TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**FINAL FROJECT**

**DỰ ÁN CUỐI KỲ MÔN NHẬP MÔN HỌC MÁY**

*Người hướng dẫn*: **GV LÊ ANH CƯỜNG**

*Người thực hiện*: **LƯƠNG PHAN HOÀN NHÂN – 52000899**

**NGÔ HOÀNG KHÔI - 52000676**

**ĐỖ DUY LINH - 51702222**

Lớp **:** 503044

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**FINAL PROJECT**

**DỰ ÁN CUỐI KỲ MÔN NHẬP MÔN HỌC MÁY**

*Người hướng dẫn*: **GV LÊ ANH CƯỜNG**

*Người thực hiện*: **LƯƠNG PHAN HOÀN NHÂN – 52000899**

**NGÔ HOÀNG KHÔI - 52000676**

**ĐỖ DUY LINH - 51702222**

Lớp **:** 503044

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

LỜI CẢM ƠN

“Lời đầu tiên, xin trân trọng gửi lời cảm ơn đến Thầy Lê Anh Cường đã hướng dẫn nhóm, thầy đã tận tình hướng dẫn nhóm trong suốt quá trình học tập cũng như hoàn thành báo cáo một cách tốt nhất.

Xin chân thành cảm ơn các thầy, cô thuộc khoa công nghệ thông tin trường đại học Tôn Đức Thắng đã tận tình giảng dạy cho chúng tôi trong thời gian học tập.

Xin cảm ơn thầy Lê Anh Cường đã đọc bài báo cáo và cho nhóm những nhận xét ý nghĩa và quý báu, chỉnh sửa những thiếu sót của chúng tôi trong bài.

Do kiến thức và khả năng lý luận của chúng em còn giới hạn và có nhiều thiếu sót, kính mong sự chỉ dẫn và đóng góp của Thầy để bài báo cáo của chúng em được hoàn thiện hơn.”

*Xin chân thành cảm ơn!*

ĐỒ ÁN ĐƯỢC HOÀN THÀNH

TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là sản phẩm đồ án của chúng em và được sự hướng dẫn của GV Lê Anh Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào chúng em xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do chúng gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Lương Phan Hoàn Nhân*

*Ngô Hoàng Khôi*

*Đỗ Duy Linh*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

MỤC LỤC

Mục lục

[LỜI CẢM ƠN i](#_Toc154246614)

[ĐỒ ÁN ĐƯỢC HOÀN THÀNH ii](#_Toc154246615)

[TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG ii](#_Toc154246616)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc154246617)

[MỤC LỤC 1](#_Toc154246618)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 3](#_Toc154246619)

[DANH MỤC HÌNH 3](#_Toc154246620)

[DANH MỤC BẢNG 4](#_Toc154246621)

[CHƯƠNG 1 - GIỚI THIỆU 5](#_Toc154246622)

[1.1 Yêu cầu. 5](#_Toc154246623)

[1.2 Đề bài. 5](#_Toc154246624)

[1.3 Mục tiêu. 8](#_Toc154246625)

[CHƯƠNG 2 - PHÂN TÍCH VẤN ĐỀ 10](#_Toc154246626)

[2.1 Phân tích thống kê trên dữ liệu, vẽ các đồ thị và tìm hiểu các đặc trưng. 10](#_Toc154246627)

[2.1.1 Phân tích thống kê trên dữ liệu. 10](#_Toc154246628)

[2.1.2 Vẽ các biểu đồ. 18](#_Toc154246629)

[2.1.3 Xác định vai trò của từng đặc trưng đối với mục tiêu. 33](#_Toc154246630)

[2.2 Ứng dụng các mô hình học máy cơ bản để giải quyết bài toán, bao gồm cả các mô hình thuộc Ensemble Learing. 36](#_Toc154246631)

[2.3 Giải quyết bài toán bằng cách sử dụng Feed Forward Neural Network và Recurrent Neural Network. 43](#_Toc154246632)

[2.4 Áp dụng các kỹ thuật tránh Overfiting. 48](#_Toc154246633)

[2.5 Phân tích các trường hợp sai, đề ra giải pháp và thực hiện giải pháp: 55](#_Toc154246634)

[TÀI LIỆU THAM KHẢO 59](#_Toc154246635)

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

## DANH MỤC HÌNH

[Hình 1 – Đọc dữ liệu từ dataset 10](#_Toc154246636)

[Hình 2 – Dùng phương thức describe() và kết quả sau khi in ra 11](#_Toc154246637)

[Hình 3 – Lấy các Numerical Features 11](#_Toc154246638)

[Hình 4 – Tính mean của các Numerical Features 12](#_Toc154246639)

[Hình 5 – Kết quả tính mean của các Numerical Features 12](#_Toc154246640)

[Hình 6 – Tính median của các Numerical Features 12](#_Toc154246641)

[Hình 7 – Kết quả tính median của các Numerical Features 13](#_Toc154246642)

[Hình 8 – Tính giá trị nhỏ nhất của các Numerical Features 13](#_Toc154246643)

[Hình 9 – Kết quả tính giá trị nhỏ nhất của các Numerical Features 13](#_Toc154246644)

[Hình 10 – Tính giá trị lớn nhất của các Numerical Features 14](#_Toc154246645)

[Hình 11 – Kết quả tính giá trị lớn nhất của các Numerical Features 14](#_Toc154246646)

[Hình 12 – Tính phân phối của các Numerical Features 14](#_Toc154246647)

[Hình 13 – Kết quả tính phân phối của các Numerical Features 15](#_Toc154246648)

[Hình 14 – Code biểu đồ histogram của các numerical features 18](#_Toc154246649)

[Hình 15 – Đồ thị histogram của h1\_spo2\_min 19](#_Toc154246650)

[Hình 16 – Đồ thị histogram của h1\_sysbp\_max 20](#_Toc154246651)

[Hình 17 – Đồ thị histogram của h1\_sysbp\_min 21](#_Toc154246652)

[Hình 18 – Đồ thị histogram của h1\_sysbp\_noninvasive\_max 22](#_Toc154246653)

[Hình 19 – Đồ thị histogram của h1\_sysbp\_noninvasive\_min 23](#_Toc154246654)

[Hình 20 – Đồ thị histogram của d1\_glucose\_max 24](#_Toc154246655)

[Hình 21 – Đồ thị histogram của d1\_glucose\_min 25](#_Toc154246656)

[Hình 22 – Đồ thị histogram của d1\_potassium\_max 26](#_Toc154246657)

[Hình 23 – Đồ thị histogram của d1\_potassium\_min 27](#_Toc154246658)

[Hình 24 – Đồ thị histogram của apache\_4a\_hospital\_death\_prob 28](#_Toc154246659)

[Hình 25 – Đồ thị histogram của apache\_4a\_icu\_death\_prob 29](#_Toc154246660)

[Hình 26 – Code vẽ biểu đồ box plot cho các Numerical Features 30](#_Toc154246661)

[Hình 27 – Kết quả biểu đồ box plot của các numerical features 30](#_Toc154246662)

[Hình 28 – Vẽ đồ thị scatter plot cho cặp numerical features trên 32](#_Toc154246663)

[Hình 29 – Đồ thị scatter plot 32](#_Toc154246664)

[Hình 30 – Code tính ma trận tương quang và vẽ biểu đồ heatmap 34](#_Toc154246665)

[Hình 31 – Biểu đồ Correlation Heatmap 35](#_Toc154246666)

[Hình 32 – Code huấn luyện các mô hình học máy cơ bản 37](#_Toc154246667)

[Hình 33 – Kết quả ứng dụng mô hình Logistic Regression 38](#_Toc154246668)

[Hình 34 – Kết quả ứng dụng mô hình Random Forest 39](#_Toc154246669)

[Hình 35 – Kết quả ứng dụng SVM 41](#_Toc154246670)

[ Hình 36 – Kết quả ứng dụng mô hình GradientBoostingClassifier 42](#_Toc154246671)

[Hình 37 – Huấn luyện mô hình Feed Forward Neural Network 44](#_Toc154246672)

[Hình 38 – Kết quả huấn luyện mô hình Feed Forward Neural Network 44](#_Toc154246673)

[Hình 39 – Huấn luyện mô hình Recurrent Neural Network 45](#_Toc154246674)

[Hình 40 – Kết quả huấn luyện mô hình Recurrent Neural Network 46](#_Toc154246675)

[Hình 41 – Đánh giá và dự đoánmô hình Feed Forward Neural Network và Recurrent Neural Network 47](#_Toc154246676)

[Hình 42 – Kết quả đánh giá và dự đoán mô hình Feed Forward Neural Network và Recurrent Neural Network 47](#_Toc154246677)

[Hình 43 – Dự đoán và đánh giá hiệu suất trên tập validation 49](#_Toc154246678)

[Hình 44 – Kết quả dự đoán trên tập validation 50](#_Toc154246679)

[Hình 45 – Áp dụng kỹ thuật Early Stopping 51](#_Toc154246680)

[Hình 46 – Kết quả của Early Stopping 52](#_Toc154246681)

[Hình 47 – Áp dụng kỹ thuật dropout 53](#_Toc154246682)

[Hình 48 – Kết quả của dropout 54](#_Toc154246683)

[Hình 49 – Thay đổi siêu tham số trong mô hình Neural Network 56](#_Toc154246684)

[Hình 50 – Kết quả sau khi thay đổi siêu tham số 57](#_Toc154246685)

## DANH MỤC BẢNG

**No table of figures entries found.**

1. GIỚI THIỆU
2. Yêu cầu.

Đưa ra một bài toán dự đoán có thể giải quyết bằng học máy (machine leanring) với các yêu cầu sau:

* Số Feature/Attribute gồm nhiều kiểu: categorial và numerical;
* Dữ liệu phải chưa được học, thực tập trên lớp và trong bài tập về nhà;

1. Phân tích thống kê trên dữ liệu, vẽ các đồ thị để hiểu bài toán, hiểu dữ liệu. Tìm hiểu các đặc trưng và đánh gía vai trò của các đặc trưng đối với mục tiêu bài toán;
2. Ứng dụng các mô hình học máy cơ bản để giải quyết bài toán, bao gồm cả các mô hình thuộc Ensemble Learing;
3. Sử dụng Feed Forward Neural Network và Reccurent Neural Network (hoặc mô thuộc loại này) để giải quyết bài toán;
4. Áp dụng các kỹ thuật tránh Overfiting trên các mô hình của câu (2) và câu (3) để giải quyết bài toán;
5. Sau khi huấn luyện xong mô hình thì muốn cải thiện độ chính xác, ta sẽ làm gì để giải quyết nó? Phân tích các trường hợp sai, đề ra giải pháp và thực hiện nó, sau đó đánh giá xem có cải tiến so với trước không.
6. Đề bài.

**Bài toán**: Dự đoán Khả năng Sống Sót của Bệnh Nhân

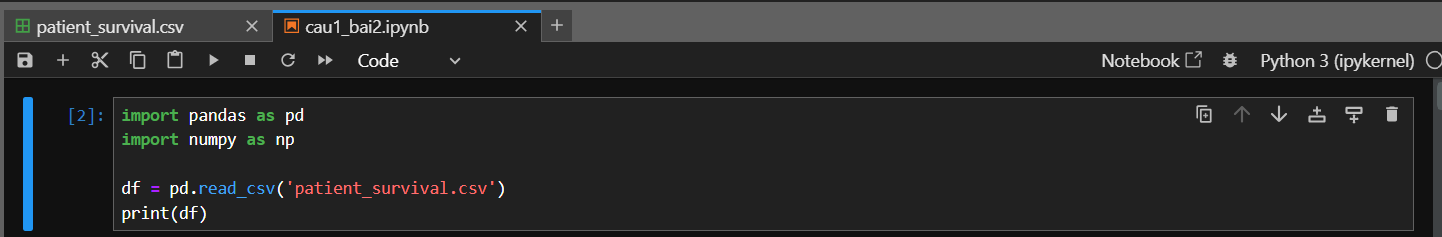
* Dựa trên dataset “patient\_survival.csv”[[1]](#footnote-1). Bài toán yêu cầu xây dựng một mô hình dự đoán khả năng sống sót của bệnh nhân dựa trên các đặc trưng (features) sau:
* **Numerical Features:**
* encounter\_id: ID duy nhất cho mỗi sự kiện gặp bệnh nhân.
* patient\_id: ID duy nhất cho mỗi bệnh nhân.
* hospital\_id: ID duy nhất cho mỗi bệnh viện.
* age: Tuổi của bệnh nhân.
* bmi: Chỉ số khối cơ thể (Body Mass Index) của bệnh nhân.
* elective\_surgery: Chỉ định phẫu thuật không khẩn cấp (0 = Không, 1 = Có).
* height: Chiều cao của bệnh nhân.
* icu\_admit\_source: Nguồn nhập viện vào ICU.
* icu\_id: ID duy nhất cho mỗi ICU.
* icu\_stay\_type: Loại lưu trú trong ICU.
* icu\_type: Loại ICU.
* pre\_icu\_los\_days: Số ngày bệnh nhân đã ở ngoại vi trước khi vào ICU.
* weight: Cân nặng của bệnh nhân.
* apache\_2\_diagnosis: Mã chẩn đoán Apache II.
* apache\_3j\_diagnosis: Mã chẩn đoán Apache III-J.
* apache\_post\_operative: Chỉ định phẫu thuật sau phẫu thuật (0 = Không, 1 = Có).
* arf\_apache: Mã hóa ARF Apache.
* gcs\_eyes\_apache: Điểm mắt trong hệ thống GCS Apache.
* gcs\_motor\_apache: Điểm cơ trong hệ thống GCS Apache.
* gcs\_unable\_apache: Không thể xác định điểm GCS Apache (0 = Không, 1 = Có).
* gcs\_verbal\_apache: Điểm nói chuyện trong hệ thống GCS Apache.
* heart\_rate\_apache: Nhịp tim của bệnh nhân khi nhập viện ICU.
* intubated\_apache: Chỉ định đã truyền ống thông gió cho bệnh nhân (0 = Không, 1 = Có).
* map\_apache: Áp lực động mạch trung bình (MAP) của bệnh nhân khi nhập viện ICU.
* resprate\_apache: Tốc độ hô hấp của bệnh nhân khi nhập viện ICU.
* temp\_apache: Nhiệt độ của bệnh nhân khi nhập viện ICU.
* ventilated\_apache: Chỉ định đã sử dụng máy thở cho bệnh nhân (0 = Không, 1 = Có).
* h1\_spo2\_max: Mức tối đa của lượng oxy trong máu của bệnh nhân trong giờ đầu
* h1\_spo2\_min: Mức tối thiểu của lượng oxy trong máu của bệnh nhân trong giờ đầu.
* h1\_sysbp\_max: Áp lực máu tối đa của bệnh nhân trong giờ đầu.
* h1\_sysbp\_min: Áp lực máu tối thiểu của bệnh nhân trong giờ đầu.
* h1\_sysbp\_noninvasive\_max: Áp lực máu không xâm lấn tối đa trong giờ đầu.
* h1\_sysbp\_noninvasive\_min: Áp lực máu không xâm lấn tối thiểu trong giờ đầu.
* d1\_glucose\_max: Đường huyết tối đa trong ngày đầu tiên.
* d1\_glucose\_min: Đường huyết tối thiểu trong ngày đầu tiên.
* d1\_potassium\_max: Mức kali tối đa trong ngày đầu tiên.
* d1\_potassium\_min: Mức kali tối thiểu trong ngày đầu tiên.
* apache\_4a\_hospital\_death\_prob: Xác suất tử vong trong bệnh viện dựa trên mô hình APACHE IVa.
* apache\_4a\_icu\_death\_prob: Xác suất tử vong trong ICU dựa trên mô hình APACHE IVa.
* **Categorical Features:**
* aids: Chỉ định bệnh nhân mắc bệnh AIDS (0 = Không, 1 = Có).
* cirrhosis: Chỉ định bệnh nhân mắc bệnh xơ gan (0 = Không, 1 = Có).
* diabetes\_mellitus: Chỉ định bệnh nhân mắc bệnh tiểu đường (0 = Không, 1 = Có).
* hepatic\_failure: Chỉ định bệnh nhân mắc suy gan (0 = Không, 1 = Có).
* immunosuppression: Chỉ định bệnh nhân mắc suy giảm miễn dịch (0 = Không, 1 = Có).
* leukemia: Chỉ định bệnh nhân mắc bệnh bạch cầu (0 = Không, 1 = Có).
* lymphoma: Chỉ định bệnh nhân mắc bệnh lymphoma (0 = Không, 1 = Có).
* solid\_tumor\_with\_metastasis: Chỉ định bệnh nhân mắc bệnh u ác tính với di căn (0 = Không, 1 = Có).
* apache\_3j\_bodysystem: Hệ thống cơ thể của bệnh nhân dựa trên mã số APACHE III-J.
* apache\_2\_bodysystem: Hệ thống cơ thể của bệnh nhân dựa trên mã số APACHE II.
* **Target Variable:**
  + hospital\_death: Biến nhị phân (0 hoặc 1) mô tả liệu bệnh nhân có sống sót hay không trong bệnh viện.

1. Mục tiêu.

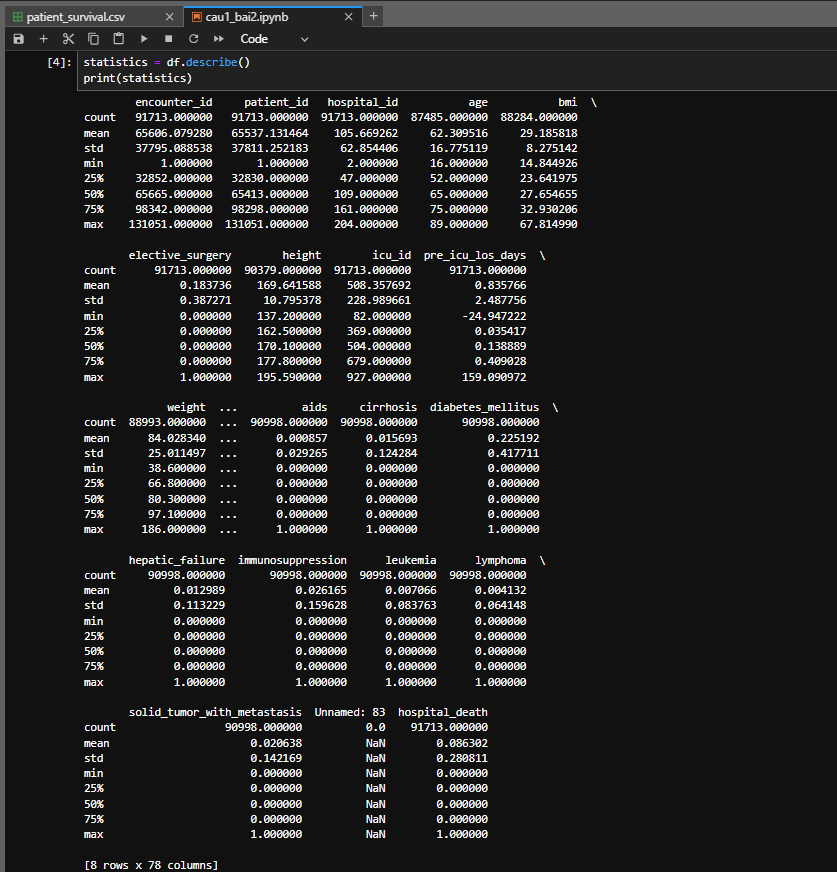
* Nhằm  xây dựng và huấn luyện một mô hình học máy hiệu quả để giải quyết bài toán dự đoán, đồng thời cải thiện độ chính xác của mô hình sau khi huấn luyện.
* Thực hiện phân tích thống kê và vẽ các đồ thị để hiểu mục tiêu bài toán và tính chất của dữ liệu. Đánh giá vai trò của các đặc trưng đối với mục tiêu bài toán.
* Áp dụng được các mô hình học máy cơ bản và mô hình thuộc Ensemble Learning nhằm giải quyết bài toán dự đoán, đồng thời đánh giá và so sánh hiệu suất của các mô hình để chọn ra mô hình tốt nhất.
* Sử dụng tốt các mô hình nơ-ron truyền thẳng và mạng nơ-ron tuần hoàn để giải quyết bài toán dự đoán, đồng thời đánh giá và so sánh hiệu suất của các mô hình nơ-ron để chọn ra mô hình tốt nhất.
* Áp dụng được các kỹ thuật tránh overfitting như chia dữ liệu thành tập huấn luyện và tập kiểm tra, sử dụng kỹ thuật chọn mô hình tối ưu, kiểm soát thừa số hóa, thay đổi độ phức tạp mô hình, sử dụng kỹ thuật Dropout, và early stopping để đảm bảo mô hình không bị overfitting.
* Phân tích các trường hợp sai của mô hình sau khi huấn luyện xong và đề ra được các giải pháp cải thiện. Thực hiện các giải pháp, đánh giá xem có cải thiện độ chính xác so với trước không và chọn ra giải pháp tốt nhất.

1. PHÂN TÍCH VẤN ĐỀ
2. Phân tích thống kê trên dữ liệu, vẽ các đồ thị và tìm hiểu các đặc trưng.
3. Phân tích thống kê trên dữ liệu.

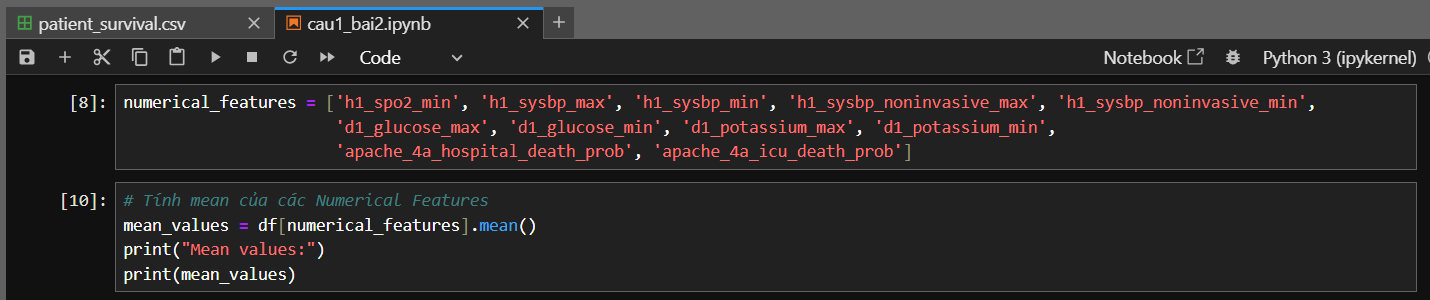
* Để phân tích thống kê trên tập dữ liệu “patient\_survival.csv” ta cần tính toán các giá trị mean, median, min, max và phân phối của các đặc trưng.
* Phương thức để thực hiện tính toán ở đây là *describe()* để tính toán các giá trị thống kê mô tả như count, mean, std, min, max và các phần vị của các đặc trưng trong DataFrame.
* Ở đây ta sử dụng **Numerical Features** để tính giá trị trung bình (mean) của các đặc trưng bằng phương thức 'mean()', tính giá trị trung vị (median) bằng phương thức 'median()', tìm giá trị tối thiểu (min) bằng phương thức 'min()', và tìm giá trị tối đa (max) bằng phương thức 'max()'.
* Cuối cùng, chúng ta tính phân phối của các đặc trưng bằng cách sử dụng phương thức 'value\_counts()' trên từng cột của DataFrame. Kết quả của phân phối được lưu trong DataFrame 'distributions'.



Hình – Đọc dữ liệu từ dataset

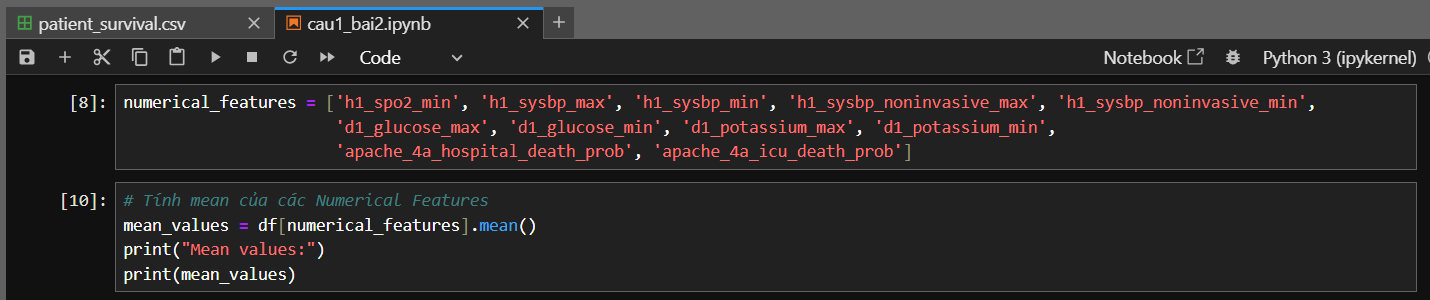


Hình – Dùng phương thức describe() và kết quả sau khi in ra



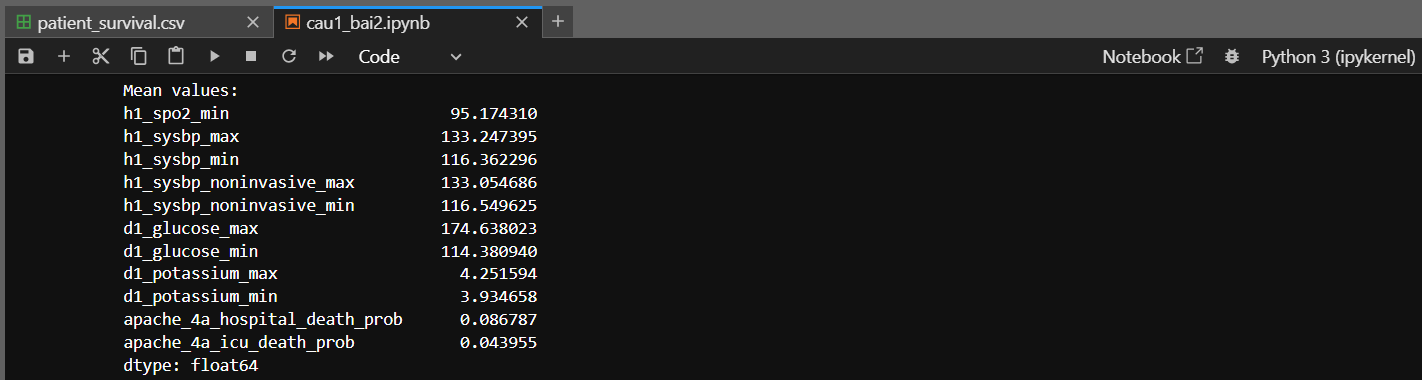
Hình – Lấy các Numerical Features

* Code tính giá trị trung bình (mean) của các Numerical Features:



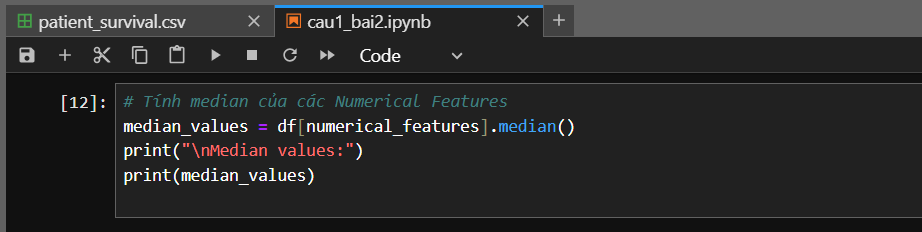
Hình – Tính mean của các Numerical Features

* Kết quả tính mean của các Numerical Features:



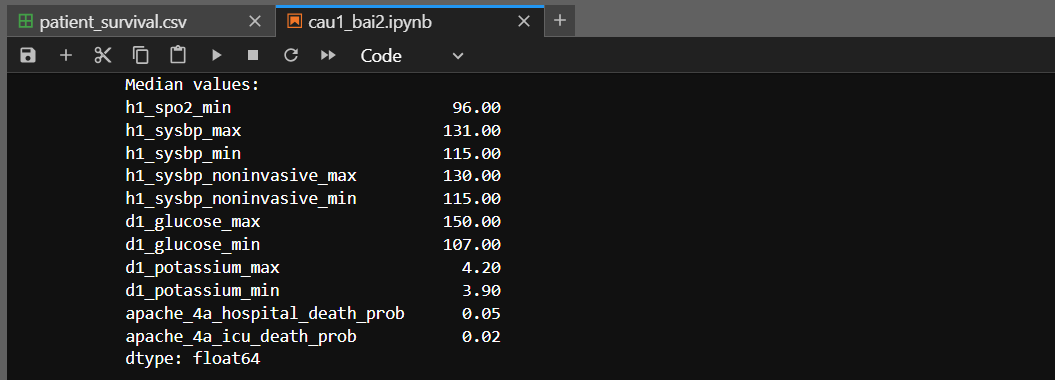
Hình – Kết quả tính mean của các Numerical Features

* Code tính giá trị trung vị (median) của các Numerical Features:



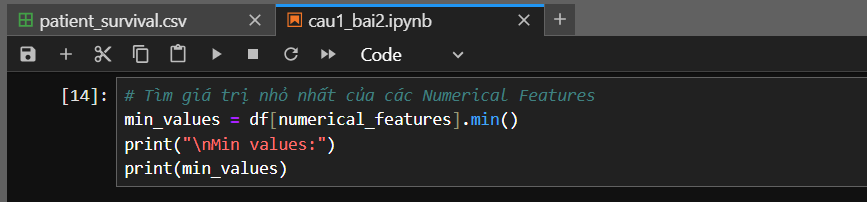
Hình – Tính median của các Numerical Features

* Kết quả tính median của các Numerical Features:



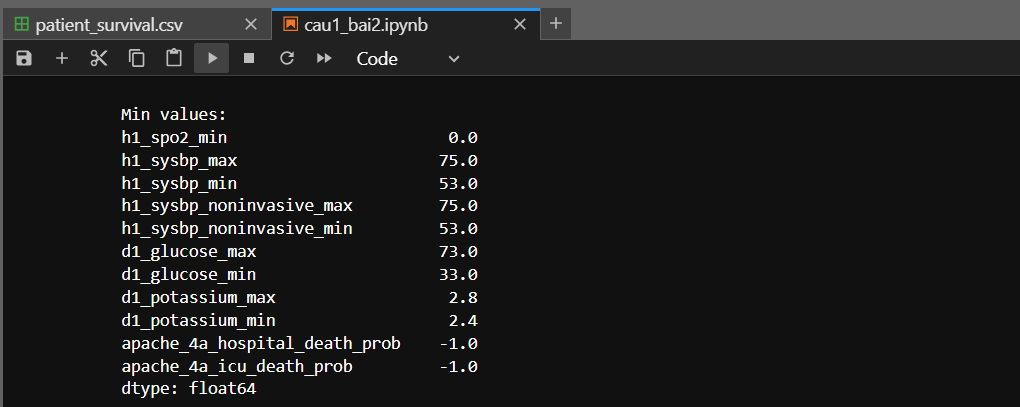
Hình – Kết quả tính median của các Numerical Features

* Tính giá trị nhỏ nhất của các Numerical Features:



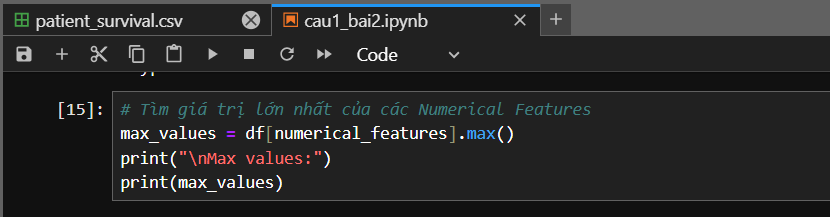
Hình – Tính giá trị nhỏ nhất của các Numerical Features

* Kết quả tính giá trị nhỏ nhất của các Numerical Features:



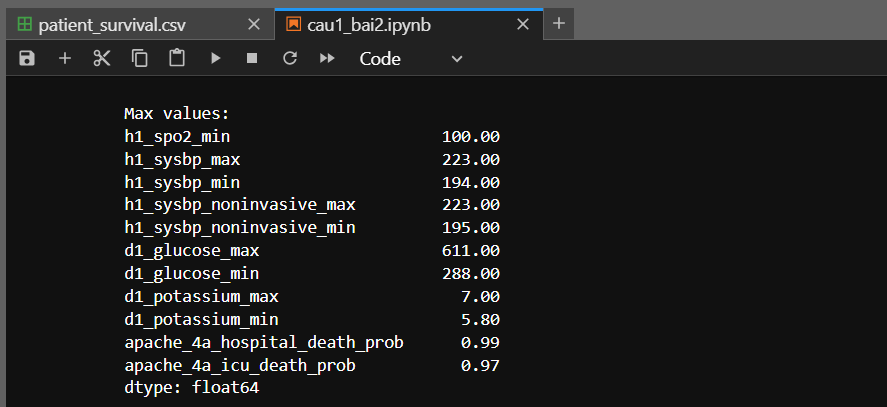
Hình – Kết quả tính giá trị nhỏ nhất của các Numerical Features

* Tính giá trị lớn nhất của các Numerical Features:



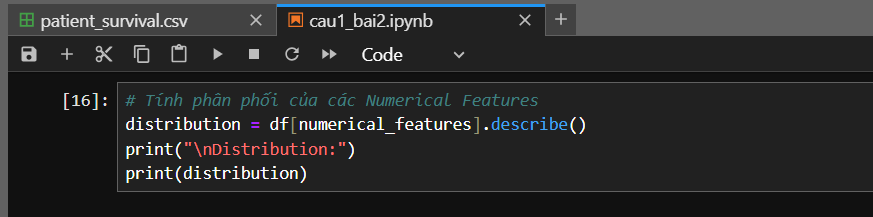
Hình – Tính giá trị lớn nhất của các Numerical Features

* Kết quả tính giá trị lớn nhất của các Numerical Features:



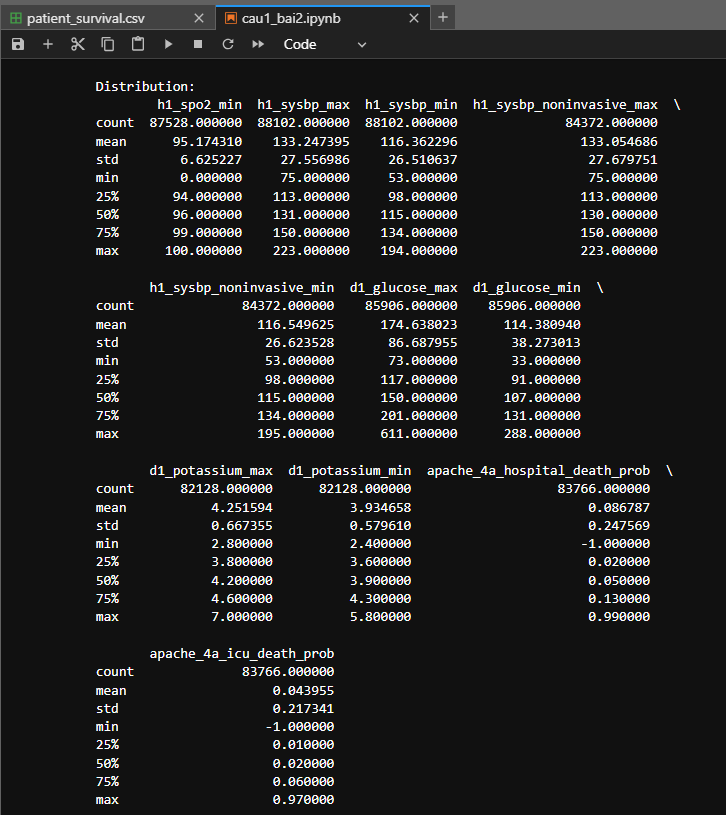
Hình – Kết quả tính giá trị lớn nhất của các Numerical Features

* Code tính phân phối của các Numerical Features:



Hình – Tính phân phối của các Numerical Features

* Kết quả tính phân phối của các Numerical Features:

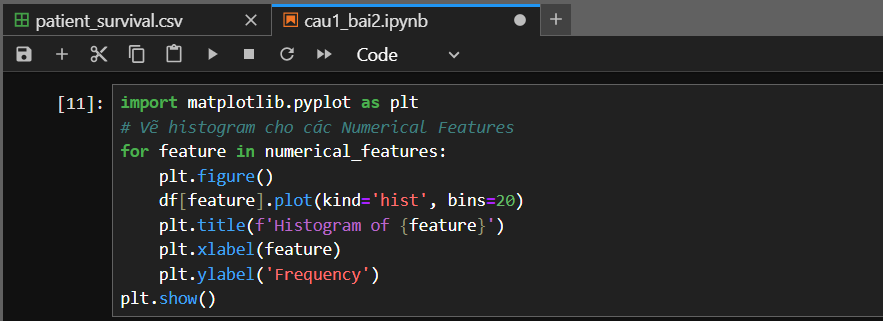


Hình – Kết quả tính phân phối của các Numerical Features

* Dựa vào các giá trị trung bình (mean), trung vị (median), giá trị nhỏ nhất (min) và giá trị lớn nhất (max) của các numerical features đã cung cấp, ta thấy:
* Các giá trị h1\_spo2\_min, h1\_sysbp\_max, h1\_sysbp\_min, h1\_sysbp\_noninvasive\_max và h1\_sysbp\_noninvasive\_min có phạm vi giữa giá trị nhỏ nhất và giá trị lớn nhất khá rộng. Điều này cho thấy sự biến động lớn trong dữ liệu của các features này.
* Đối với feature d1\_glucose\_max và d1\_glucose\_min, giá trị trung bình và trung vị khá gần nhau, trong khi giá trị lớn nhất và nhỏ nhất có sự chênh lệch lớn. Điều này có thể cho thấy có một số điểm dữ liệu ngoại lai (outliers) gây ảnh hưởng đến giá trị lớn nhất và nhỏ nhất.
* Feature d1\_potassium\_max và d1\_potassium\_min có phạm vi giữa giá trị nhỏ nhất và giá trị lớn nhất không quá rộng, và giá trị trung bình và trung vị khá gần nhau. Điều này cho thấy dữ liệu trong các features này không có sự biến động lớn.
* Apache\_4a\_hospital\_death\_prob và apache\_4a\_icu\_death\_prob là hai features có giá trị nhỏ nhất là -1.0, đây có thể là một biểu diễn cho giá trị thiếu hoặc không xác định trong dữ liệu. Ngoài ra, giá trị lớn nhất của hai features này là 0.99 và 0.97, cho thấy sự biến động từ gần 0 đến gần 1.
* Tóm lại, các features có sự biến động lớn trong dữ liệu như h1\_spo2\_min, h1\_sysbp\_max, h1\_sysbp\_min, h1\_sysbp\_noninvasive\_max, d1\_glucose\_max, và d1\_glucose\_min có thể cần được kiểm tra kỹ hơn để loại bỏ các điểm dữ liệu ngoại lai và xử lý các giá trị thiếu hoặc không xác định. Trong khi đó, các features khác có phạm vi giá trị hẹp hơn và không có sự biến động lớn, có thể ít cần phải xử lý hơn.
* Dựa vào thông tin về phân phối của các numerical features, ta thấy được rằng:
* Đối với các features liên quan đến huyết áp (h1\_sysbp\_max, h1\_sysbp\_min, h1\_sysbp\_noninvasive\_max, h1\_sysbp\_noninvasive\_min), phân phối có vẻ gần với phân phối chuẩn, với giá trị trung bình (mean) và giá trị trung vị (median) không quá chênh lệch nhau. Đồng thời, độ lệch chuẩn (standard deviation) cũng không quá lớn, cho thấy dữ liệu phân bố tương đối đều quanh các giá trị trung bình.
* Feature h1\_spo2\_min có phân phối gần với phân phối chuẩn, nhưng có độ lệch chuẩn (standard deviation) nhỏ hơn so với các features khác. Điều này có nghĩa là dữ liệu có sự tập trung chặt chẽ hơn quanh giá trị trung bình.
* Đối với các features liên quan đến mức đường huyết (d1\_glucose\_max, d1\_glucose\_min), phân phối có xu hướng lệch phải (right-skewed), tức là có sự tập trung chủ yếu ở các giá trị thấp hơn giá trị trung vị. Điều này được thể hiện qua giá trị trung vị (median) lớn hơn giá trị trung bình (mean), và độ lệch chuẩn (standard deviation) lớn hơn.
* Feature d1\_potassium\_max và d1\_potassium\_min có phân phối gần với phân phối chuẩn, với độ lệch chuẩn (standard deviation) không quá lớn. Điều này cho thấy dữ liệu phân bố tương đối đều quanh giá trị trung bình.
* Apache\_4a\_hospital\_death\_prob và apache\_4a\_icu\_death\_prob có phân phối gần với phân phối chuẩn, nhưng có giá trị nhỏ nhất và lớn nhất là -1 và 0.99, tương ứng. Điều này cho thấy có một số điểm dữ liệu ngoại lai hoặc giá trị thiếu hoặc không xác định trong dữ liệu.
* Tóm lại, các numerical features có phân phối gần với phân phối chuẩn và có độ lệch chuẩn không quá lớn cho thấy dữ liệu có sự tập trung tương đối đều quanh giá trị trung bình. Tuy nhiên, cần xử lý các giá trị thiếu hoặc không xác định (-1) và kiểm tra các điểm dữ liệu ngoại lai (outliers) để đảm bảo tính chính xác của phân tích dữ liệu.

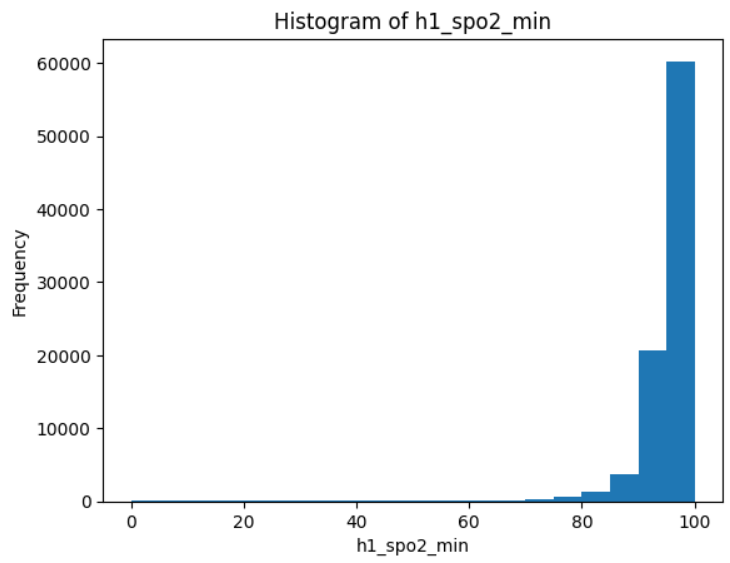
1. Vẽ các biểu đồ.

* **Vẽ đồ thị histogram:**
* Dựa vào dữ liệu của các numerical features đã lấy như ở hình 3 để vẽ biểu đồ histogram:



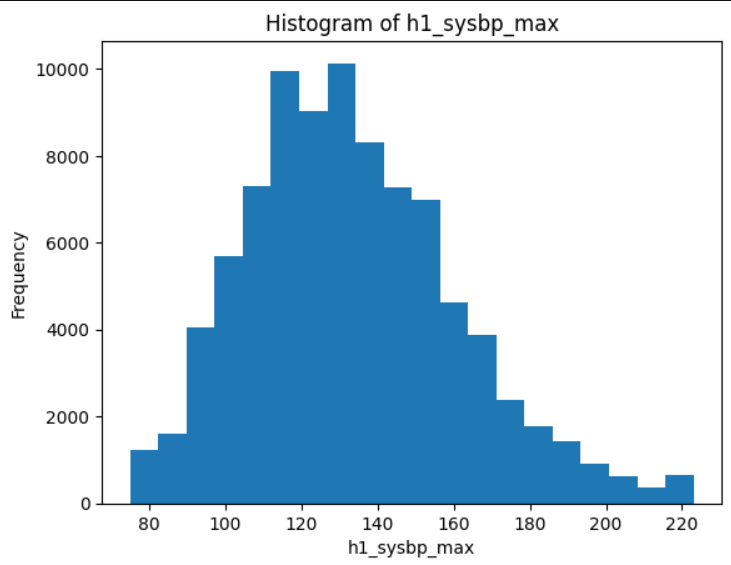
Hình – Code biểu đồ histogram của các numerical features

* Kết quả đồ thị histogram cho từng features:
* Đồ thị histogram hiển thị mức tối thiểu của lượng oxy trong máu của bệnh nhân trong giờ đầu:



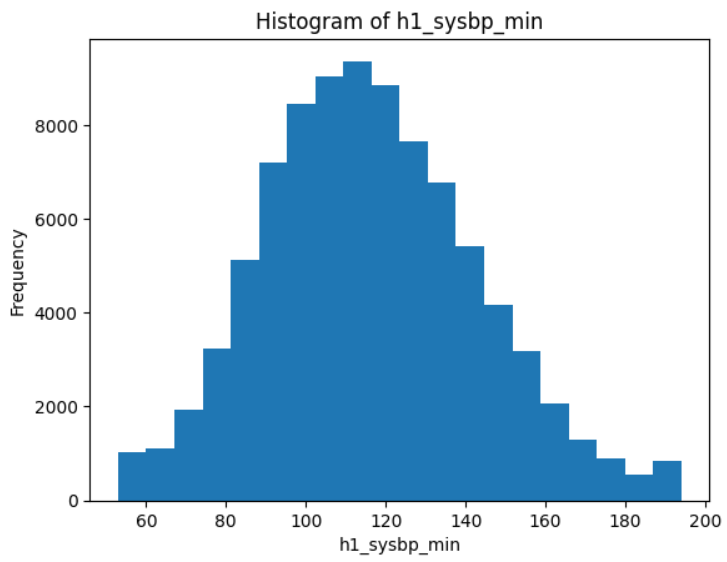
Hình – Đồ thị histogram của h1\_spo2\_min

* Đồ thị histogram hiển thị áp lực máu tối đa của bệnh nhân trong giờ đầu:



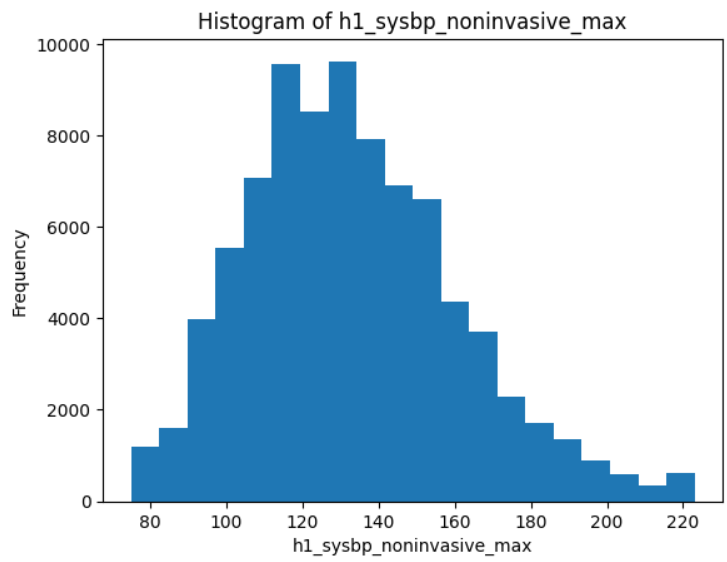
Hình – Đồ thị histogram của h1\_sysbp\_max

* Đồ thị histogram hiển thị áp lực máu tối thiểu của bệnh nhân trong giờ đầu:



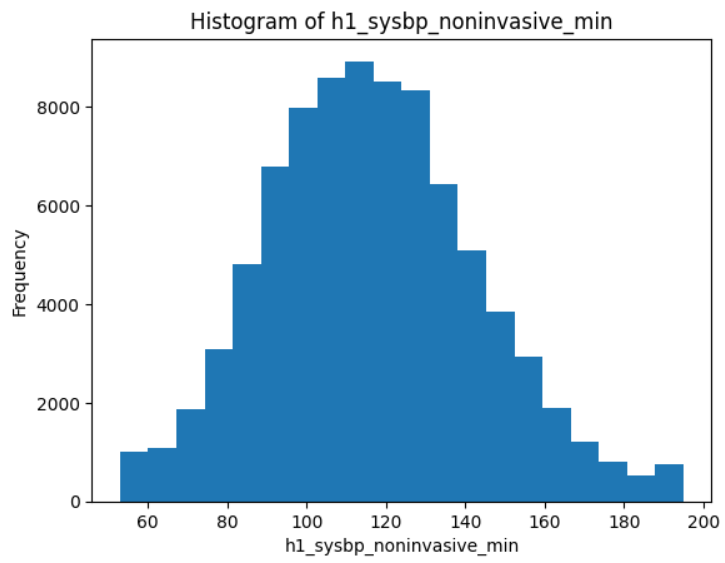
Hình – Đồ thị histogram của h1\_sysbp\_min

* Đồ thị histogram hiển thị áp lực máu không xâm lấn tối đa trong giờ đầu.



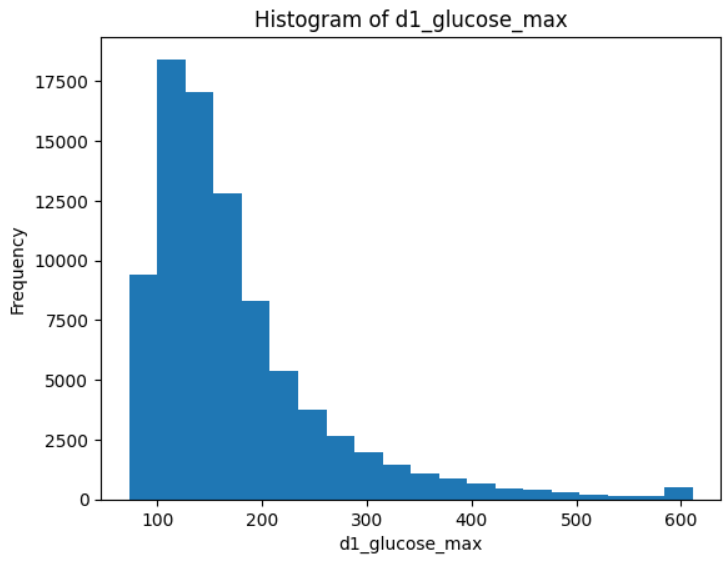
Hình – Đồ thị histogram của h1\_sysbp\_noninvasive\_max

* Đồ thị histogram hiển thị áp lực máu không xâm lấn tối thiểu trong giờ đầu:



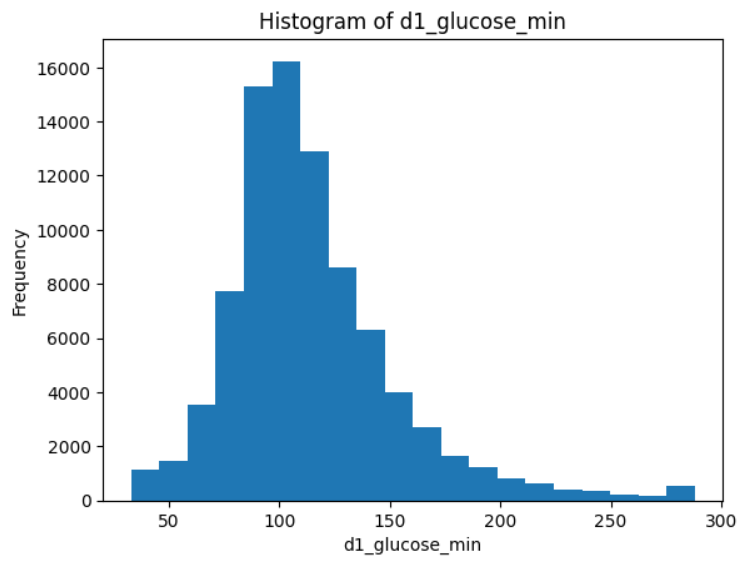
Hình – Đồ thị histogram của h1\_sysbp\_noninvasive\_min

* Đồ thị histogram hiển thị đường huyết tối đa trong ngày đầu tiên:



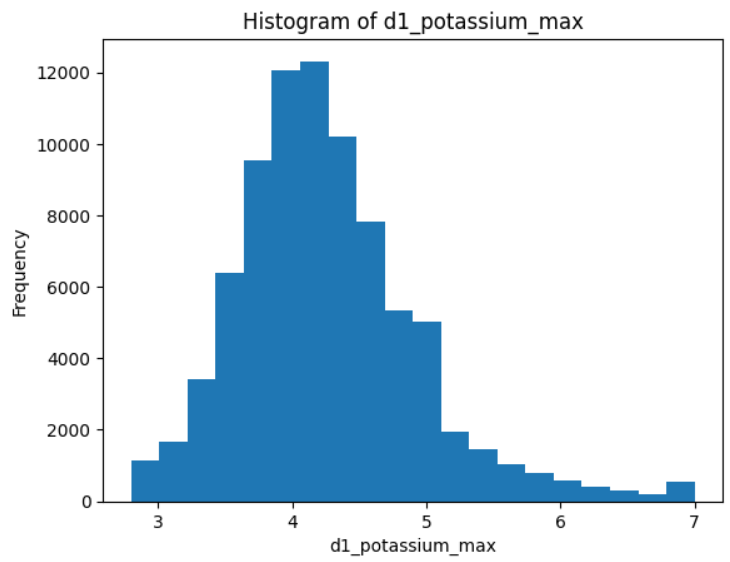
Hình – Đồ thị histogram của d1\_glucose\_max

* Đồ thị histogram hiển thị đường huyết tối thiểu trong ngày đầu tiên:



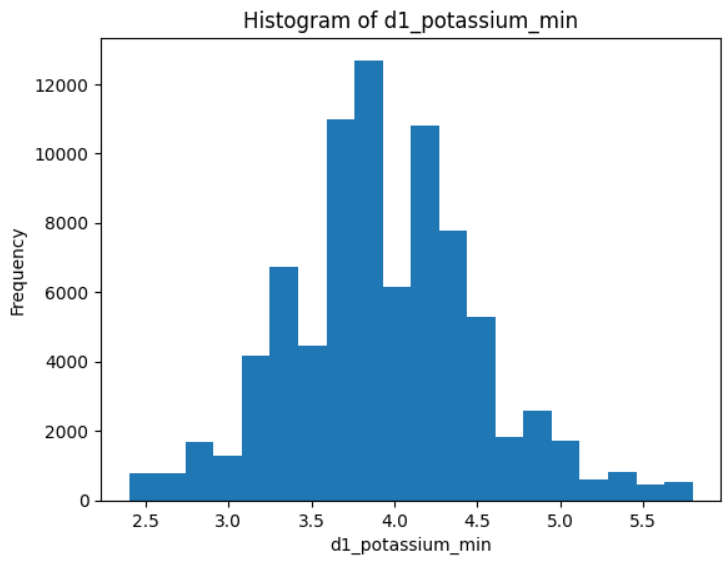
Hình – Đồ thị histogram của d1\_glucose\_min

* Đồ thị histogram hiển thị mức kali tối đa trong ngày đầu tiên:



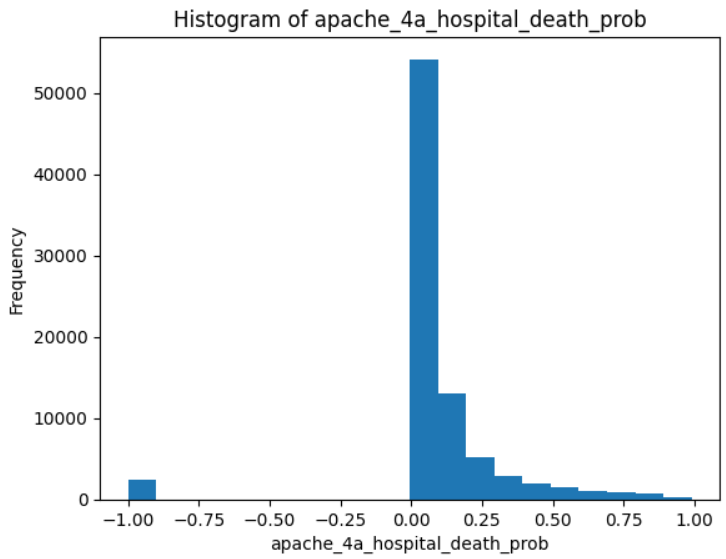
Hình – Đồ thị histogram của d1\_potassium\_max

* Đồ thị histogram hiển thị mức kali tối thiểu trong ngày đầu tiên:



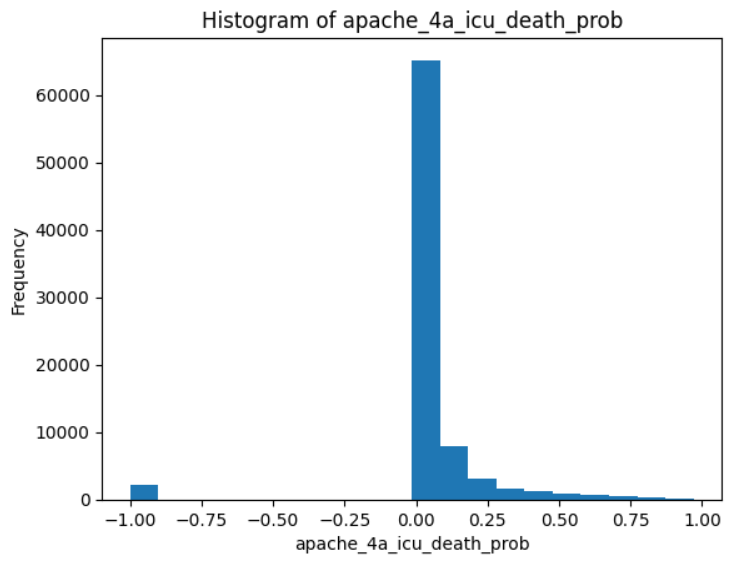
Hình – Đồ thị histogram của d1\_potassium\_min

* Đồ thị histogram hiển thị xác suất tử vong trong bệnh viện dựa trên mô hình APACHE Iva:



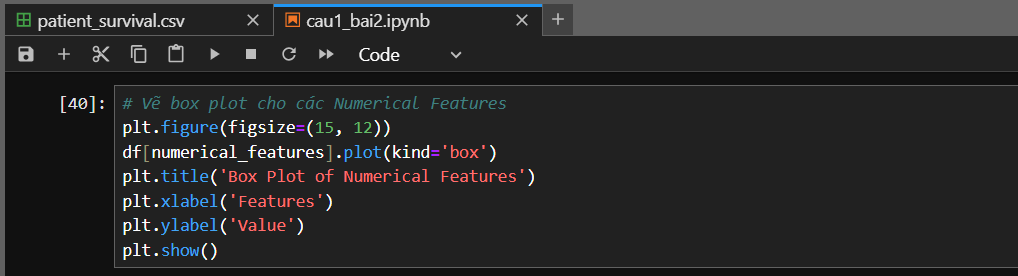
Hình – Đồ thị histogram của apache\_4a\_hospital\_death\_prob

* Đồ thị histogram hiển thị đồ thị histogram hiển thị đường huyết tối đa trong ngày đầu tiên:



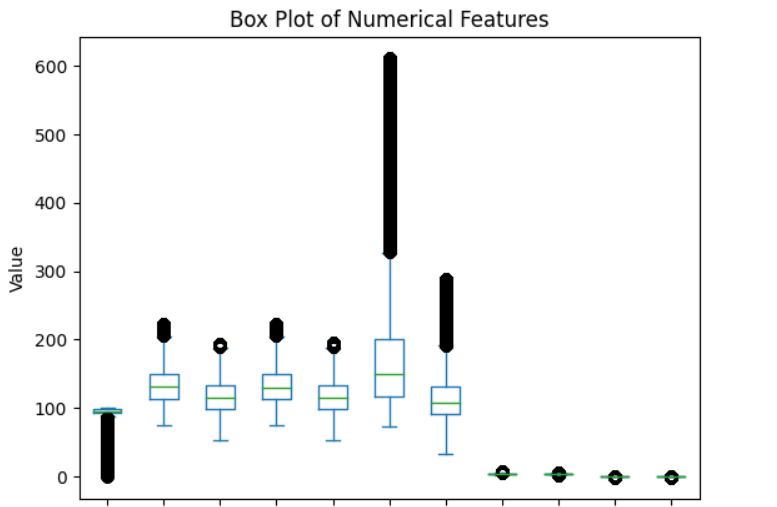
Hình – Đồ thị histogram của apache\_4a\_icu\_death\_prob

* **Vẽ đồ thị box plot:**
* Biểu đồ box plot cho phép chúng ta nhìn thấy phân bố, tập trung và sự biến động của các đặc trưng số trong tập dữ liệu. Các giá trị ngoại lai có thể là những giá trị đặc biệt hoặc sai sót và cần được xem xét kỹ hơn.
* Ta sử dụng lại các Features đã lấy ở hình 3 để vẽ biểu đồ minh họa:



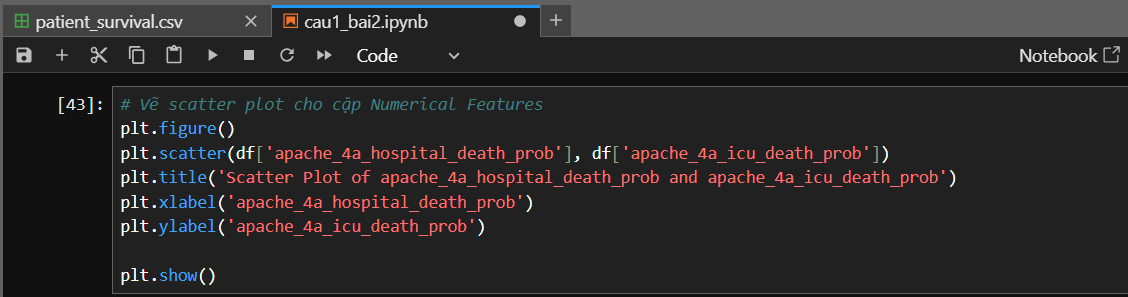
Hình – Code vẽ biểu đồ box plot cho các Numerical Features

* Kết quả in ra biểu đồ box plot:



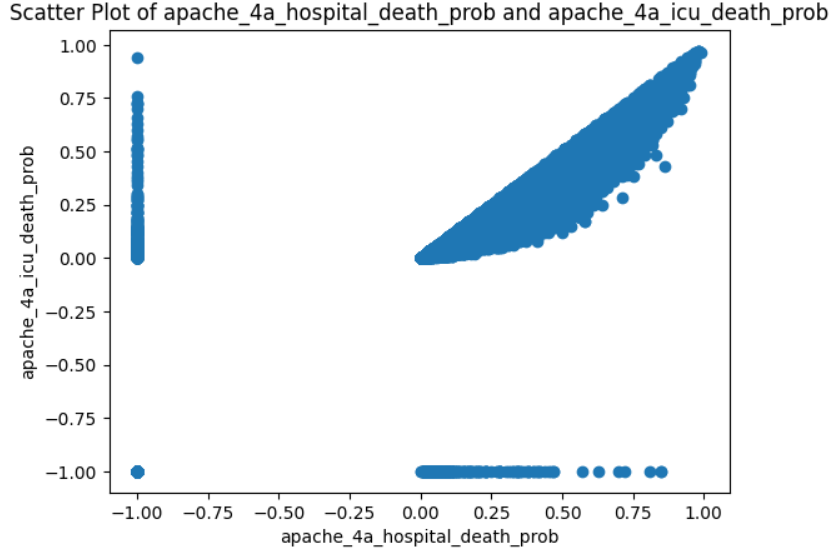
Hình – Kết quả biểu đồ box plot của các numerical features

* Sau khi vẽ biểu đồ box plot cho các numerical features được chỉ định, chúng ta thấy:
* 'h1\_spo2\_min': Biểu đồ cho thấy phân bố giá trị tối thiểu của lượng oxy trong máu của bệnh nhân trong giờ đầu. Có một số giá trị ngoại lai nằm ngoài khoảng giá trị thông thường.
* 'h1\_sysbp\_max' và 'h1\_sysbp\_min': Biểu đồ cho thấy phân bố giá trị tối đa và tối thiểu của áp lực máu của bệnh nhân trong giờ đầu. Có một số giá trị ngoại lai ở cả hai biểu đồ.
* 'h1\_sysbp\_noninvasive\_max' và 'h1\_sysbp\_noninvasive\_min': Biểu đồ cho thấy phân bố giá trị tối đa và tối thiểu của áp lực máu không xâm lấn của bệnh nhân trong giờ đầu. Có một số giá trị ngoại lai ở cả hai biểu đồ.
* 'd1\_glucose\_max' và 'd1\_glucose\_min': Biểu đồ cho thấy phân bố giá trị tối đa và tối thiểu của đường huyết của bệnh nhân trong ngày đầu tiên. Có một số giá trị ngoại lai ở cả hai biểu đồ.
* 'd1\_potassium\_max' và 'd1\_potassium\_min': Biểu đồ cho thấy phân bố giá trị tối đa và tối thiểu của mức kali của bệnh nhân trong ngày đầu tiên. Có một số giá trị ngoại lai ở cả hai biểu đồ.
* 'apache\_4a\_hospital\_death\_prob' và 'apache\_4a\_icu\_death\_prob': Biểu đồ cho thấy phân bố xác suất tử vong trong bệnh viện và xác suất tử vong trong ICU dựa trên mô hình APACHE IVa. Có một số giá trị ngoại lai trong cả hai biểu đồ.
* **Vẽ đồ thị scatter plot:**
* Để vẽ đồ thị scatter plot, ta sử dụng hàm plot() với loại biểu đồ 'scatter' để vẽ biểu đồ scatter plot. Chúng ta chỉ định cột 'apache\_4a\_hospital\_death\_prob' làm trục x và cột 'apache\_4a\_icu\_death\_prob' làm trục y trong biểu đồ scatter plot sau:



Hình – Vẽ đồ thị scatter plot cho cặp numerical features trên

* Kết quả đồ thị scatter plot giữa 2 cặp tương quan apache\_4a\_hospital\_death\_prob và apache\_4a\_icu\_death\_prob:

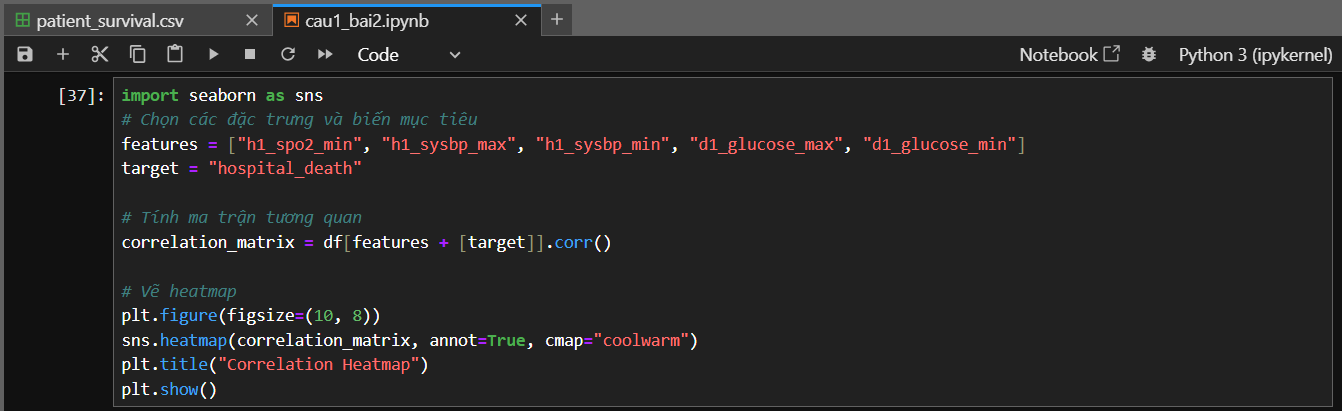


Hình – Đồ thị scatter plot

* Sau khi vẽ biểu đồ scatter plot cho 'apache\_4a\_hospital\_death\_prob' và 'apache\_4a\_icu\_death\_prob', chúng ta có thể thấy được rằng:
* Biểu đồ scatter plot cho phép chúng ta quan sát mối quan hệ giữa xác suất tử vong trong bệnh viện (Apache IVa Hospital Death Probability) và xác suất tử vong trong ICU (Apache IVa ICU Death Probability) của bệnh nhân.
* Dựa trên biểu đồ scatter plot, chúng ta có thể thấy có một sự tương quan dương giữa hai đặc trưng này. Điều này có nghĩa là khi giá trị xác suất tử vong trong bệnh viện tăng lên, giá trị xác suất tử vong trong ICU cũng có xu hướng tăng lên.
* Tuy nhiên, chúng ta cũng có thể thấy rằng có một phần dữ liệu phân tán rải rác và không theo một mô hình tuyến tính rõ ràng. Điều này có thể chỉ ra rằng mối quan hệ giữa hai đặc trưng có thể phức tạp hơn và có thể bị ảnh hưởng bởi các yếu tố khác.
* Ngoài ra, biểu đồ scatter plot cũng cho phép chúng ta nhìn thấy có tồn tại một số giá trị ngoại lai (outliers) trong dữ liệu. Những điểm ngoại lai này có thể là những điểm đặc biệt hoặc không tuân theo xu hướng chung của dữ liệu.

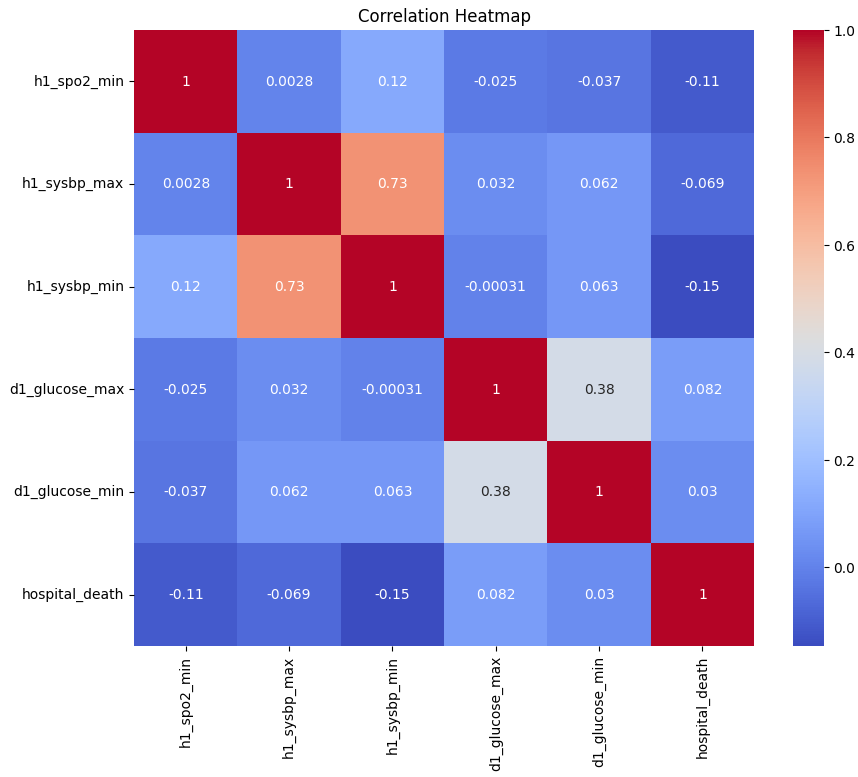
1. Xác định vai trò của từng đặc trưng đối với mục tiêu.

* Để xác định vai trò của từng đặc trưng đối với mục tiêu bài toán, bạn có thể tính toán hệ số tương quan (correlation coefficient) giữa các đặc trưng số (numerical features) và mục tiêu bài toán. Hệ số tương quan có thể giúp định lượng mức độ tương quan tuyến tính giữa hai biến. Sau đó vẽ heatmap của ma trận tương quan giữa các biến và biến mục tiêu để xác định vai trò của từng đặc trưng đối với mục tiêu.



Hình – Code tính ma trận tương quang và vẽ biểu đồ heatmap

* Đồ thị heatmap dựa vào hệ số tương quan (correlation):

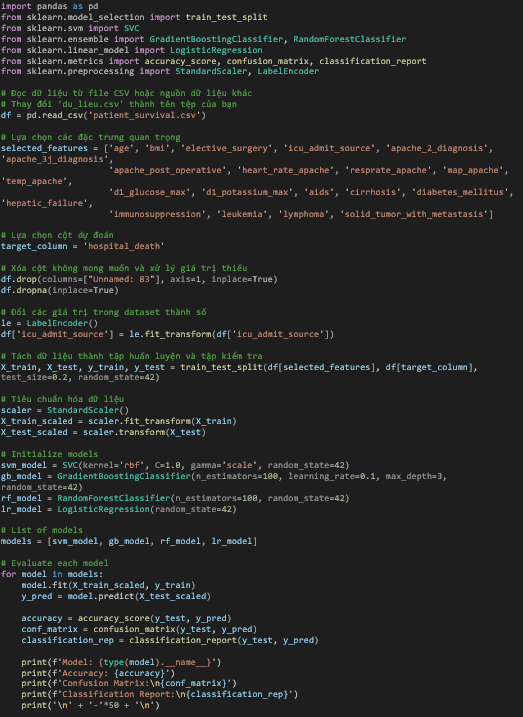


Hình – Biểu đồ Correlation Heatmap

* Khi xem heatmap trên, ta sẽ nhìn thấy các ô có màu đỏ đậm thể hiện mức tương quan cao, trong khi các ô có màu xanh nhạt hơn thể hiện mức tương quan thấp hoặc không có tương quan, cụ thể:
* Các đặc trưng có màu đỏ đậm (bằng 1) với biến mục tiêu có tương quan mạnh. Điều này có thể cho thấy mối quan hệ tuyến tính thuận (vì là giá trị dương) giữa các đặc trưng này và khả năng sống sót của bệnh nhân.
* Các đặc trưng có màu xanh nhạt (gần 0) với biến mục tiêu có tương quan yếu hoặc không có tương quan tuyến tính. Điều này cho thấy rằng các đặc trưng này có ít hoặc không có ảnh hưởng tuyến tính đáng kể đến khả năng sống sót của bệnh nhân.
* Sự tương quan giữa các cặp đặc trưng cũng có thể được nhìn thấy từ heatmap. Các ô có màu đậm hơn giữa các đặc trưng cho thấy một mức độ tương quan cao giữa chúng.

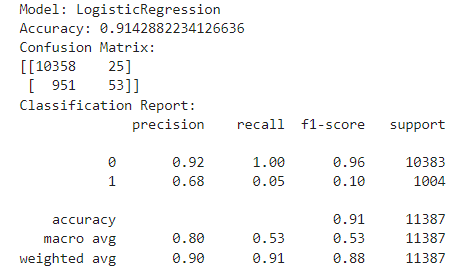
1. Ứng dụng các mô hình học máy cơ bản để giải quyết bài toán, bao gồm cả các mô hình thuộc Ensemble Learing.

Dưới đây là đoạn code áp dụng những mô hình học máy cơ bản như Logistic Regression, Random Forest, GradientBoostingClassifier, SVM để giải quyết bài toán dự đoán khả năng sống sót của bệnh nhân. Các bước thực hiện như trong hình dưới đây:



Hình – Code huấn luyện các mô hình học máy cơ bản

* **Kết quả, nhận xét:**
* **Logistic Regression:**

****

Hình – Kết quả ứng dụng mô hình Logistic Regression

* **Accuracy (Độ chính xác):** 0.914, tức là khoảng 91.4% của các dự đoán là đúng trên tổng số dự đoán.
* Số lượng true negatives (TN): 10,358
* Số lượng false positives (FP): 25
* Số lượng false negatives (FN): 951
* Số lượng true positives (TP): 53
* **Precision**:
* 0 (sống sót): 0.92, tức là 92% các dự đoán là chính xác.
* 1 (không sống sót): 0.68, tức là 68% các dự đoán là chính xác.
* **Recall**:
* 0: 1.00, tức là 100% các trường hợp thực tế là sống sót được mô hình nhớ lại.
* 1: 0.05, tức là mô hình chỉ nhớ lại khoảng 5% các trường hợp thực tế là không sống sót.
* **F1-score:**

0: 0.96, tổng hợp giữa chính xác và độ nhớ.

1: 0.10, tổng hợp giữa chính xác và độ nhớ.

* **Support:**

Số lượng thực tế mỗi lớp.

* Macro Avg: Là giá trị trung bình của các metric tính toán cho mỗi lớp mà không quan tâm đến tỷ lệ phân phối của các lớp.
* Weighted Avg: Là giá trị trung bình của các metric tính toán cho mỗi lớp với trọng số dựa trên số lượng mẫu trong mỗi lớp.
* **Random Forest:**

A screenshot of a computer

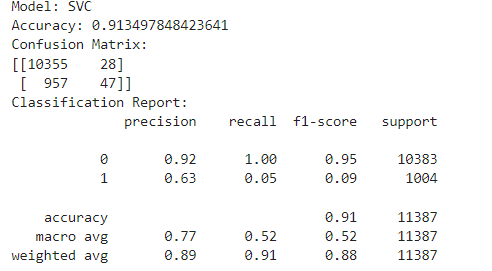
Description automatically generated

Hình – Kết quả ứng dụng mô hình Random Forest

* **Accuracy (Độ chính xác):** 0.914, tức là khoảng 91.4% của các dự đoán là đúng trên tổng số dự đoán.
* **Confusion Matrix (Ma trận nhầm lẫn)**
* Số lượng true negatives (TN): 10,319
* Số lượng false positives (FP): 64
* Số lượng false negatives (FN): 874
* Số lượng true positives (TP): 130
* **Precision**:
* 0 (sống sót): 0.92, tức là 92% các dự đoán là chính xác.
* 1 (không sống sót): 0.67, tức là 67% các dự đoán là chính xác.
* **Recall:**
* 0: 0.99 tức là 99% các trường hợp thực tế là sống sót được mô hình nhớ lại.
* 1: 0.13 tức là mô hình chỉ nhớ lại khoảng 13% các trường hợp thực tế là không sống sót.
* F1-**score**:
* 0: 0.96, tổng hợp giữa chính xác và độ nhớ.
* 1: 0.22, tổng hợp giữa chính xác và độ nhớ.
* **Support:**

Số lượng thực tế mỗi lớp.

* Macro Avg: Là giá trị trung bình của các metric tính toán cho mỗi lớp mà không quan tâm đến tỷ lệ phân phối của các lớp.
* Weighted Avg: Là giá trị trung bình của các metric tính toán cho mỗi lớp với trọng số dựa trên số lượng mẫu trong mỗi lớp.
* **SVM**

****

Hình – Kết quả ứng dụng SVM

* Accuracy (Độ chính xác): Là tỷ lệ phần trăm của các dự đoán đúng trên tổng số các mẫu. Trong trường hợp này, độ chính xác là 91.35%, tức là mô hình SVM đúng 91.35% các dự đoán trên tập kiểm tra.
* Confusion Matrix (Ma trận nhầm lẫn): Là một ma trận bốn ô chứa số lượng các dự đoán đúng và sai cho cả hai lớp (0 và 1). Trong trường hợp này:
* True Positive (TP): 47. Số lượng bệnh nhân sống sót mà mô hình dự đoán đúng.
* True Negative (TN): 10355. Số lượng bệnh nhân không sống sót mà mô hình dự đoán đúng.
* False Positive (FP): 28. Số lượng bệnh nhân không sống sót mà mô hình dự đoán sai (false alarm).
* False Negative (FN): 957. Số lượng bệnh nhân sống sót mà mô hình dự đoán sai.
* Classification Report (Báo cáo phân loại): Cung cấp các thông số đánh giá chi tiết như Precision, Recall và F1-Score cho cả hai lớp (0 và 1).
* Precision (Độ chính xác): Là tỷ lệ của số lượng True Positive trên tổng số các dự đoán dương (TP + FP). Trong trường hợp này, Precision cho lớp 1 là 0.63, tức là khi mô hình dự đoán dương, chỉ có khoảng 63% là đúng.
* Recall (Độ nhớ): Là tỷ lệ của số lượng True Positive trên tổng số thực sự dương (TP + FN). Recall cho lớp 1 là 0.05, tức là mô hình chỉ bắt được khoảng 5% của tất cả các trường hợp thực sự dương.
* F1-Score: Là trung bình điều hòa giữa Precision và Recall. F1-Score giúp đánh giá cân bằng giữa Precision và Recall.
* Macro Avg và Weighted Avg: Cung cấp giá trị trung bình của Precision, Recall và F1-Score. Macro Avg tính trung bình không quan trọng đối với kích thước của các lớp, trong khi Weighted Avg tính trung bình dựa trên kích thước của các lớp.
* Trong trường hợp này, mô hình có độ chính xác cao nhưng hiệu suất yếu đối với lớp hospital\_death=1. Precision và Recall cho lớp này đều thấp, có thể yêu cầu điều chỉnh mô hình hoặc sử dụng mô hình khác để cải thiện dự đoán trên lớp thiểu số.
* **GradientBoostingClassifier**

**A screenshot of a computer

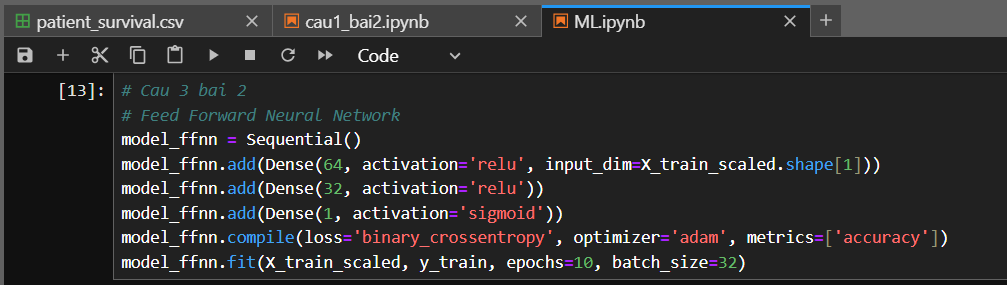
Description automatically generated**

* Hình – Kết quả ứng dụng mô hình GradientBoostingClassifier
* Accuracy (Độ chính xác): Là tỷ lệ phần trăm của các dự đoán đúng trên tổng số các mẫu. Trong trường hợp này, độ chính xác là 91.82%, tức là mô hình Gradient Boosting đúng 91.82% các dự đoán trên tập kiểm tra.
* Confusion Matrix (Ma trận nhầm lẫn): Là một ma trận bốn ô chứa số lượng các dự đoán đúng và sai cho cả hai lớp (0 và 1). Trong trường hợp này:
* True Positive (TP): 159. Số lượng bệnh nhân sống sót mà mô hình dự đoán đúng.
* True Negative (TN): 10297. Số lượng bệnh nhân không sống sót mà mô hình dự đoán đúng.
* False Positive (FP): 86. Số lượng bệnh nhân không sống sót mà mô hình dự đoán sai (false alarm).
* False Negative (FN): 845. Số lượng bệnh nhân sống sót mà mô hình dự đoán sai.
* Classification Report (Báo cáo phân loại): Cung cấp các thông số đánh giá chi tiết như Precision, Recall và F1-Score cho cả hai lớp (0 và 1).
* Precision (Độ chính xác): Là tỷ lệ của số lượng True Positive trên tổng số các dự đoán dương (TP + FP). Trong trường hợp này, Precision cho lớp 1 là 0.65, tức là khi mô hình dự đoán dương, khoảng 65% là đúng.
* Recall (Độ nhớ): Là tỷ lệ của số lượng True Positive trên tổng số thực sự dương (TP + FN). Recall cho lớp 1 là 0.16, tức là mô hình chỉ bắt được khoảng 16% của tất cả các trường hợp thực sự dương.
* F1-Score: Là trung bình điều hòa giữa Precision và Recall. F1-Score giúp đánh giá cân bằng giữa Precision và Recall.
* Cả bốn mô hình đều có độ chính xác (accuracy) khá cao, nhưng hiệu suất của chúng đối với lớp hospital\_death=1 không cao, với giá trị Recall và F1-Score thấp. Điều này có thể chỉ ra khả năng không cân bằng giữa các lớp trong tập dữ liệu, nơi mà lớp hospital\_death=1 (sống sót) chiếm một phần nhỏ so với lớp hospital\_death=0 (không sống sót).

1. Giải quyết bài toán bằng cách sử dụng Feed Forward Neural Network và Recurrent Neural Network.

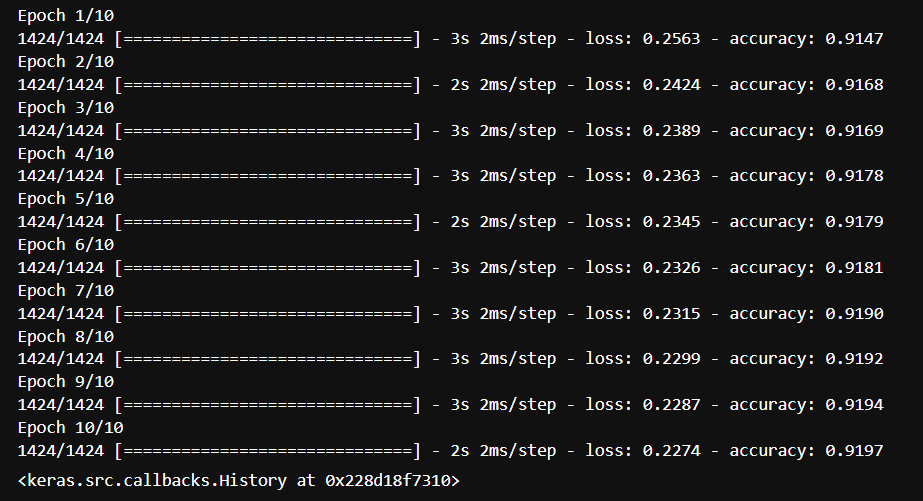
Để sử dụng Feed Forward Neural Network (mạng nơ-ron truyền thẳng) và Recurrent Neural Network (mạng nơ-ron tuần hoàn) để giải quyết bài toán, ta cần thực hiện:

* **Huấn luyện mô hình Feed Forward Neural Network:**
* Code huấn luyện mô hình Feed Forward Neural Network:



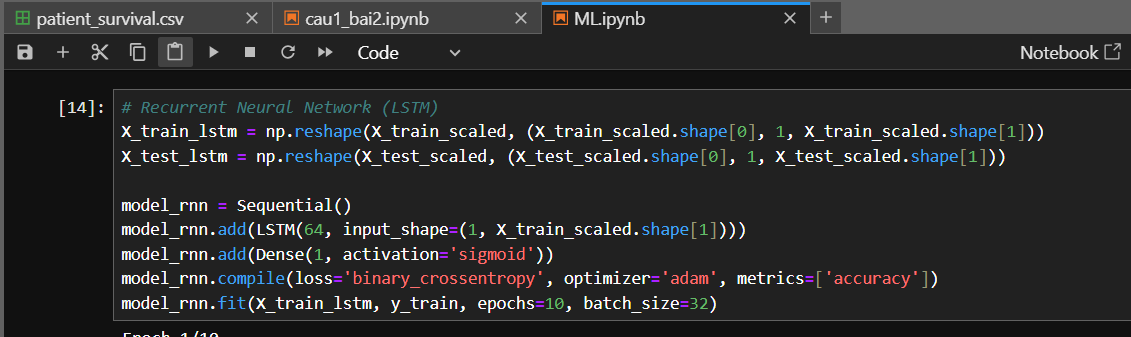
Hình – Huấn luyện mô hình Feed Forward Neural Network

* Kết quả huấn luyện:



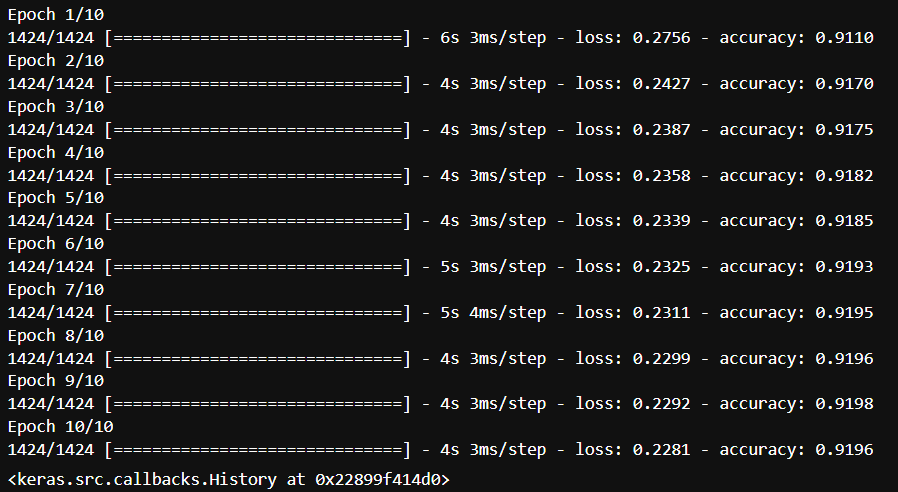
Hình – Kết quả huấn luyện mô hình Feed Forward Neural Network

* Dựa vào kết quả huấn luyện mô hình Feed Forward Neural Network, ta có thể nhận thấy:
* Mô hình đã được huấn luyện trong 10 epochs.
* Thời gian huấn luyện của mỗi epoch dao động từ 2 đến 3 giây.
* Giá trị mất mát (loss) giảm dần qua các epoch, từ 0.2563 ở epoch đầu tiên xuống 0.2274 ở epoch cuối cùng. Điều này cho thấy mô hình đang học từ dữ liệu huấn luyện và giảm thiểu sai số.
* Độ chính xác (accuracy) tăng dần qua các epoch, từ 0.9147 ở epoch đầu tiên lên 0.9197 ở epoch cuối cùng. Điều này cho thấy mô hình đạt được độ chính xác khá cao trên dữ liệu huấn luyện.
* Mô hình đạt được độ chính xác khoảng 91.97% trên tập huấn luyện cuối cùng.
* **Huấn luyện mô hình Recurrent Neural Network:**
* Code huấn luyện mô hình Recurrent Neural Network:



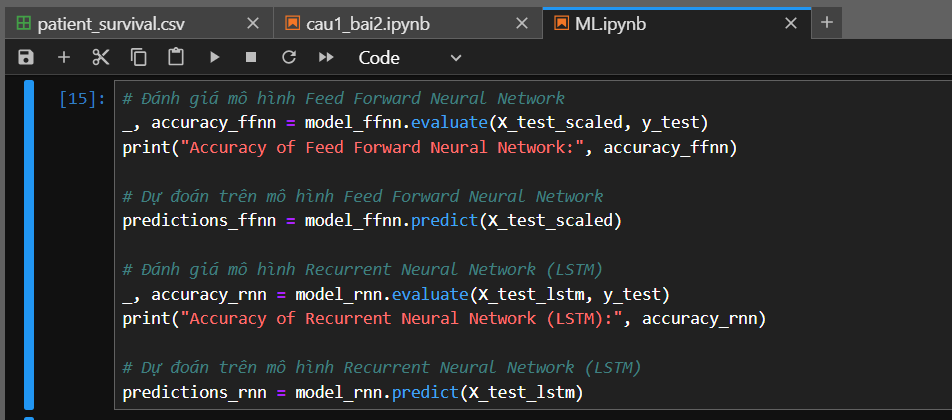
Hình – Huấn luyện mô hình Recurrent Neural Network

* Kết quả huấn luyện:



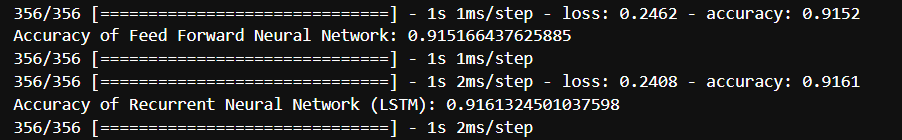
Hình – Kết quả huấn luyện mô hình Recurrent Neural Network

* Dựa vào kết quả huấn luyện mô hình Recurrent Neural Network, ta có thể thấy:
* Mô hình đã được huấn luyện trong 10 epochs.
* Thời gian huấn luyện của mỗi epoch dao động từ 4 đến 6 giây.
* Giá trị mất mát (loss) giảm dần qua các epoch, từ 0.2756 ở epoch đầu tiên xuống 0.2281 ở epoch cuối cùng. Điều này cho thấy mô hình đang học từ dữ liệu huấn luyện và giảm thiểu sai số.
* Độ chính xác (accuracy) tăng dần qua các epoch, từ 0.9110 ở epoch đầu tiên lên 0.9196 ở epoch cuối cùng. Điều này cho thấy mô hình đạt được độ chính xác khá cao trên dữ liệu huấn luyện.
* Mô hình đạt được độ chính xác khoảng 91.96% trên tập huấn luyện cuối cùng.
* **Đánh giá, dự đoán mô hình Feed Forward Neural Network và Recurrent Neural Network:**
* Code đánh giá và dự đoán 2 mô hình:



Hình – Đánh giá và dự đoánmô hình Feed Forward Neural Network và Recurrent Neural Network

* Kết quả đánh giá và dự đoán:

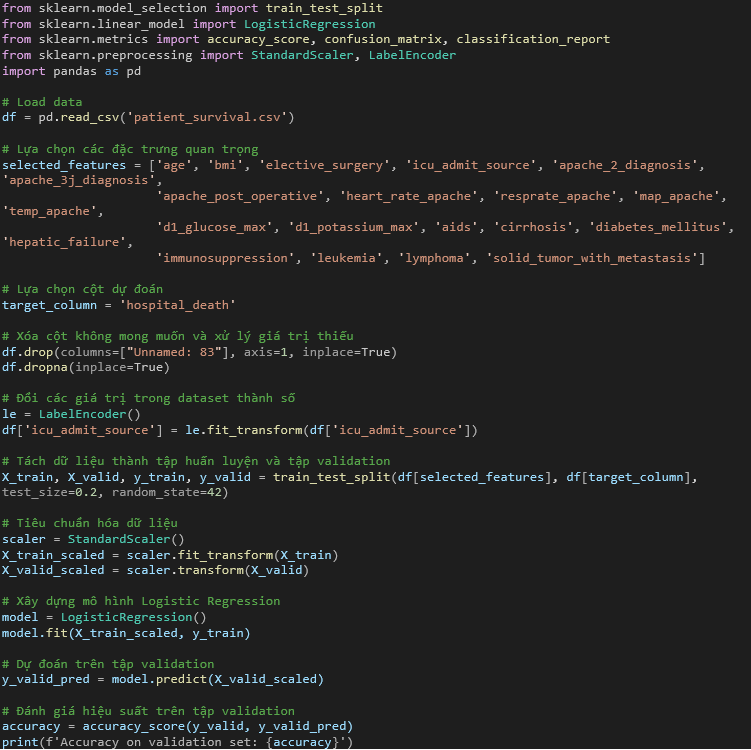


Hình – Kết quả đánh giá và dự đoán mô hình Feed Forward Neural Network và Recurrent Neural Network

* Dựa vào kết quả đánh giá và dự đoán trên tập kiểm tra của hai mô hình, ta có thể nhận xét và đánh giá hiệu suất của chúng như sau:
* Feed Forward Neural Network:
* Độ chính xác trên tập kiểm tra: 0.915166437625885 (91.52%)
* Giá trị mất mát trên tập kiểm tra: 0.2462
* Thời gian dự đoán: Khoảng 1ms/instance
* Mô hình này đạt được độ chính xác cao, đạt khoảng 91.52% trên tập kiểm tra, và có mất mát thấp.
* Thời gian dự đoán của mô hình này cũng khá nhanh.
* Recurrent Neural Network:
* Độ chính xác trên tập kiểm tra: 0.9161324501037598 (91.61%)
* Giá trị mất mát trên tập kiểm tra: 0.2408
* Thời gian dự đoán: Khoảng 2ms/instance
* Mô hình này cũng đạt được độ chính xác cao, đạt khoảng 91.61% trên tập kiểm tra và có mất mát thấp hơn so với mô hình Feed Forward Neural Network.
* Tuy nhiên, mô hình Recurrent Neural Network mất thời gian dự đoán hơn so với mô hình Feed Forward Neural Network với khoảng 2ms/instance.
* Cả hai mô hình đều đạt được độ chính xác khá cao trên tập kiểm tra, với lần lượt là khoảng 91.52% cho Feed Forward Neural Network và 91.61% cho Recurrent Neural Network (LSTM). Sự khác biệt giữa hai mô hình về độ chính xác là không đáng kể.
* Giá trị mất mát cũng tương đối thấp cho cả hai mô hình, chỉ trong khoảng từ 0.2408 đến 0.2462. Điều này cho thấy cả hai mô hình đều đạt được khả năng dự đoán tốt trên tập kiểm tra.

1. Áp dụng các kỹ thuật tránh Overfiting.

* **Áp dụng kỹ thuật chia tập huấn luyện thành tập huấn luyện và tập validation:**



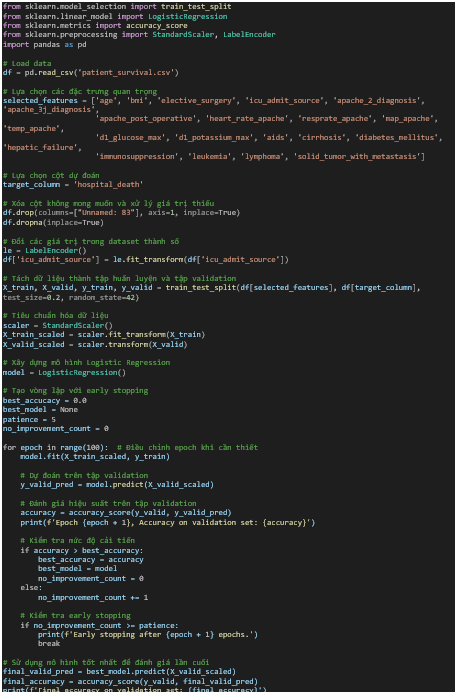
Hình – Dự đoán và đánh giá hiệu suất trên tập validation

* Kết quả sau khi áp dụng kỹ thuật chia tập huấn luyện:



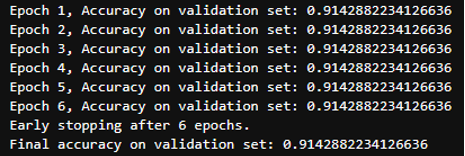
Hình – Kết quả dự đoán trên tập validation

* **Accuracy (Độ chính xác):** 0.914, tức là khoảng 91.4% của các dự đoán là đúng trên tổng số dự đoán.
* **Áp dụng Early Stopping:**
* Đoạn code sau đây, sử dụng kỹ thuật early stopping để dừng quá trình huấn luyện khi hiệu suất trên tập validation không cải thiện.

****

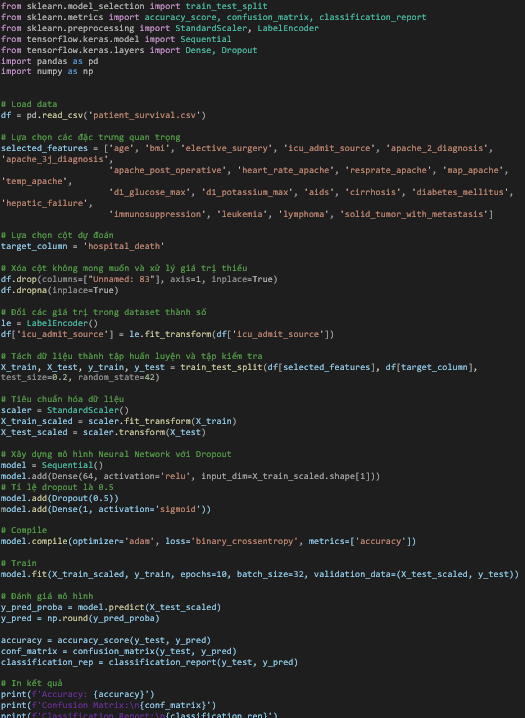
Hình – Áp dụng kỹ thuật Early Stopping

* Kết quả sau khi áp dụng kỹ thuật Early Stopping:



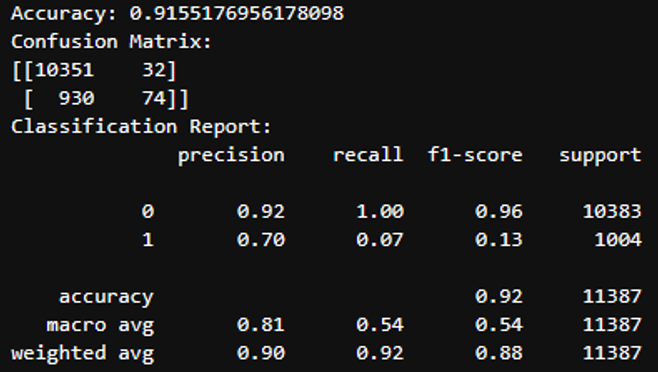
Hình – Kết quả của Early Stopping

* Accuracy (Độ chính xác): 0.914, tức là khoảng 91.4% của các dự đoán là đúng trên tổng số dự đoán.
* Epochs: Dừng quá trình huấn luyện sau 6 lần epochs.
* **Áp dụng dropout để giảm overfitting.**



Hình – Áp dụng kỹ thuật dropout

* Kết quả sau khi áp dụng Dropout:

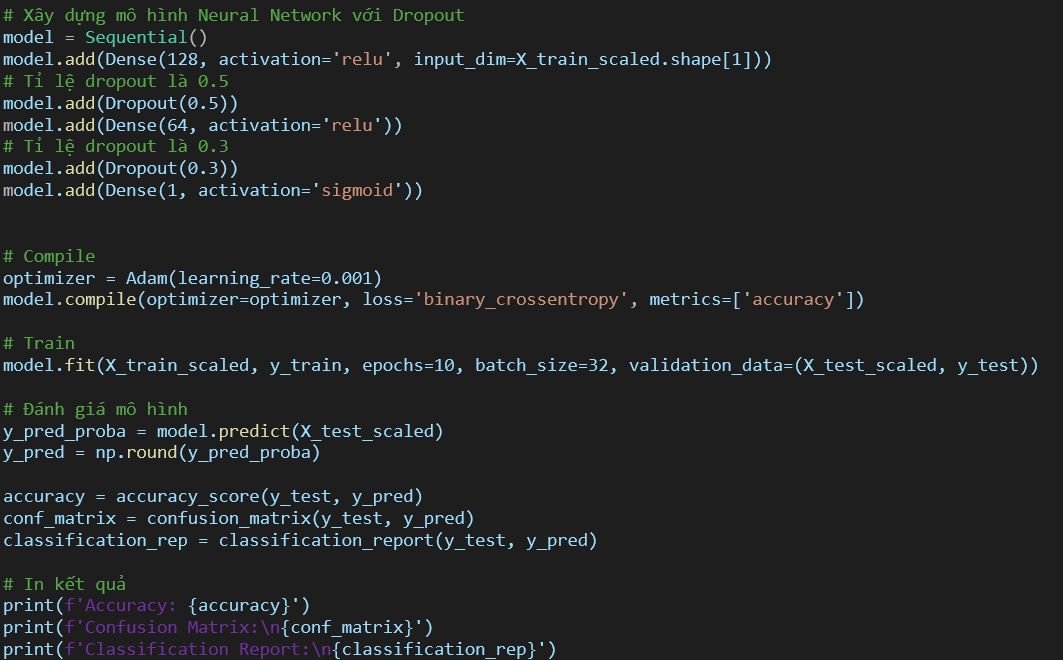


Hình – Kết quả của dropout

* Dựa vào kết quả kỹ thuật dropout để giảm tình trạng overfitting, ta thấy được:
* Accuracy (Độ chính xác): Là tỷ lệ phần trăm của các dự đoán đúng trên tổng số các mẫu. Trong trường hợp này, độ chính xác là 91.55%, tức là mô hình Neural Network đúng 91.55% các dự đoán trên tập kiểm tra.
* Confusion Matrix (Ma trận nhầm lẫn): Là một ma trận bốn ô chứa số lượng các dự đoán đúng và sai cho cả hai lớp (0 và 1). Trong trường hợp này:
* True Positive (TP): 74. Số lượng bệnh nhân sống sót mà mô hình dự đoán đúng.
* True Negative (TN): 10351. Số lượng bệnh nhân không sống sót mà mô hình dự đoán đúng.
* False Positive (FP): 32. Số lượng bệnh nhân không sống sót mà mô hình dự đoán sai (false alarm).
* False Negative (FN): 930. Số lượng bệnh nhân sống sót mà mô hình dự đoán sai.
* Classification Report (Báo cáo phân loại): Cung cấp các thông số đánh giá chi tiết như Precision, Recall và F1-Score cho cả hai lớp (0 và 1).
* Precision (Độ chính xác): Là tỷ lệ của số lượng True Positive trên tổng số các dự đoán dương (TP + FP). Trong trường hợp này, Precision cho lớp 1 là 0.7, tức là khi mô hình dự đoán dương, chỉ có khoảng 70% là đúng.
* Recall (Độ nhớ): Là tỷ lệ của số lượng True Positive trên tổng số thực sự dương (TP + FN). Recall cho lớp 1 là 0.07, tức là mô hình chỉ bắt được khoảng 7% của tất cả các trường hợp thực sự dương.
* F1-Score: Là trung bình điều hòa giữa Precision và Recall. F1-Score giúp đánh giá cân bằng giữa Precision và Recall.
* Macro Avg và Weighted Avg: Cung cấp giá trị trung bình của Precision, Recall và F1-Score. Macro Avg tính trung bình không quan trọng đối với kích thước của các lớp, trong khi Weighted Avg tính trung bình dựa trên kích thước của các lớp.
* Trong trường hợp này, mô hình có độ chính xác cao nhưng hiệu suất yếu đối với lớp hospital\_death=1. Precision và Recall cho lớp này đều thấp, có thể yêu cầu điều chỉnh mô hình hoặc sử dụng mô hình khác để cải thiện dự đoán trên lớp thiểu số.

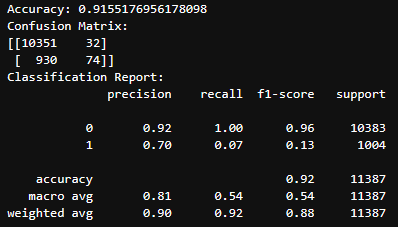
1. Phân tích các trường hợp sai, đề ra giải pháp và thực hiện giải pháp:

* Sau khi huấn luyện xong mô hình trong Machine learning, nếu muốn cải thiện độ chính xác của mô hình vừa huấn luyện, ta cần áp dụng một số phương pháp và kỹ thuật sau:
* Tăng kích thước dữ liệu huấn luyện (Data Augmentation): Một lượng dữ liệu lớn hơn thường giúp mô hình học tốt hơn và có khả năng tổng quát hóa tốt hơn.
* Kiểm soát tránh Overfitting: Nếu mô hình quá phức tạp và bắt đầu overfitting (hiệu suất trên dữ liệu kiểm tra kém hơn so với dữ liệu huấn luyện), cần xem xét việc giảm độ phức tạp của mô hình bằng cách thêm các lớp dropout, giảm số lượng tham số, hoặc sử dụng kỹ thuật regularization.
* Tinh chỉnh siêu tham số (Hyperparameter Tuning): Các siêu tham số như learning rate, số lượng lớp và đơn vị ẩn, kích thước batch, v.v., có thể ảnh hưởng đến hiệu suất của mô hình. Sử dụng kỹ thuật tìm kiếm siêu tham số (grid search, random search) để tìm giá trị tốt nhất.
* Sử dụng mô hình tiên đoán (Ensemble): Kết hợp nhiều mô hình có thể cải thiện độ chính xác. Có thể sử dụng các kỹ thuật như bagging (ví dụ: Random Forest) hoặc boosting (ví dụ: Gradient Boosting) để kết hợp các mô hình khác nhau.
* Kiểm tra lại dữ liệu (Data Inspection): việc kiểm tra lại dữ liệu có thể giúp phát hiện và loại bỏ các ngoại lệ (outliers) hoặc dữ liệu không chính xác, từ đó cải thiện hiệu suất của mô hình.
* Sử dụng mô hình tiên đoán chính xác hơn: Thử nghiệm các mô hình máy học khác nhau hoặc thậm chí chuyển sang mô hình sâu hơn và phức tạp hơn nếu cần thiết.
* Kiểm tra lại và hiểu rõ về đối tượng mô hình: Đôi khi, việc hiểu rõ về đối tượng mô hình và đặc điểm của dữ liệu có thể đưa ra sự hiểu biết quan trọng để cải thiện mô hình.
* **Thay đổi siêu tham số trong mô hình Neural Network sử dụng dropout ở câu 4:**

****

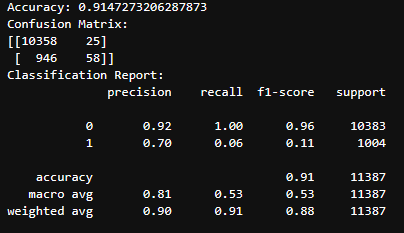
Hình – Thay đổi siêu tham số trong mô hình Neural Network

* Đây là siêu tham số cũ:
* Số lượng neurals: 64
* Tỉ lệ dropout: 0,5
* Learning rate: default
* Epochs: 10
* Kết quả sau khi thay đổi siêu tham số:

****

Hình – Kết quả sau khi thay đổi siêu tham số

* Siêu tham số mới:
* Số lượng neurals: 128
* Tỉ lệ dropout: 0,3
* Learning rate: 0,001
* Epochs: 20



* Accuracy (Độ chính xác): Thấp hơn so với mô hình sử dụng các siêu tham số cũ.
* Confusion Matrix (Ma trận nhầm lẫn): Là một ma trận bốn ô chứa số lượng các dự đoán đúng và sai cho cả hai lớp (0 và 1). Trong trường hợp này:
* True Positive (TP): Thấp hơn so với mô hình sử dụng các siêu tham số cũ.
* True Negative (TN): Cao hơn so với mô hình sử dụng các siêu tham số cũ.
* False Positive (FP): Cao hơn so với mô hình sử dụng các siêu tham số cũ.
* False Negative (FN): Cao hơn
* Classification Report (Báo cáo phân loại): Cung cấp các thông số đánh giá chi tiết như Precision, Recall và F1-Score cho cả hai lớp (0 và 1).
* Precision (Độ chính xác): Ngang bằng với mô hình sử dụng các siêu tham số cũ.
* Recall (Độ nhớ): Thấp hơn hơn so với mô hình sử dụng các siêu tham số cũ.
* Tổng quan sau khi thay đổi 4 siêu tham số trong mô hình là: Số lượng neurals, epochs, learning rate và tỉ lệ drop out thì mô hình có độ chính xác thấp hơn khi sử dụng các siêu tham số cũ.

# TÀI LIỆU THAM KHẢO

1. [Trần Trung Trực. (17/10/2020). Optimizer- Hiểu sâu về các thuật toán tối ưu ( GD,SGD,Adam,..). Viblo.](https://viblo.asia/p/optimizer-hieu-sau-ve-cac-thuat-toan-toi-uu-gdsgdadam-Qbq5QQ9E5D8#_tong-quan-8)
2. [Machine Learning cơ bản. (Jan 12, 2017). Bài 7: Gradient Descent (phần 1/2).](https://machinelearningcoban.com/2017/01/12/gradientdescent/)
3. [Machine Learning cơ bản. (Jan 16, 2017). Bài 8: Gradient Descent (phần 2/2).](https://machinelearningcoban.com/2017/01/16/gradientdescent2/)
4. [Tek4.vn. Thuật toán Adagrad và vấn đề hiệu chỉnh Learning Rate.](https://tek4.vn/khoa-hoc/machine-learning-co-ban/thuat-toan-adagrad-va-van-de-hieu-chinh-learning-rate)
5. [Tek4.vn. Thuật toán RMSprop.](https://tek4.vn/khoa-hoc/machine-learning-co-ban/thuat-toan-rmsprop)
6. [Tek4.vn. Thuật toán tối ưu ADAM.](https://tek4.vn/khoa-hoc/machine-learning-co-ban/thuat-toan-toi-uu-adam)
7. [Tek4.vn. Momentum.](https://tek4.vn/khoa-hoc/machine-learning-co-ban/momentum)
8. [ML Collective. (1/5/2022). Continual Learning and Test in Production - "Designing ML Systems" Reading Group.](https://www.youtube.com/watch?v=94afCm4igHQ)
9. [serodriguez68. Chapter 9 - Continual learning and test in production.](https://github.com/serodriguez68/designing-ml-systems-summary/blob/main/09-continual-learning-and-test-in-production.md)
10. [Đắm mình vào học sâu. 11.11. Adagrad.](https://d2l.aivivn.com/chapter_optimization/adagrad_vn.html)
11. [Đắm mình vào học sâu. 11.12. RMSProp.](https://d2l.aivivn.com/chapter_optimization/rmsprop_vn.html)
12. [Đắm mình vào học sâu. 11.14. Adam.](https://d2l.aivivn.com/chapter_optimization/adam_vn.html)
13. [admin. (30/05/2023). Neural Network là gì? Neural Network có ứng dụng như thế nào trong đời sống?](https://www.bkns.vn/neural-network-la-gi.html)
14. [Trần Trung Trực. (6/11/2020). Các phương pháp tránh Overfitting.](https://viblo.asia/p/cac-phuong-phap-tranh-overfitting-gDVK24AmlLj)

1. [patient\_survival.csv. MITISHA AGARWAL. (2 years ago). Patient Survival Prediction.](https://www.kaggle.com/datasets/mitishaagarwal/patient) [↑](#footnote-ref-1)