Dự án:

Dự đoán giá bảo hiểm y tế sử dụng mô hình hồi quy tuyến tính

1. Giới thiệu:

* Báo cáo này tập trung vào dự án "Dự đoán giá bảo hiểm y tế" sử dụng mô hình hồi quy tuyến tính trong lĩnh vực Trí Tuệ Nhân Tạo (AI). Mục tiêu của dự án là xây dựng một mô hình có khả năng dự đoán giá bảo hiểm y tế dựa trên các biến đầu vào.
* Mục tiêu nghiên cứu: Nghiên cứu này nhằm xác định mối quan hệ giữa các biến đầu vào như tuổi, giới tính, khu vực, chỉ số thể trọng, với giá bảo hiểm y tế.

1. Nghiên cứu:
2. Thu thập và kiểm tra dữ liệu:

- Bộ dữ liệu được lấy trên Kaggle.

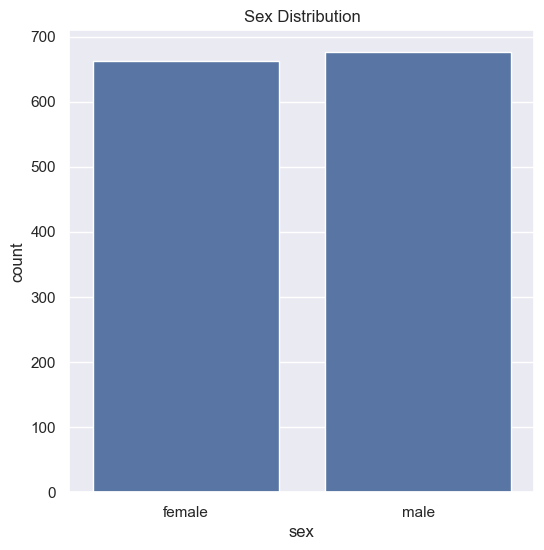
- Bộ dữ liệu bao gồm 1338 hàng và 7 cột, bao gồm :”tuổi”, ‘giới tính”, “BMI (chỉ số thể trọng ), ”số lượng con cái” , ”tình trạng hút thuốc”, ”khu vực sống”,”giá y tế”.

- Bộ dữ liệu không có giá trị trống.

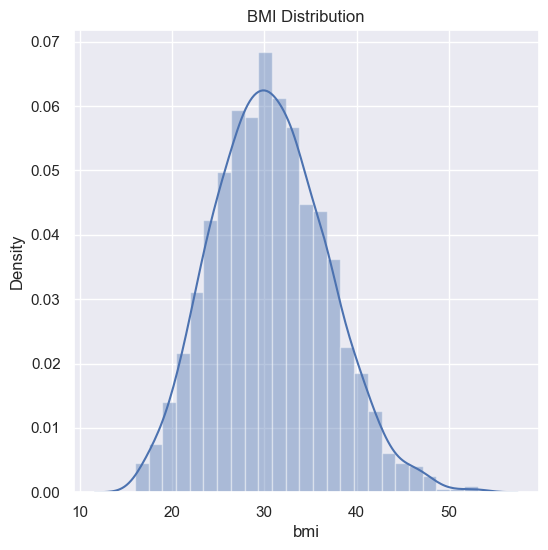
1. Phân tích dữ liệu:

- Phân bố về giá trị độ tuổi: 

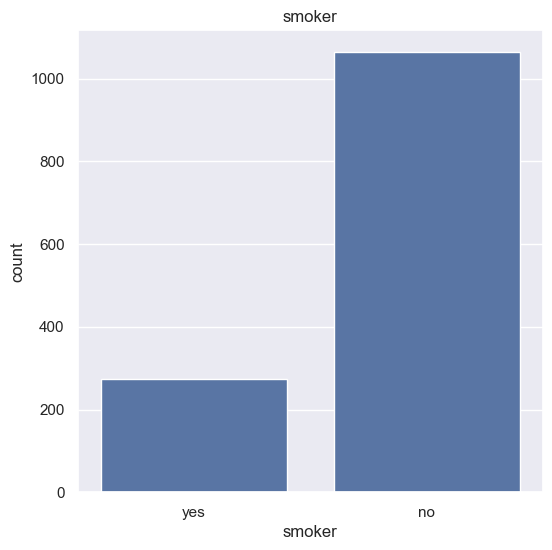
- Phân tích về giới tính:



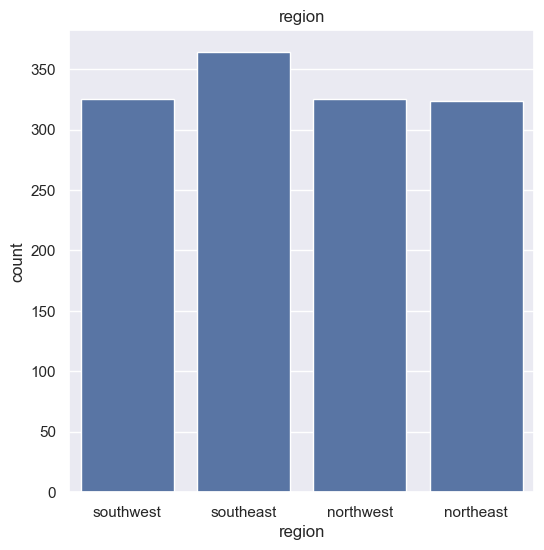
- Phân bố chỉ số thể trọng:



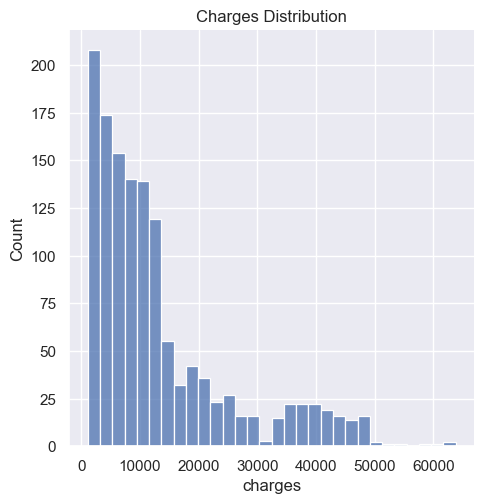
- Về số người hút thuốc:



- Phân bố về khu vực:



- Phân bố về giá bh y tế:



1. Xử lý dữ liệu:

- Vì máy tính không thể đọc được dữ liệu dạng chữ, nên ta sẽ mã hóa các tính năng phân loại:

* Với giới tính: nam sẽ được mã hóa thành 0, nữ được mã hóa thành 1
* Với hút thuốc: có hút thuốc sẽ được mã hóa thành 0, không hút thuốc sẽ được mã hóa thành 1
* Với khu vực: vùng Đông Nam sẽ đực mã hóa thành 0, Tây Nam thành 1, Đông Bắc thành 2, Tây Bắc thành 3

- Phân tách giữa các tính năng và mục tiêu cần dự đoán (giá bảo hiểm)

- Phân chia dữ liệu thành bộ huấn luyện và bộ kiểm thử để đảm bảo độc lập và đánh giá hiệu suất mô hình

1. Thuật toán

* **Hàm Hồi Quy:** Sử dụng hàm hồi quy tuyến tính dạng:

Y= β0+β1.X1+β2.X2+ β3.X3+ β4.X4+ β5.X5+ β6.X6 +ϵ Trong đó, Y là giá bảo hiểm y tế, X1,X2,X3,X4,X5​,X6​​ là các biến độc lập tượng trưng cho các tính năng trong bộ dữ liệu, β0,β1,β2, β3, β4, β5, β6​ là các hệ số hồi quy cần được ước lượng, và ϵ là sai số ngẫu nhiên.

1. Kết quả:

- Dự đoán đối với bộ dữ liệu huấn luyện, ta có chỉ số R2 là: 0.751505643411174, độ chính xác là ~ 75%

- Dự đoán đối với bộ dữ liệu kiểm thử, ta có chỉ số R2 là: 0.7447273869684077, độ chính xác là ~ 74%

- Từ những kết quả này, ta hoàn toàn có thể dự đoán xấp xỉ được giá bảo hiểm y tế với bộ dữ liệu đầu vào mới. VD: một người là nam, 31 tuổi, có BMI là 25.74, không hút thuốc, không có con, đến từ vùng đông nam thì có chi phí bảo hiểm y tế xấp xỉ là 3760 USD.

File code: https://github.com/KainuMan/Suffamore-Proj./tree/main/main/Proj1-Logistic%20Regression

-