**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT**

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue circle with white text

Description automatically generated

**BÁO CÁO**

**BÁO CÁO BÀI TẬP GIỮA KÌ**

**MÔN HỌC: TRÍ TUỆ NHÂN TẠO**

**ĐỀ TÀI**

**DỰ BÁO CHI PHÍ BẢO HIỂM SỨC KHỎE DỰA TRÊN CÁC YẾU TỐ ẢNH HƯỞNG BẰNG MÔ HÌNH HỒI QUY TUYẾN TÍNH**

**SINH VIÊN THỰC HIỆN**

Nguyễn Thị Bảo Ngọc 2221050172

Lớp DCCTCLC67A

**CÁN BỘ GIẢNG DẠY**

**TS. Nguyễn Hoàng Long**

**Hà Nội, 2024**

**MỤC LỤC**

[**MỤC LỤC** 1](#_Toc181317656)

[**DANH MỤC HÌNH ẢNH** 2](#_Toc181317657)

[I. TÓM TẮT ĐỀ TÀI 4](#_Toc181317658)

[II. MỤC TIÊU 5](#_Toc181317659)

[III. PHƯƠNG PHÁP NGHIÊN CỨU 5](#_Toc181317660)

[1. Sử dụng mô hình hồi quy tuyến tính đa biến 5](#_Toc181317661)

[2. Kết hợp với ngôn ngữ Python 6](#_Toc181317662)

[3. Sử dụng môi trường Jupyter notebook: 7](#_Toc181317663)

[IV. CÁC BƯỚC THỰC HIỆN 8](#_Toc181317664)

[1. Tìm kiếm dữ liệu 8](#_Toc181317665)

[2. Import các thư viện cần thiết 8](#_Toc181317666)

[3. Đọc tệp dữ liệu 9](#_Toc181317667)

[4. Xử lý dữ liệu 9](#_Toc181317668)

[5. Trực quan hóa dữ liệu 14](#_Toc181317669)

[6. Chọn biến độc lập tốt 22](#_Toc181317670)

[7. Chia tệp dữ liệu 23](#_Toc181317671)

[8. Xây dựng mô hình 24](#_Toc181317672)

[9. Kiểm tra các giả định 26](#_Toc181317673)

[V. KẾT LUẬN 28](#_Toc181317674)

# **DANH MỤC HÌNH ẢNH**

[Hình 1 - Import các thư viện 8](#_Toc181317630)

[Hình 2 - Hiển thị dữ liệu 9](#_Toc181317631)

[Hình 3 - Kiểm tra giá trị trùng lặp 9](#_Toc181317632)

[Hình 4 - Tìm giá trị trùng lặp 9](#_Toc181317633)

[Hình 5 - Xóa giá trị trùng lặp 10](#_Toc181317634)

[Hình 6 - Thông tin dữ liệu 10](#_Toc181317635)

[Hình 7 - Đếm và kiểm tra dữ liệu có đồng nhất hay không 11](#_Toc181317636)

[Hình 8 - Kiểm tra khoảng trắng trong cột dữ liệu object 11](#_Toc181317637)

[Hình 9 - Xử lý các dữ liệu bị thiếu 12](#_Toc181317638)

[Hình 10 - Chỉ số của các cột dữ liệu 12](#_Toc181317639)

[Hình 11 - Tính chỉ số z\_score để tìm giá trị ngoại lai 13](#_Toc181317640)

[Hình 12 - Xóa các dữ liệu ngoại lai 13](#_Toc181317641)

[Hình 13 - Biểu đồ pairplot thể hiện mối tương quan giữa các biến 14](#_Toc181317642)

[Hình 14 Biểu đồ barplot tương quan giữa Age và Charges 15](#_Toc181317643)

[Hình 15 Biểu đồ barplot tương quan giữa Age, Sex với Charges 16](#_Toc181317644)

[Hình 16 Biểu đồ barplot tương quan giữa smoker với charges 17](#_Toc181317645)

[Hình 17 Biểu đồ barplot tương quan giữa Region với Charges 18](#_Toc181317646)

[Hình 18 - Biểu đồ scatter thể hiện mối tương quan 3 biến 20](#_Toc181317647)

[Hình 19 Biểu đồ scatter tương quan giữa bmi, sex với charges 21](#_Toc181317648)

[Hình 21 Kiểm tra tính đa cộng tuyến 22](#_Toc181317649)

[Hình 20 - Hàm Forward Selection 23](#_Toc181317650)

[Hình 22 - Chia tệp dữ liệu 24](#_Toc181317651)

[Hình 23 - Kết quả mô hình 25](#_Toc181317652)

[Hình 24 - Kiểm tra tính chuẩn tắc 26](#_Toc181317653)

[Hình 25 - KIểm tra tính đồng nhất 26](#_Toc181317654)

[Hình 27 - Kiểm tra hiện tượng fitting 27](#_Toc181317655)

1. **TÓM TẮT ĐỀ TÀI**

Đề tài này tập trung vào xem xét các yếu tố như tuổi, giới tính, chỉ số BMI, số lượng trẻ em, việc hút thuốc và khu vực cư trú của nước Mỹ ở 4 khu vực Đông Nam, Đông Bắc, Tây Nam, Tây Bắc ảnh hưởng đến mức phải trả tiền cho bảo hiểm sức khỏe vì chúng có mối liên hệ với rủi ro bệnh tật và chi phí y tế. Dưới đây là cách mỗi yếu tố có thể ảnh hưởng:

* Tuổi: Người cao tuổi thường có nguy cơ mắc các bệnh liên quan đến tuổi tác cao hơn, do đó, có thể đòi hỏi chi phí bảo hiểm cao hơn.
* Giới tính: Có những bệnh lý ảnh hưởng đặc biệt đến nam giới hoặc nữ giới, do đó, giới tính có thể ảnh hưởng đến mức độ rủi ro và chi phí bảo hiểm.
* Chỉ số BMI: Người có chỉ số BMI cao hơn có nguy cơ cao hơn mắc các bệnh như tiểu đường, bệnh tim và các vấn đề sức khỏe khác, dẫn đến chi phí bảo hiểm cao hơn.
* Số lượng trẻ em: Việc bảo hiểm càng nhiều người (bao gồm trẻ em) sẽ tăng chi phí bảo hiểm do có nhiều người phải được bảo vệ.
* Người hút thuốc: Người hút thuốc thường có nguy cơ cao hơn mắc các bệnh lý liên quan đến thuốc lá, nên bảo hiểm có thể tăng giá để bù đắp cho rủi ro này.
* Khu vực cư trú: Chi phí y tế và mức độ rủi ro bệnh tật có thể biến đổi tùy theo khu vực, dẫn đến sự khác biệt trong mức độ chi phí bảo hiểm.

Sử dụng mô hình hồi quy tuyến tính đa biến cho phép chúng ta đồng thời đánh giá tác động của nhiều yếu tố này đến chi phí bảo hiểm sức khỏe. Kết quả từ mô hình này sẽ cung cấp thông tin quan trọng giúp công ty bảo hiểm hiểu rõ hơn về rủi ro và các yếu tố ảnh hưởng đến chi phí bảo hiểm, từ đó họ có thể điều chỉnh chiến lược và chính sách bảo hiểm một cách hiệu quả hơn để bảo vệ khách hàng.

1. **MỤC TIÊU**

Mục tiêu chính của báo cáo này là khám phá những yếu tố ảnh hưởng đến chi phí bảo hiểm sức khỏe, cung cấp thông tin quan trọng giúp công ty bảo hiểm hiểu rõ hơn về rủi ro và các yếu tố ảnh hưởng đến chi phí bảo hiểm, từ đó họ có thể điều chỉnh chiến lược và chính sách bảo hiểm một cách hiệu quả hơn để bảo vệ khách hàng, cụ thể:

* Xây dựng mô hình hồi quy tuyến tính dự đoán chi phí bảo hiểm
* Chạy mô hình hồi quy tuyến tính cho ra kết quả.
* So sánh được tệp dữ liệu dự báo với tệp dữ liệu gốc.
* Tính toán các chỉ số và xem chỉ số nào phù hợp nhất.
* Đưa ra kết luận.

1. **PHƯƠNG PHÁP NGHIÊN CỨU**
2. **Sử dụng mô hình hồi quy tuyến tính đa biến**

Hồi quy tuyến tính đa biến là một kỹ thuật thống kê khám phá cách một số biến độc lập ảnh hưởng đến một biến phụ thuộc – giống như việc tìm hiểu cách các nguyên liệu khác nhau trong một công thức nấu ăn, ảnh hưởng đến hương vị của món ăn.

Trong hồi quy tuyến tính đa biến, ta dự đoán kết quả (biến phụ thuộc) dựa trên các giá trị của hai hoặc nhiều yếu tố (biến độc lập):

A group of blue symbols

Description automatically generated

Trong lĩnh vực máy học, các thuật toán được sử dụng để phân tích các tập dữ liệu lớn và dự đoán giá trị của các biến không xác định dựa trên dữ liệu đã biết. Thuật toán được huấn luyện trên dữ liệu đã biết hoặc có nhãn trước khi được áp dụng để dự đoán giá trị của dữ liệu chưa biết. Trong thực tế, quá trình này phức tạp hơn so với ví dụ đơn giản về hồi quy tuyến tính. Điều này yêu cầu phải điều chỉnh hoặc biến đổi các giá trị dữ liệu để đáp ứng các giả định quan trọng sau đây:

Mối quan hệ tuyến tính: Để xác định mối quan hệ giữa biến độc lập và biến phụ thuộc trong hồi quy tuyến tính đa biến, các biểu đồ phân tán được sử dụng để kiểm tra liệu chúng có tuân theo một mô hình tuyến tính hay không. Nếu không, có thể cần áp dụng các biến đổi phi tuyến tính như căn bậc hai hoặc log để xác định mối quan hệ.

Phần dư độc lập: Phần dư được sử dụng để đo lường độ chính xác của dự đoán. Chúng biểu thị sự chênh lệch giữa dữ liệu thực tế và giá trị dự đoán từ mô hình. Việc đảm bảo phần dư độc lập quan trọng để mô hình hồi quy đa biến hoạt động hiệu quả.

Tính chuẩn: các kỹ thuật vẽ đồ thị như các biểu đồ Q-Q xác định xem các phần dư có phân bố bình thường hay không. Các phần dư sẽ nằm dọc theo một đường chéo ở trung tâm của đồ thị. Nếu phần dư không được chuẩn hóa, bạn có thể kiểm tra dữ liệu cho các điểm ngoại lai ngẫu nhiên hoặc các giá trị không phải là điển hình. Loại bỏ các điểm ngoại lai hoặc thực hiện phép biến đổi phi tuyến tính có thể khắc phục sự cố.

Phương sai không đổi: phương sai không đổi giả định rằng phần dư có phương sai cố định hoặc độ lệch chuẩn so với trung bình cho mọi giá trị của x. Nếu không, kết quả phân tích có thể không chính xác. Nếu giả định này chưa được đáp ứng, bạn có thể phải thay đổi biến phụ thuộc. Bởi vì phương sai xuất hiện tự nhiên trong các tập dữ liệu lớn, giải thích được lý do phải thay đổi quy mô của biến phụ thuộc. Ví dụ: thay vì sử dụng quy mô dân số để dự đoán số trạm cứu hỏa trong một thành phố, có thể sử dụng kích thước dân số để dự đoán số trạm cứu hỏa trên mỗi người.

1. **Kết hợp với ngôn ngữ Python**

Python là một ngôn ngữ lập trình phổ biến được sử dụng trong web development (server-side), software development, mathematics, system scripting. Do có thể chạy gần như mọi kến trúc hệ thống, hỗ trợ các mô-đun và packages khác nhau, cho phép sử dụng mô đun chương trình và tái sử dụng mã.

Các thư viện thường được sử dụng trong python:

* Numpy: Hỗ trợ tính toán các mảng nhiều chiều có kích thước lớn với các hàm đã được tối ưu áp dụng lên các mảng nhiều chiều đó. Đặc biệt hữu ích khi thực hiện các hàm liên quan tới đại số tuyến tính.
* Pandas: dùng để làm việc với các tập dữ liệu, có chức năng phân tích làm sạch, khám phá và thao tác dữ liệu
* Matplotlib: được sử dụng để vẽ đồ thị cấp thấp trong python, đóng vai trò như một tiện ích trực quan hóa. Hiện thị dữ liệu dưới dạng nhiều biểu đồ khác nhau như biểu đồ cột, đường.
* Seaborn: sử dụng Methplotlib để vẽ đồ thị. Sẽ được sử dụng để hình dung sự phân bố ngẫu nhiên.

1. **Sử dụng môi trường Jupyter notebook:**

Jupyter Notebook là một nền tảng tính toán khoa học mã nguồn mở, bạn có thể sử dụng để tạo và chia sẻ các tài liệu có chứa code trực tiếp, phương trình, trực quan hóa dữ liệu và văn bản tường thuật Với Jupyter Notebook, người dùng có thể đưa dữ liệu, code, hình ảnh, công thức, video,.. vào trong cùng một file, giúp cho việc trình bày trở nên dễ dàng hơn. Bạn có thể vừa trình chiếu vừa chạy code để tương tác trên đó, cốt lõi của việc này là Markdown.

Một số tính năng quan trọng bao gồm:

* Mã tương tác cho phép bạn viết và chạy code python từng ô (cell). Bạn có thể thực hiện tính toán và xử lý dữ liệu tại vị trí bạn cần thực hiện.
* Hiển thị kết quả và visualizations : khi đã thực hiện, kết quả sẽ được hiển thị trực tiếp bên dưới. Cũng như có thể tạo biểu đồ và trực quan hóa dữ liệu, khiến bạn dễ dàng trong nghiên cứu của mình.
* Hỗ trợ markdown và code trong cùng một notebook, tạo báo cáo nghiên cứu đẹp mắt dễ đọc.
* Tích hợp các công cụ và thư viện phổ biến như Numpy, Pandas, Matplotlib,…

1. **CÁC BƯỚC THỰC HIỆN**
2. **Tìm kiếm dữ liệu**

Dữ liệu được lấy từ bộ dữ liệu trên Kaggle có tên là “Medical Insurance Cost Prediction” do một tài khoản có tên là M Rahul Vyas. Đây là link trên trang Kaggle “https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction”. Bộ dữ liệu bảo hiểm y tế bao gồm nhiều yếu tố khác nhau ảnh hưởng đến chi phí y tế, chẳng hạn như tuổi tác, giới tính, BMI, tình trạng hút thuốc, số trẻ em và khu vực. Mục đích của nó là làm sáng tỏ các yếu tố then chốt góp phần làm tăng chi phí bảo hiểm, hỗ trợ công ty đưa ra các quyết định sáng suốt hơn liên quan đến giá cả và đánh giá rủi ro.

Bộ dữ liệu bao gồm 7 cột:

* Age
* Sex
* BMI (Body Mass Index)
* Children
* Smoker
* Region
* Charges

1. **Import các thư viện cần thiết**

A screen shot of a computer code

Description automatically generated

Hình 1 - Import các thư viện

1. **Đọc tệp dữ liệu**

A screenshot of a computer

Description automatically generated

Hình 2 - Hiển thị dữ liệu

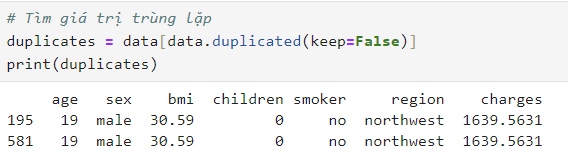
1. **Xử lý dữ liệu**

* **Loại bỏ các giá trị trùng lặp**

**A screenshot of a computer code

Description automatically generated**

Hình 3 - Kiểm tra giá trị trùng lặp

****

Hình 4 - Tìm giá trị trùng lặp

Nhật xét: có vẻ như khi thu thập dữ liệu đã vô tình nhập thành 2 lần, để đảm bảo có 1 bộ dữ liệu đẹp em quyết định xóa dữ liệu bị trùng đi.

A screenshot of a computer

Description automatically generated

Hình 5 - Xóa giá trị trùng lặp

**Nhận xét:** Tệp dữ liệu sau khi xóa dữ liệu trùng lặp còn 1337 dòng.

* **Sửa lỗi cấu trúc**

**A screenshot of a computer

Description automatically generated**

Hình 6 - Thông tin dữ liệu

**Nhận xét:** Có 3 cột dữ liệu ở dạng danh mục, em sẽ đếm tần suất và kiểm tra xem các dữ liệu có đồng nhất hay không.

**A screenshot of a computer code

Description automatically generated**

Hình 7 - Đếm và kiểm tra dữ liệu có đồng nhất hay không

Nhận xét: Có thể thấy các dữ liệu đã được đồng nhất.

A screenshot of a computer code

Description automatically generated

Hình 8 - Kiểm tra khoảng trắng trong cột dữ liệu object

* **Xử lý các dữ liệu bị thiếu**

**A screenshot of a computer code

Description automatically generated**

Hình 9 - Xử lý các dữ liệu bị thiếu

Nhận xét: Các cột dữ liệu không bị thiếu.

* **Lọc các dữ liệu ngoại lệ**

**A screenshot of a computer

Description automatically generated**

Hình 10 - Chỉ số của các cột dữ liệu

**Nhận xét:** Nhìn vào các hệ số mean, max thì có vẻ như biến bmi, charges có ngoại lai, để chắc chắn hơn em sẽ tính thêm chỉ số z\_score.

A screenshot of a computer program

Description automatically generated

Hình 11 - Tính chỉ số z\_score để tìm giá trị ngoại lai

Nhận xét: Đúng như dự đoán thì cả 2 biến “bmi” và “charges” đều có ngoại lai. Tiếp theo ta sẽ tiến hành xóa các dữ liệu ngoại lai này để cho ra một mô hình có chất lượng tốt.

A screenshot of a computer code

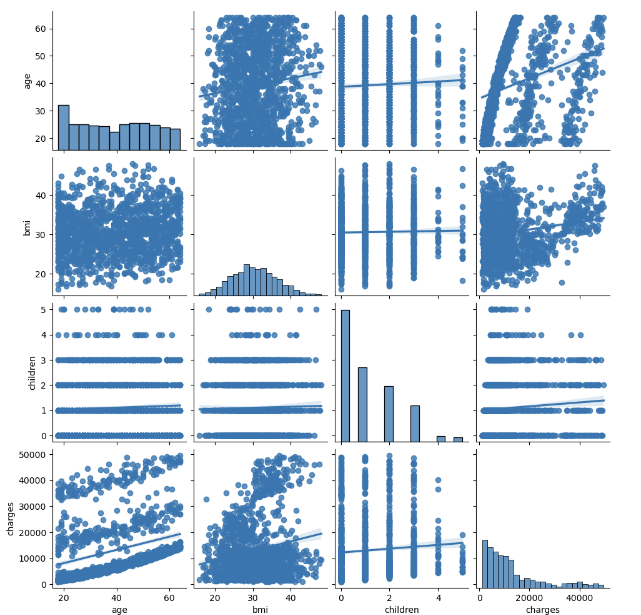
Description automatically generated

Hình 12 - Xóa các dữ liệu ngoại lai

Sau khi tiến hành xóa các dữ liệu ngoại lai, tệp dữ liệu ban đầu bây giờ còn 1326 dòng.

1. **Trực quan hóa dữ liệu**

* **Tìm mối tương quan giữa các biến qua biểu đổ pairplot**



Hình 13 - Biểu đồ pairplot thể hiện mối tương quan giữa các biến

**Nhận xét:** Nhìn vào biểu đồ pairplot, ta có thể thấy:

* Đầu tiên khi sử dụng pairplot có thể thấy ở các đường chéo các cột đã không còn thấy giá trị bất thường cho thấy việc xử lý ngoại lai ở trên đã khá tốt.
* Bmi, Age có mối tương quan dương với charge, giữa smoker và children với charges có sự tương quan yếu.
* Nhưng để có cái nhìn rõ hơn, em sẽ sử dụng 1 số dạng biểu đồ khác để thể hiện mối tương quan giữa các biến độc lập với biến phục thuộc "charges".
* **Tìm mối tương quan giữa các biến qua biểu đổ barplot**

**A bar graph with blue rectangles

Description automatically generated**

Hình 14 Biểu đồ barplot tương quan giữa Age và Charges

**Nhận xét:**

Nhìn vào biểu đồ ta thấy giữa các nhóm tuổi với charges có mối tương quan dương với nhau, điều này có thể giải thích bởi các yếu tố sau:

* Chi phí y tế tăng theo độ tuổi: Thường thì người cao tuổi có xu hướng sử dụng dịch vụ y tế nhiều hơn, do có nguy cơ cao hơn về các vấn đề sức khỏe và cần kiểm tra, chăm sóc định kỳ.
* Bệnh lý và điều trị đắt đỏ: Các bệnh lý thường phát hiện và phát triển nhiều khi tuổi tác tăng, và điều trị cho những bệnh lý này thường đắt đỏ, dẫn đến chi phí y tế tăng.
* Yếu tố lối sống và rủi ro: Người ở độ tuổi cao hơn có khả năng gặp các vấn đề sức khỏe liên quan đến lối sống (như bệnh tim mạch, tiểu đường) cao hơn, điều này cũng là một yếu tố làm cho chi phí y tế tăng.
* Yếu tố văn hóa và xã hội: Các yếu tố văn hóa và xã hội như lối sống, chế độ ăn uống, và môi trường sống cũng ảnh hưởng đến sức khỏe và chi phí y tế ở mọi độ tuổi.

A graph of a graph

Description automatically generated with medium confidence

Hình 15 Biểu đồ barplot tương quan giữa Age, Sex với Charges

**Nhận xét:**

* Nhìn vào biểu đồ, giá trị "charges" ở cả nam và nữ ở các độ tuổi đều có xu hướng tăng nhanh, ngoài các yếu tố ảnh hưởng chung đã giải thích theo độ tuổi ở phía trên thì nguyên nhân giá trị trung bình của "charges" ở nam luôn cao hơn nữ là do:
* Hành vi nguy hiểm: Nam giới thường có xu hướng tham gia vào các hoạt động rủi ro hơn như hút thuốc, uống rượu, lái xe quá tốc độ... Điều này làm tăng nguy cơ mắc các bệnh liên quan như ung thư phổi, bệnh gan, tai nạn giao thông.
* Các bệnh lý đặc trưng: Một số bệnh lý thường gặp ở nam giới như bệnh tim mạch, ung thư tuyến tiền liệt có thể đòi hỏi các phương pháp điều trị phức tạp và tốn kém hơn.
* Kiểm tra sức khỏe định kỳ: Nam giới thường ít đi khám sức khỏe định kỳ hơn nữ giới, dẫn đến việc phát hiện bệnh ở giai đoạn muộn hơn, khi chi phí điều trị đã cao hơn.
* Yếu tố sinh học: Một số nghiên cứu cho thấy rằng nam giới có hệ miễn dịch yếu hơn nữ giới, dễ mắc bệnh hơn và tuổi thọ trung bình thấp hơn.

A graph showing the average charge by smoker group

Description automatically generated

Hình 16 Biểu đồ barplot tương quan giữa smoker với charges

**Nhận xét:**

Những người hút thuốc lại có mức chi trả chi phí bảo hiểm cao hơn những người không hút vì:

* Nguy cơ sức khỏe cao hơn: Người hút thuốc có nguy cơ mắc nhiều bệnh mãn tính cao hơn đáng kể so với người không hút thuốc, như bệnh tim mạch, ung thư đặc biệt là ung thư phổi, có liên quan trực tiếp đến việc hút thuốc
* Chi phí điều trị cao hơn: Các bệnh liên quan đến thuốc lá thường đòi hỏi chi phí điều trị cao hơn, bao gồm:
* Thuốc men: Người hút thuốc thường phải dùng nhiều loại thuốc để điều trị các bệnh liên quan.
* Viện phí: Chi phí nhập viện, phẫu thuật và các thủ tục y tế khác cho người hút thuốc thường cao hơn.
* Tuổi thọ trung bình thấp hơn: Người hút thuốc thường có tuổi thọ trung bình thấp hơn so với người không hút thuốc, điều này cũng ảnh hưởng đến mức phí bảo hiểm.

A graph of a number of blue rectangular objects

Description automatically generated

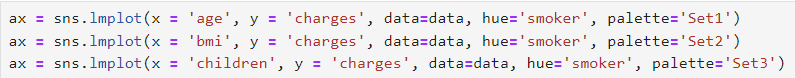
Hình 17 Biểu đồ barplot tương quan giữa Region với Charges

**Nhận xét:**

Nhìn chung, giá trị trung bình của "charges" ở các khu vực đều cao, cao nhất là khu vực Đông Nam, sau đó là Đông Bắc, khu vực Tây Nam và Tây Bắc có vẻ có mức giá trị trung bình ngang nhau. Nguyên nhân dẫn đến chi phí phải chi trả bảo hiểm cảu khu vực Đông Nam cao là do:

* Sức khỏe kém: Người dân Đông Nam Mỹ thường gặp các vấn đề sức khỏe như béo phì, tiểu đường, thói quen ăn uống không lành mạnh, ít vận động và tỷ lệ hút thuốc cao ở một số khu vực Đông Nam cũng góp phần làm tăng gánh nặng bệnh tật, kéo theo chi phí bảo hiểm tăng.
* Hệ thống y tế: Cơ sở vật chất y tế chưa đồng đều, sự cạnh tranh giữa các nhà cung cấp dịch vụ y tế thấp, dẫn đến việc các cơ sở y tế có thể tăng giá dịch vụ, và cuối cùng, các công ty bảo hiểm phải chi trả nhiều hơn.
* Yếu tố kinh tế - xã hội: Thu nhập thấp, giáo dục hạn chế, chính sách y tế chưa đồng bộ cũng góp phần làm tăng chi phí.
* **Tìm mối tương quan giữa các biến qua biểu đổ scatter**

Tiếp theo để xem khi thêm nhân tố “smoker” vào biểu đồ để xem mối tương quan có ảnh hưởng gì không, em quyết định dung biểu đồ scratter.



A graph showing different types of age

Description automatically generated

A graph of different colored dots

Description automatically generated

**A graph of smoke and smoke

Description automatically generated with medium confidence**

Hình 18 - Biểu đồ scatter thể hiện mối tương quan 3 biến

**Nhận xét:**

* Có thể thấy mối tương quan mạnh giữa 2 biến “age” và “smoker” với “charges”, tuổi càng cao thì tể lệ hút thuốc càng nhiều do đó kéo theo chi phí phải trả cho bảo hiểm cũng cao. Điều này cũng dễ hiểu, người trưởng thành phải đối mặt nhiều với nhiều áp lực cuộc sống công việc nên sẽ tìm đến các loại thuốc lá để trấn an tinh thần, cũng có vài người sử dụng thuốc lá như thói quen hằng ngày. Dù là nguyên nhân tác động nào thì khi hút thuốc sẽ gây ra rất nhiều căn bệnh nguy hiểm, chính ví vậy chi phí bảo hiểm phải chi trả cho những trường hợp này cũng khá là cao.
* Tương tự cũng có thể thấy mối tương quan mạnh giữa 3 biến BMI, smoker với charger, số người hút thuốc tỉ lệ thuận với chỉ số BMI đồng thời tỉ lệ thuận với Charges. Những người hút thuốc thường có nguy cơ mắc bệnh liên quan đến béo phì đồng nghĩa với việc chi phí bảo hiểm cho những người này cũng tăng.
* Ta có thể thấy giữa smoker và children với charges có sự tương quan yếu.

A close-up of a text

Description automatically generated

A diagram of red and blue dots

Description automatically generated

Hình 19 Biểu đồ scatter tương quan giữa bmi, sex với charges

**Nhận xét:**

Nhìn biểu đồ có thể thấy, chỉ số bmi ở nam giới lẫn nữ giới đều có mối tương quan dương với chi phải mức bảo hiểm phải chi trả, nhưng ở nam giới lại có mối tương quan mạnh hơn nữ là do ở nam giới thường có lối sống không lành mạnh như hút thuốc, rượu bia dẫn đến các bệnh béo phì, tim mạch, tiểu đường.. điều đó đồng nghĩa với việc chi phí bảo hiểm phải chi trả ở nhóm người này cũng tăng cao.

* **Kiểm tra tính đa cộng tuyến**

**A screenshot of a computer

Description automatically generated**

Hình 21 Kiểm tra tính đa cộng tuyến

**Nhận xét:**

* Ta thấy các biến độc lập có chỉ số tương quan rất thấp và hầu như không có.
* Nên chúng ta có thể kết luận mô hình này tốt và không có sự tương quan giữa các biến độc lập.

1. **Chọn biến độc lập tốt**

Ở đây, em sẽ viết hàm Forward Selection để có thể lựa chọn các đặc trưng phù hợp với mô hình.

A screenshot of a computer

Description automatically generated

Hình 20 - Hàm Forward Selection

**Nhận xét:**

Qua quá trình chọn lựa biến theo phương pháp forward selection, chúng ta nhận thấy:

* Biến smoker, age và bmi: Khi lần lượt đưa các biến này vào mô hình, chỉ số Adj R-squared tăng đáng kể, cho thấy các biến này có mối quan hệ tuyến tính mạnh với biến phụ thuộc (charges). Điều này đồng nghĩa với việc hút thuốc, tuổi tác và chỉ số khối cơ thể là những yếu tố có ảnh hưởng lớn đến chi phí bảo hiểm y tế.
* Biến children, region và sex: Việc đưa thêm các biến này vào mô hình chỉ làm tăng nhẹ chỉ số Adj R-squared, cho thấy ảnh hưởng của các biến này đến chi phí bảo hiểm là không đáng kể so với các biến đã chọn trước đó.

**Kết luận:** Dựa trên kết quả phân tích, chúng ta có thể loại bỏ các biến region và sex khỏi mô hình cuối cùng. Mô hình tối ưu nhất sẽ bao gồm các biến smoker, age và bmi.

1. **Chia tệp dữ liệu**

A computer code with text

Description automatically generated with medium confidence

Hình 22 - Chia tệp dữ liệu

Lựa chọn các biến độc lập là các cột có dữ liệu liên tục và biến phụ thuộc sẽ là cột “charges” dựa vào các cột kia để có thể dự báo được.

Tiến hành chia tệp dữ liệu thành 2 mảng X và y với tỷ lệ là 30% cho dữ liệu test và 70% cho dữ liệu training. Random\_state là tham số cần khai báo để người khác có thể sao chép công việc đang làm thường sẽ để là 42.

1. **Xây dựng mô hình**

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Hình 23 - Kết quả mô hình

**Nhận xét:** Ta thấy chỉ số Adjusted R^2có sai số khoảng 30% ở mức tạm chấp nhận được. Để chắc chắn hơn chúng ta sẽ tiến hành kiểm tra các giả định và hiện tượng fitting của mô hình.

1. **Kiểm tra các giả định**

* **Kiểm tra tính chuẩn tắc**

A screenshot of a computer screen

Description automatically generated

Hình 24 - Kiểm tra tính chuẩn tắc

**Nhận xét:** Qua biểu đồ ta thấy dữ liệu không đạt tính chuẩn tắc, giá trị dữ liệu dư còn có sai số khá cao và giá trị dư có sự phân bố không tốt lắm, vì rất nhiều dữ liệu có mức chênh lệch khá cao.

1. **Kiểm tra tính đồng nhất**

A screen shot of a computer screen

Description automatically generated

Hình 25 - KIểm tra tính đồng nhất

**Nhận xét:** Qua biểu đồ thấy dữ liệu không đồng nhất, do các giá trị dư đang bị lệch về nhiều vị trí sai số khác, với độ sai số cao và cục bộ, không đồng đều.

1. **Kiểm tra hiện tượng fitting**

**A screenshot of a computer program

Description automatically generated**

Hình 27 - Kiểm tra hiện tượng fitting

**Nhận xét:**

* RMSE vẫn ở mức khá cao thể hiện bình phương sai số của mô hình, với mô hình này tỉ lệ chênh lệch dư giữa giá trị thực tế và giá trị huấn luyện cao nên RMSE vẫn ở mức khá cao.
* Ta thấy Adjusted R^2của tập dữ liệu huấn luyện (0.75) gần bằng Adjusted R^2 của tập dữ liệu thực tế test (0.73), nên mô hình không bị overfitting hay underfitting do chênh lệnh rất thấp.

1. **KẾT LUẬN**

Dựa trên phân tích dữ liệu về các yếu tố ảnh hưởng đến chi phí bảo hiểm cũng phản ánh được sự quan trọng của sức khỏe con người và ảnh hưởng nghiêm trọng nhất từ việc hút thuốc lá. Để duy trì sức khỏe tốt và tránh chi phí y tế không cần thiết, việc không hút thuốc lá và duy trì một chế độ ăn uống khoa học là rất quan trọng. Còn đối với công ty bảo hiểm, họ có thể điều chỉnh chiến lược và chính sách bảo hiểm một cách hiệu quả hơn để bảo vệ khách hàng.

* Điều chỉnh chi phí bảo hiểm theo nhóm rủi ro: Công ty có thể đề xuất mức phí bảo hiểm khác nhau cho từng nhóm rủi ro dựa trên các yếu tố ảnh hưởng. Ví dụ, người hút thuốc có thể phải trả mức phí cao hơn so với người không hút thuốc.
* Thúc đẩy hành vi lành mạnh: Công ty có thể áp dụng chính sách khuyến khích khách hàng duy trì một lối sống lành mạnh để giảm rủi ro và chi phí bảo hiểm. Ví dụ, có thể cung cấp ưu đãi cho người duy trì chỉ số khối cơ thể (bmi) trong mức lý tưởng.
* Tích hợp công nghệ và dữ liệu: Sử dụng công nghệ để theo dõi và đánh giá các yếu tố ảnh hưởng đến rủi ro và chi phí bảo hiểm. Công ty có thể sử dụng dữ liệu để tối ưu hóa quy trình định giá bảo hiểm và điều chỉnh chiến lược dựa trên thông tin mới nhất.
* Tạo ra sản phẩm bảo hiểm linh hoạt: Công ty có thể cung cấp các sản phẩm bảo hiểm linh hoạt cho khách hàng, cho phép họ lựa chọn mức bảo hiểm và điều kiện phù hợp với nhu cầu và khả năng tài chính của họ.