**TỐI ƯU BÀI TOÁN DỰ BÁO KHẢ NĂNG PHẢI SỬ DỤNG ICU CỦA BỆNH NHÂN**

**Hà Kiều Anh1, Trần Thành Công2, Trương Nguyễn Duy Phương3**

1Lớp CNTTVB2, Đại học Bách Khoa Hà Nội, anhhakieuanh@gmail.com

2Lớp CNTTVB2, Đại học Bách Khoa Hà Nội, thanhcong18121993@gmail.com

3Lớp CNTTVB2, Đại học Bách Khoa Hà Nội, finance@almbank.vn

***TÓM TẮT***

**Giới thiệu*:*** *Tình hình Covid-19 trên toàn* thế *giới diễn ra càng ngày càng phức tạp, các biến chủng mới khiến cho việc truyền nhiễm bệnh xảy ra càng khó lường dẫn đến số lượng người mắc Covid-19 gia tăng nhanh* chóng*. Để có thể chuẩn bị hệ thống y tế tốt hơn và tránh việc quá tải, không thể sắp xếp được giường ICU cho các bệnh nhân nguy cấp thì việc thu thập thông tin chính xác và dự đoán ca bệnh là cực kỳ cần thiết*

***Phương pháp:*** *Sử dụng thuật toán học máy Random Forest, cải tiến trong việc dán nhãn và training.*

***Kết quả:*** *Chúng tôi*

**Từ khoá *-*** *Dự đoán ICU, random forest, supervised machine learning.*

1. **Giới thiệu**

Với hơn 169 triệu ca mắc và 3,5 triệu ca tử vong [1] tính từ lúc bùng phát dịch đến đầu tháng 5 năm 2021, COVID-19 đại dịch đã nhanh chóng trở thành tình trạng khẩn cấp nghiêm trọng về sức khỏe toàn cầu [2], đặt nặng lên khả năng chăm sóc sức khỏe của hệ thống y tế. Theo báo cáo tại Mỹ, một trong những nước chịu hậu quả nặng nề từ dịch bệnh, gánh nặng đối với hệ thống chăm sóc sức khỏe bắt nguồn từ cả tỷ lệ mắc Covid-19 cao và thực tế là 20% đến 30% bệnh nhân gặp phải tình trạng trung bình nặng cùng với nguy cơ suy giảm chức năng đa cơ quan, thời gian mắc bệnh và nhập viện kéo dài, và dẫn tới tử vong [3]. Hơn nữa, từ 5% đến 12% tổng số bệnh nhân được chẩn đoán mắc COVID-19 và lên đến 33% bệnh nhân nhập viện yêu cầu chăm sóc quan trọng hỗ trợ trong đơn vị chăm sóc đặc biệt (ICU) [3- 5]. Những ước tính chỉ ra rằng tỷ lệ chuyển ICU của bệnh nhân COVID-19 nhập viện là đáng kể cao hơn so với tỷ lệ chuyển giao ICU là 11% được báo cáo cho các bệnh nhân nhập viện khác [6, 7].

Hơn nữa, nhu cầu chăm sóc ICU vì bệnh Covid 19 thậm chí có thể cao hơn ở các nhóm nguy cơ cao, chẳng hạn như những người lớn tuổi hơn [3] hoặc những người có bệnh đi kèm từ trước [8]. Ví dụ: trên 75% trong số những bệnh nhân COVID-19 được nhận vào ICU có một hoặc nhiều bệnh lý đi kèm từ trước [9]. Bệnh nhân COVID-19 được nhận vào các đơn vị không thuộc ICU thường gặp tình trạng xấu đi nhanh chóng về mặt lâm sàng [10] và do đó cần có đánh giá lâm sàng thường xuyên. Tuy nhiên, với nguồn lực bị kéo mỏng, việc đánh giá thường xuyên rất khó và có thể làm tăng nguy cơ tiếp xúc giữa các nhân viên tuyến đầu.

Việc quản lý hiệu quả những nguồn lực và nhân sự, tối ưu nhân sự và sử dụng hiệu quả các nguồn lực của bệnh viện là cần thiết .Vì vậy cần thiết xác định xem liệu nguy cơ suy giảm sức khỏe lâm sàng của một cá nhân có đảm bảo chuyển ICU hay không có thể yêu cầu phân tích những thay đổi theo thời gian trong tình trạng của bệnh nhân và các chỉ số chính sắp xảy ra biến chứng của COVID-19. Các phương pháp học máy được giám sát có thể hữu ích để (a) phân tích và giải thích các giá trị lâm sàng và xét nghiệm của bệnh nhân và những thay đổi theo thời gian của họ, và (b) định lượng nguy cơ suy giảm lâm sàng và nhu cầu chuyển ICU.

Mục tiêu chính của nghiên cứu này là phát triển một bộ phân loại học máy có giám sát mới cho việc *dự đoán nguy cơ chuyển ICU trong vòng 24 giờ* tới cho bệnh nhân COVID-19. Chúng tôi đã áp dụng thuật toán Random Forest (RF) [11] - phương pháp tiếp cận được chứng minh là đầy hứa hẹn trong việc phân tích dữ liệu lâm sàng phức tạp [11], có tính tổng quát hóa mô hình cao [11], và có thể làm sáng tỏ tương tác bậc cao giữa các biến mà không ảnh hưởng đến độ chính xác của dự đoán [12].

1. **Thành phần và phương pháp**

Chúng tôi sử dụng Dataset từ Bệnh viện Sírio-Libanês, São Paulo và Brasilia tại Brazil. Tất cả dữ liệu được ẩn danh theo các khuyến nghị và thông lệ quốc tế. Dữ liệu đã được làm sạch và chia tỷ lệ theo Min Max Scaler. Bộ dữ liệu chứa 1925 bản ghi được thu thập từ 385 bệnh nhân, mỗi bản ghi chính xác 5 lần (tương ứng với 5 khoảng thời gian nhập viện của bệnh nhân). Trong số 231 cột của nó, có một "ICU" là mục tiêu đào tạo và dự đoán của chúng tôi, và 230 cột khác đều có thể được sử dụng làm input. ICU có giá trị nhị phân là 1 hoặc 0.



Chúng tôi thực hiện bài toán đưa ra bằng việc training mô hình học máy với các features từ dữ liệu Dataset các bệnh nhân Covid-19. Phương pháp của chúng tôi sẽ được thực hiện như sau:

1. Chuẩn bị dữ liệu: Data thô sẽ được xử lý bằng mean imputation (phương pháp thay thế lặp - thực hiện việc thay thế các giá trị còn thiếu bằng một số trung bình tính trên những giá trị còn lại) và Feature Engineering (quá trình trích chọn đăng trưng - biến đổi dữ liệu thành các đặc trưng đóng vai trò là đầu vào cho các mô hình học máy. Các đặc trưng được xử lý tốt sẽ nâng cao hiệu suất của mô hình. Các đặc trưng cũng ảnh hưởng rất bởi vấn đề cần giải quyết)

2. Chia dữ liệu: Dữ liệu sẽ được dùng 70% để training, 30% để test.

3.Mô hình: Decision Trees Random Forest

4.Metrics: Precision-recall auc và ROC auc là các metrics chính, recall và accuracy là thứ yếu.

5. Điều chỉnh siêu tham số (Hyperparameter tunning): Nevergrad, gaussian mixture, lagrangian relaxation and manual inputs.

1. **Mô hình bài toán và phân tích dữ liệu**
   1. ***Mô hình bài toán***

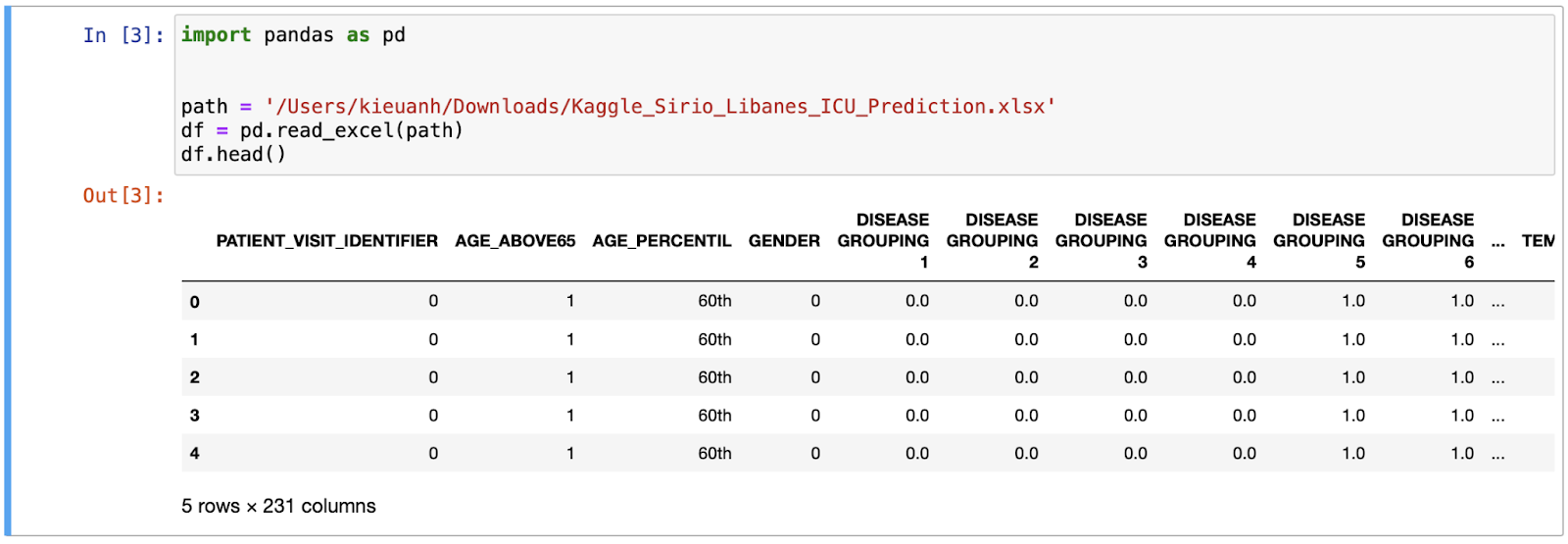
* *Dataset:* Dữ liệu ẩn danh từ Bệnh viện Sírio-Libanês, São Paulo và Brasilia tại Brazil. Tất cả dữ liệu được ẩn danh theo các khuyến nghị và thông lệ quốc tế. Dữ liệu đã được làm sạch và chia tỷ lệ theo Min Max Scaler.
* *Input dữ liệu đầu vào:* Dữ liệu đầu vào gồm các thông tin:

1. Thông tin nhân khẩu học của bệnh nhân
2. Các nhóm bệnh trước đây của bệnh nhân
3. Kết quả xét nghiệm máu
4. Chỉ số sinh tồn
5. Khí máu

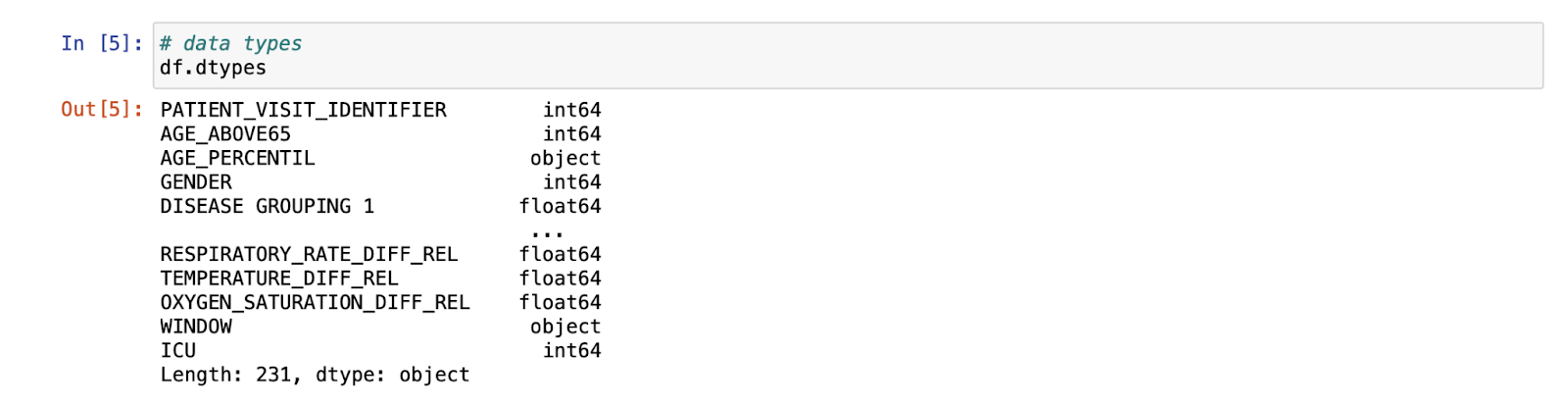
Các chỉ số được đo từ lúc vào bệnh viện , được chia thành 5 khoảng

|  |  |
| --- | --- |
| Khoảng thời gian nhập viện | Mô tả |
| 0-2 | Từ 0 - 2 giờ sau khi nhập viện |
| 2-4 | Từ 2 đến 4 giờ sau khi nhập viện |
| 4-6 | Từ 4 đến 6 giờ sau khi nhập viện |
| 6-12 | Từ 6 đến 12 giờ sau khi nhập viện |
| Trên 12 | Trên 12 giờ kể từ khi nhập viện |

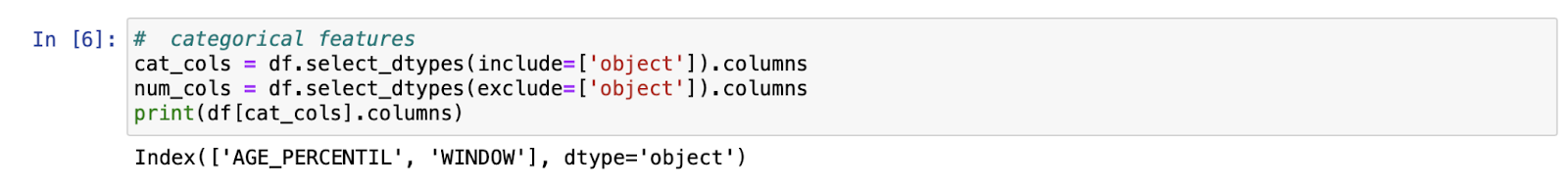
* *Output đầu ra:* Xác định bệnh nhân có thể sẽ phải sử dụng ICU hay không?
  1. ***Phân tích dữ liệu***
* *Mục đích:* Đưa ra cái nhìn tổng quan toàn cảnh đối với Dataset. Khám phá ra biến số tác động nhiều nhất đến việc bệnh nhân vào ICU (output mà bài toán đang cố gắng giải quyết).
* *Nhận xét chung:* Dữ liệu gồm 231 thông tin đặc trưng về từng ca bệnh Covid. Là tập hợp dữ liệu y tế, dữ liệu được coi là nhạy cảm. Do đó nhiều thông tin đặc trưng đã được xử lý với tên của chúng đã được thay đổi, tuy nhiên không ảnh hưởng nhiều đến sự khách quan của dữ liệu.
* *Phân tích:*



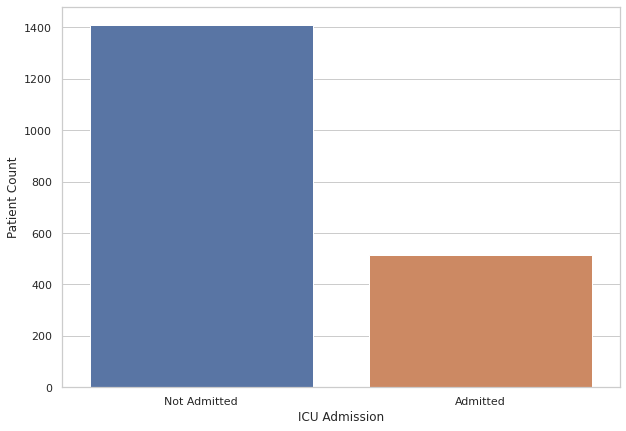
Có thể thấy rằng, có nhiều trường thông tin đặc trưng khiến cho tập hợp dữ liệu rộng, nhiều dữ liệu nhiễu; do đó cần thực hiện quá trình trích chọn đặc trưng (Feature Engineering).



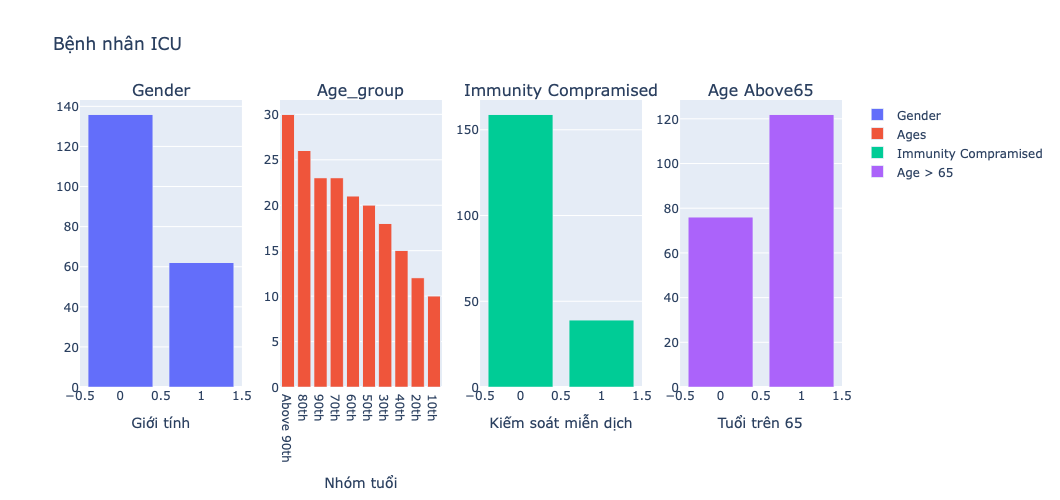
Trong dataset, có 2 trường thông tin đặc trưng mang tính dán nhãn, phân loại là ‘AGE\_PERCENTIL' và ‘WINDOW’:



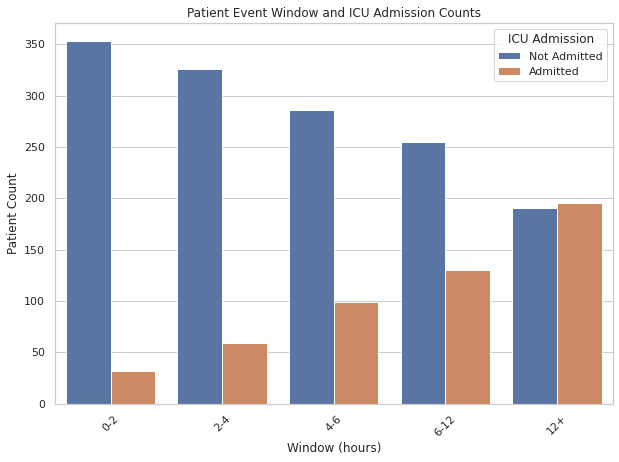
* *Thống kê mô phỏng*

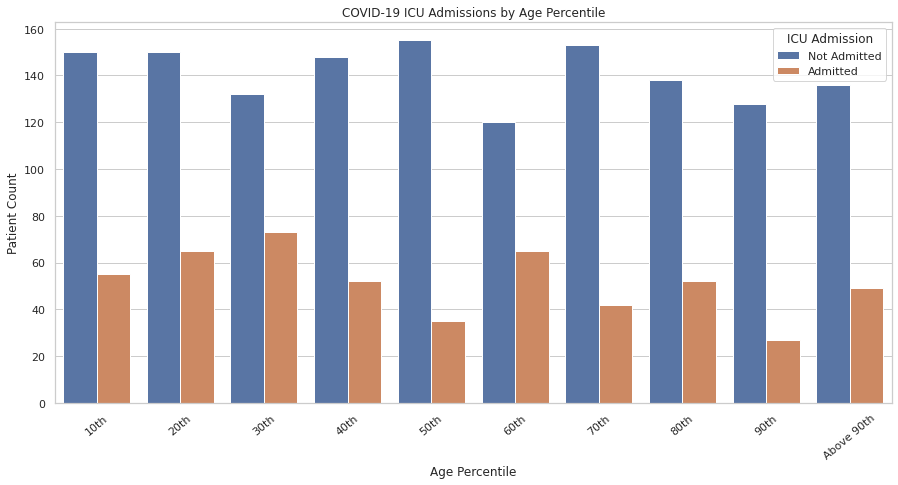


Trong số tất cả các bệnh nhân, khoảng 26,7% được nhận vào ICU



Có thể thấy rằng các cá nhân trên 65 tuổi dường như được nhận vào ICU thường xuyên hơn, chứng thực này cũng đã được tuyên bố bởi các quan chức y tế công cộng về việc bệnh nhân lớn tuổi có nguy cơ mắc bệnh nghiêm trọng hơn [3]. Ngoài ra các đặc trưng khác không có mối tương quan rõ ràng với việc vào ICU.





Có một mối tương quan tích cực rõ ràng giữa Window và ICU Admission, do đó nó được đưa vào mô hình. Ngược lại, dường như không có mối quan hệ rõ ràng giữa phần trăm tuổi (age percentile) và việc nhận vào ICU.

1. **Cơ sở lý thuyết** 
   1. ***Thuật toán Decision Tree***
2. *Tổng quan về thuật toán Decision Tree*

Thể hiện của thuật toán decision tree thể hiện dưới dạng hàm là đầu vào với các vector với giá trị của các thuộc tính và trả lại là một đầu ra duy nhất. Học bằng cây quyết định là một phương pháp thông dụng trong khai phá dữ liệu. Khi đó, cây quyết định mô tả một cấu trúc cây, trong đó, các lá đại diện cho các phân loại còn cành đại diện cho các kết hợp của các thuộc tính dẫn tới phân loại đó. Một cây quyết định có thể được học bằng cách chia tập hợp nguồn thành các tập con dựa theo một kiểm tra giá trị thuộc tính. Quá trình này được lặp lại một cách đệ quy cho mỗi tập con dẫn xuất. Quá trình đệ quy hoàn thành khi không thể tiếp tục thực hiện việc chia tách được nữa, hay khi một phân loại đơn có thể áp dụng cho từng phần tử của tập con dẫn xuất. Một bộ phân loại rừng ngẫu nhiên (random forest) sử dụng một số cây quyết định để có thể cải thiện tỉ lệ phân loại. Cây quyết định cũng là một phương tiện có tính mô tả dành cho việc tính toán các xác suất có điều kiện.

Cây quyết định có thể được mô tả như là sự kết hợp của các kỹ thuật toán học và tính toán nhằm hỗ trợ việc mô tả, phân loại và tổng quát hóa một tập dữ liệu cho trước.

Dữ liệu được cho dưới dạng các bản ghi có dạng: (x, y) = (x1, x2, x3..., xk, y)

Biến phụ thuộc (dependant variable) y là biến mà chúng ta cần tìm hiểu, phân loại hay tổng quát hóa. x1, x2, x3... là các biến sẽ giúp ta thực hiện công việc đó.

Decision Trees gồm 3 phần chính: 1 node gốc (root node), những node lá (leaf nodes) và các nhánh của nó (branches). Node gốc là điểm bắt đầu của cây quyết định và cả hai node gốc và node chứa câu hỏi hoặc tiêu chí để được trả lời. Nhánh biểu diễn các kết quả của kiểm tra trên nút. Ví dụ câu hỏi ở node đầu tiên yêu cầu câu trả lời là “yes” hoặc là “no” thì sẽ có 1 node con chịu trách nhiệm cho phản hồi là “yes”, 1 node là “no”. Hay nói cách khác, một cây quyết định đúng sai tương ứng với một khẳng định rằng thuộc tính mục tiêu là đúng nếu và chỉ nếu các thuộc tính đầu vào thỏa mãn rằng sẽ tạo nên một đường trên cây dẫn đến giá trị cuối cùng là đúng.

Ta có thể viết như sau

Mỗi một Path là kết hợp của các kiểm tra của các giá trị thuộc tính. Vì vaayj, toàn bộ biểu thức có thể đưa về dạng chuẩn hội, điều này có nghĩa rằng bất kì hàm nào trong logic mệnh đề cũng có thể biểu diễn thành một cây quyết định

Ví dụ:

[Text(167.4, 181.2, 'LACTATE\_MEAN <= 0.188\nentropy = 1.0\nsamples = 1347\nvalue = [665, 682]\nclass = 1'),

Text(83.7, 108.72, 'BLOODPRESSURE\_SISTOLIC\_MAX <= -0.097\nentropy = 0.782\nsamples = 379\nvalue = [88, 291]\nclass = 1'),

Text(41.85, 36.23999999999998, '\n (...) \n'),

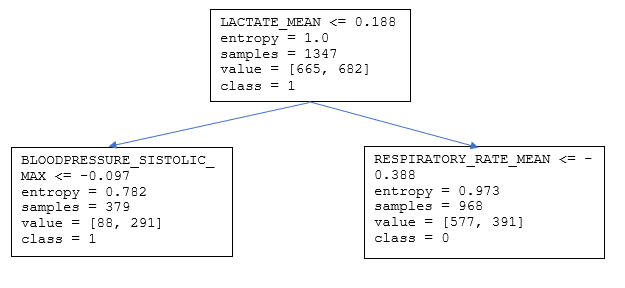
Text(125.55000000000001, 36.23999999999998, '\n (...) \n'),

Text(251.10000000000002, 108.72, 'RESPIRATORY\_RATE\_MEAN <= -0.388\nentropy = 0.973\nsamples = 968\nvalue = [577, 391]\nclass = 0'),

Text(209.25, 36.23999999999998, '\n (...) \n'),

Text(292.95, 36.23999999999998, '\n (...) \n')]

// Thể hiện ở hình bên dưới



Ta có thể sử dụng chỉ số trung bình của xét nghiệm lactate trong máu như nút đầu tiên, nếu chỉ số này nhỏ hơn 0.188 ta xem xét bệnh nhân có thể phải dùng đến giường ICU. Sau đó nếu chỉ số BLOODPRESSURE SISTOLIC MAX trong máu nhỏ hơn -0.097, ta tiếp tục đi dần xuống theo các nhánh của cây

1. *Chi tiết giải thuật*

Vấn đề nảy sinh khi áp dụng thuật toán với bài toán phân loại người có thể sẽ cần sử dụng đến ICU đó là có quá nhiều thuộc tính và nếu sử dụng thuật toán vét cạn, ta không thể đủ thời gian và nguồn lực để xử lý tìm kiếm với 2253 cách kết hợp các chỉ số xét nghiệm và thông tin liên quan như trong bài.

Áp dụng công thức chung của AI:

Với thuật toán học cây quyết định, giải thuật tìm kiếm tham lam chia để trị được áp dụng: Luôn luôn kiểm tra thuộc tính quan trọng nhất đầu tiên. Hàm heuristic trong trường hợp này là hàm xác định tính quan trọng của thuộc tính.

Ví dụ:

|  |  |
| --- | --- |
| LACTATE MEAN <= 0.188 | |
| 379 mẫu | 968 mẫu |
| 88 mẫu không cần dùng ICU  291 mẫu cần dùng | 577 mẫu không cần dùng  391 mẫu cần dùng |

**Giải thuật tổng quan như sau (Tham khảo – Artification Intelligence: A modern approach)**

Function DECISION-TREE-LEARNING (examples, attributes, parent\_examples) returns a tree

If examples is empty then return PLURALITY-VALUE(parent\_examples)

Else if all examples have the same classification then return the classification

Else if attributes is empty then return PLURALITY-VALUE(examples)

Else

For each value k of A do:

Add a branch to tree with label (A = vk) and subtree *subtree*

Return *tree*

1. *Lựa chọn thuộc tính*

Đến đây, phần khó nhất chính là làm sao để có thể chọn được các thuộc tính để đưa vào dựa theo yêu cầu về tính quan trọng. Thuật toán tìm kiếm của việc học theo cây quyết định được thiết kế nhằm tìm kiếm một giá trị “xấp xỉ” gần với chiều sâu tối thiểu của cây tối ưu. Ý tưởng là tìm kiếm những thuộc tính mà tốt nhất trong việc phân hoạt chính xác các ví dụ được học. Như trong ví dụ, chỉ số trung bình của xét nghiệm LACTATE máu có thể không phân loại chính xác hoàn toàn nhưng có thể phân loại 2 nhóm ở mức “tạm” tốt.

Để có thể đánh giá mức độ quan trọng của thông tin, ta dùng khái niệm entropy thông tin (Shannon and Weaver, 1949). Entropy là phép đo sự không chắc chắn của một biến ngẫu nhiên. Ví dụ như với một xúc sắc luôn lật ngửa, sẽ không có sự không chắc chắn và lúc này entropy sẽ bằng không, hay nói cách khác sẽ không thu được thông tin gì từ việc quan sát giá trị này.

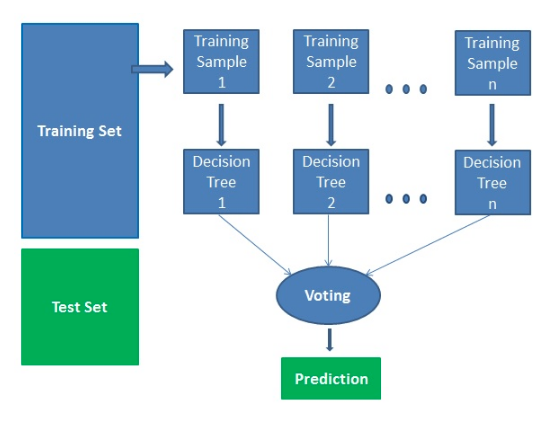
Hàm entropy thông tin cho biến ngẫu nhiên liên tục được định nghĩa như sau:

Shannon đã xây dựng định nghĩa về entropy thỏa mãn các giả định: Entropy phải tỷ lệ thuận liên tục với các xác suất xuất hiện của các phần tử ngẫu nhiên trong tín hiệu. Thay đổi nhỏ trong xác suất phải dẫn đến thay đổi nhỏ trong entropy. Nếu các phần tử ngẫu nhiên đều có xác suất xuất hiện bằng nhau, việc tăng số lượng phần tử ngẫu nhiên phải làm tăng entropy. Có thể tạo các chuỗi tín hiệu theo nhiều bước, và entropy tổng cộng phải bằng tổng có trọng số của entropy của từng bước.

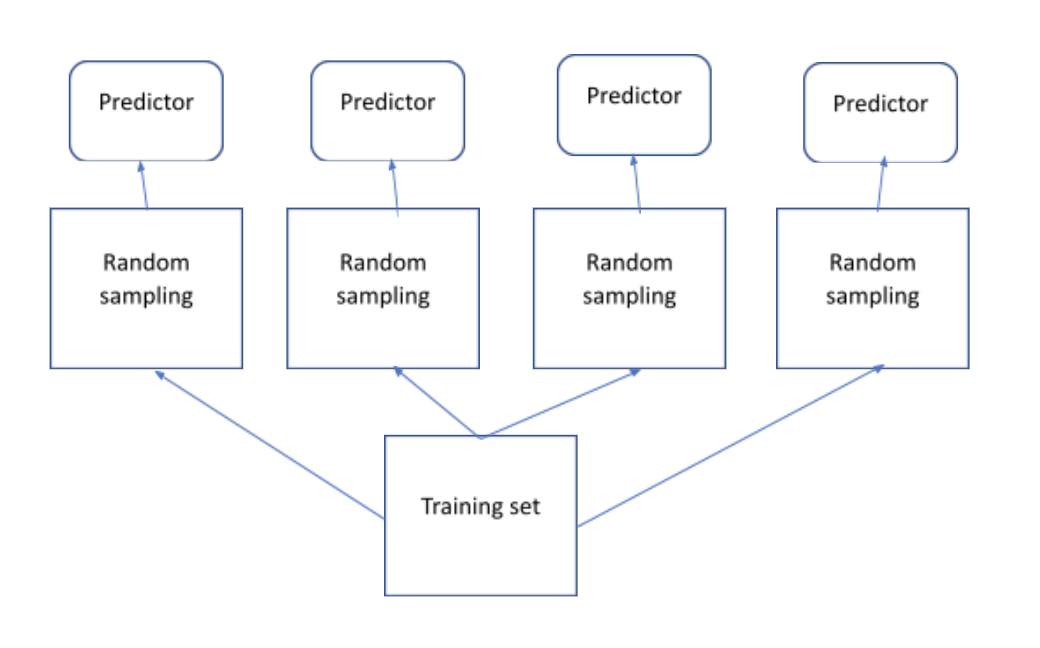
Như trong ví dụ, LACTATE\_MEAN với entropy = 1 sẽ được chọn mà thuộc tính đầu tiên, sau đó là BLOODPRESSURE\_SISTOLIC\_MAX với entropy = 0.782 và RESPIRATORY\_RATE\_MEAN với entropy = 0.973.

* 1. ***Thuật toán Random forest***

Thuật toán Random Forest được phát triển dựa trên decision trees, ý tưởng đơn giản là: để dự báo chính xác hơn, ta sẽ sử dụng cùng lúc nhiều decision trees và lựa chọn theo phương án xuất hiện nhiều nhất. Random forests tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu. Nó cũng cung cấp một chỉ báo khá tốt về tầm quan trọng của tính năng.

******

1. *Kỹ thuật bagging:* Với kỹ thuật này, tập dữ liệu training ban đầu sẽ được chia nhỏ thành các tập mẫu ngẫu nhiên, mỗi tập mẫu sẽ được đưa vào training để tạo ra một model riêng được thể hiện như dưới hình sau:



1. *Random forest là sự kết hợp của decision tree và kỹ thuật bagging*

Thuật toán random forest hoạt động theo bốn bước:

1. Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho.

2. Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây.

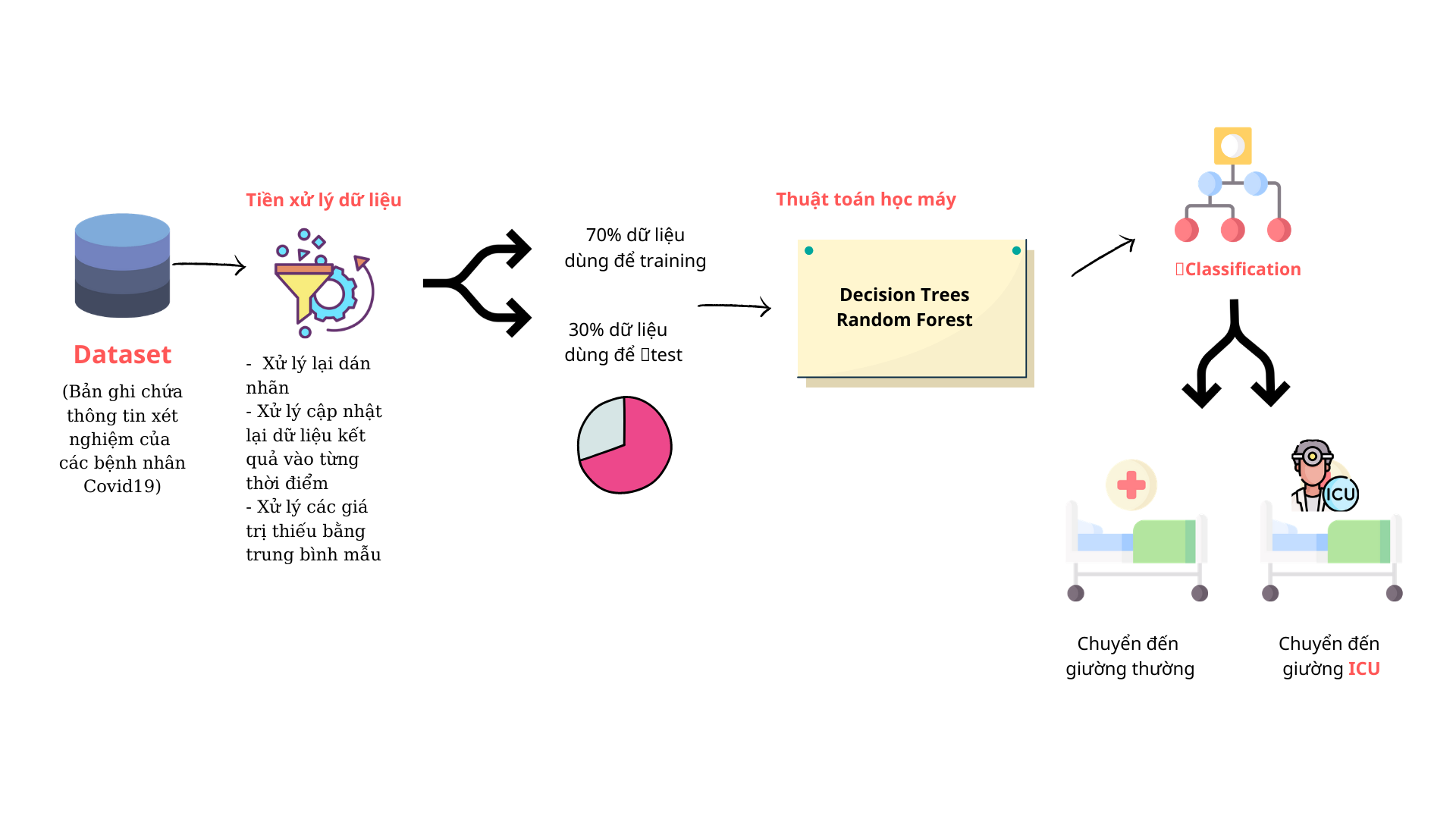
3. Hãy bỏ phiếu cho mỗi kết quả dự đoán.

4. Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.

Bằng việc kết hợp decision tree với bagging, ta có được một thuật toán học không quá phức tạp nhưng rất hiệu quả và mạnh mẽ

1. **Mô hình đề xuất**

Từ các cơ sở đã phân tích như trên, chúng tôi đưa ra mô hình đề xuất cho bài toán dự đoán khả năng nằm ICU của bệnh nhân Covid theo minh hoa như hình dưới đây:



1. **Điểm mới cải tiến**

Sau khi tham khảo Notebook trên Kaggle với chủ đề tương tự đề tài, nhóm thấy rằng các bài đang tập trung vào model, mà chưa đánh giá chính xác tính hợp lý và logic của dữ liệu, có hiện tượng overfitting, chưa phù hợp với yêu cầu của bệnh viện và giải quyết được bài toán triệt để. Nhóm có đề xuất các điểm như sau.

* 1. ***Đánh giá***
* *Training các mô hình tại từng thời điểm:*

Do tính chất của việc xét nghiệm (sẽ có những xét nghiệm được thực hiện ngay khi người bệnh bắt đầu vào bệnh viện) và nhu cầu của người dùng (bệnh viện muốn biết thông tin ngay khi có những kết quả xét nghiệm đầu tiên). Nhóm đề xuất chạy các mô hình tương ứng với các khoảng thời gian: **ngay từ khi vào viện, sau 2h, sau 4h, sau 6h và sau 12h.**

* *Cần đánh nhãn lại cho các bản ghi người bệnh có cần sử dụng ICU:*

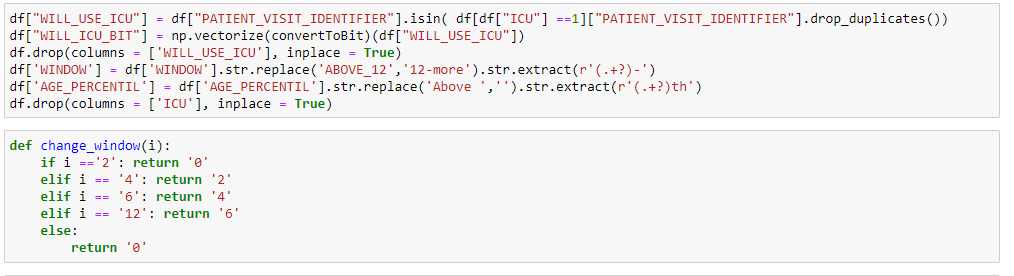
Hiện các bản ghi của cùng một người bệnh sẽ ghi dấu người bệnh cần dùng ICU vào khoảng thời gian khi người này bắt đầu sử dụng. Cụ thể, với người bệnh nhập viện, sau 6 h bắt đầu cần sử dụng ICU thì nhãn ICU sẽ được đánh dấu từ đây, các bản ghi từ khi nhập viện vẫn đánh dấu là 0. Mục tiêu của đề bài đặt ra là có thể nhận biết được bệnh nhân có cần sử dụng ICU hay không ngay từ khi bệnh nhân bắt đầu vào viện và có được những thông tin đầu tiên. Do đó cần phải đánh nhãn lại.

* *Cập nhật lại thông tin cho các bản ghi:*

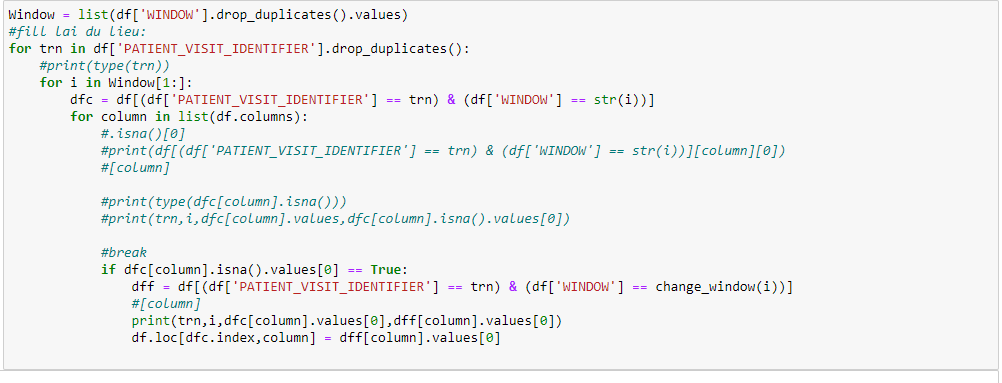
Hiện trong dữ liệu có rất nhiều trường xét nghiệm bị trống tại một số khoảng thời gian, một trong số những lý do đó là việc xét nghiệm không thực hiện thường xuyên mà theo những khoảng thời gian nhất định. Do đó để có thể có kết quả training tốt nhất, các bản ghi cần được ghi với kết quả xét nghiệm của lần thực hiện gần nhất nếu có, thay vì để trống.

* 1. ***Thực nghiệm***

Bước 1: Chuyển đổi dữ liệu về dạng phù hợp và dán lại nhãn ICU



Bước 2: Cập nhật lại dữ liệu theo hướng tại mỗi thời điểm (Window) các chỉ số của bệnh nhân là chỉ số được đo tại thời điểm gần nhất



Dữ liệu trước khi xử lý:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PATIENT\_VISIT\_IDENTIFIER | AGE\_ABOVE65 | AGE\_PERCENTIL | GENDER | POTASSIUM\_MEDIAN | POTASSIUM\_MEAN | POTASSIUM\_MIN | POTASSIUM\_MAX | WINDOW | ICU |
| 2 | 0 | 10th | 0 | -0.7037 | -0.7037 | -0.7037 | -0.7037 | 0-2 | 0 |
| 2 | 0 | 10th | 0 |  |  |  |  | 2-4 | 0 |
| 2 | 0 | 10th | 0 |  |  |  |  | 4-6 | 0 |
| 2 | 0 | 10th | 0 |  |  |  |  | 6-12 | 0 |
| 2 | 0 | 10th | 0 | -0.59259 | -0.59259 | -0.59259 | -0.59259 | ABOVE\_12 | 1 |

Dữ liệu sau khi xử lý

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PATIENT\_VISIT\_IDENTIFIER | AGE\_ABOVE65 | AGE\_PERCENTIL | PLATELETS\_DIFF | POTASSIUM\_MEDIAN | POTASSIUM\_MEAN | POTASSIUM\_MIN | POTASSIUM\_MAX | WINDOW | WILL\_ICU\_BIT |
| 2 | 0 | 10 | -1 | -0.7037 | -0.7037 | -0.7037 | -0.7037 | 0 | 1 |
| 2 | 0 | 10 | -1 | -0.7037 | -0.7037 | -0.7037 | -0.7037 | 2 | 1 |
| 2 | 0 | 10 | -1 | -0.7037 | -0.7037 | -0.7037 | -0.7037 | 4 | 1 |
| 2 | 0 | 10 | -1 | -0.7037 | -0.7037 | -0.7037 | -0.7037 | 6 | 1 |
| 2 | 0 | 10 | -1 | -0.59259 | -0.59259 | -0.59259 | -0.59259 | 12 | 1 |

* 1. ***Thực hiện chạy mô hình:***

Nhóm thực hiện chạy mô hình vào các khoảng thời gian sau:

c.1/ Xây dựng các hàm để thực hiện training



Nhóm sẽ chia dữ liệu thành 2 phần với tỷ lệ 70% dữ liệu để training, 30% dữ liệu để test. Các hệ số được dùng để đánh giá hiệu suất của mô hình gồm: Accuracy, Precision, Recall, ROC Auc, PR Auc.

Cài đặt tham số cho mô hình training



Các tham số gồm:

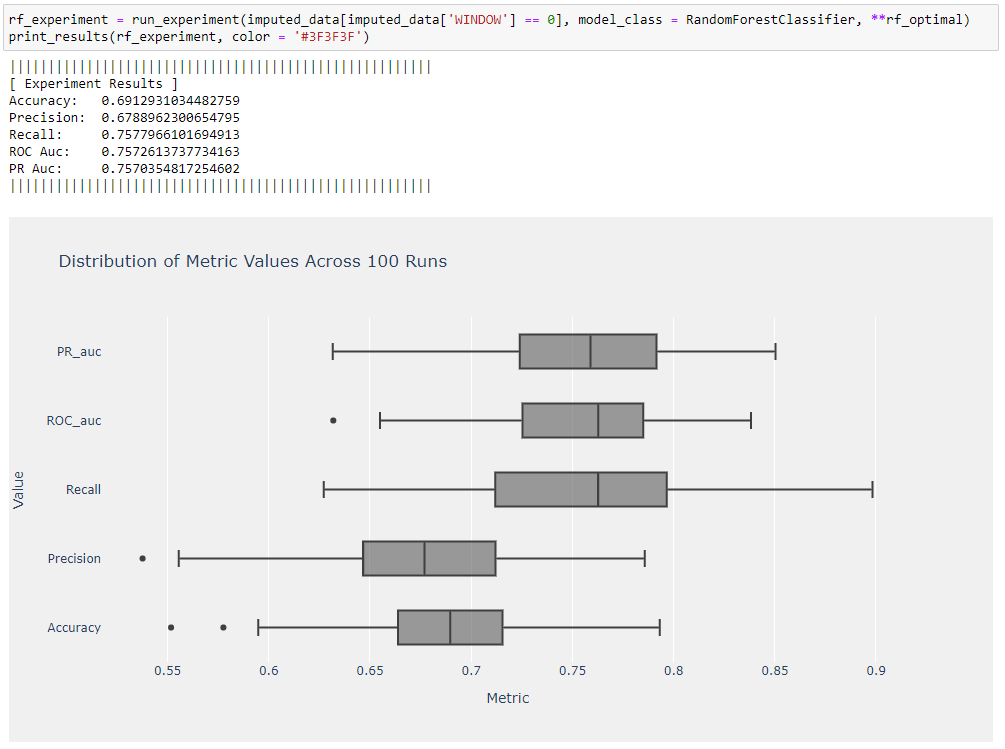
Số lượng cây: 2100

Độ sâu lớn nhất của 1 cây: 27

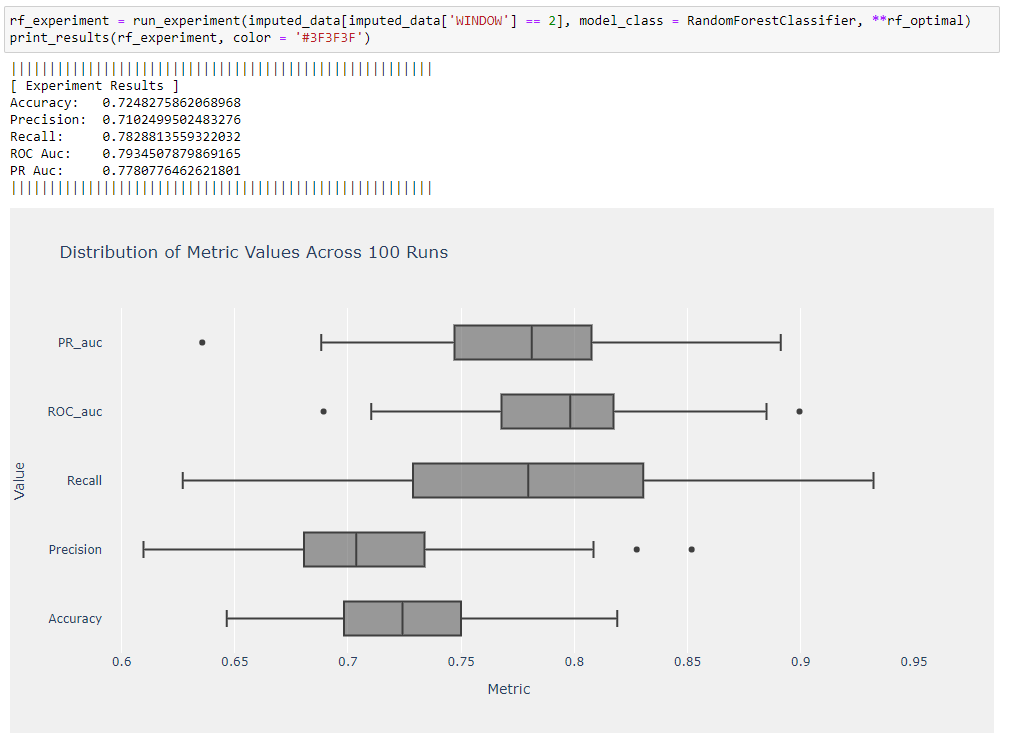
Số lượng thuộc tính được xem xét mở mỗi lần chia tách: 15%

c.2/ Thực hiện training

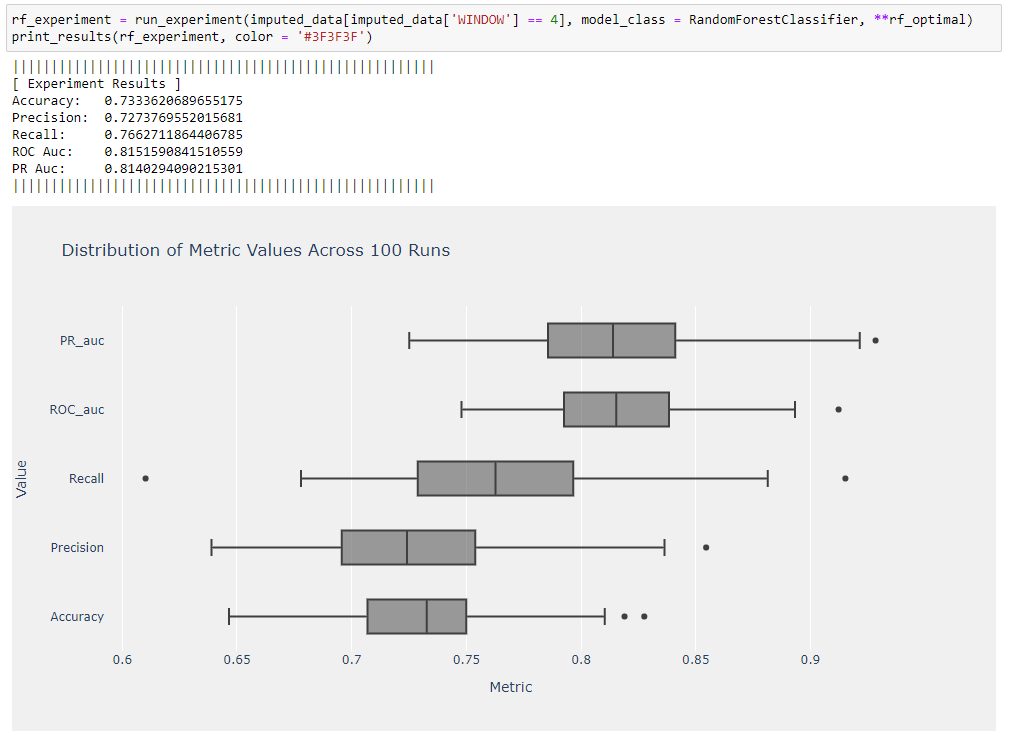
1. Tại thời điểm bệnh nhân bắt đầu vào viện



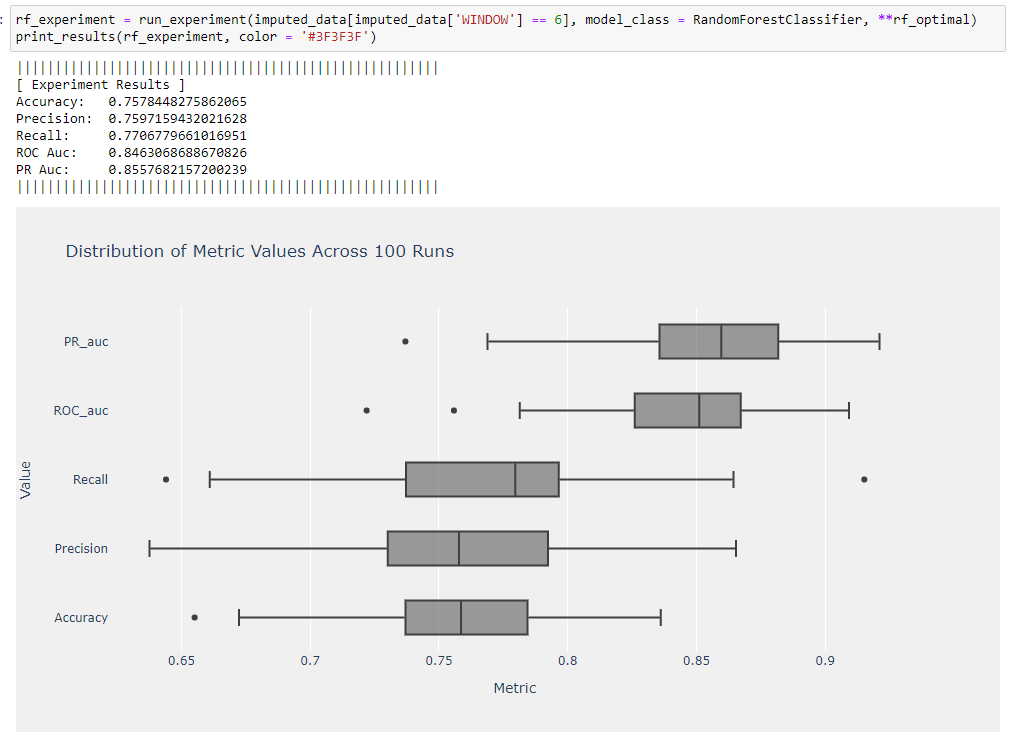
1. Tại thời điểm sau 2 giờ vào viện



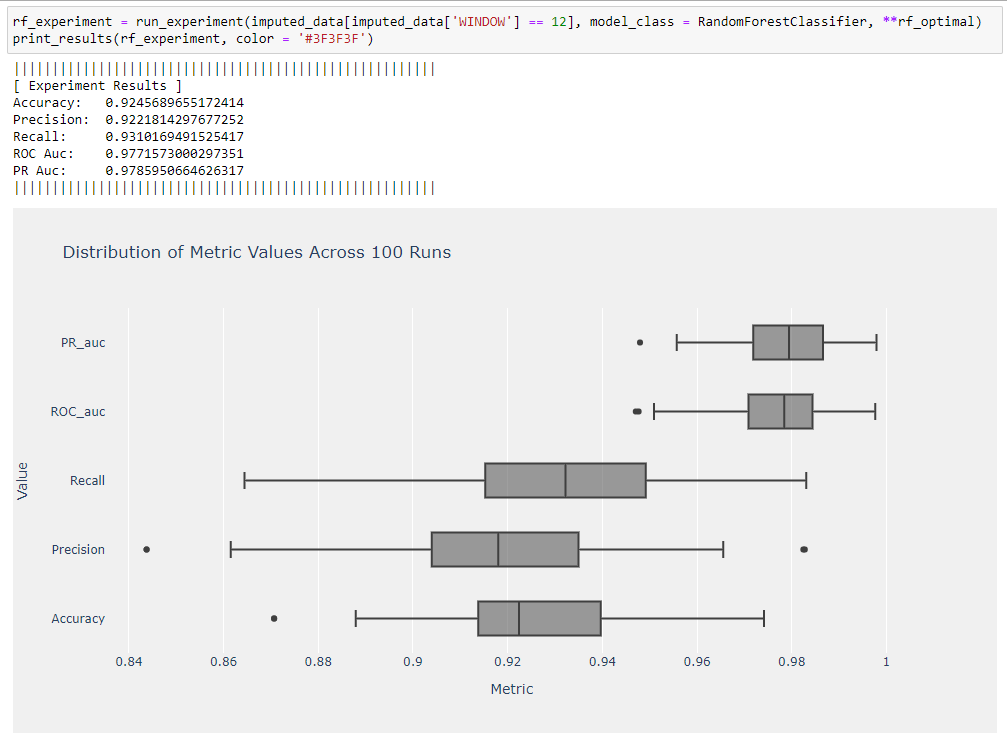
1. Tại thời điểm sau 4h vào viện



1. Tại thời điểm sau 6h vào viện

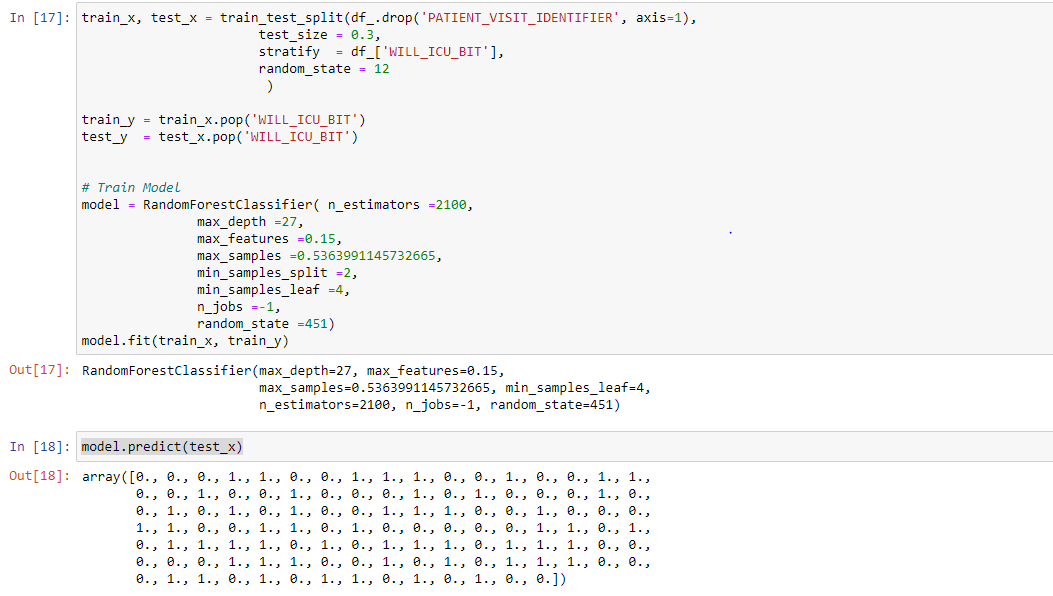


1. Tại thời điểm sau 12h vào viện



Qua việc thực hiện chạy theo từng khoảng thời gian, ta thấy được rằng hiệu quả của mô hình tăng dần theo thời gian vào viện, điều này là phù hợp với thực tế.

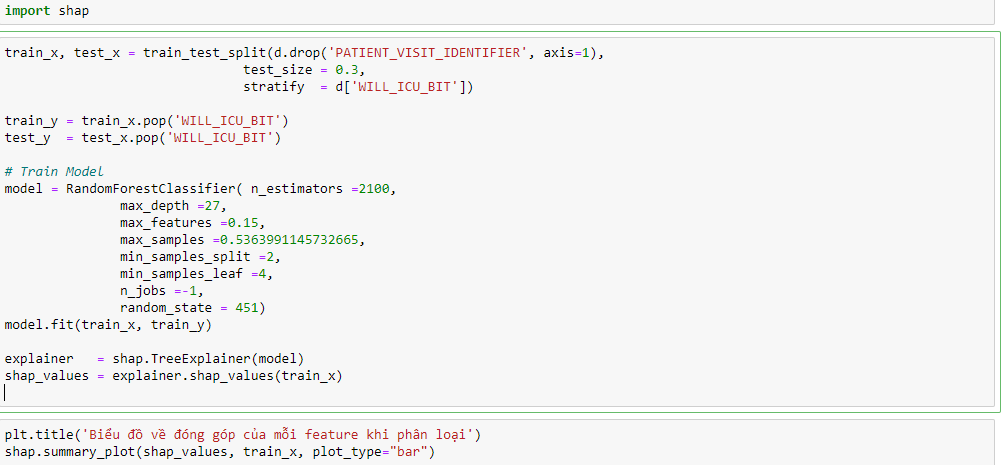
Kết quả sử dụng mô hình phân loại cho các bệnh nhân tại thời điểm 12h, kiểm tra lại với dữ liệu dùng để test, kết quả thu được mô hình phân loại đúng cho 103/116 mẫu.

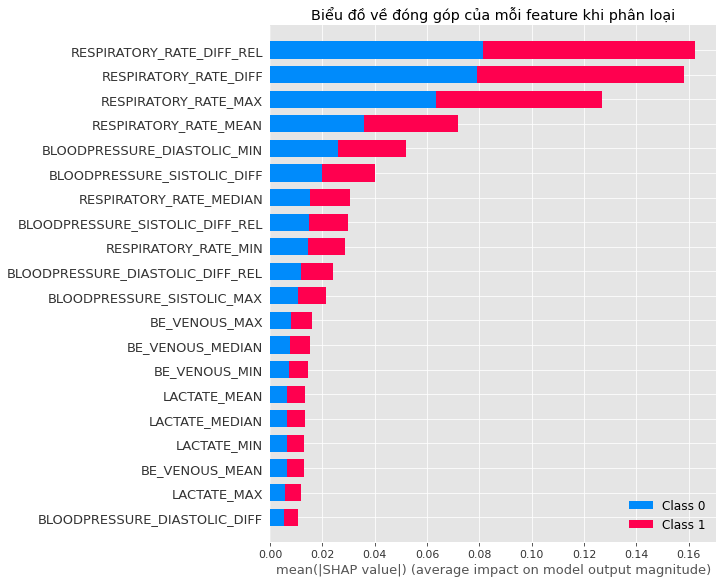


|  |  |  |  |
| --- | --- | --- | --- |
| NO | Predict | Reality | Check |
| 0 | 0 | 0 | TRUE |
| 1 | 0 | 0 | TRUE |
| 2 | 0 | 0 | TRUE |
| 3 | 1 | 1 | TRUE |
| 4 | 1 | 0 | FALSE |
| 5 | 0 | 0 | TRUE |
| 6 | 0 | 0 | TRUE |
| 7 | 1 | 1 | TRUE |
| 8 | 1 | 1 | TRUE |
| 9 | 1 | 1 | TRUE |
| 10 | 0 | 0 | TRUE |
| 11 | 0 | 0 | TRUE |
| 12 | 1 | 1 | TRUE |
| 13 | 0 | 0 | TRUE |
| 14 | 0 | 0 | TRUE |
| 15 | 1 | 1 | TRUE |
| 16 | 1 | 1 | TRUE |
| 17 | 0 | 0 | TRUE |
| 18 | 0 | 0 | TRUE |
| 19 | 1 | 1 | TRUE |
| 20 | 0 | 0 | TRUE |
| 21 | 0 | 0 | TRUE |
| 22 | 1 | 1 | TRUE |
| 23 | 0 | 0 | TRUE |
| 24 | 0 | 0 | TRUE |
| 25 | 0 | 0 | TRUE |
| 26 | 1 | 1 | TRUE |
| 27 | 0 | 0 | TRUE |
| 28 | 1 | 1 | TRUE |
| 29 | 0 | 0 | TRUE |
| 30 | 0 | 1 | FALSE |
| 31 | 0 | 0 | TRUE |
| 32 | 1 | 1 | TRUE |
| 33 | 0 | 1 | FALSE |
| 34 | 0 | 0 | TRUE |
| 35 | 1 | 1 | TRUE |
| 36 | 0 | 0 | TRUE |
| 37 | 1 | 1 | TRUE |
| 38 | 0 | 0 | TRUE |
| 39 | 1 | 1 | TRUE |
| 40 | 0 | 0 | TRUE |
| 41 | 0 | 0 | TRUE |
| 42 | 1 | 1 | TRUE |
| 43 | 1 | 0 | FALSE |
| 44 | 1 | 1 | TRUE |
| 45 | 0 | 0 | TRUE |
| 46 | 0 | 1 | FALSE |
| 47 | 1 | 1 | TRUE |
| 48 | 0 | 0 | TRUE |
| 49 | 0 | 1 | FALSE |
| 50 | 0 | 1 | FALSE |
| 51 | 1 | 1 | TRUE |
| 52 | 1 | 1 | TRUE |
| 53 | 0 | 0 | TRUE |
| 54 | 0 | 0 | TRUE |
| 55 | 1 | 1 | TRUE |
| 56 | 1 | 1 | TRUE |
| 57 | 0 | 0 | TRUE |
| 58 | 1 | 1 | TRUE |
| 59 | 0 | 0 | TRUE |
| 60 | 0 | 0 | TRUE |
| 61 | 0 | 0 | TRUE |
| 62 | 0 | 0 | TRUE |
| 63 | 0 | 0 | TRUE |
| 64 | 1 | 1 | TRUE |
| 65 | 1 | 1 | TRUE |
| 66 | 0 | 0 | TRUE |
| 67 | 1 | 1 | TRUE |
| 68 | 0 | 0 | TRUE |
| 69 | 1 | 1 | TRUE |
| 70 | 1 | 1 | TRUE |
| 71 | 1 | 1 | TRUE |
| 72 | 1 | 1 | TRUE |
| 73 | 0 | 1 | FALSE |
| 74 | 1 | 1 | TRUE |
| 75 | 0 | 0 | TRUE |
| 76 | 1 | 0 | FALSE |
| 77 | 1 | 1 | TRUE |
| 78 | 1 | 1 | TRUE |
| 79 | 0 | 0 | TRUE |
| 80 | 1 | 1 | TRUE |
| 81 | 1 | 1 | TRUE |
| 82 | 1 | 1 | TRUE |
| 83 | 0 | 0 | TRUE |
| 84 | 0 | 0 | TRUE |
| 85 | 0 | 0 | TRUE |
| 86 | 0 | 0 | TRUE |
| 87 | 0 | 0 | TRUE |
| 88 | 1 | 1 | TRUE |
| 89 | 1 | 1 | TRUE |
| 90 | 1 | 1 | TRUE |
| 91 | 0 | 0 | TRUE |
| 92 | 0 | 0 | TRUE |
| 93 | 1 | 1 | TRUE |
| 94 | 0 | 0 | TRUE |
| 95 | 1 | 1 | TRUE |
| 96 | 0 | 1 | FALSE |
| 97 | 1 | 1 | TRUE |
| 98 | 1 | 1 | TRUE |
| 99 | 1 | 0 | FALSE |
| 100 | 0 | 0 | TRUE |
| 101 | 0 | 0 | TRUE |
| 102 | 0 | 1 | FALSE |
| 103 | 1 | 1 | TRUE |
| 104 | 1 | 1 | TRUE |
| 105 | 0 | 0 | TRUE |
| 106 | 1 | 1 | TRUE |
| 107 | 0 | 0 | TRUE |
| 108 | 1 | 1 | TRUE |
| 109 | 1 | 1 | TRUE |
| 110 | 0 | 0 | TRUE |
| 111 | 1 | 1 | TRUE |
| 112 | 0 | 1 | FALSE |
| 113 | 1 | 1 | TRUE |
| 114 | 0 | 0 | TRUE |
| 115 | 0 | 0 | TRUE |

c.3/ Phân tích

Thực hiện đánh giá tác động của các feature khi chạy mô hình phân loại





Qua biểu đồ trên, ta có thể thấy rằng tần số hô hấp (Respiratory rate) và áp lực máu (Blood Pressure) là những nhân tố quan trọng nhất ảnh hưởng đến việc bệnh nhân covid có cần dùng đến giường ICU hay không. Do đó nhóm khuyến nghị các chỉ số này cần được đo lường sớm và thường xuyên cho các bệnh nhân mắc covid 19

**Tài liệu tham khảo**

1. World Health Organization. WHO COVID-19 Dashboard. Available online: https://who.sprinklr.com/ (accessed on 9 April 2020).
2. Fauci, A.S.; Lane, H.C.; Redfield, R.R. Covid-19—Navigating the Uncharted. N. Engl. J. Med. 2020, 382, 1268–1269.
3. CDC COVID-19 Response Team. Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19)—United States, February 12–March 16, 2020. MMWR Morb. Mortal. Wkly. Rep. 2020, 69, 343–346.
4. Myers, L.C.; Parodi, S.M.; Escobar, G.J.; Liu, V.X. Characteristics of Hospitalized Adults With COVID-19 in an Integrated Health Care System in California. JAMA 2020.
5. Richardson, S.; Hirsch, J.S.; Narasimhan, M.; Crawford, J.M.; McGinn, T.; Davidson, K.W.; Cookingham, J. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA 2020.
6. Howell, E. Active Bed Management by Hospitalists and Emergency Department Throughput. Ann. Intern. Med. 2008, 149, 804.
7. Moriarty, J.P.; Schiebel, N.E.; Johnson, M.G.; Jensen, J.B.; Caples, S.M.; Morlan, B.W.; Huddleston, J.M.; Heubner, M.; Naessens, J.M. Evaluating implementation of a rapid response team: Considering alternative outcome measures. Int. J. Qual. Health Care 2014, 26, 49–57.
8. Verity, R.; Okell, L.C.; Dorigatti, I.; Winskill, P.; Whittaker, C.; Imai, N.; Cori, A.; Fu, H.; Baguelin, M.; Dighe, A.; et al. Estimates of the severity of coronavirus disease 2019: A model-based analysis. Lancet Infect. Dis. 2020, 20, 669–677. [CrossRef]
9. CDC COVID-19 Response Team. Preliminary Estimates of the Prevalence of Selected Underlying Health Conditions Among Patients with Coronavirus Disease 2019—United States, February 12–March 28, 2020. MMWR Morb. Mortal. Wkly. Rep. 2020, 69, 382–386.
10. National Center for Immunization and Respiratory Diseases (NCIRD). Interim Clinical Guidance for Management of Patients with Confirmed Coronavirus Disease (COVID-19). 2020. Available online: https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html (accessed on 9 April 2020).
11. Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32.
12. Basu, S.; Kumbier, K.; Brown, J.B.; Yu, B. Iterative random forests to discover predictive and stable high-orderinteractions. Proc. Natl. Acad. Sci. USA 2018, 115, 1943–1948.