ đó ta có thể nhận biết và điều chỉnh.

III. PHƯƠNG PHÁP MÁY HỌC

A. Mô hình máy học

1. Ridge & Lasso Regression

Ridge và Lasso Regression là hai trong số các thuật toán Regression (Hồi quy). Hai thuật toán này là biến thể của Linear Regression (Hồi quy tuyến tính) được sử dụng để xử lý vấn đề về overfitting với tập dữ liệu lớn.

1.1 Ridge Regression

Ridge Regression (L2 regularization) là thuật toán tìm cách để khớp với số liệu bằng cách tối thiểu RSS. Thuật toán được thêm hệ số tổng bình phương của các hệ số nhằm mục tiêu tối ưu hóa. Điều này sẽ giúp kiểm soát độ phức tạp của thuật toán. Do đó, ta có hàm hồi quy Ridge:

$Objective = RSS + \alpha * (tổng\ bình\ phương\ của\ các\ hệ\ số)$

Trong đó, α (alpha) là tham số cân bằng mức độ được đưa ra để giảm thiểu RSS so với việc giảm thiểu tổng bình phương của các hệ số. Giá trị của α là siêu tham số của Ridge, có nghĩa là chúng không được thuật toán tự động tìm giá trị, thay vào đó chúng phải được tìm theo cách thủ công.

Hàm cost function cho Ridge Regression như sau:

$$\boldsymbol{cost(w) = \sum_{i=1}^{n}(y_i - w^T x_i)^2 + \sum_{i=1}^{k} \lambda w_i^2}$$

1.2 Lasso Regression

Lasso Regression (L1 Regularization) là một biến thể khác với hồi quy Ridge. Thay vì thêm hệ số tổng bình phương của các hệ số thì thuật toán này sẽ thêm hệ số tổng giá trị tuyệt đối của các hệ số nhằm mục tiêu tối ưu hóa. Do đó, ta có hàm hồi quy Lasso:

$Objective = RSS + \alpha * (tổng\ trị\ tuyệt\ đối\ của\ các\ hệ\ số)$

Trong đó, α (alpha) hoạt động tương tự như của ridge, thuật toán sẽ cân bằng RSS và độ lớn của các hệ số. Giống như Ridge, α có thể nhận nhiều giá trị khác nhau và được tìm theo cách thủ công. Hàm cost function cho Lasso Regression như sau:

$$\boldsymbol{cost(w) = \sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda \sum_{i=1}^{k} |w_i|}$$

2. Support Vector Regression

Support Vector Regression (SVR) là mô hình cung cấp một sự linh hoạt để xác định mức độ lỗi có thể chấp nhận được trong mô hình và sẽ tìm một đường thích hợp (hoặc là một siêu phẳng ở các kích thước cao hơn) để phù hợp với dữ liệu. SVR sử dụng nguyên tắc tương tự như SVM, nhưng chủ yếu với các vấn đề hồi quy. Có nghĩa là tìm một hàm gần đúng ánh xạ từ miền đầu vào tạo thành các số thực trên cơ sở một mẫu huấn luyện.

Hàm giả thuyết cho SVR:

$$\boldsymbol{Y = wx + b}$$

Khi đó, ta có phương trình của đường ranh giới quyết định:

$$\boldsymbol{wx + b = +\epsilon}$$
$$\boldsymbol{wx + b = -\epsilon}$$

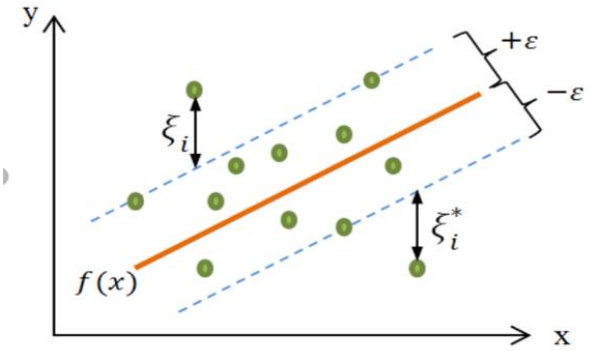
Xét bài toán tối ưu có ràng buộc:

$$\boldsymbol{(w, b) = argmin}\left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i\right]$$

Với bất kì siêu phẳng nào của SVR đều phải thoả điều kiện:

$$\boldsymbol{-\epsilon < Y - wx + b < +\epsilon}$$

Do đó, ở đây chỉ lấy những điểm nằm trong đường ranh giới quyết định và có tỷ lệ lỗi ít nhất hoặc nằm trong Margin of Tolerance (tạm dịch là phạm vi dung sai). Từ đó sẽ tạo điều kiện thuận lợi để đào tạo một mô hình phù hợp hơn.

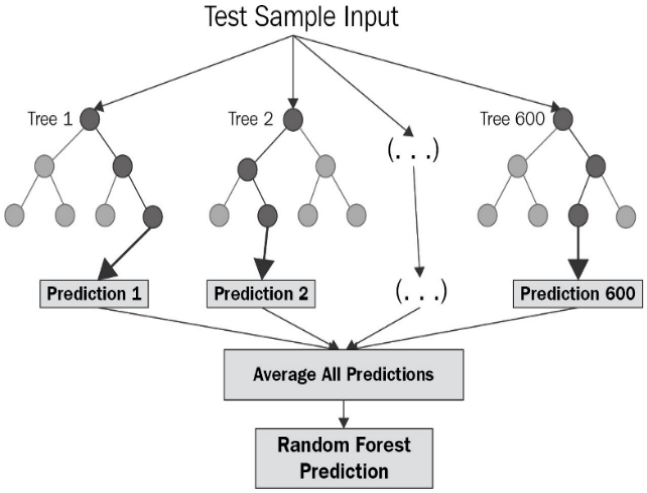


Hình 3.2 Mô hình Support Vector Regression

3. Random Forest Regression

Random Forest là một thuật toán học máy có giám sát sử dụng phương pháp học máy theo nhóm để phân loại và hồi quy. Random là ngẫu nhiên, Forest là rừng, nên ở thuật toán Random Forest mình sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.

Mọi cây quyết định đều có phương sai cao, nhưng khi chúng ta kết hợp tất cả chúng lại với nhau song song thì phương sai của kết quả sẽ thấp vì mỗi cây quyết định được đào tạo hoàn hảo trên dữ liệu mẫu cụ thể và do đó kết quả đầu ra không phụ thuộc vào một cây quyết định mà nhiều cây quyết định. Trong trường hợp có vấn đề phân loại, kết quả cuối cùng được thực hiện bằng cách sử dụng bộ phân loại biểu quyết đa số. Trong trường hợp của một bài toán hồi quy, đầu ra cuối cùng là giá trị trung bình của tất cả các đầu ra. Phần này được gọi là Aggregation.



Hình 3.3 Mô hình Random Forest Regression

B. Công cụ sử dụng

Nền tả sử dụng: Google Colab.

Thư viện sử dụng:

- Pandas: được sử dụng để thao tác và phân tích dữ liệu thông qua các phép toán và cấu trúc dữ liệu trên bảng số và chuỗi thời gian.
- Numpy: bổ sung hỗ trợ cũng như chứa các hàm toán học cấp cao để hoạt động trên các ma trận và mảng nhiều chiều lớn.
- Matplotlib: là thư viện vẽ biểu đồ cho phép lập sơ đồ 2d và sắp xếp các biểu đồ thanh, biểu đồ, v.v.
- Scipy: được xây dựng dựa trên thư viện NumPy, cung cấp thao tác mảng N chiều thuận tiện và nhanh chóng, gồm các gói con (submodule) cho đại số tuyến tính, tối ưu hóa, tích hợp và thống kê.
- Seaborn: là một thư viện trực quan hóa dữ liệu.
- Sklearn: cung cấp một tập các công cụ xử lý các bài toán machine learning và statistical modeling.

Công cụ khác:

- GridSearchCV: lấy một từ điển mô tả các tham số có thể được thử trên một mô hình để huấn luyện nó. Lưới tham số được định nghĩa như một từ điển, mà các khóa là các tham số và các giá trị là cài đặt cần kiểm tra.

C. Các phương pháp đánh giá

1. MAE, MSE, RMSE

- Mean Absolute Error (MAE) là trung bình độ sai lệch tuyệt đối giữa giá trị thực tế với giá trị dự đoán.

Công thức: $\boldsymbol{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$

- Mean Squared Error (MSE) hay sai số bình phương trung bình là trung bình tổng bình phương sai số giữa giá trị thực tế với giá trị dự đoán.

Công thức: $\boldsymbol{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$

- Root Mean Squared Error (RMSE) là căn bậc hai của sai số toàn phương trung bình MSE

Công thức: $\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i)^2}{n}}$

2. R2 score

- $TSS(total\ sum\ squared) = \sum(y_i - \bar{y})^2 (SS_{TOT})$
  Tổng bình phương tất cả sai lệch giữa y_i và giá trị trung bình.
- $ESS(explained\ sum\ of\ squared) = \sum(\hat{y}_i - \bar{y})^2 (SS_{REG})$
  Tổng bình phương các sai lệch giữa giá trị dự đoán của biến phụ thuộc y và giá trị trung bình => đo độ chính xác của hàm hồi quy.
- $RSS(residual\ sum\ of\ squared) = \sum(e_i^2) (SS_{ERR})$
  Tổng bình phương sai số.
- Từ quan hệ: *TSS = ESS + RSS, chia 2 vế cho TSS ta được hệ số xác định* R^2

Công thức: $\boldsymbol{R^2 = 1 - \frac{RSS}{TSS}}$

R^2 tỉ lệ thuận với mức giải thích của biến độc lập với sự thay đổi của biến phụ thuộc. R^2 cũng thể hiện độ phù hợp của mô hình hồi quy với tập dữ liệu.

IV. CÁC THỬ NGHIỆM TINH CHỈNH MÔ HÌNH

Sau khi chạy ba mô hình máy học bao gồm: Random Forest Regression, Support Vector Regression, Ridge Regression và Lasso Regression với tham số mặc định, nhận thấy:

Thuật toán Random Forest Regression (0.96) đã đem lại kết quả cao nhất trong ba thuật toán, và các giá trị sai số cũng là thấp nhất.

Trái ngược lại thì Lasso Regression đem lại kết quả khá thấp (0.7), thấp nhất trong bốn thuật toán với cái giá trị đo MAE, MSE, RMSE khá cao.

Hai mô hình là Support Vector Regression và Ridge Regression đạt kết quả tương đối lần lượt là 0.86 và 0.81.

Mô hình	Kết quả độ đo R2	Kết quả độ đo MAE	Kết quả độ đo MSE	Kết quả độ đo RMSE
Random Forest Regression	0.96	1.22	3.75	1.94
Support Vector Regression	0.86	2.4	12.57	3.55
Ridge Regression	0.81	3.03	16.98	4.12
Lasso Regression	0.7	3.88	26.53	5.15

Bảng 4.1 Kết quả chạy mô hình Máy học với tham số mặc định

Chúng tôi tiếp tục tinh chỉnh tham số của mỗi phương pháp với GridSearchCV để tăng hiệu suất mô hình và nhận được kết quả thể hiện ở bảng 4.2 cùng với bộ tham số thu được như sau:

- Thuật toán Random Forest Regression với tham số 'bootstrap': False, 'max_depth': 15, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 133.
- Thuật toán Support Vector Regression với tham số 'C': 100.0, 'gamma': 1, 'kernel': 'rbf'.
- Thuật toán Ridge Regression với 'alpha': 0.01.
- Thuật toán Lasso Regression với 'alpha': 0.001.

Mô hình	Kết quả độ đo R2	Kết quả độ đo MAE	Kết quả độ đo MSE	Kết quả độ đo RMSE
Random Forest Regression	0.96	1.28	3.93	1.98
Support Vector Regression	0.91	1.61	7.82	2.77
Ridge Regression	0.81	3.03	16.95	4.12
Lasso Regression	0.81	3.03	16.95	4.12

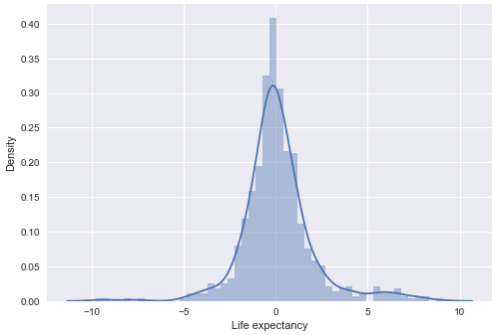
Bảng 4.2 Kết quả mô hình sau khi tinh chỉnh tham số

Ba mô hình Support Vector Regression, Ridge Regression và Lasso Regression đã được cải thiện so với khi chạy mô hình với tham số mặc định. Đặc biệt là mô hình Lasso Regression đã tăng lên đáng kể nhất (từ 0.7 lên 0.81) và Support Vector Regression (từ 0.86 lên 0.91), các giá trị sai số cũng giảm đi đáng kể so với ban đầu.

Thuật toán Random Forest (từ 0.958 xuống 0.956) tuy có giảm xuống nhưng không quá đáng kể và vẫn là thuật toán mang lại kết tốt nhất, đạt độ tin cậy cao nhất so với các thuật toán học máy còn lại.

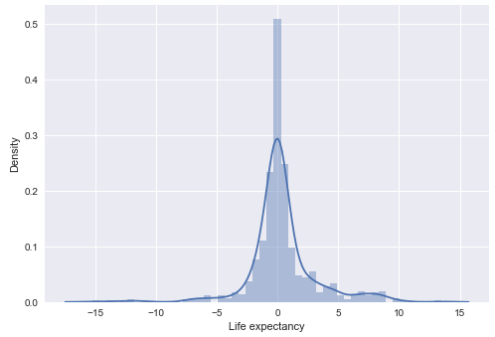
V. PHÂN TÍCH LỖI, HƯỚNG PHÁT TRIỂN

A. Phân tích lỗi



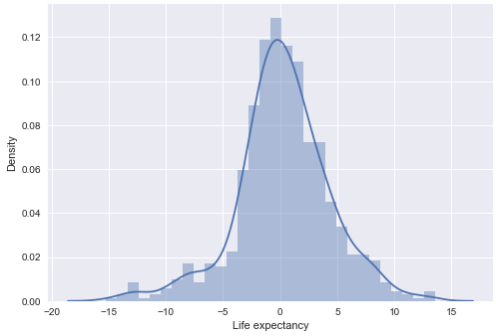
Hình 4.1 Random Forest Regression

Với mô hình Random Forest, giá trị dự đoán nằm sát với các tập các giá trị thực tế, mức sai lệch so với tập giá trị thực đa phần nằm trong khoảng (0;2.5), các dự đoán có độ sai lệch cao chiếm tỉ lệ thấp. Điều này cho thấy kết quả dự đoán của mô hình là đáng tin cậy hơn so với ba mô hình còn lại.

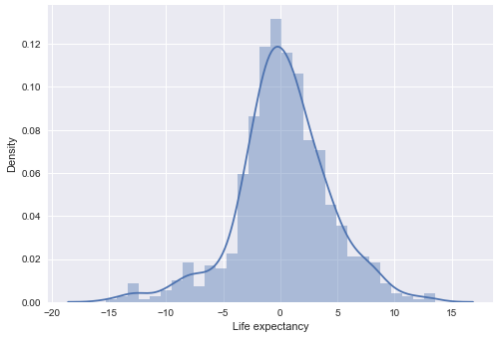


Hình 4.2 Support Vector Regression

Với mô hình Support Vector Regression, mức sai lệch tuyệt đối phần lớn nằm trong khoảng (0;2.5). Tuy nhiên, ở mô hình này xuất hiện các giá trị dự đoán có mức sai lệch rất lớn so với giá trị thực tế như trên hình. Các giá trị này làm giảm độ tin cậy của mô hình.



Hình 4.3 Ridge Regression



Hình 4.4 Lasso Regression

Với hai mô hình Ridge Regression và Lasso Regression, khả năng dự đoán của cả hai mô hình gần như là tương đương nhau với độ sai lệch tuyệt đối trong khoảng (0;5). Giống như mô hình Support Vector Regression, cả hai mô hình này đều dự đoán được các giá trị có độ sai lệch là rất lớn, và số lượng các giá trị này là tương đối nhiều so với hai môn hình còn lại.

Nhìn chung, qua các biểu đồ phân phối dự đoán dễ dàng thấy được mô hình tốt nhất với bộ dữ liệu là Random Forest, thứ hai là Support Vector Regression, cuối cùng là Ridge Regression và Lasso Regression và hai mô hình này có thể thay thế cho nhau.

Đối với mô hình Random Forest, chúng tôi thấy các giá trị hàm mất mát là khá thấp, tuy nhiên với ba mô hình còn lại (SVR, Ridge và Lasso) cái giá trị hàm mất mát là cao hơn so với mô hình Random Forest đặc biệt là độ đo MSE và RMSE. Các lỗi này có thể bắt nguồn từ việc các hàm của các thuật toán này không thực sự tối ưu với bộ dữ liệu (Ridge và Lasso). Ngoài ra, việc loại bỏ một vài thuộc tính (biến độc lập) có độ quan trọng tương đối (GDP, Percentage expenditure) cũng khiến cho khả năng dự đoán của mô hình bị giảm đi. Mặt khác, bộ dữ liệu có nhiều giá trị bị khuyết, việc điền khuyết bằng giá trị trung bình của thuộc tính phần nào làm thay đổi phân phối giá trị của thuộc tính khiến khả năng dự đoán của mô hình cũng bị giảm.

Các lỗi trong dữ liệu cũng có thể dẫn đến dự đoán sai. Vì mô hình học trên các dữ liệu có sẵn nên có thể có các yếu tố khác hoặc điểm khác mà tuổi thọ phụ thuộc vào.

B. *Hướng phát triển*

Sử dụng phương pháp phân tích các thuộc tính của bài báo cáo để phân tích các bộ dữ liệu khác.

Các yếu tố khác phụ thuộc vào tuổi thọ có thể được đưa vào mô hình bằng cách sử dụng tập dữ liệu lớn hơn.

Phân tích thêm các thuộc tính quốc gia và trình trạng quốc gia như: phát triển hoặc đang đang phát triển để so sánh mức tuổi thọ trung bình giữa các quốc gia.

Thực hiện sử dụng các mô hình máy học nâng cao hơn như các mô hình học kết hợp, các thuật toán máy học khác, mô hình học sâu (deep learning), hay thử nhiều bộ siêu tham số khác để tìm ra phương pháp tối ưu hơn cho bộ dữ liệu.

Các tổ chức phi chính phủ, các bộ phận doanh nghiệp và chính phủ có thể sử dụng mô hình và phân tích này để đề xuất các kế hoạch và chính sách trong tương lai liên quan đến chăm sóc sức khỏe, nâng cao chất lượng sống và tuổi thọ của người dân.

VI. KẾT LUẬN

A. *Kết quả đạt được*

Hoàn thành việc phân tích mức độ của các yếu tố ảnh hưởng đến tuổi thọ.

Xây dựng được các mô hình máy học và cho kết quả đạt độ chính xác trên 80%

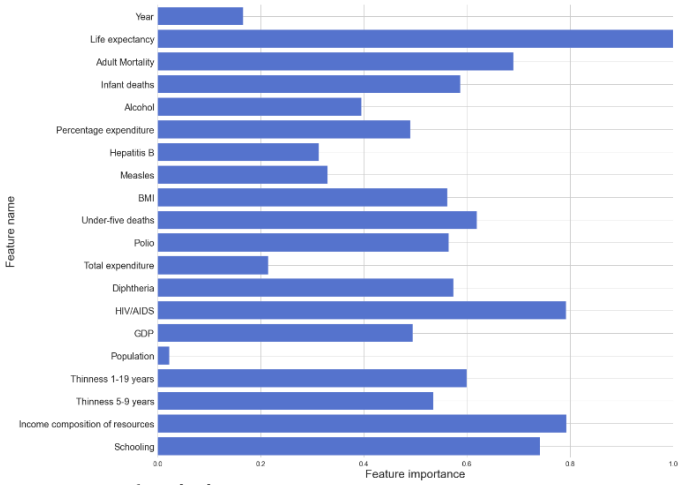
Đã xây dựng được mô hình Random Forest Regression với độ chính xác 96%. Mô hình phát triển hiện tại cho thấy các mô hình được xây dựng phù hợp với bộ dữ liệu và có thể dùng trong thực tế.

B. *Khó khăn gặp phải*

Tìm kiếm nguồn dữ liệu uy tín.

Kiến thức đòi hỏi nâng cao nên gặp một số khó khăn nhất định trong việc lựa chọn mô mình để giải quyết các vấn đề trong thực tiễn.

Thời gian nghiên cứu bị hạn chế nên chưa thể tìm ra được giải pháp tối ưu hơn cho bài toán.



Hình 5. Biểu đồ tầm quan trọng của các thuộc tính với thuộc tính tuổi thọ (Life expectancy)

VII. TÀI LIỆU THAM KHẢO

[1] Life expectancy and healthy life expectancy. Accessed: June 12, 2022. [Online].Available:https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-life-expectancy-and-healthy-life-expectancy

[2] V. Malpe and P. Tugaonkar, "Machine LearningTrends in Medical Sciences," 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), pp. 495-499, August 2018.

[3] Kaggle. Life Expectancy (WHO). Accessed: June 12, 2022. [Online]. Available: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?datasetId=12603

[4] Life expectancy and life tables. Accessed: June 12, 2022 https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/healthy-life-expectancy-(hale)

Bảng phân công công việc

Họ tên	MSSV	Nhiệm vụ
Nguyễn Hoàng Minh	20521609	Xử lý bộ dữ liệu, Viết code phân tích thuộc tính, Hỗ trợ làm slide, Làm báo cáo word.
Tạ Nhật Minh	20521614	Tinh chỉnh đánh giá mô hình, Phân tích lỗi, Hỗ trợ làm slide, Làm báo cáo word.
Nguyễn Minh Tiến	20522010	Thuyết trình, Tìm hiểu hai thuật toán, Viết code chạy model, Làm báo cáo word.
Nguyễn Thiện Thuật	20521998	Làm slide, Tìm hiểu hai thuật toán, Viết code chạy model, Làm báo cáo word.