**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

# PHÂN TÍCH DỮ LIỆU VỀ TỬ VONG DO UNG THƯ Ở HOA KỲ



| Sinh viên thực hiện: | | |
|---|---|---|
| STT | Họ tên | MSSV |
| 1 | Trần Quang Linh | 18520997 |
| 2 | Trần Trung Hiếu | 18520754 |



**TP. HỒ CHÍ MINH – 12/2020**

MỤC LỤC

# 1. GIỚI THIỆU

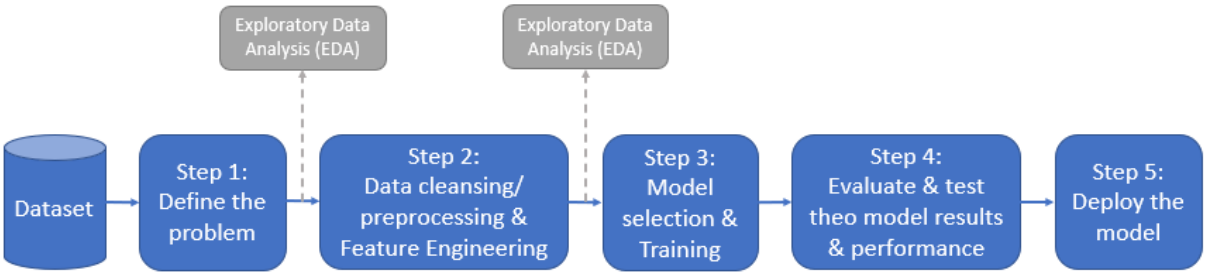
Khoa học dữ liệu đang đóng một vai trò ngày càng quan trọng trong cuộc sống con người với rất nhiều ứng dụng từ các lĩnh vực kinh tế, kĩ thuật và đặc biệt là y tế. Trong đồ án này, nhóm em áp dụng các kiến thức đã học để phân tích dữ liệu, xây dựng mô hình máy học dự đoán tỉ lệ người chết do ung thư tại các hạt của Hoa Kỳ.

Dựa vào quy trình phân tích dữ liệu được học trên lớp, nhóm em đã áp dụng module preprocessing trong thư viện sklearn và phương pháp Log-transformation để tiền xử lí dữ liệu, thư viện matplotlib và seaborn để phân tích trực quan và module linear_model trong thư viện sklearn để xây dựng, đánh giá mô hình. Ngoài ra, nhóm em còn thực hiện Feature Engineering để lựa chọn, kết hợp các thuộc tính giúp tăng độ chính xác cho mô hình.

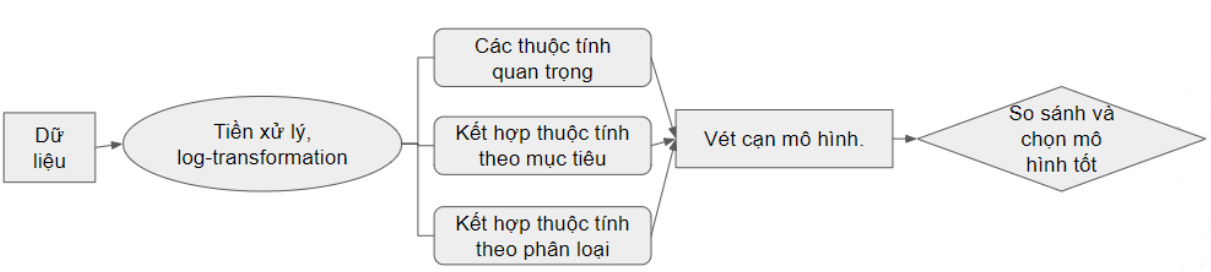
Kết quả đạt được khá khả quan với mô hình tốt nhất có hệ số Root Mean Square Error (RMSE) bằng 14, R2-score Test bằng 0.743 và Mean 5-Fold Validation bằng 0.7.

# 2. NỘI DUNG

Đây là quy trình mà nhóm sử dụng để giải quyết bài toán này. Các mục phía dưới theo như các bước của quy trình này:




Hình 1: Quy trình chung PTDL.




Hình 2: Quy trình cụ thể cho bài toán.

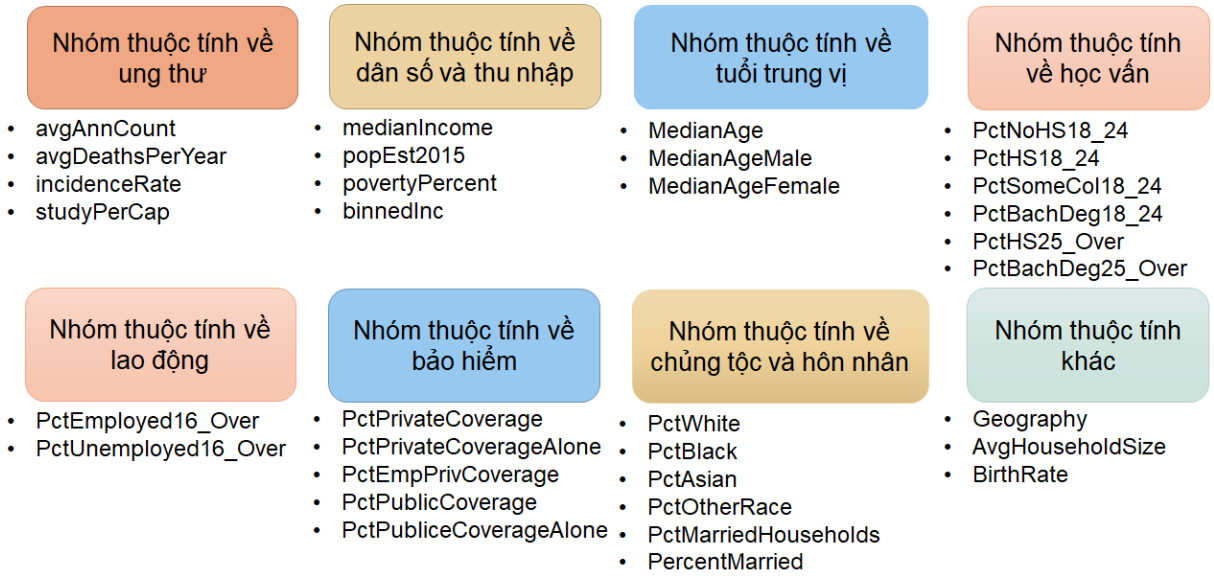
Từ dữ liệu gốc, nhóm em sẽ thực hiện 3 cách tiếp cận khác nhau để tạo ra 3 bộ dữ liệu phục vụ đánh giá kết quả và so sánh để lựa chọn ra mô hình tốt nhất.

- Lựa chọn thuộc tính theo độ tương quan: bộ dữ liệu số 0 bao gồm 13 thuộc tính, bao gồm thuộc tính TARGET_deathRate và 12 thuộc tính có độ tương quan so với biến phụ thuộc cao nhất.

- Kết hợp các thuộc tính theo phân loại: bộ dữ liệu số 1 bao gồm 12 thuộc tính, bao gồm thuộc tính TARGET_deathRate, một số thuộc tính cũ và một số thuộc tính mới được tạo ra bằng cách kết hợp các thuộc tính theo phân loại.
- Kết hợp các thuộc tính theo mục tiêu: bộ dữ liệu số 2 bao gồm 13 thuộc tính, bao gồm thuộc tính TARGET_deathRate, một số thuộc tính cũ và một số thuộc tính mới được tạo ra bằng cách kết hợp các thuộc tính theo mục tiêu.

## 2.1. Bộ dữ liệu 'Cancer'

Bộ dữ liệu Cancer này được giảng viên cung cấp, được tập hợp từ 3 bộ dữ liệu nhỏ hơn (death, cancer_data_notes và incd). Bộ dữ liệu này có 3047 điểm dữ liệu và 34 thuộc tính. Trong đó thuộc tính cần dự đoán là TARGET-deathRate thể hiện số người chết do ung thư trên 100.000 người. 33 thuộc tính còn lại được phân loại thành 8 nhóm sau đây:




Hình 3: 8 nhóm thuộc tính.

Một số thuộc tính được mô tả trong bảng dưới:

Bảng 1: Mô tả thông tin các thuộc tính.

| STT | Tên thuộc tính | Thông tin thuộc tính | Miền giá trị | Kiểu dữ liệu |
|---|---|---|---|---|
| 1 | avgAnnCount | Số người trung bình được dự đoán mắc bệnh ung thư hằng năm | [6, 38150] | Float64 |
| 2 | avgDeathsPerYear | Số lượng người chết trung bình vì ung thư hằng năm | [3, 14010] | Int64 |

| | | | | |
|---|---|---|---|---|
| 3 | incidenceRate | Số người trung bình trên 100.000 người được chuẩn đoán ung thư | [201.3, 1206.9] | Float64 |
| 4 | medIncome | Thu nhập trung bình mỗi hạt | [22640, 125635] | Int64 |
| 5 | popEst2015 | Dân số của hạt đó | [827, 10170292] | Int64 |
| 6 | povertyPercent | Tỉ lệ người nghèo | [3.2, 47.4] | Float64 |
| 7 | studyPerCap | Số thử nghiệm lâm sàng liên quan đến ung thư trên đầu người trên mỗi hạt | [0, 9762.3] | Float64 |
| 8 | binnedInc | Thu nhập bình quân đầu người tính theo thập phân vị | 10 giá trị | Object |
| 9 | MedianAge | Tuổi trung vị | [22.3, 624] | Float64 |
| 10 | Geography | Tên hạt | 3047 giá trị | Object |
| 11 | PercentMarried | Phần trăm công dân kết hôn | [23.1, 72.5] | Float64 |
| 12 | PctNoHS18_24 | Phần trăm cư dân quận trong độ tuổi 18-24 đạt được trình độ dưới trung học phổ thông | [0, 64.1] | Float64 |
| 13 | PctBachDeg18_24 | Phần trăm cư dân quận trong độ tuổi 18-24 có bằng cử nhân | [0, 51.8] | Float64 |
| 14 | PctHS25_Over | Phần trăm cư dân quận trong độ tuổi trên 25 đạt được trình độ trung học phổ thông | [7.5, 54.8] | Float64 |
| 15 | PctBachDeg25_Over | Phần trăm cư dân quận trong độ tuổi trên 25 có bằng cử nhân | [2.5, 42.2] | Float64 |
| 16 | PctEmployed16_Over | Phần trăm công dân trên 16 tuổi có việc làm | [17.6, 80.1] | Float64 |
| 17 | PctUnemployed16_Over | Phần trăm công dân trên 16 tuổi thất nghiệp | [0.4, 29.4] | Float64 |
| 18 | PctPrivate Coverage | Phần trăm cư dân quận có bảo hiểm y tế tư nhân | [22.3, 92.3] | Float64 |
| 19 | PctEmpPriv Coverage | Phần trăm cư dân có bảo hiểm y tế từ công ty cấp | [13.5, 70.7] | Float64 |
| 20 | PctPublic Coverage | Phần trăm cư dân có bảo hiểm y tế từ Chính phủ cấp | [11.2, 65.1] | Float64 |

| | | | | | |
|---|---|---|---|---|---|
| | 21 | PctWhite | Phần trăm của cư dân người da trắng | [10.2, 100] | Float64 |
| | 22 | PctBlack | Phần trăm người da đen | [0, 85.9] | Float64 |
| | 23 | PctMarriedHouseholds | Phần trăm hộ gia đình đã kết hôn | [22.9, 78.0] | Float64 |
| 24 | TARGET_deathRate | | Số người trung bình trên 100.000 người bị chết do ung thư | [59.7, 362.8] | Float64 |

Còn lại 10 thuộc tính là MedianAgeMale, MedianAgeFemale, AvgHouseholdSize, PctHS18_24, PctSomeCol18_24, PctPrivateCoverageAlone, PctPublicCoverageAlone, PctAsian, PctOtherRace, BirthRate nhưng vì giới hạn số lượng trong báo cáo nên nhóm em không thể trình bày đầy đủ ở bảng 1.

➔ Vấn đề đặt ra: Dựa vào các biến độc lập được cung cấp mà đưa ra dự đoán giá trị cho biến phụ thuộc (số người chết do ung thư trên 100.000 người).

## 2.2. Tiền xử lí dữ liệu

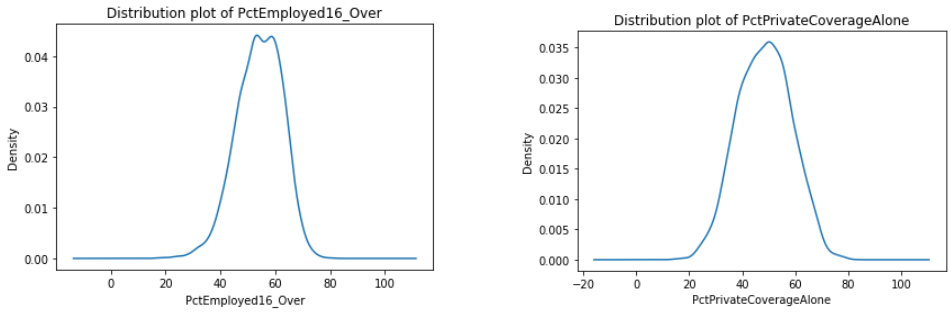
### 2.2.1. Xử lí dữ liệu thiếu

Qua thống kê, nhóm em nhận thấy có 3 thuộc tính bị thiếu dữ liệu: PctSomeCol18_24 với 2285 giá trị bị thiếu, PctEmployed16_Over với 152 giá trị bị thiếu và PctPrivateCoverageAlone với 609 giá trị bị thiếu.

Đối với thuộc tính PctSomeCol18_24, nhóm em chọn phương án loại bỏ thuộc tính này khỏi bộ dữ liệu vì 2 lý do sau[1]:

- Số lượng giá trị bị thiếu là quá lớn (2285/3047).
- Hệ số tương quan của biến này với biến phụ thuộc thấp (-0.18).

Đối với 2 thuộc tính còn lại thì tổng số điểm dữ liệu thiếu chiếm 23.46% so với bộ dữ liệu nên nhóm em tiến hành thay thế giá trị bị thiếu. Nhìn vào hình ảnh phân bố dữ liệu của 2 thuộc tính ta thấy rằng dữ liệu phân bố tương đối chuẩn (không bị lệch). Vì vậy nhóm em quyết định thay thế giá trị bị thiếu bằng giá trị trung bình.




Hình 4: Đồ thị phân phối của 2 thuộc tính PctEmployed16_Over và PctSomeCol18_24.

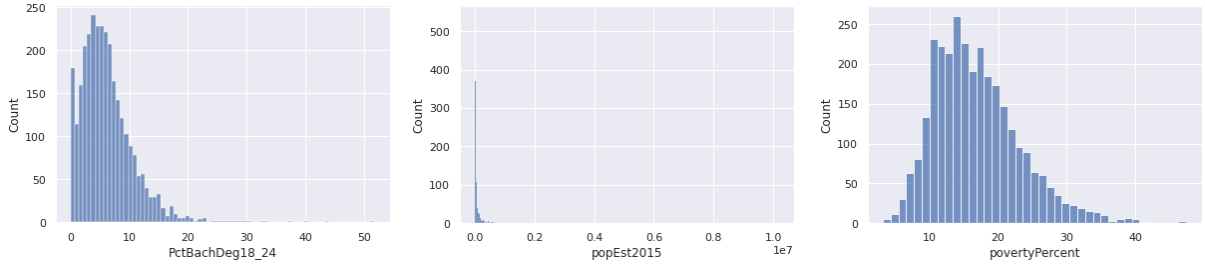
### 2.2.2. Xử lí các thuộc tính thuộc loại biến phân loại

Có 2 thuộc tính là Geography và binnedInc thuộc kiểu dữ liệu Object trong pandas. Đối với thuộc tính Geography thì đây là thuộc tính để chỉ tên hạt (quận) nên có 3047 giá trị khác nhau. Vì vậy nhóm em quyết định chuyển các giá trị của cột này thành index trong dataframe.

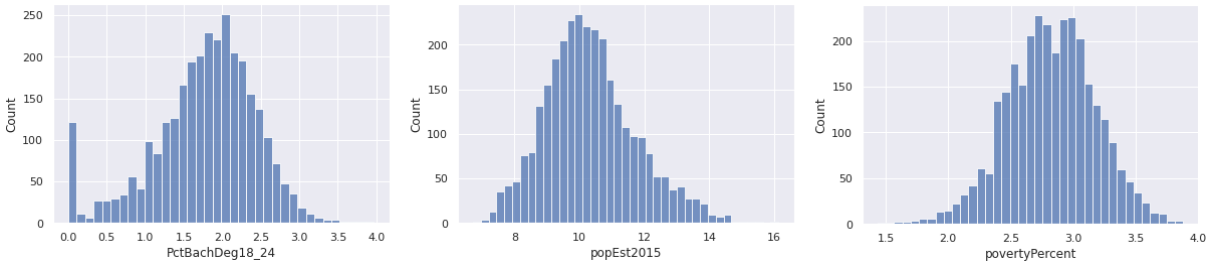
Đối với thuộc tính còn lại là binnedInc thì có 10 nhóm giá trị tương ứng với 10 nhóm thu nhập của các hạt. Thuộc tính này đã được xử lí bằng kĩ thuật binning. Vì vậy nhóm em dùng giá trị trung bình của mỗi nhóm để thay thế cho nhóm đó vì đây là cách phổ biến và dễ dàng nhất để xử lí binning [5].

### 2.2.3. Chuẩn hóa dữ liệu

Nhóm em áp dụng phương pháp Log-transformation [6] để chuẩn hóa cho tất cả các thuộc tính (ngoại trừ biến phụ thuộc TARGET_deathRate). Việc chuẩn hóa dữ liệu sẽ giúp tiết kiệm tài nguyên tính toán cũng như làm cho phân bố của dữ liệu có dạng chuẩn hơn, giảm độ lệch phân bố dữ liệu.




Hình 5: Phân bố dữ liệu của PctBachDeg18_24, popEst2015 và povertyPercent trước khi áp dụng Log-transformation.




Hình 6: Phân bố dữ liệu của PctBachDeg18_24, popEst2015 và povertyPercent sau khi chuẩn hóa Log-transformation.

### 2.2.4. Feature Engineering

Như đã đề cập trước đó, nhóm em thực hiện 3 hướng xử lí như sau:

- Lựa chọn thuộc tính theo độ tương quan: lựa chọn thuộc tính TARGET_deathRate và 12 thuộc tính khác có tương quan so với thuộc tính TARGET_deathRate cao nhất để tạo nên bộ dữ liệu số 0 bao gồm 13 thuộc tính.
- Kết hợp các thuộc tính [2] theo phân loại: dựa vào các nhóm đã được giới thiệu ở phần 2.1 (Hình 3), kết hợp những thuộc tính cùng nằm trong 1 nhóm (+, -, *, /). Ví dụ: cộng 2 thuộc tính ở nhóm giáo dục, chia 2 thuộc tính ở nhóm lao động, … Bằng cách này, từ 4 nhóm thuộc tính: giáo dục, bảo hiểm, lao động và ung thư thì nhóm em đã tạo ra 5 thuộc tính mới: DeathperDiagnose, Education, PctPrivatevsPublic, UnemployvsEmploy và PctEmvPrivatevsPublic. Qua bước này nhóm em có bộ dữ liệu số 1 bao gồm 12 thuộc tính. (Chi tiết xem trong phụ lục).
- Kết hợp các thuộc tính theo mục tiêu: vét cạn tất cả các sự kết hợp giữa 2 thuộc tính (+, -, *, /) để tạo ra thuộc tính mới có độ tương quan so với biến phụ thuộc cao hơn. Khi kết hợp xong, nhóm em có được bộ dữ số 2 có 13 thuộc tính. (Chi tiết xem trong phụ lục).

### 2.3. Phân tích thăm dò

Trong phần này, nhóm em giới thiệu về các phân tích thăm dò[3][4] tiến hành trên bộ dữ liệu số 2.

#### 2.3.1. Các thống kê mô tả

Bảng 2: Thông tin mô tả một vài thuộc tính bộ dữ liệu số 2.

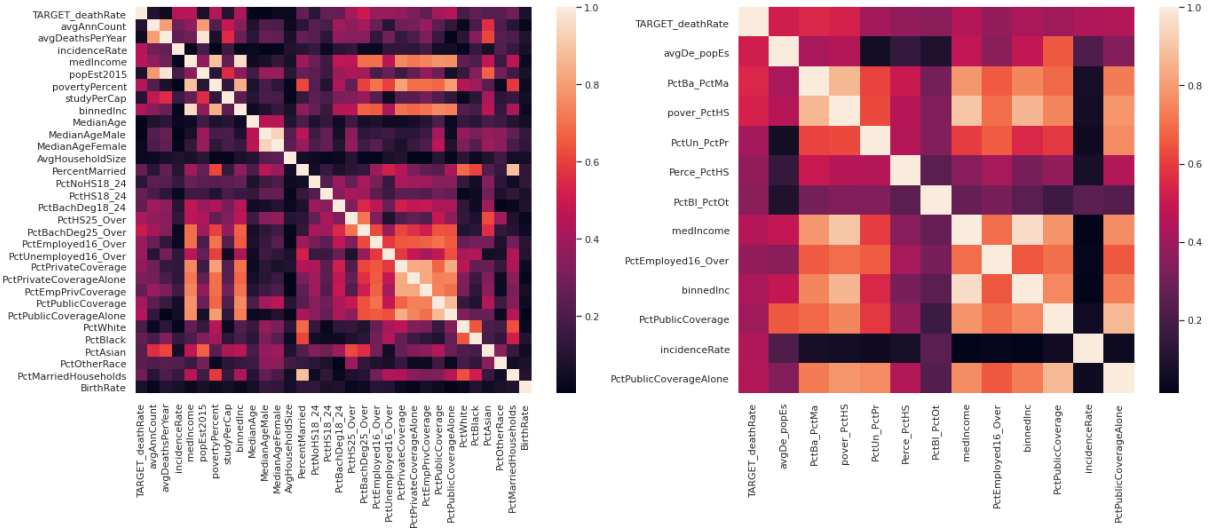
| Thông tin | TARGET_ deathRate | avgDe_ popEs | PctBl_ PctOt | Med Income | BinnedInc |
|---|---|---|---|---|---|
| **count** | 3047 | 3047 | 3047 | 3047 | 3047 |
| **mean** | 178.66 | -6.09 | 0.73 | 10.72 | 10.74 |
| **std** | 27.75 | 0.29 | 1.3 | 0.23 | 0.3 |
| **min** | 59.7 | -7.59 | -3.52 | 10.02 | 10.25 |
| **25%** | 161.2 | -6.24 | -0.08 | 10.56 | 10.56 |
| **50%** | 178.1 | -6.04 | 0.46 | 10.71 | 10.74 |
| **75%** | 195.2 | -5.89 | 1.52 | 10.86 | 10.87 |

| | | | | | |
|---|---|---|---|---|---|
| **max** | 362.8 | -5.11 | 4.46 | 11.74 | 11.44 |

Ở bộ dữ liệu số 2, trước khi kết hợp các thuộc tính lại với nhau thì tất cả các thuộc tính đều có khoảng biến thiên (Range) nhỏ và độ lệch chuẩn (Std) luôn nhỏ hơn 1 nhờ vào Log-transformation. Sau khi kết hợp các thuộc tính, trong số các thuộc tính mới được tạo ra sẽ có một số thuộc tính có độ lệch chuẩn lớn hơn 1. Ví dụ: thuộc tính PctBl_ PctOt có độ lệch chuẩn bằng 1.3.

### 2.3.2. Độ tương quan Pearson Correlation

Do tất cả các thuộc tính đều là định lượng nên nhóm em sử dụng độ đo Pearson Correlation để tính toán độ tương quan giữa các thuộc tính với nhau. Kết quả thu được được thể hiện như sau:



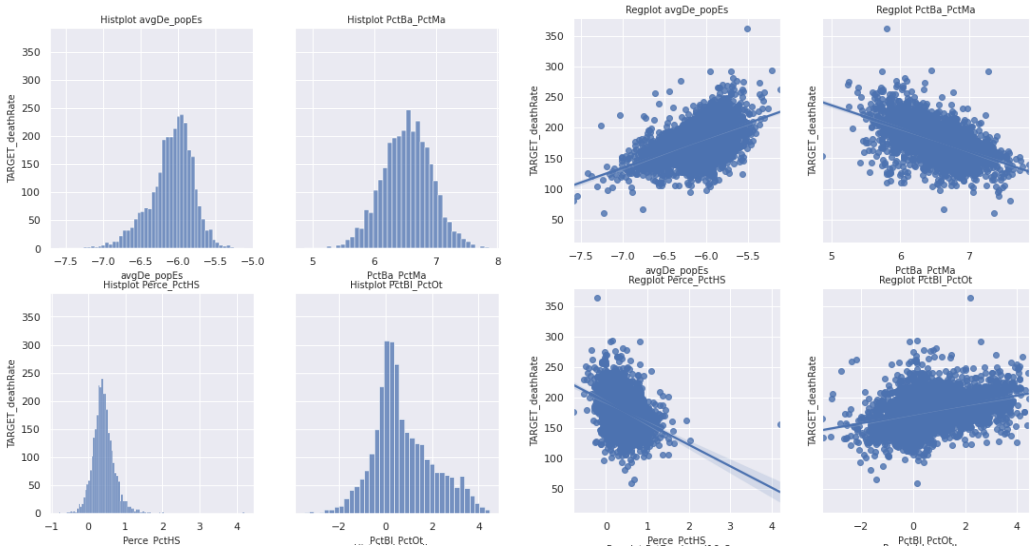

Hình 7: Độ tương quan các thuộc tính của bộ dữ liệu gốc và bộ dữ liệu số 2.

Ở bộ dữ liệu gốc, độ tương quan giữa các biến độc lập so với biến phụ thuộc TARGET-deathRate thường rất thấp (hình bên trái). Trong số các nhóm thuộc tính, chỉ có duy nhất nhóm thuộc tính bảo hiểm là có mức độ tương quan giữa các thuộc tính với nhau cao.

Sau khi thực hiện Feature Engineering như đã đề cập ở phần 2.2.4 để tạo ra bộ dữ liệu số 2 thì độ tương quan giữa các biến độc lập so với biến phụ thuộc và cả độ tương quan giữa các biến độc lập với nhau đều được cải thiện rõ rệt (hình bên phải).

### 2.3.3. Phân tích trực quan

Nhóm em sử dụng các đồ thị như Histogram, Regplot, Boxplot để xem độ phân phối, độ tương quan cũng như độ trải của tập dữ liệu sau khi Feature Engineering theo mục tiêu.

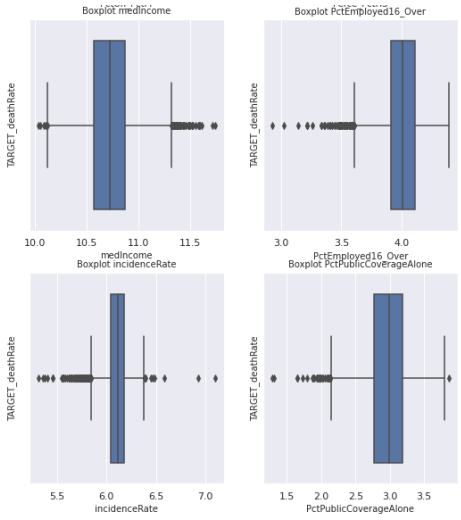



Hình 8: Histogram và Regplot của một vài thuộc tính trên bộ dữ liệu số 2.

Nhìn vào hình Histogram bên trái, ta có thể thấy phân bố của một vài thuộc tính trên cho thấy phân bố dữ liệu đã trở nên phân bố chuẩn hơn, ít có trường hợp dữ liệu lệch trái hay lệch phải. Cùng với đó là khoảng biến thiên (Range) của các thuộc tính đã giảm rất nhiều sau khi sử dụng chuẩn hóa.

Đối với các Regplot, phân bố của các điểm dữ liệu trên là phân bố tập trung nhiều vào trung tâm và có ít điểm ngoại lai. Thêm vào đó thì mức độ tương quan giữa các thuộc tính với biến phụ thuộc này cũng khá cao, có thể nhìn thấy rõ chiều hướng tương quan thuận hay tương quan nghịch. Phương pháp chuẩn hóa Log-transformation đã thể hiện sự hiệu quả trong việc làm giảm độ lệch của phân bố dữ liệu.

Các trực quan trên cho thấy những thuộc tính này phù hợp cho việc xây dựng mô hình hồi quy.




Hình 9: Boxplot của một vài thuộc tính trên bộ dữ liệu số 2.

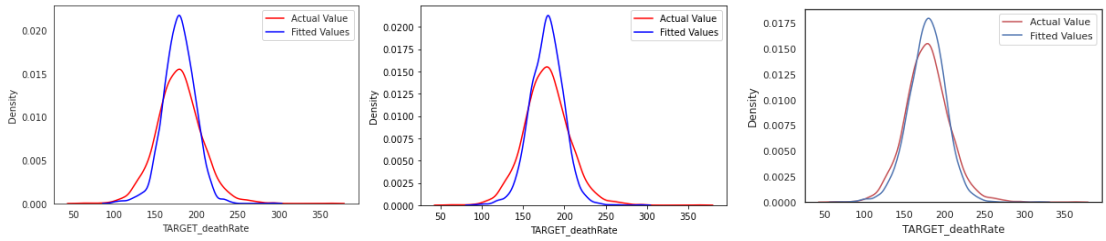
Thông qua một vài thuộc tính trong bộ dữ liệu số 2, ta có thể thấy trên Boxplot là mặc dù vẫn còn tồn tại các điểm outlier nhưng số lượng không quá nhiều nên nhóm em quyết định giữ nguyên outlier mà không cần loại bỏ chúng.

## 2.4. Chọn mô hình và huấn luyện

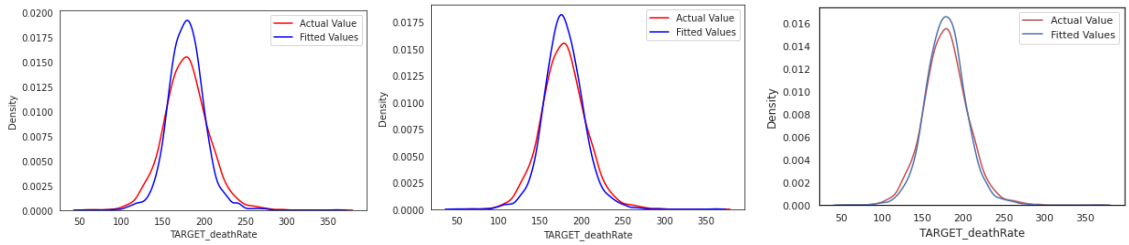
### 2.4.1. Chọn mô hình phù hợp.

Sau khi đã có 3 bộ dữ liệu (số 0, số 1 và số 2) đã sạch và phù hợp cho việc huấn luyện mô hình thì nhóm em tiến hành chọn thuật toán và huấn luyện mô hình.

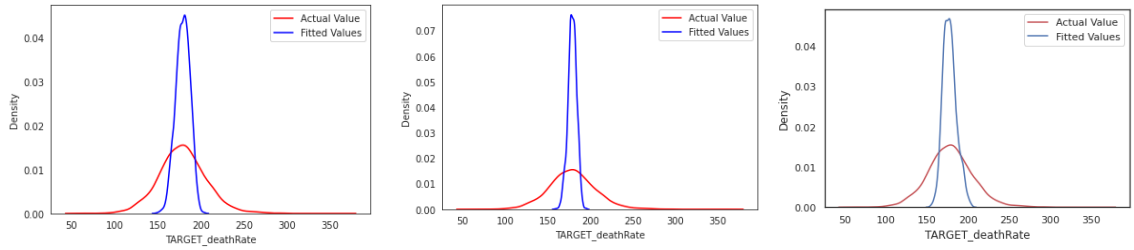
3 thuật toán mà nhóm em chọn để thực nghiệm là Polynomial Regression[7], SVM[8] và Linear Regression. Dựa vào đồ thị phân phối của giá trị dự đoán so với giá trị thực tế và hệ số R2-score để đánh giá xem liệu thuật toán có phù hợp cho việc huấn luyện mô hình hay không.




Hình 10: Distribution plot trên Linear Regression (bộ dữ liệu số 0, 1 và số 2).




Hình 11: Distribution plot trên Polynomial Regression (bộ dữ liệu số 0, 1 và số 2).




Hình 12: Distribution plot trên Support Vector Machine (bộ dữ liệu số 1 vs số 2).

Bảng 3: Tổng hợp kết quả hệ số R2-score trên 3 thuật toán.

| **Mô hình \ Bộ dữ liệu** | **Bộ dữ liệu số 0** | **Bộ dữ liệu số 1** | **Bộ dữ liệu số 2** |
|---|---|---|---|
| Linear Regression | 0.48 | 0.51 | 0.70 |
| **Polynomial Regression** | **0.65** | **0.75** | **0.83** |
| SVM | 0.24 | 0.16 | 0.25 |

Xét về bộ dữ liệu, bộ dữ liệu số 2 luôn cho kết quả tốt nhất trong cả 3 thuật toán. Điều này cho thấy việc kết hợp các thuộc tính theo mục tiêu giúp ích rất nhiều cho mô hình trong quá trình dự đoán.

Xét về mặt thuật toán, kết quả thu được trên Polynomial Regression là cao nhất trên cả 3 bộ dữ liệu, phù hợp để đưa vào huấn luyện mô hình. Linear Regression thấp hơn Polynomial Regression nhưng kết quả vẫn ở mức chấp nhận được, có thể chọn Linear Regression để huấn luyện. Thuật toán SVM cho kết quả rất thấp trên cả 3 bộ dữ liệu. Điều này chứng tỏ SVM không phù hợp để đưa vào huấn luyện.

Như vậy Polynomial Regression và Linear Regression sẽ là 2 thuật toán được nhóm em triển khai huấn luyện mô hình.

### 2.4.2. Huấn luyện mô hình

Nhóm em tiến hành huấn luyện mô hình bằng cách vét cạn tất cả các tập thuộc tính và huấn luyện bằng Polynomial Regression từ bậc 2 tới 4 và cả Linear Regression. Đối với từng mô hình có được, nhóm em quan tâm tới những tiêu chí sau đây:

- Root Mean Square Error (RMSE)
- R2-score trên tập Train và tập Test (test_size bằng 0.3)
- Trung bình và độ lệch chuẩn của cross_val_score sau khi thực hiện cross-validation với K bằng 5.

Sau hơn 3 giờ thực hiện vét cạn và huấn luyện mô hình, số lượng mô hình được tạo ra được thể hiện trong bảng sau:

Bảng 4: Thống kê số lượng mô hình sau khi vét cạn.

| | Linear Regression | Polynomial Feature (bậc = 2) | Polynomial Feature (bậc = 3) | Polynomial Feature (bậc = 4) |
|---|---|---|---|---|
| Bộ dữ liệu số 0 | 4095 | 4095 | 4095 | 4095 |
| Bộ dữ liệu số 1 | 2047 | 2047 | 2047 | 2047 |
| Bộ dữ liệu số 2 | 4095 | 4095 | 4095 | 4095 |

## 2.5. Đánh giá kết quả

Bảng 5: Mô hình dự đoán tốt nhất.

| | Mô hình | Thuộc tính | RMSE | R2-score Train | R2-score Test | Mean 5-Fold Validat-ion | Std 5-Fold Validat-ion |
|---|---|---|---|---|---|---|---|
| Bộ dữ liệu số 0 | Polynomial Regression (bậc = 3) | incidenceRate, binnedInc, PctBachDeg25_Over, PctPublicCoverage, PctPublicCoverageAlone | 19.03 | 0.534 | 0.518 | 0.474 | 0.041 |
| Bộ dữ liệu số 1 | Polynomial Regression (bậc = 3) | incidenceRate, binnedInc, DeathperDiagnose, UnemployvsEmploy, Education | 16.09 | 0.665 | 0.659 | 0.586 | 0.085 |
| Bộ dữ liệu số 2 | Polynomial Regression (bậc = 2) | avgDe_popEs, PctBa_PctMa, PctUn_PctPr, Perce_PctHS, PctBl_PctOt, PctEmployed16_Over, PctPublicCoverage, incidenceRate, PctPublicCoverageAlone | 14.00 | 0.745 | 0.743 | 0.700 | 0.026 |

Với bộ dữ liệu số 0, mô hình cho kết quả tốt nhất là mô hình có các thuộc tính incidenceRate, PctPublicCoverageAlone, PctBachDeg25_Over, PctPublicCoverage và binnedInc. Tuy nhiên, kết quả này vẫn khá thấp (giá trị Mean 5-Fold Validation bằng 0.474). Không có mô hình nào đạt Mean 5-Fold Validation trên 0.5. Như vậy cách tiếp cận bằng phương pháp lựa chọn thuộc tính theo độ tương quan không phù hợp so với bài toán này.

Chọn ngưỡng Mean 5-Fold Validation lớn hơn 0.5, ở bộ dữ liệu số 1 nhóm em có 479 mô hình. Trong đó số lượng các mô hình Polynomial Regression bậc 2, 3, 4 lần lượt là 251, 223 và 4. Mô hình tốt nhất là Polynomial Regression với bậc bằng 3 được huấn luyện từ các thuộc tính sau đây: incidenceRate, binnedInc, UnemployvsEmploy, DeathperDiagnose và Education. Các giá trị RMSE, R2-score Train, R2-score Test và Mean 5-Fold Validation đều ở mức tương đối. Giá trị Std 5-Fold Validation bằng 0.085. Tuy mô hình này là tốt nhất ở bộ dữ liệu số 1 nhưng kết quả chỉ ở mức tương đối tốt, không quá cao. Do đó, việc kết hợp các thuộc tính theo phân loại đem lại hiệu quả ở mức tương đối.

Ở bộ dữ liệu số 2 có 3711 mô hình mà Mean 5-Fold Validation lớn hơn 0.5. Trong đó có 1314 mô hình Linear Regression, số lượng mô hình Polynomial Regression bậc 2, 3, 4 lần lượt là 1377, 987 và 33. Mô hình tốt nhất là Polynomial Regression với bậc bằng 2 được huấn luyện từ 9 thuộc tính sau: avgDe_popEs, PctBa_PctMa, PctUn_PctPr, Perce_PctHS, PctEmployed16_Over, PctPublicCoverage, incidenceRate, PctBl_PctOt, PctPublicCoverageAlone. Ngoại trừ Std 5-Fold Validation, các giá trị còn lại ở mô hình này như RMSE, R2-score Train, R2-score Test và Mean 5-Fold Validation đều cao hơn so với tất cả các mô hình mà nhóm em đã tạo ra. Vì vậy, việc kết hợp các thuộc tính theo mục tiêu đem lại hiệu quả ấn tượng nhất trong số 3 cách tiếp cận mà nhóm em đã thực nghiệm.

Nhóm em quyết định lựa chọn mô hình tốt nhất khi huấn luyện trên bộ dữ liệu số 2 để dự đoán giá trị cho biến phụ thuộc (tỉ lệ người chết do ung thư).

## 3. KẾT LUẬN

Trong quá trình làm việc, nhóm em đã tiến hành các việc như sau:

- Từ dữ liệu gốc, tiến hành tiền xử lý dữ liệu:
  - + Xử lý dữ liệu bị thiếu: xóa thuộc tính PctSomeCol18_24, thay thế giá trị bị thiếu bằng giá trị trung bình trong thuộc tính PctEmployed16_Over và PctSomeCol18_24.
  - + Xử lý thuộc tính thuộc loại biến phân loại: đưa thuộc tính Geography làm index cho dataframe, thay thế giá trị trung bình mỗi nhóm con trong thuộc tính BinnedInc.
  - + Chuẩn hóa dữ liệu: áp dụng Log-transformation.
- Sử dụng Feature Engineering để kết hợp các thuộc tính.
- Huấn luyện mô hình: áp dụng Linear Regression và Polynomial Regression với bậc lần lượt bằng 2,3,4.

Nhóm em đã thực nghiệm và phân tích kết quả trên cả 3 cách tiếp cận: lựa chọn các thuộc tính quan trọng, kết hợp các thuộc tính theo mục tiêu và kết hợp các thuộc tính theo phân loại. Kết quả cho thấy rằng cách tiếp cận bằng cách kết hợp các thuộc tính theo mục tiêu có hiệu quả cao nhất. Điều này mang lại cho nhóm em mô hình dự đoán thuộc tính TARGET_deathRate với hệ số Mean 5-Fold Validation bằng 0.7.

# TÀI LIỆU THAM KHẢO

**[1]** Hyun Kang, The prevention and handling of the missing data, 2013.

**[2]** 7 Feature Engineering Techniques in Machine Learning.

Link: https://www.analyticsvidhya.com/blog/2020/10/7-feature-engineering-techniques-machine-learning/# (6/12/2020)

**[3]** What is EDA.

Link: https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm. (8/12/2020)

**[4]** Significance of Exploratory Data Analysis(EDA)

Link: http://www.jeannjoroge.com/significance-of-exploratory-data-anaysis/ (8/12/2020)

**[5]** Paul T. von Hippel, David J. Hunter, McKalie Drown, Better estimates from binned income data:Interpolated CDFs and mean-matching, 2017.

**[6]** Changyong FENG, Hongyue WANG, Naiji LU, Tian CHEN, Hua HE, Ying LU, Xin M. TU, Log-transformation and its implications for data analysis, 2014.

**[7]** Eva Ostertagova, Modelling using polynomial regression, 2012.

**[8]** DU Shu-xin, WU Tie-jun, Support Vector Machines for Regression, 2003

# PHỤ LỤC

Bảng 1: Bảng phân công công việc

| STT | Thành viên | Nhiệm vụ |
|---|---|---|
| 1 | Trần Quang Linh | Phân tích dữ liệu bộ số 0, 1. Viết slide báo cáo, thuyết trình. |
| 2 | Trần Trung Hiếu | Phân tích dữ liệu bộ số 2. Viết báo cáo. |

## 1. Bộ dữ liệu số 0

Bộ dữ liệu này là các thuộc tính quan trọng (có độ tương quan với biến phụ thuộc) được lấy ra từ qua trình phân tích thăm dò. Bộ dữ liệu số 0 có 13 thuộc tính (1 biến phụ thuộc và 12 biến độc lập)

Bảng 2: Bộ dữ liệu số 0

| STT | Tên thuộc tính | Độ tương quan so với thuộc tính TARGET_ deathRate |
|---|---|---|
| 1 | TARGET_deathRate | 1 |
| 2 | medIncome | -0.4522 |
| 3 | povertyPercent | 0.4382 |
| 4 | binnedInc | -0.4256 |
| 5 | PctHS25_Over | 0.4119 |
| 6 | PctBachDeg25_Over | -0.4964 |
| 7 | incidenceRate | 0.4436 |
| 8 | PctEmployed16_Over | -0.3813 |
| 9 | PctUnemployed16_Over | 0.3808 |
| 10 | PctPrivateCoverage | -0.3651 |
| 11 | PctPrivateCoverageAlone | -0.3098 |

| | | |
|---|---|---|
| 12 | PctPublicCoverage | 0.4052 |
| 13 | PctPublicCoverageAlone | 0.4483 |

**2. Bộ dữ liệu số 1**

Bộ dữ liệu này được tạo ra từ bộ dữ liệu Cancer. Có tổng cộng 12 thuộc tính (1 biến phụ thuộc và 11 biến độc lập).

Bảng 3: Bộ dữ liệu số 1

| **STT** | **Tên thuộc tính** | **Cách kết hợp từ bộ dữ liệu Cancer** | **Độ tương quan so với thuộc tính TARGET_ deathRate** |
|---|---|---|---|
| 1 | TARGET_deathRate | TARGET_deathRate | 1 |
| 2 | medIncome | medIncome | -0.4522 |
| 3 | povertyPercent | povertyPercent | 0.4382 |
| 4 | binnedInc | binnedInc | -0.4256 |
| 5 | PctPrivateCoverageAlone | PctPrivateCoverageAlone | -0.3098 |
| 6 | PctPublicCoverageAlone | PctPublicCoverageAlone | 0.4483 |
| 7 | incidenceRate | incidenceRate | 0.4436 |
| 8 | PctPrivatevsPublic | PctPrivateCoverage / PctPublicCoverage | -0.4219 |
| 9 | PctEmvPrivatevsPublic | PctEmpPrivCoverage / PctPublicCoverage | -0.3550 |
| 10 | UnemployvsEmploy | PctUnemployed16_Over / PctEmployed16_Over | 0.4056 |
| 11 | Education | PctHS25_Over / PctBachDeg25_Over | 0.5000 |
| 12 | DeathperDiagnose | avgDeathsPerYear / avgAnnCount | 0.2838 |

### 3. Bộ dữ liệu số 2

Bộ dữ liệu này được tạo ra từ bộ dữ liệu Cancer. Có tổng cộng 13 thuộc tính (1 biến phụ thuộc và 12 biến độc lập)

Bảng 4: Bộ dữ liệu số 2

| STT | Tên thuộc tính | Cách kết hợp từ bộ dữ liệu Cancer | Độ tương quan so với thuộc tính TARGET_ deathRate |
|---|---|---|---|
| 1 | TARGET_deathRate | TARGET_deathRate | 1 |
| 2 | binnedInc | binnedInc | -0.425 |
| 3 | PctEmployed16_Over | PctEmployed16_Over | -0.381 |
| 4 | incidenceRate | incidenceRate | 0.443 |
| 5 | PctPublicCoverage | PctPublicCoverage | 0.405 |
| 6 | PctPublicCoverageAlone | PctPublicCoverageAlone | 0.448 |
| 7 | medIncome | medIncome | -0.452 |
| 8 | avgDe_popEs | avgDeathsPerYear - popEst2015 | 0.519 |
| 9 | PctBa_PctMa | PctBachDeg25_Over + PctMarriedHouseholds | -0.547 |
| 10 | pover_PctHS | povertyPercent + PctHS25_Over | 0.531 |
| 11 | PctUn_PctPr | PctUnemployed16_Over - PctPrivateCoverage | 0.413 |
| 12 | Perce_PctHS | PercentMarried - PctHS18_24 | -0.371 |
| 13 | PctBl_PctOt | PctBlack – PctOtherRace | 0.353 |