|  |  |
| --- | --- |
| http://www.hcmus.edu.vn/images/stories/logo-khtn2009_transparency.png | **Toán Ứng Dụng Thống Kê** |

Lớp Cử nhân Tài năng khóa 2019 Ngày thực hiện: 25 tháng 01 năm 2022

**Giảng viên:** Nguyễn Đình Thúc

**Họ và tên sinh viên:** Chung Hoàng Tuấn Kiệt   
**Mã số sinh viên:** 19120553

***Đề tài:* Phân tích dữ liệu ung thư**

1. **Giới thiệu**

Mục tiêu: Kiểm định mối quan hệ giữa các thuộc tính trong file *data\_cntn19.xlsx* với các trường thông tin:

*ImmunoPhenotype* - Trạng thái của thuốc gồm 3 giá trị: Đã tới tế bào (Infiltrated), Chưa tới tế bào (NO\_IF), Chưa rõ (Desert)

*PFS* - Số ngày sống tính theo tuần của

*Max* *FrameShift* *Expression* - Số lượng tối đa của các biểu hiện

*Sum* *FrameShift* *Expression* - Các biểu hiện của ung thư

Đồ án được thực hiện với mục tiêu đánh giá sự tương quan và tìm ra mối quan hệ giữa các trường dữ liệu với *PFS.*

Quá trình thực nghiệm:

Kiểm tra mối tương quan giữa từng trường dữ liệu trên đối với giá trị *PFS* và trực quan hóa kết quả bằng đồ thị.

Kiểm tra mối quan hệ trong tổ hợp các trường dữ liệu với giá trị *PFS* (cụ thể là *Max FrameShift Expression* và *Max FrameShift Expression*) đối với từng loại của *ImmunoPhenotype.*

Kiểm tra mối tương quan của toàn bộ các trường dữ liệu với giá trị *PFS* sử dụng hồi quy rừng (RandomForestRegressor).

Thông qua thực nghiệm, tiến hành tính toán các hệ số và đánh giá mối tương quan, từ đó rút ra kết luận.

1. **Phân tích dữ liệu**
   1. **Tóm tắt:**

Dữ liệu ung thư bao gồm 203 mẫu với 4 thuộc tính:

- **ImmunoPhenotype**: tình trạng kháng thể của thuốc đối với tế bào ung thư có 4 giá trị phân loại : Infiltrated (kháng thể đã xâm nhập khối u), NO\_IF (kháng thể không xâm nhập khối u), Desert (kháng thể chưa phát hiện khối u), Excluded (kháng thể đã phát hiện khối u nhưng chưa xâm nhập).

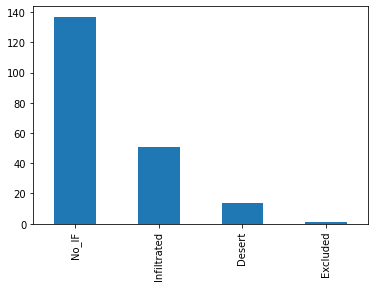
- **sum\_frameshift\_expression**: tổng số biểu hiện của bệnh.

- **max\_frameshift\_expression**: số biểu hiện tối đa của bệnh.

- **PFS**: số ngày sống của bệnh nhân mắc ung thư, tính theo đơn vị tuần.

Số lượng từng trường hợp trong trường dữ liệu ImmunoPhenotype**:**

|  |  |
| --- | --- |
| ImmunoPhenotype | Số lượng |
| Desert | 14 |
| Excluded | 1 |
| Infiltrated | 51 |
| No\_IF | 137 |

****

***Hình 1****: Thống kê số liệu cho mỗi trường hợp*

**Scatter chart

Description automatically generated**

***Hình 2****: Phân loại các trường hợp của thuốc đối với tế bào ung thư*

Hình 2 thể hiện tập giá trị của trường *FPS*  đối với từng trường hợp trong trường dữ liệu *ImmunoPhenoType.*

Qua đó cho thấy những người có*ImmunoPhenotype* là *Infiltrated* có số *PFS* tương đối cao hơn so với các trường hợp còn lại (Có thể vượt ngưỡng giá trị 60 tuần, nhưng hầu hết có giá trị dao động bé hơn 30 tuần), tập giá trị của Desert trong khi Desert hầu hết là dưới 15 tuần chỉ có 1 giá trị vượt ngưỡng 60 tuần, NO\_IF thì chiếm số lượng mẫu lớn nhất (nhiều gấp 2.7 lần so với số lượng mẫu Infiltrated và 9.7 lần so với số lượng mẫu Desert) nhưng tập giá trị chỉ tập trung dưới 30 tuần, duy nhất một trường hợp xâp xỉ 40 tuần. Số lượng người mang giá trị Excluded chỉ có một nên sẽ không được xét đến.

* 1. **Đánh giá độ tương quan đối với các trường dữ liệu:**
     1. **Quan hệ giữa 2 trường dữ liệu sum\_frameshift\_expression và FPS:**

1. **Nhận xét và đánh giá:**

**Chart, scatter chart

Description automatically generated**

Hình vẽ bao gồm các điểm dữ liệu phân tán và đường đường thẳng (màu đỏ) tượng trưng cho mô hình hồi quy tuyến tính giữa hai thuộc tính. Ta dễ thấy rằng, chỉ có một số ít các trường hợp có *PFS* > 30 đều có giá trị *sum***\_***frameshift***\_***expression* < 200. Các trường hợp có *PFS <* 30 đa phần có *sum***\_***frameshift***\_***expression <* 400. Các trường hợp khác hầu như đều có giá trị *sum***\_***frameshift***\_***expression <* 10.

1. **Số liệu của mô hình:**

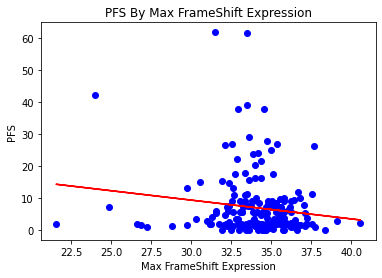
**Score:** 0.0035 (quá bé)

**Hệ số góc:**

Đường thẳng trên không thể hiện mối quan hệ tuyến tính giữa 2 trường dữ liệu trên. Các điểm dữ liệu phân tán một các rời rạc và giá trị của *score* chỉ đạt 0.0035. Có thể kết luận 2 trường dữ liệu này không có quan hệ tuyến tính. Hệ số góc âm cho thấy *PFS* có xu hướng giảm khi *sum***\_***frameshift***\_***expression* tăng,

* + 1. **Quan hệ giữa 2 trường dữ liệu max\_frameshift\_expression và FPS:**

1. **Trực quan hóa:**

****

Ta thấy, *PFS*có tập giá trị từ 0 đến 60 tuần, trong đó chỉ có 3 điểm dữ liệu có giá trị vượt ngưỡng 40 tuần, 3 điểm dữ liệu trong khoảng từ 30 đến 40 tuần và đa phần còn lại có giá trị < 30 tuần.

1. **Số liệu và đánh giá:**

**Score:** 0.021 (quá bé)

**Hệ số góc:**

Đường thẳng trên không thể hiện mối quan hệ tuyến tính giữa 2 trường dữ liệu trên. Các điểm dữ liệu phân tán một các rời rạc và giá trị của *score* chỉ đạt 0.021. Có thể kết luận 2 trường dữ liệu này không có quan hệ tuyến tính. Hệ số góc là âm cho thấy *PFS* có xu hướng giảm khi *max\_frameshift\_expression* tăng.

* + 1. **Quan hệ giữa 2 trường dữ liệu sum\_frameshift\_expression, max\_frameshift\_expression so với FPS:**

**Score:** 0.021

**Hệ số góc:** 0.049 và -2.17

Hệ số score nhỏ cho thấy *PFS* không có mối quan hệ tuyến tính đối với 2 trường dữ liệu trên,

* 1. **Đánh giá độ tương quan theo phân loại ImmunoPhenoType:**

Sử dụng *linear* *model* để đánh giá mối tương quan với tập 2 trường dữ liệu (*sum\_frameshift\_expression* và *max\_frameshift\_expression*) với *PFS* trong từng phân loại cụ thể.

* + 1. **Desert:**

**Score:** 0.122

Mặc dù Score lớn hơn trường hợp No-IF nhưng số điểm dữ liệu chưa đủ lớn và hệ số Score chưa đủ lớn để đánh giá mối quan hệ tuyến tính giữa các trường dữ liệu xét riêng trong trường hợp này. Với giá trị hiện tại (0.122) thì có thể kết luận tập dữ liệu không có mối quan hệ tuyến tính nếu xét riêng trong trường hợp này.

* + 1. **Infiltrated:**

**Score:** 0.282

Giá trị của Score lớn nhất trong các loại cho thấy mô hình có thể đánh giá một cách chính xác hơn. Các trường dữ liệu nếu xét riêng vói trường hợp *ImmunoPhenoType = Infiltrated* sẽ cho giá trị ổn định hơn.

* + 1. **No\_IF:**

**Score:** 0.004

Mặc dù có số điểm dữ liệu lớn nhất trong các loại nhưng dữ liệu không có mối quan hệ tuyến tính nên Score đạt được quá nhỏ (0.004). Dữ liệu mà chúng ta đang xét riêng với trường hợp này không có mối quan hệ tuyến tính.

1. **Random Forest Model:**

Ta xét đến mối quan hệ của toàn bộ các trường dữ liệu so với *PFS.* Sử dụng *RandomForestRegressor* (Hồi Quy Rừng Ngẫu Nhiên) xây dựng mô hình dự đoán **PFS** dựa trên các thuộc tính còn lại, chúng ta dùng hàm R-Squared để tính avarage R2Square dựa vào hàm r2\_square của hệ thống để đánh giá mối tương quan trong dữ liệu.

* 1. **Thêm trường dữ liệu:**

Đối với *ImmunoPhenoType=Excluded* sẽ được bỏ qua do chỉ có 1 mẫu dữ liệu duy nhất. Chúng ta sẽ đổi giá trị của cột *ImmunoPhenoType* hiện tại thành 3 trường dữ liệu khác: *Desert, Infiltrated* và *No*-*IF*. Đối với mỗi trường dữ liệu nếu mẫu dữ liệu đang xét thuộc loại đó thì tại ô đó sẽ mang giá trị là 1, ngược lại mang giá trị là 0.

* 1. **Huấn luyện mô hình:**

Hyper paramters:

n\_testimators = 100  
 random\_state = 100  
 max\_depth = 2

Số mẫu dữ liệu: 202

Input gồm các trường dữ liệu: *sum\_frameshift\_expression, max\_frameshift\_expression, Desert, Infiltrated* và *No-IF.*

Output: Giá trị *PFS.*

Model: RandomForestRegressor

Mô hình sẽ tính R2\_Score bằng cách lấy giá trị từ 100 cây để đưa ra quyết định. Mô hình sẽ predict *PFS* khi chúng ta truyền vào đủ 5 tham số trong input. Chúng ta triển khai hàm R2Score bằng cách lấy gí trị trung bình của 100 lần (bằng với *n\_estimators*), mỗi lần đều sẽ train mô hình với tham số trên, chỉ đổi tham số random\_state cho mô hình. Sau *n\_estimators* lần sẽ tính được giá trị trung bình R2Score của mô hình.

* 1. **Số liệu và đánh giá:**

**R2Score:** 0.188

**RMSE (Root Mean Square Error):** 8.594

Giá trị RMSE không nhỏ, cho thấy dữ liệu bất ổn định. R2-Score bằng 0.188 là một giá trị quá thấp để kết luận các thuộc tính đầu vào có mối quan hệ rõ ràng với *PFS*.

1. **Kết luận**

Thông qua hình ảnh và số liệu trong quá trình thực nghiệm, ta chưa thể kết luận mối quan hệ giữa các trường dữ liệu so với *PFS.* Do số mẫu dữ liệu chưa đủ nhiều để sử dụng các mô hình Neural Network để đánh giá. Trong tập dữ liệu có vấn đề imbalance (do Excluded chỉ có 1 mẫu dữ liệu). Bên cạnh đó, Các thuộc tính không có mối quan hệ tuyến tính với *PFS*. Khi phân tích dữ liệu, các số liệu đánh giá cho thấy các trường dữ liệu không có mối quan hệ tuyến tính với *PFS* do đó phương với tập dữ liệu hiện tại chúng ta không tìm ra được mối tương quan trong chúng.

**Tài liệu tham khảo**

1. **Mô hình RandomForestRegressor**

[*RandomForestRegressor*](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html)