MỤC LỤC

LỜI NÓI ĐẦU 3
TỔNG QUAN 4
DỰ ĐOÁN TUỔI THỌ SỬ DỤNG HỒI QUY TUYẾN TÍNH 5
Bước 1: Đọc hiểu dữ liệu 5
Bước 2: Xử lí dữ liệu 5
Bước 3: Phân tích dữ liệu 6
Bước 4: **Xây dựng mô hình sử dụng hồi quy tuyến tính** 8
XÂY DỰNG MÔ HÌNH SỬ DỤNG THƯ VIỆN SCIKIT-LEARN . 10
KẾT LUẬN 11
TÀI LIỆU THAM KHẢO 12

LỜI NÓI ĐẦU

Tuổi thọ con người bấy lâu luôn là một vấn đề cần được phân tích, thống kê, và trong những năm qua, các nghiên cứu dự đoán về tuổi thọ đang được sử dụng rộng rãi trong các dịch vụ y tế, chăm sóc sức khỏe và các dịch vụ liên quan đến vấn đề lương hưu, bởi các cơ quan chính phủ cũng như các cơ quan tư nhân. Ở nhiều quốc gia, đó là một vấn đề tranh luận mang tính chính trị về những quy định ở tuổi nghỉ hưu và cách quản lý các vấn đề tài chính liên quan đến vấn đề công cộng. Dự đoán tuổi thọ cung cấp các giải pháp liên quan đến những vấn đề này ở nhiều nước phát triển, cũng như các nước đang phát triển. Với sự tiến bộ trong các kỹ thuật phân tích, dự đoán mang tính hệ thống, chính xác, hiệu quả và định hướng kết quả trong lĩnh vực Khoa học dữ liệu đã giúp phát triển các mô hình dự đoán chính xác, các dự đoán về tuổi thọ đang trở nên nổi bật hơn trong nhu cầu của các cơ quan chính phủ và các cơ quan tư nhân trong việc hoạch định chính sách của họ.

Tuổi thọ con người đã tăng lên nhanh chóng kể từ thời đại Khai sáng. Vào đầu thế kỷ 19, tuổi thọ bắt đầu tăng ở các nước công nghiệp hóa trong khi nó vẫn ở mức thấp ở phần còn lại của thế giới. Sức khỏe tốt, tuổi thọ cao ở các nước giàu và sức khỏe xấu, tuổi thọ thấp ở những quốc gia còn nghèo đói, lạc hậu. Trong những thập kỷ qua, sự chênh lệch toàn cầu này đã giảm. Không có quốc gia nào trên thế giới có tuổi thọ thấp hơn các quốc gia có tuổi thọ cao nhất vào năm 1800. Nhiều quốc gia mà cách đây không lâu đã bị ảnh hưởng bởi sức khỏe kém đang bắt kịp nhanh chóng. Kể từ năm 1900, tuổi thọ trung bình toàn cầu đã tăng hơn gấp đôi và hiện nay là trên 70 tuổi.

Mọi thứ đều có "thời hạn sử dụng" , mọi sinh vật đều có thời gian sinh tồn và con người cũng không ngoại lệ. Với sự phát triển không ngừng trong Máy học và Khoa học dữ liệu, chúng ta có thể dự đoán khá chính xác tuổi thọ của con người với những thông số thiết yếu. Bài báo cáo này sẽ khám phá các thông số ảnh hưởng đến tuổi thọ của các dân cư sống ở các quốc gia đang phát triển và phát triển với sự giúp đỡ của các mô hình học máy.

TỔNG QUAN

Mặc dù đã có rất nhiều nghiên cứu được thực hiện trong quá khứ về các yếu tố ảnh hưởng đến tuổi thọ khi xem xét các biến số nhân khẩu học, thành phần thu nhập và tỷ lệ tử vong. Người ta thấy rằng ảnh hưởng của tiêm chủng và chỉ số phát triển con người đã không được tính đến trong quá khứ. Ngoài ra, một số nghiên cứu trước đây đã được thực hiện xem xét nhiều hồi quy tuyến tính dựa trên bộ dữ liệu một năm cho tất cả các quốc gia. Do đó, điều này tạo động lực để giải quyết cả hai yếu tố đã nêu trước đây bằng cách xây dựng một mô hình hồi quy dựa trên mô hình hiệu ứng hỗn hợp và nhiều hồi quy tuyến tính trong khi xem xét dữ liệu từ giai đoạn 2000 đến 2015 cho tất cả các quốc gia. Tiêm chủng quan trọng như viêm gan B, bại liệt và bạch hầu cũng sẽ được xem xét. Tóm lại, nghiên cứu này sẽ tập trung vào các yếu tố tiêm chủng, yếu tố tử vong, yếu tố kinh tế, yếu tố xã hội và các yếu tố liên quan đến sức khỏe khác. Vì các quan sát bộ dữ liệu này dựa trên các quốc gia khác nhau, nên sẽ dễ dàng hơn cho một quốc gia để xác định yếu tố dự đoán đang góp phần làm giảm giá trị tuổi thọ.

Kho dữ liệu của Đài quan sát Y tế Toàn cầu (GHO) thuộc Tổ chức Y tế Thế giới (WHO) theo dõi tình trạng sức khỏe cũng như nhiều yếu tố liên quan khác cho tất cả các quốc gia Các bộ dữ liệu được cung cấp cho công chúng với mục đích phân tích dữ liệu sức khỏe. Bộ dữ liệu liên quan đến tuổi thọ, các yếu tố sức khỏe của 193 quốc gia đã được thu thập từ cùng một trang web lưu trữ dữ liệu của WHO và dữ liệu kinh tế tương ứng của nó được thu thập từ trang web của Liên Hợp Quốc. Trong số tất cả các loại yếu tố liên quan đến sức khỏe, chỉ có những yếu tố quan trọng đó được chọn đại diện hơn. Người ta đã quan sát thấy rằng trong 15 năm qua, đã có một sự phát triển rất lớn trong lĩnh vực y tế dẫn đến cải thiện tỷ lệ tử vong của con người đặc biệt là ở các nước đang phát triển so với 30 năm qua. Do đó, trong bài báo cáo, ta xem xét sử dụng dữ liệu từ năm 2000-2015 cho 193 quốc gia để phân tích. Khi kiểm tra trực quan ban đầu của dữ liệu cho thấy một số giá trị còn thiếu. Tệp dữ liệu của Kaggle bao gồm 22 Cột và 2938 hàng có nghĩa là 20 biến dự đoán.

DỰ ĐOÁN TUỔI THỌ SỬ DỤNG HỒI QUY TUYẾN TÍNH

<u>Bước 1</u>: Đọc hiểu dữ liệu

Bắt đầu bằng việc tải dữ liệu.

```
In [2]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns;
        from sklearn.linear_model import LinearRegression

In [3]: life_exp = pd.read_csv('Life Expectancy Data.csv')

In [4]: life_exp.head()
```

Out[4]:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	GD
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8.16	65.0	0.1	584.2592
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	8.18	62.0	0.1	612.6965
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8.13	64.0	0.1	631.7449
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	8.52	67.0	0.1	669.9590
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	68.0	0.1	63.5372

5 rows × 22 columns

<u>Bước 2</u>: Xử lí dữ liệu

Chuyển đổi kiểu dữ liệu của cột Status (chuyển thành giá trị số 1 – phát triển, 0 – đang phát triển):

```
from sklearn import preprocessing
label_encoder = preprocessing.LabelEncoder()
life_exp['Status']= label_encoder.fit_transform(life_exp['Status'])
life_exp.head()
```

Xử lí, loại bỏ các giá trị **null:**

```
print(life_exp.isna().sum())
print(life_exp.shape)
```

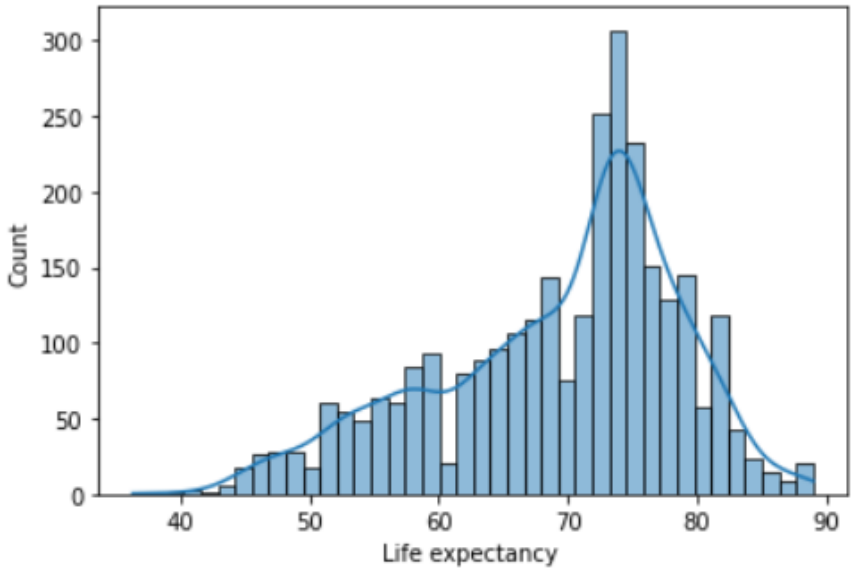
```
life_exp = life_exp.dropna()
```

```
print(life_exp.isna().sum())
print(life_exp.shape)
```

<u>Bước 3</u>: Phân tích dữ liệu

Xem phân bố tuổi thọ sử dụng **hisplot**.

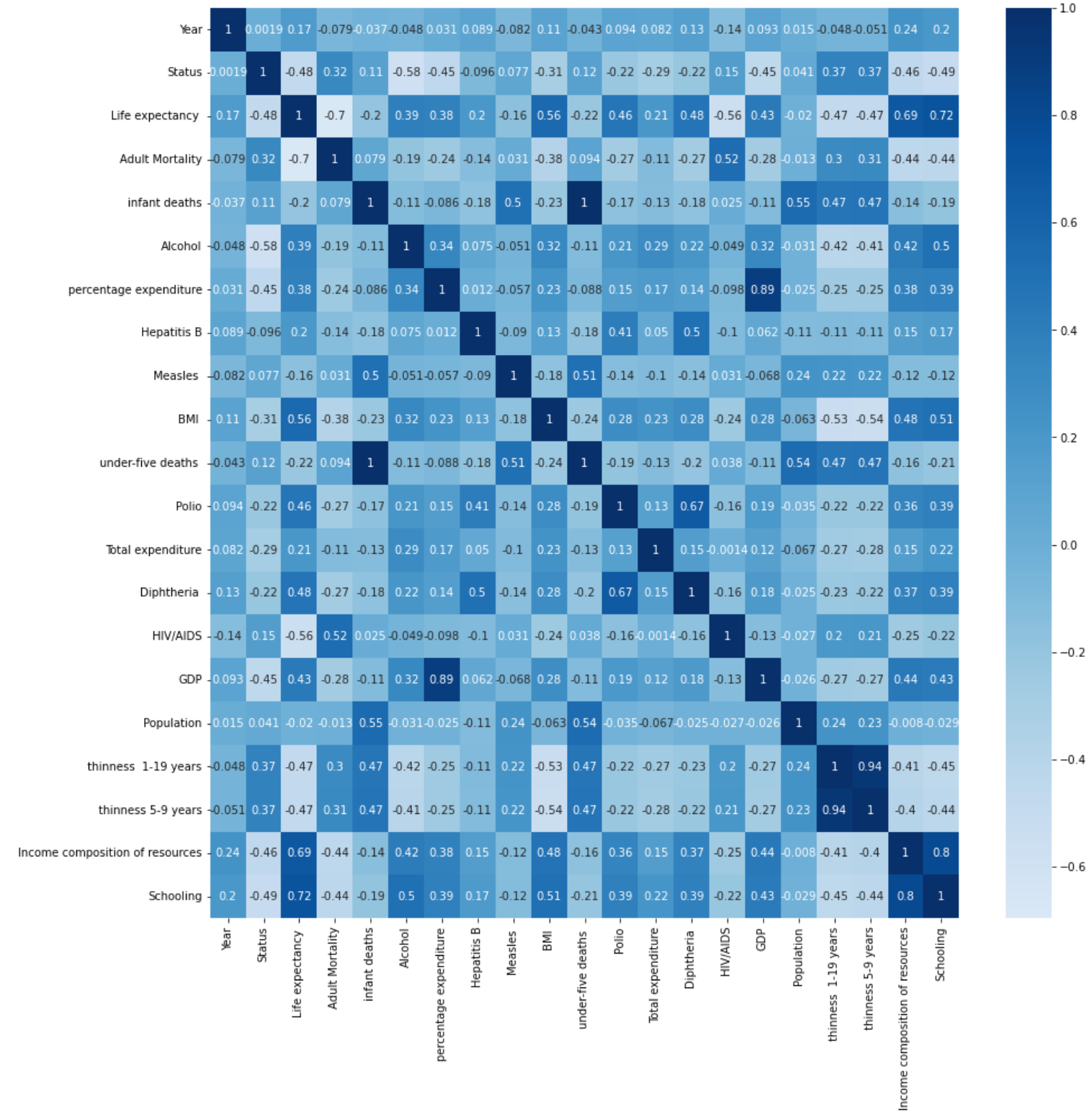
Phân bố tuổi thọ phần lớn nằm trong khoảng từ 45 đến 90 năm, với tuổi thọ trung bình là 69 năm.



Bản đồ nhiệt tương quan là một đại diện đồ họa của một ma trận tương quan đại diện cho mối tương quan giữa các biến khác nhau. Điều này giúp hiểu được sự phụ thuộc tuyến tính của các biến so với nhau. Mối tương quan luôn được tính toán giữa hai biến, và nó có phạm vi [-1, 1].

- Giá trị tương quan gần bằng không có nghĩa là hai biến không có mối tương quan.
- Giá trị tương quan tuyệt đối gần 1 (hoặc -1) có nghĩa là hai biến có mối tương quan cao.

Sử dụng bản đồ nhiệt trong python để trực quan hóa phụ thuộc của các khả năng đối với tuổi thọ.

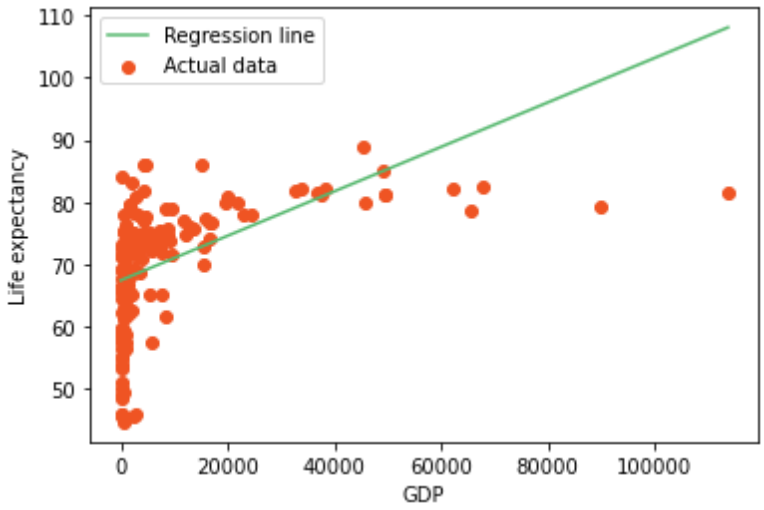


Những thông tin chi tiết sau có thể rút ra dựa trên **Heatmap**.

- Tuổi thọ có mối tương quan đáng kể đối với tỉ lệ tử vong ở người trưởng thành, số năm đi học, thu nhập của các nguồn lực, HIV/AIDS và GDP
- Tuổi thọ và tỉ lệ tử vong ở người trưởng thành có mối tương quan nghịch cao.
- Tuổi thọ và số năm đi học có mối tương quan thuận cao.
- Tuổi thọ có mối tương quan tích cực với BMI.
- GDP cũng có mối tương quan tích cực với tuổi thọ, có thể suy ra rằng khi GDP của đất nước tăng lên, tuổi thọ cũng tăng lên.

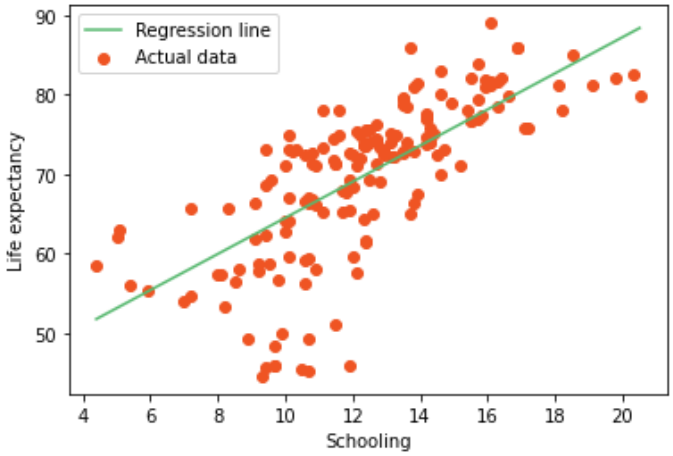
<u>Bước 4</u>: Xây dựng mô hình sử dụng hồi quy tuyến tính

Dự đoán tuổi thọ dựa vào GDP.



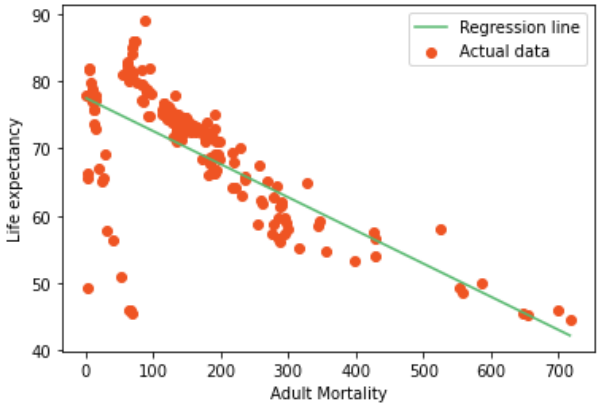
Dựa vào đồ thị hồi quy tuyến tính ta thấy GDP có ảnh hưởng tương đối đối với tuổi thọ. Nhưng ta có thể thấy dữ liệu có điểm ngoại lai, cũng như các dữ liệu không tập trung vào đường hồi quy, mô hình dự đoán tuổi thọ dựa trên GDP có thể thiếu tính chính xác.

Dự đoán tuổi thọ dựa vào dữ liệu Schooling.



Ta nhận thấy, số năm đi học có mối tương quan thuận cao với tuổi thọ. Trình độ giáo dục phổ thông cao có thể giúp ta tiếp thu nhiều kiến thức về sức khoẻ, từ đó xây dựng và áp dụng các thói quen sống lành mạnh, góp phần nâng cao tuổi thọ.

Dự đoán tuổi thọ dựa vào Adult Mortality.



Tỷ lệ tử vong ở người trưởng thành được thể hiện trong xác suất những người đến tuổi 15 đến 60 tuổi tử vong trên 1.000 người. Dựa vào đồ thị hồi quy ta thấy nó có mối tương quan nghịch cao với tuổi thọ, có thể hiểu là tỉ lệ tử vong ở người trưởng thành càng thấp, thì tuổi thọ càng cao (hiển nhiên).

XÂY DỰNG MÔ HÌNH SỬ DỤNG THƯ VIỆN SCIKIT-LEARN

Kiểm tra độ chính xác của mô hình: (mô hình này đạt độ chính xác 86.66%)

```
#Check how well the model works (kiểm tra độ chính xác của model 22 biến)
lr  = LinearRegression()
lr_model = lr.fit(x_train, y_train)
r2_score = lr_model.score(x_test, y_test)
print(r2_score)
```

```
0.8666411611001144
```

```
# Linear Regression
lr1 = LinearRegression()
lr2 = LinearRegression()
lr3 = LinearRegression()

# Reshape your data either using array.reshape(-1, 1)
x_train1 = x_train1.reshape(-1,1)
x_train2 = x_train2.reshape(-1,1)
x_train3 = x_train3.reshape(-1,1)

x_test1 = x_test1.reshape(-1,1)
x_test2 = x_test2.reshape(-1,1)
x_test3 = x_test3.reshape(-1,1)

# Build model
lr1_model = lr1.fit(x_train1, y_train)
lr2_model = lr2.fit(x_train2, y_train)
lr3_model = lr3.fit(x_train3, y_train)

# Predict life expectancy based on x_test.
y_preds1 = lr1_model.predict(x_test1)
y_preds2 = lr2_model.predict(x_test2)
y_preds3 = lr3_model.predict(x_test3)

# Check how well the model works
r2_score1 = lr1_model.score(x_train1, y_train)
r2_score2 = lr2_model.score(x_train2, y_train)
r2_score3 = lr3_model.score(x_train3, y_train)
```

Biến lr chứa một đối tượng LinearRegression trong bộ thư viện Scikit-learn.

Reshape lại dữ liệu.

Fit() thực hiện tính toán tối ưu hóa các tham số B0 và B1.

Dự đoán tuổi thọ dựa vào biến x_test.Sau đó là kiểm tra độ chính xác của mô hình.

Sử dụng dataframe để so sánh các số liệu giữa giá trị thực và giá trị dự đoán.

Out[32]:

	y	y_preds1	y_preds2	y_preds3
0	57.5	69.404254	69.277520	56.489407
1	81.7	68.960003	78.370669	73.447139
2	75.6	70.597491	70.186835	71.278126
3	56.6	67.734778	64.048959	56.390815
4	61.4	67.653277	69.959506	63.341514
...	...	...	...	...
160	81.3	80.876449	78.370669	74.433053
161	62.0	67.827456	53.137180	63.292218
162	56.5	67.415081	61.093686	63.489401
163	77.5	69.029560	74.051423	77.144319
164	57.3	67.671118	59.957042	63.982358

KẾT LUẬN

Nghiên cứu bắt đầu với thông tin, dữ liệu về tuổi thọ và các yếu tố ảnh hưởng đến nó. Tiếp tục quan sát các nhân tố ảnh hưởng và xem xét các mối tương quan giữa tuổi thọ với các nhân tố đó để rút ra kết luận. Cuối cùng là sử dụng và thực hiện thuật toán Hồi quy tuyến tính để dự đoán tuổi thọ.

Trong dự án trên, ta đã xây dựng mô hình dự đoán tuổi thọ bằng thuật toán Hồi quy tuyến tính thông qua 3 yếu tố GDP, số năm đi học và tỉ lệ tử vong ở người trưởng thành. Trong đó các nhân tố như số năm đi học, tỉ lệ tử vong ở người trưởng thành có mối tương quan mạnh với tuổi thọ, đóng vai trò quan trọng trong việc dự đoán; còn GDP cũng có mối tương quan, nhưng không ảnh hưởng quá mạnh. Theo tính toán, mô hình đưa ra dự đoán trên tất cả các biến có độ chính xác là 86.66%, dự đoán trên GDP: 0.19, schooling: 0.51, Adult Mortality: 0.49

Hơn nữa, dự án còn cho thấy rõ ràng tỷ lệ tử vong ở người lớn, HIV / AIDS, BMI, trình độ học vấn và chỉ số thu nhập là những yếu tố tác động nhiều nhất đến tuổi thọ, qua đó nó nhấn mạnh tầm quan trọng của sức khỏe, giáo dục và các đặc điểm kinh tế đối với tuổi thọ. Nhưng vẫn còn một số ảnh hưởng đến tuổi thọ khác như các đặc điểm môi trường và địa lý mà trong bộ dữ liệu hiện tại chưa đề cập, và có thể các nhân tố này sẽ được cập nhật trong tương lai.

TÀI LIỆU THAM KHẢO

1. https://ourworldindata.org/life-expectancy#:~:text=The%20United%20Nations%20estimate%20a,life%20expectancy%20of%2072.3%20years.

2. https://www.enjoyalgorithms.com/blog/life-expectancy-prediction-using-linear-regression

3. https://www.kaggle.com/code/mathchi/life-expectancy-who-with-several-ml-techniques/notebook