ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

BÁO CÁO PROJECT 01
REGRESSION

Bộ môn: Nhập môn học máy
Giảng viên hướng dẫn: Nguyễn Tiến Huy

2020 - 2021

MỤC LỤC

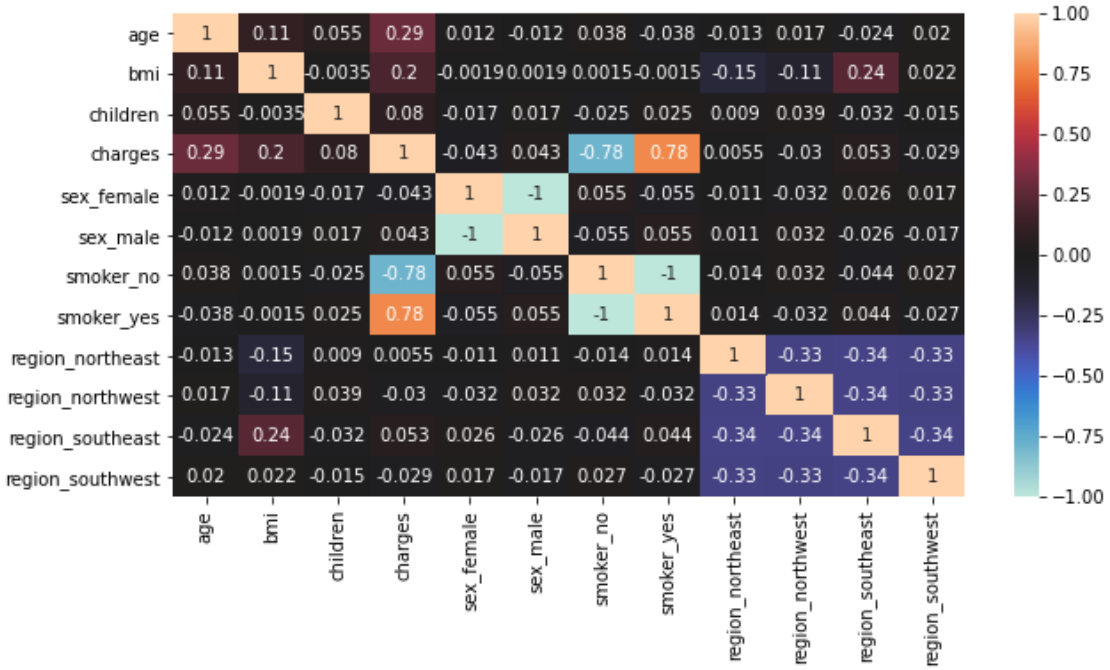
I. THÔNG TIN NHÓM VÀ MỨC ĐỘ ĐÓNG GÓP

STT	Họ và tên	MSSV	Tự đánh giá	Mức độ đóng góp
1	Du Chí Nhân	18120492	100%	100%
2	Phạm Minh Sỹ	18120540	100%	100%
3	Phan Văn Võ Quyền	18120529	100%	100%
4	Lê Thị Như Quỳnh	18120530	100%	100%
5	Lê Hoàng Phương Nhi	18120496	100%	100%

II. NỘI DUNG CHI TIẾT

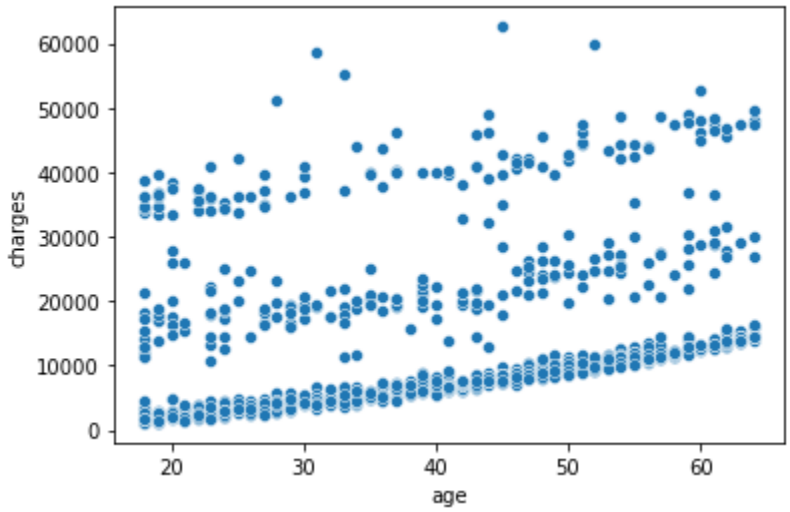
1. Các thông tin hữu ích

Vẽ biểu đồ Heatmap để xem xét độ tương quan giữa các thuộc tính:

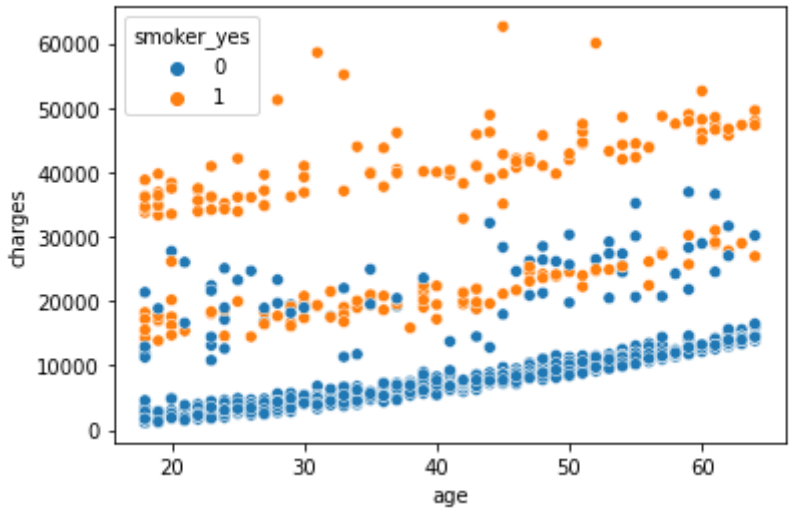


→ Từ biểu đồ Heatmap trên thì nhóm thấy được các thuộc tính ***age, bmi, smoker*** có mối tương quan khá chặt chẽ với thuộc tính ***charges:***

- Ảnh hưởng của thuộc tính ***age*** đến thuộc tính ***charges:***



→ Qua biểu đồ ta thấy được chi phí sử dụng dịch vụ y tế có xu hướng tăng theo độ tuổi, nó cũng chia thành 3 cụm riêng biệt. Để hiểu rõ hơn từng cụm, chúng ta xem xét kết hợp với các thuộc tính khác nữa.

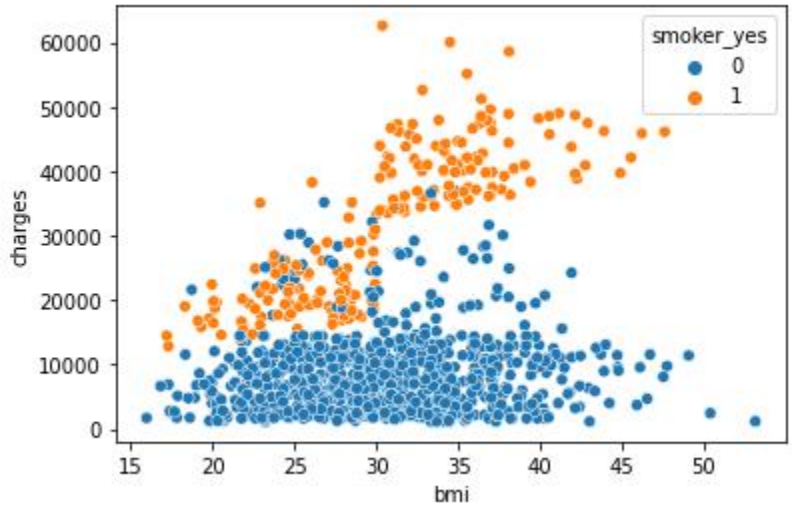


→ Sau khi kết hợp với thuộc tính ***smoker*** thì ta thấy được 3 cụm riêng biệt đó là:

- ✓ Cụm có chi phí thấp gồm những người không hút thuốc
- ✓ Cụm có chi phí cao gồm những người hút thuốc
- ✓ Cụm có chi phí trung bình gồm những người hút thuốc và những người không hút thuốc

Và có thể thấy từng cụm có quan hệ tuyến tính, tăng dần về những người lớn tuổi

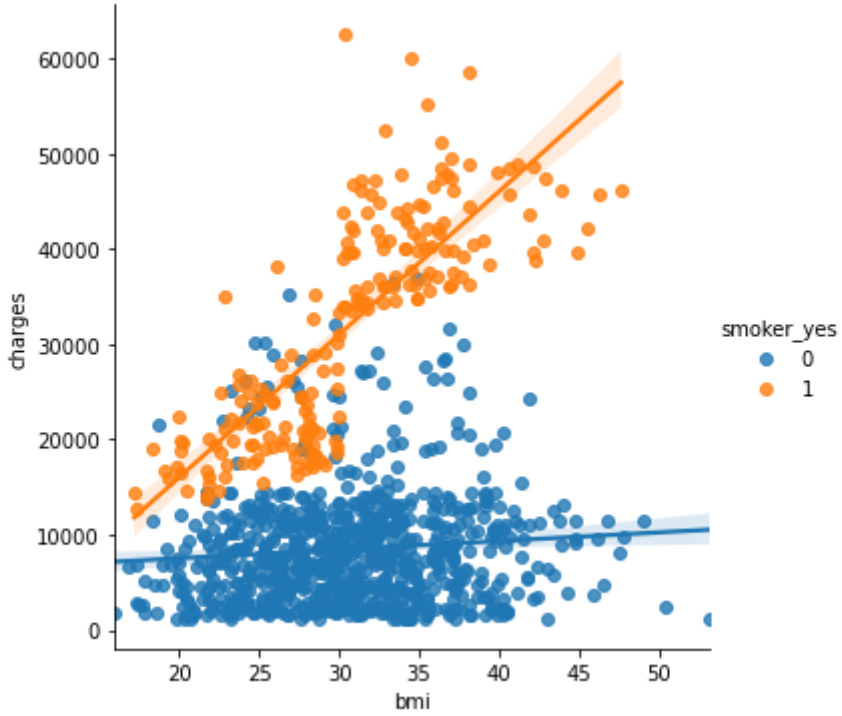
- Ảnh hưởng của thuộc tính ***bmi*** đến thuộc tính ***charges*** và tác động của thuộc tính ***smoker*** lên mối quan hệ đó:



→ Từ biểu đồ ta thấy rằng:

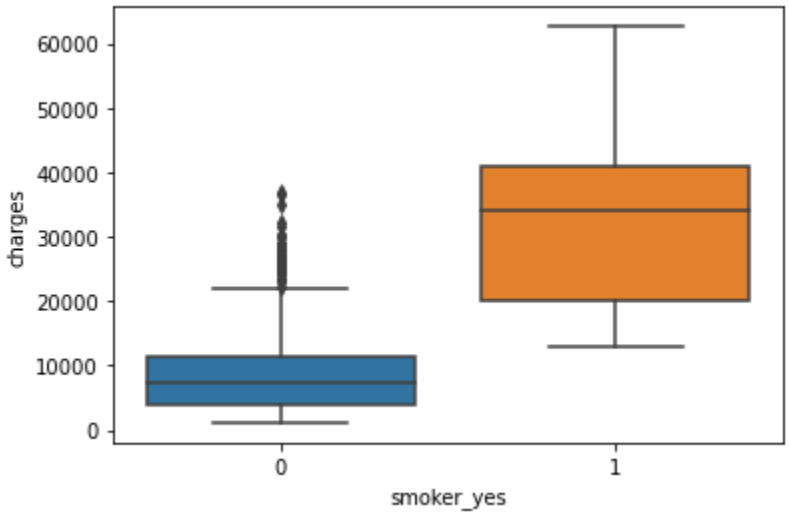
- ✓ Chi phí của những người không hút thuốc có xu hướng tăng lên một chút khi chỉ số bmi tăng
- ✓ Chi phí của những người hút thuốc có xu hướng tăng mạnh khi chỉ số bmi tăng

Để xem xét rõ ràng hơn, ta vẽ thêm hai đường hồi quy tương ứng với người hút thuốc và người không hút thuốc:



→ Ta thấy được đường hồi quy ứng với những người hút thuốc có độ dốc lớn hơn so với những người không hút thuốc

- Ảnh hưởng của thuộc tính ***smoker*** đến thuộc tính ***charges***:



→ Qua biểu đồ này và các biểu đồ trên ta thấy được thuộc tính ***smoker*** có ảnh hưởng rất lớn đến chi phí sử dụng dịch vụ y tế. Những người hút thuốc phải trả cao hơn nhiều so với những người không hút thuốc

2. Cài đặt các thuật toán học máy

- **Linear Regression:**
  - Chia dữ liệu train.csv thành 2 tập: train và validation
  - Kiểm tra xem liệu có insight nào chưa tìm thấy hoặc những insight rút ra có bị sai hay không
    - ✓ Ý tưởng kiểm tra: Bỏ đi một thuộc tính của dữ liệu sau đó huấn luyện trên dữ liệu train và đánh giá bằng dữ liệu validation. Nếu giá trị R^2 thay đổi một khoảng nhỏ hơn giá trị epsilon định trước thì coi như thuộc tính không quan trọng (Ở đây epsilon = 0.01)
  - Các bước thực hiện:
    - ✓ Bước 1: Mã hóa các thuộc tính categorical thành dạng số và xây dựng thêm thuộc tính ***smoke_bmi = bmi * smoker***
    - ✓ Bước 2: Chuẩn hóa các thuộc tính
    - ✓ Bước 3: Huấn luyện và đem mô hình đi dự đoán kết quả
- **Random Forest Regressor:**
  - Chia dữ liệu train.csv thành 2 tập: train và validation
  - Kiểm tra xem liệu có insight nào chưa tìm thấy hoặc những insight rút ra có bị sai hay không
    - ✓ Ý tưởng kiểm tra: Bỏ đi một thuộc tính của dữ liệu sau đó huấn luyện trên dữ liệu train và đánh giá bằng dữ liệu validation. Nếu giá trị R^2

thay đổi một khoảng nhỏ hơn giá trị epsilon định trước thì coi như thuộc tính không quan trọng (Ở đây epsilon = 0.01)

- Các bước thực hiện:
  - ✓ Bước 1: Mã hóa các thuộc tính categorical thành dạng số và xây dựng thêm thuộc tính ***smoke_bmi = bmi * smoker***
  - ✓ Bước 2: Chuẩn hóa các thuộc tính
  - ✓ Bước 3: Huấn luyện và đem mô hình đi dự đoán kết quả

3. Kết quả đạt được

- **Linear Regression:**
  - Rút ra được 4 thuộc tính bao gồm: ***sex, bmi, children, region*** không có tác động đến chi phí cá nhân. Có 3 thuộc tính tác động đến chi phí y tế cá nhân là: ***age, smoker, smoke_bmi***

```
Prev score: 0.8602592522357086
Epsilon: 0.01
--------------------------------------------------
Without 'age':
New score: 0.7595965703187375
Difference: 0.10066268191697114
--------------------------------------------------
Without 'sex':
New score: 0.8598900615903153
Difference: 0.0003691906453933136
--------------------------------------------------
Without 'bmi':
New score: 0.8603182772642691
Difference: 5.9025028560477644e-05
--------------------------------------------------
Without 'children':
New score: 0.8597391121354709
Difference: 0.0005201401002377093
--------------------------------------------------
Without 'region':
New score: 0.8602235301414888
Difference: 0.000664277905780164
--------------------------------------------------
Without 'smoke_bmi':
New score: 0.7538700980226231
Difference: 0.10638915421308559
--------------------------------------------------
====> Unsignificant feature list:  ['sex', 'bmi', 'children', 'region']
```

  - Sử dụng các kết luận rút ra được, xây dựng tập dữ liệu có 3 thuộc tính age, smoker, smoke_bmi
  - Giá trị R^2:
    - ✓ Trên tập dữ liệu train.csv: 0.8320841714457461

✓ Trên tập dữ liệu test.csv: 0.8478662021721436

➢ Phân tích thuộc tính:

```
intercept: -2234.445279282376
age : 271.05533839972236
smoker_yes : -21192.084451965744
smoke_bmi : 1471.271383996227
```

➔ Ý nghĩa:

+ Mỗi 1 tuổi sẽ làm tăng xấp xỉ 271$ cho chi phí y tế cá nhân
+ Với những người hút thuốc, mỗi 1 điểm chỉ số BMI sẽ làm tăng 1471$ chi phí y tế cá nhân

- **Random Forest Regressor:**

```
Prev score: 0.8393667365591864
Epsilon: 0.01
-------------------------------------------------
Without 'age':
New score: 0.7138170618367885
Difference: 0.12554967472239786
-------------------------------------------------
Without 'sex':
New score: 0.8363605928432583
Difference: 0.0030061437159281112
-------------------------------------------------
Without 'bmi':
New score: 0.8054085546234012
Difference: 0.03395818193578515
-------------------------------------------------
Without 'children':
New score: 0.8237504465496766
Difference: 0.015616290009509814
-------------------------------------------------
Without 'smoker':
New score: 0.8364914835745116
Difference: 0.0028752529846747255
-------------------------------------------------
Without 'region':
New score: 0.8255357626404731
Difference: 0.01383097391871324
-------------------------------------------------
Without 'smoke_bmi':
New score: 0.8348653856588737
Difference: 0.004501350900312673
-------------------------------------------------
====> Unsignificant feature list:  ['sex', 'smoker', 'smoke_bmi']
```

- Sử dụng các kết luận rút ra được, xây dựng tập dữ liệu có 4 thuộc tính quan trọng: ***age, children, bmi, region***
- Giá trị R^2:
  - ✓ Trên tập dữ liệu train.csv: 0.9729337862533126
  - ✓ Trên tập dữ liệu test.csv: 0.8408823507138184

III. TÀI LIỆU THAM KHẢO

- https://www.kaggle.com/davsmith33/insurance-premium-predictions
- https://www.kaggle.com/kushshah95/improvised-accuracy-predicting-insurance-premium