**QQPHÂN TÍCH TỔNG QUAN VỀ BỘ DATASET Patient Survival Prediction**

**I. Thông tin tổng quan dataset:**

**1.1 Mô tả tổng quan**

* **Tên Dataset: Patient Survival Prediction**
* **Mục tiêu: Xây dựng mô hình dự đoán khả năng tử vong tại bệnh viện sau khi bệnh nhân nhập viện, dựa trên các thông số lâm sàng và thông tin ban đầu.**

**1.2. Quy mô dataset**

* **Số lượng mẫu:** 91,713
* **Số lượng đặc trưng (biến):** 85(bao gồm cả biến mục tiêu)

**1.3. Loại dữ liệu và kiểu dữ liệu:**

Bộ dữ liệu chứa tổng cộng 85 cột, phân loại cụ thể như sau:

| **Kiểu dữ liệu** | **Số lượng cột** | **Ví dụ các biến tiêu biểu** |
| --- | --- | --- |
| Số thực (float64) | **71** | age, bmi, temp\_apache, heart\_rate\_apache, apache\_4a\_hospital\_death\_prob, d1\_glucose\_max |
| Số nguyên (int64) | **7** | encounter\_id, patient\_id, hospital\_id, elective\_surgery, icu\_id, apache\_post\_operative, hospital\_death |
| Phân loại (object) | **7** | ethnicity, gender, icu\_admit\_source, icu\_stay\_type, icu\_type, apache\_3j\_bodysystem, apache\_2\_bodysystem |

Giải thích :

* Các biến kiểu float64 và int64 thường liên quan trực tiếp đến các chỉ số y khoa đo được (ví dụ: nhiệt độ cơ thể, huyết áp, BMI, nhịp tim…).
* Các biến kiểu object phản ánh đặc điểm phân loại bệnh nhân (giới tính, dân tộc, loại ICU nhập viện, nguồn nhập viện, và nhóm bệnh lý liên quan đến APACHE).

**1.4. Các nhóm đặc trưng chính**

Nhóm đặc trưng trong bộ dữ liệu có thể chia thành các loại chính sau đây:

* **Thông tin cá nhân và nhân khẩu học**:
  + age (tuổi)
  + gender (giới tính)
  + ethnicity (dân tộc)
  + height, weight, bmi (chỉ số khối cơ thể)
* **Thông tin về ICU (Intensive Care Unit)**:
  + icu\_admit\_source (nguồn nhập ICU: cấp cứu, phẫu thuật...)
  + icu\_stay\_type (loại lưu trú ICU: cấp cứu, theo dõi)
  + icu\_type (loại ICU: tim mạch, phẫu thuật thần kinh...)
  + pre\_icu\_los\_days (số ngày nằm viện trước khi nhập ICU)
* **Các chỉ số APACHE (Acute Physiology and Chronic Health Evaluation)**:
  + apache\_4a\_hospital\_death\_prob, apache\_4a\_icu\_death\_prob (xác suất tử vong dự đoán)
  + apache\_2\_diagnosis, apache\_3j\_diagnosis (chẩn đoán theo phân loại APACHE)
* **Các chỉ số sinh lý (trong 24h đầu)**:
  + Huyết áp tâm thu và tâm trương (sysbp, diasbp)
  + Nhịp tim (heartrate)
  + Nhiệt độ cơ thể (temp)
  + Tần số thở (resprate)
  + SpO2 (độ bão hòa oxy máu)
* **Các chỉ số GCS (Glasgow Coma Scale)**:
  + gcs\_eyes\_apache, gcs\_motor\_apache, gcs\_verbal\_apache
  + gcs\_unable\_apache
* **Các chỉ số huyết học và sinh hóa**:
  + Đường máu (d1\_glucose\_max/min)
  + Kali máu (d1\_potassium\_max/min)
* **Thông tin về bệnh lý nền và bệnh lý đặc biệt**:
  + diabetes\_mellitus (đái tháo đường)
  + cirrhosis (xơ gan)
  + hepatic\_failure (suy gan)
  + immunosuppression (suy giảm miễn dịch)
  + solid\_tumor\_with\_metastasis (u ác tính di căn)

**II. MÔ TẢ CHI TIẾT CÁC BIẾN QUAN TRỌNG VÀ Ý NGHĨA**

**2.1. Các đặc trưng nhân khẩu học**

* **Tuổi (age):**
  + Trung bình: **62.3 tuổi**
  + Độ tuổi phổ biến (50% mẫu): từ **52–75 tuổi**
  + Biến rất quan trọng: Tuổi càng cao, nguy cơ tử vong càng lớn.
* **Giới tính (gender):**
  + Nam (M): **53.95%**
  + Nữ (F): **46.05%**
  + Có thể ảnh hưởng đến tỷ lệ tử vong do các yếu tố bệnh lý nền và khả năng chống chịu bệnh khác nhau.
* **Chỉ số BMI (bmi):**
  + Trung bình: **29.2** (hơi cao, nằm trong vùng thừa cân nhẹ)
  + Khoảng giữa phổ biến: từ **23.6 đến 32.9**
  + BMI quá thấp hoặc quá cao đều liên quan đến nguy cơ tử vong cao hơn.

**2.2. Các chỉ số tiên lượng APACHE**

* **apache\_4a\_hospital\_death\_prob (Xác suất tử vong dự đoán theo APACHE IV):**
  + Trung bình: **8.68%**
  + Giá trị từ **0.02% đến 13%** ở đa số mẫu (25%-75%)
  + Biến này phản ánh trực tiếp và rất mạnh mẽ nguy cơ tử vong dựa trên các yếu tố lâm sàng trong ICU.

**2.3. Các chỉ số sinh lý**

* **Nhịp tim trung bình (heart\_rate\_apache):**
  + Trung bình: **99.7 lần/phút** (cao, phản ánh tình trạng cấp tính của nhiều bệnh nhân)
  + Khoảng phổ biến: **86–120 lần/phút**
* **Áp lực động mạch trung bình (map\_apache):**
  + Trung bình: **88.0 mmHg**
  + Khoảng phổ biến: **54–125 mmHg**
  + Chỉ số MAP thấp hoặc cao bất thường đều liên quan đến nguy cơ tử vong tăng lên.
* **Nhiệt độ cơ thể (temp\_apache):**
  + Trung bình: **36.4°C**
  + Đa số bệnh nhân dao động từ **36.2°C đến 36.7°C**
  + Nhiệt độ bất thường có thể báo hiệu nhiễm trùng hoặc các vấn đề khác trong ICU.

**2.4. Các chỉ số huyết học & sinh hóa**

* **Glucose máu ngày đầu (d1\_glucose\_max):**
  + Trung bình: **174.6 mg/dL** (cao hơn mức bình thường, phản ánh stress cấp tính hoặc tiểu đường)
  + Khoảng phổ biến: **117–201 mg/dL**
  + Glucose bất thường liên quan trực tiếp đến nguy cơ tử vong cao hơn.
* **Kali máu ngày đầu (d1\_potassium\_max):**
  + Trung bình: **4.25 mEq/L**
  + Khoảng phổ biến: **3.8–4.6 mEq/L**
  + Kali cao hoặc thấp bất thường đều là các chỉ dấu quan trọng, phản ánh tình trạng sinh lý nguy hiểm.

**2.5. Điểm Glasgow Coma Scale (GCS)**

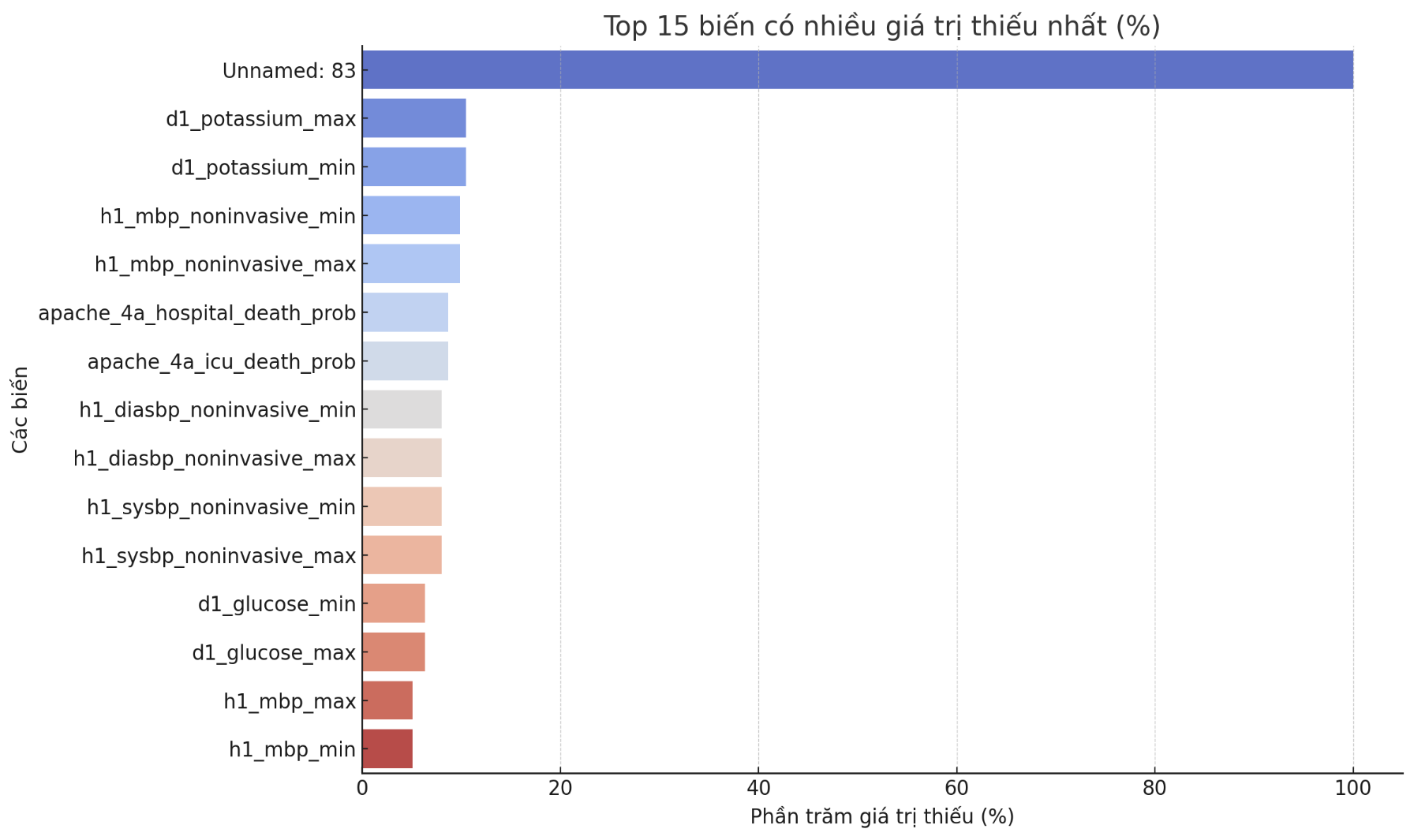
* **GCS mở mắt (gcs\_eyes\_apache):**
  + Trung bình: **3.47** (thang điểm tối đa là 4)
  + Phản ánh mức độ tỉnh táo và nhận thức của bệnh nhân.
* **GCS vận động (gcs\_motor\_apache):**
  + Trung bình: **5.47** (thang điểm tối đa là 6)
  + Quan trọng trong đánh giá tổn thương thần kinh và tiên lượng sống sót.
* **GCS lời nói (gcs\_verbal\_apache):**
  + Trung bình: **3.99** (thang điểm tối đa là 5)
  + Điểm thấp phản ánh khả năng giao tiếp và mức độ tỉnh táo giảm, nguy cơ tử vong cao hơn.

**2.6. Các bệnh lý nền quan trọng**

* **Tiểu đường (diabetes\_mellitus):**
  + Có tiểu đường: **22.5%**
  + Đây là yếu tố nguy cơ quan trọng ảnh hưởng đến thời gian hồi phục và tỷ lệ tử vong.
* **Suy gan (hepatic\_failure):**
  + Có suy gan: **1.3%**
  + Tỷ lệ thấp nhưng rất nguy hiểm, bệnh nhân có bệnh lý này có nguy cơ tử vong rất cao.
* **U đặc ác tính có di căn (solid\_tumor\_with\_metastasis):**
  + Có khối u ác tính di căn: **2.06%**
  + Bệnh lý nghiêm trọng làm tăng mạnh nguy cơ tử vong trong ICU.

**2.7. Biến mục tiêu**

* **Tử vong tại bệnh viện (hospital\_death):**
  + Tỷ lệ sống sót: **91.37%**
  + Tỷ lệ tử vong: **8.63%**
  + Mất cân bằng rõ rệt, đòi hỏi kỹ thuật cân bằng dữ liệu đặc biệt khi xây dựng mô hình.

 **III. Thống kê về giá trị thiếu (Missing Values):**

Dựa vào phân tích dữ liệu, những thông tin quan trọng về vấn đề giá trị thiếu trong dataset như sau:

**3.1. Tổng quan giá trị thiếu**

* Dataset có tổng cộng **85 đặc trưng** với rất nhiều giá trị thiếu.
* Đặc biệt, 1 cột (Unnamed: 83) bị thiếu hoàn toàn (100%), nên cần loại bỏ ngay khỏi dataset.
* **3.2. Các biến có tỷ lệ thiếu nhiều nhất**
* Dưới đây là các biến đáng chú ý nhất (Top 15 biến thiếu nhiều nhất):

| **Biến** | **Tỷ lệ thiếu (%)** | **Ý nghĩa biến và nhận xét** |
| --- | --- | --- |
| Unnamed: 83 | 100.0% | Hoàn toàn trống, cần loại bỏ khỏi dữ liệu |
| d1\_potassium\_max/min | 10.45% | Kali máu ngày đầu ICU, rất quan trọng về mặt lâm sàng (cần xử lý bằng imputation cẩn thận) |
| h1\_mbp\_noninvasive\_max/min | 9.90% | Huyết áp động mạch trung bình (MAP) không xâm lấn trong 1 giờ đầu nhập viện. Thiếu dữ liệu này gây khó khăn đáng kể vì đây là chỉ số quan trọng để đánh giá tình trạng bệnh nhân sớm |
| apache\_4a\_hospital\_death\_prob, apache\_4a\_icu\_death\_prob | 8.67% | Chỉ số tiên lượng tử vong ICU và bệnh viện theo APACHE IV. Biến rất quan trọng, cần xử lý tốt missing values |
| h1\_diasbp\_noninvasive\_max/min (tâm trương), h1\_sysbp\_noninvasive\_max/min (tâm thu) | ~8.0% | Các chỉ số huyết áp không xâm lấn ban đầu, cần xử lý cẩn thận vì rất quan trọng |
| d1\_glucose\_max/min | 6.33% | Chỉ số glucose máu ngày đầu nhập viện, cần được xử lý kỹ lưỡng |
| h1\_mbp\_max/min | 5.05% | Chỉ số MAP giờ đầu tiên nhập viện, ít thiếu hơn nhưng vẫn cần xử lý tốt |

**3.3. Ý nghĩa và ảnh hưởng của các biến thiếu giá trị nhiều nhất**

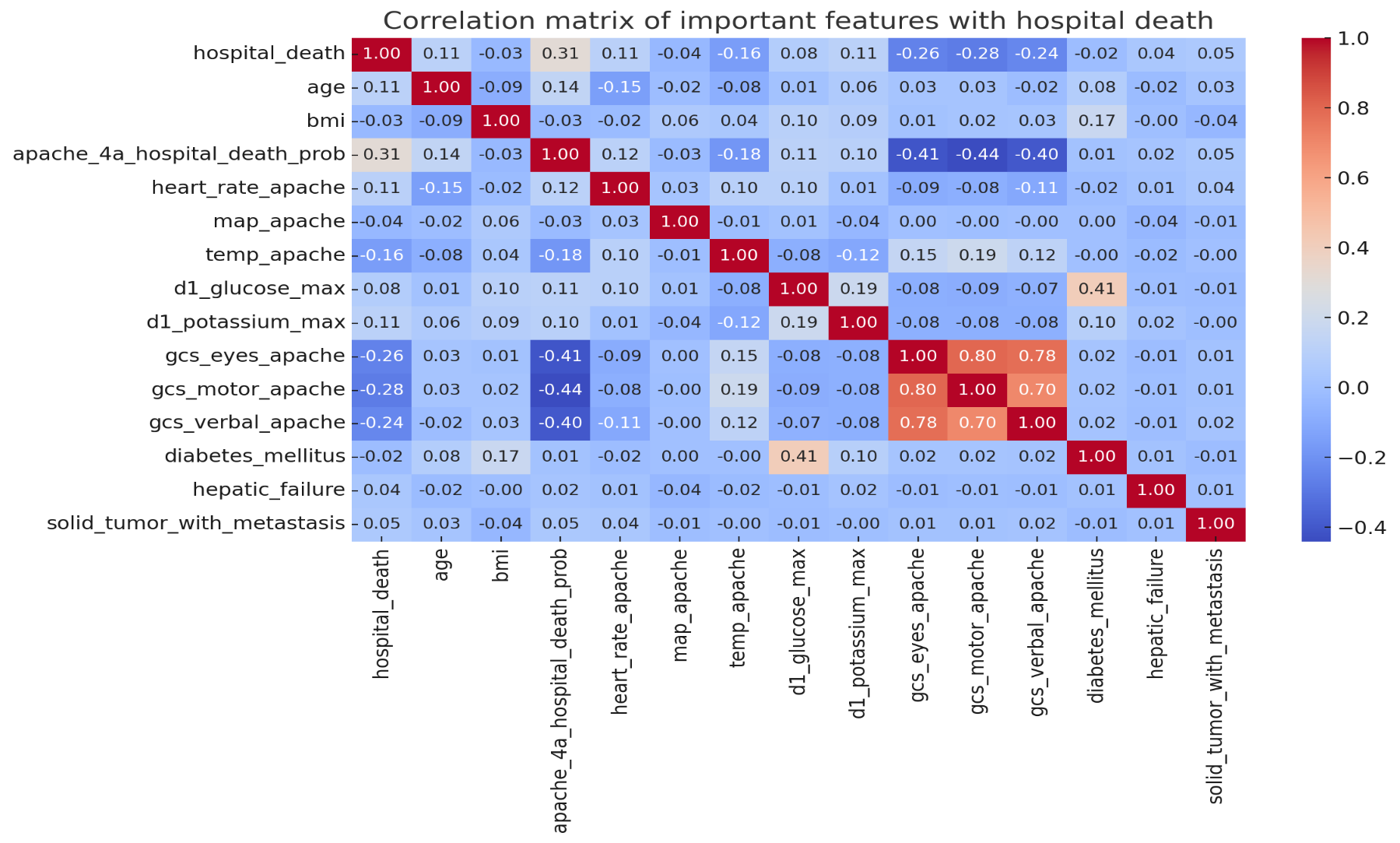
* **Các biến sinh lý & huyết áp ban đầu (h1\_\*, d1\_\*)**:
  + Đây là những chỉ số lâm sàng rất quan trọng, ảnh hưởng trực tiếp đến đánh giá tình trạng bệnh nhân và tiên lượng tử vong.
  + Nếu thiếu, cần áp dụng các kỹ thuật imputation hợp lý (mean/median, KNN, iterative imputation).
* **Chỉ số tiên lượng APACHE IV (apache\_4a\_\*)**:
  + Đây là các chỉ số tiên lượng đã được chuẩn hóa và có khả năng dự đoán rất mạnh, cần ưu tiên xử lý thiếu giá trị. Có thể dùng kỹ thuật dự đoán (predictive imputation) nếu phù hợp.
* **Các chỉ số sinh hóa (d1\_potassium\_\*, d1\_glucose\_\*)**:
  + Các biến này phản ánh tình trạng sinh lý của bệnh nhân rất rõ ràng, nếu thiếu có thể gây khó khăn trong việc dự đoán chính xác mô hình. Imputation dựa trên trung bình hoặc KNN là chiến lược phù hợp nhất.

**3.4. Các phương pháp đề xuất xử lý missing values**

Để xử lý hiệu quả, nên sử dụng các kỹ thuật sau:

* **Loại bỏ** biến Unnamed: 83 hoàn toàn.
* **Đối với biến sinh lý và APACHE IV**:
  + **Iterative Imputer** hoặc **KNNImputer**: Phù hợp để giữ lại tối đa tính đa dạng dữ liệu và giảm mất mát thông tin.
  + **Mean/Median Imputation**: Nhanh gọn, nhưng cần kiểm tra kỹ càng vì có thể gây lệch thông tin.
* **Đối với biến nhị phân hoặc categoric thiếu ít** (VD: giới tính, dân tộc):
  + Thay thế bằng giá trị phổ biến nhất (mode).

**IV. Phân tích biến mục tiêu (hospital\_death)** **& tương quan các biến khác :**

**4.1. Phân tích biến mục tiêu (hospital\_death)**

* Biến mục tiêu là dạng **nhị phân**:
  + **0 (sống sót):** **91.37%**
  + **1 (tử vong):** **8.63%**
* Tình trạng mất cân bằng dữ liệu nghiêm trọng giữa hai lớp, ảnh hưởng lớn đến việc huấn luyện mô hình.
* **4.2. Phân tích tương quan biến mục tiêu với các biến quan trọng**

Bảng dưới đây thể hiện tương quan của các biến chính với khả năng tử vong tại bệnh viện:

| **Biến số** | **Tương quan với tử vong (hospital\_death)** | **Mức độ tương quan** | **Ý nghĩa và nhận xét** |
| --- | --- | --- | --- |
| apache\_4a\_hospital\_death\_prob | **0.311** | Cao | Xác suất tử vong dự đoán theo APACHE IV cao, phản ánh chính xác khả năng tử vong thực tế. |
| d1\_potassium\_max | **0.112** | Vừa | Kali máu cao hơn mức bình thường liên quan trực tiếp đến nguy cơ tử vong. |
| age | **0.111** | Vừa | Tuổi càng cao, nguy cơ tử vong càng tăng lên đáng kể. |
| heart\_rate\_apache | **0.108** | Vừa | Nhịp tim cao phản ánh tình trạng nguy cấp, liên quan đến nguy cơ tử vong cao hơn. |
| d1\_glucose\_max | **0.082** | Yếu | Glucose máu ngày đầu cao hơn mức bình thường, liên quan nhẹ đến nguy cơ tử vong. |
| solid\_tumor\_with\_metastasis | **0.051** | Yếu | U đặc ác tính di căn làm tăng đáng kể nguy cơ tử vong nhưng ít phổ biến. |
| hepatic\_failure | **0.039** | Yếu | Suy gan ảnh hưởng rõ rệt đến nguy cơ tử vong, dù tỷ lệ bệnh nhân ít. |
| diabetes\_mellitus | **-0.016** | Rất yếu (âm) | Tiểu đường có tương quan âm rất nhẹ, không có ý nghĩa thống kê rõ ràng. |
| bmi | **-0.031** | Yếu (âm) | Chỉ số BMI có tương quan nghịch yếu, cần phân tích sâu hơn để khẳng định. |
| map\_apache (áp lực động mạch trung bình) | **-0.041** | Yếu (âm) | Huyết áp trung bình thấp hơn có xu hướng tăng nhẹ nguy cơ tử vong. |
| temp\_apache | **-0.159** | Trung bình (âm) | Nhiệt độ cơ thể thấp liên quan đến tình trạng nghiêm trọng và tăng nguy cơ tử vong. |
| gcs\_verbal\_apache | **-0.241** | Khá mạnh (âm) | Điểm lời nói thấp (GCS thấp) liên quan chặt chẽ tới nguy cơ tử vong cao. |
| gcs\_eyes\_apache | **-0.260** | Khá mạnh (âm) | Mở mắt khó khăn phản ánh giảm nhận thức, tăng nguy cơ tử vong rõ rệt. |
| gcs\_motor\_apache | **-0.282** | Khá mạnh (âm) | Khả năng vận động thấp, phản ánh tổn thương thần kinh nặng, liên quan mạnh nhất đến nguy cơ tử vong. |

**4.3. Các nhận xét và phát hiện quan trọng nhất**

* **Chỉ số APACHE IV (apache\_4a\_hospital\_death\_prob)** có tương quan mạnh nhất với tỷ lệ tử vong (0.311), đóng vai trò là biến tiên lượng mạnh mẽ nhất.
* **Các điểm GCS** (GCS motor, verbal, eyes) đều có tương quan nghịch mạnh với nguy cơ tử vong. Điểm GCS càng thấp (càng nghiêm trọng), nguy cơ tử vong càng tăng lên.
* **Chỉ số kali máu, tuổi và nhịp tim** là các biến sinh lý có liên quan vừa phải nhưng rõ ràng đến tử vong.
* **Nhiệt độ cơ thể thấp** cũng có liên quan khá rõ rệt với nguy cơ tử vong tăng lên.
* Các bệnh lý nền nghiêm trọng như **u ác tính di căn và suy gan** tuy ít gặp nhưng có tác động rõ rệt.
* Các yếu tố như tiểu đường (diabetes\_mellitus) và BMI chưa thể hiện rõ nét trong mối tương quan với tỷ lệ tử vong và cần phân tích sâu hơn.

**4.4. Ý nghĩa của phân tích này đối với mô hình học máy**

Phân tích tương quan này giúp:

* Xác định các biến quan trọng nhất cần ưu tiên khi xây dựng mô hình.
* Hỗ trợ lựa chọn đặc trưng (feature selection) hiệu quả, giảm độ phức tạp mô hình.
* Giúp hiểu sâu về dữ liệu và cải thiện độ diễn giải kết quả mô hình ML.

**V. Mối tương quan và ý nghĩa của các yếu tố với mục tiêu:**

* **Tuổi (age):** Liên quan mật thiết với nguy cơ tử vong.
* **Chỉ số BMI:** Quá cao hoặc quá thấp đều có thể liên quan tiêu cực đến sức khỏe.
* **Điểm APACHE (Acute Physiology and Chronic Health Evaluation):** Các chỉ số như apache\_4a\_hospital\_death\_prob cực kỳ quan trọng, thể hiện rõ nguy cơ tử vong.
* **GCS score:** Phản ánh tình trạng thần kinh, liên quan trực tiếp đến nguy cơ tử vong tại ICU.
* **Chỉ số huyết động (nhịp tim, huyết áp, MAP):** Ảnh hưởng trực tiếp tới nguy cơ tử vong sớm (đặc biệt trong 24h đầu).
* **Chỉ số sinh hóa (glucose, potassium):** Ảnh hưởng đáng kể tới nguy cơ tử vong, do phản ánh trực tiếp tình trạng sinh lý bệnh nhân.
* **Bệnh nền (ung thư, suy gan, tiểu đường):** Tăng mạnh nguy cơ tử vong.

**VI. Gợi ý các kỹ thuật tiền xử lý:**

**6.1. Xử lý giá trị thiếu (Missing Values)**

Dữ liệu chứa nhiều biến có giá trị thiếu đáng kể (từ 5%–10%), đặc biệt ở các biến sinh lý và chỉ số APACHE IV. Cần xử lý giá trị thiếu bằng các kỹ thuật sau đây:

* **Iterative Imputer (Multiple Imputation):**
  + **Biến phù hợp:** apache\_4a\_hospital\_death\_prob, apache\_4a\_icu\_death\_prob, d1\_glucose\_max/min, d1\_potassium\_max/min, heart\_rate\_apache, map\_apache.
  + **Ưu điểm:** Giữ được mối quan hệ giữa các biến, phù hợp với dữ liệu y khoa.
  + **Nhược điểm:** Cần thời gian tính toán nhiều.
* **KNN Imputer (K-Nearest Neighbors Imputation):**
  + **Biến phù hợp:** Các biến sinh lý, như nhịp tim, nhiệt độ, huyết áp (temp\_apache, map\_apache, d1\_\*, h1\_\*).
  + **Ưu điểm:** Giữ cấu trúc phân phối dữ liệu rất tốt, chính xác với biến số liên tục.
  + **Nhược điểm:** Cần lựa chọn số neighbors phù hợp.
* **Mean/Median Imputation (trung vị/trung bình):**
  + **Biến phù hợp:** Các biến ít thiếu hơn (dưới 2%) như age, bmi.
  + **Ưu điểm:** Đơn giản, nhanh, dễ áp dụng.
  + **Nhược điểm:** Có thể làm giảm biến động thực sự, không giữ được phân phối ban đầu.
* **Loại bỏ cột hoàn toàn trống (Unnamed: 83):**  
  Loại bỏ ngay, không cần phân tích thêm.

**6.2. Xử lý mất cân bằng dữ liệu (Class Imbalance)**

Biến mục tiêu (hospital\_death) bị mất cân bằng nghiêm trọng (8.63% tử vong vs. 91.37% sống sót), cần xử lý:

* **SMOTE (Synthetic Minority Over-sampling Technique):**
  + Tạo thêm mẫu cho lớp thiểu số (tử vong) để cân bằng dữ liệu.
  + Ưu điểm: Dễ triển khai, cải thiện hiệu suất mô hình rõ rệt.
  + Nhược điểm: Có thể gây overfitting nếu dùng quá mức.
* **Class weights trong mô hình (Logistic Regression, Random Forest, XGBoost):**
  + Điều chỉnh trọng số lớp thiểu số cao hơn.
  + Ưu điểm: Không thay đổi dữ liệu, giảm nguy cơ overfitting.
  + Nhược điểm: Hiệu quả tùy thuộc vào mô hình.
* **Under-sampling (chọn lọc):**
  + Giảm số lượng mẫu lớp đa số (sống sót).
  + Ưu điểm: Đơn giản, mô hình huấn luyện nhanh.
  + Nhược điểm: Mất thông tin quan trọng, có thể giảm độ chính xác mô hình.

**Khuyến nghị:**  
Áp dụng kết hợp SMOTE và class weights để tối ưu hiệu suất mô hình.

**6.3. Feature Encoding (Mã hóa biến phân loại)**

Các biến phân loại như gender, ethnicity, icu\_admit\_source, icu\_type cần xử lý:

* **One-hot encoding:**
  + Biến phù hợp: ethnicity, icu\_type, icu\_admit\_source.
  + Ưu điểm: Đơn giản, dễ triển khai.
  + Nhược điểm: Tăng số lượng biến, cần feature selection thêm.
* **Label encoding:**
  + Biến phù hợp: Biến nhị phân như gender (Nam/Nữ).
  + Ưu điểm: Đơn giản, không tăng số lượng biến.
  + Nhược điểm: Chỉ phù hợp khi không có nhiều mức độ.

**6.4. Feature Scaling (Chuẩn hóa dữ liệu)**

Rất cần thiết để đảm bảo hiệu suất mô hình tốt nhất, đặc biệt cho các mô hình nhạy cảm với khoảng giá trị (như Logistic Regression, Neural Networks):

* **StandardScaler:**
  + Các biến phù hợp: age, heart\_rate\_apache, map\_apache, temp\_apache, d1\_glucose\_max/min, apache\_4a\_hospital\_death\_prob.
  + Đảm bảo các biến có giá trị trung bình = 0 và độ lệch chuẩn = 1.
* **RobustScaler:**
  + Các biến phù hợp: bmi, d1\_potassium\_max/min (khi dữ liệu có ngoại lệ/outliers).
  + Chịu được ngoại lệ, không bị ảnh hưởng mạnh bởi các giá trị bất thường.

**6.5. Feature Selection (Chọn lọc đặc trưng)**

Chọn lọc đặc trưng giúp giảm số lượng biến, cải thiện mô hình và khả năng tổng quát hóa:

* **Tree-based selection (Random Forest, XGBoost):**
  + Đánh giá tầm quan trọng đặc trưng (feature importance).
  + Lựa chọn các biến quan trọng nhất, loại bỏ biến ít liên quan.
* **Recursive Feature Elimination (RFE):**
  + Phù hợp với Logistic Regression và các mô hình tuyến tính.
  + Loại bỏ từng biến một dựa trên performance validation.

**VII. Gợi ý các mô hình Machine Learning:**

**7.1. Các mô hình đề xuất**

Dưới đây là các mô hình Machine Learning được đề xuất để dự đoán nguy cơ tử vong trong bệnh viện dựa trên bộ dữ liệu Patient Survival Prediction:

**a. Logistic Regression (Baseline)**

* Là mô hình tuyến tính đơn giản, dễ triển khai và giải thích.
* Thích hợp làm mô hình baseline để đánh giá sơ bộ chất lượng dữ liệu và độ khả thi của dự án.
* Thường có hiệu suất hạn chế khi gặp dữ liệu có tương tác phi tuyến tính phức tạp.

**b. Random Forest**

* Là thuật toán học máy dựa trên cây quyết định, hiệu quả trong việc giảm thiểu hiện tượng overfitting.
* Có thể xử lý hiệu quả các đặc trưng phi tuyến tính, tương tác giữa các biến.
* Hiệu suất cao, ổn định, nhưng việc diễn giải mô hình ở mức chi tiết sẽ khó khăn hơn mô hình tuyến tính.

**c. XGBoost**

* Một thuật toán boosting mạnh mẽ, thường cho kết quả tốt hơn Random Forest nhờ kỹ thuật tăng cường (boosting).
* Xử lý tốt dữ liệu mất cân bằng và khả năng mở rộng trên dữ liệu lớn.
* Tốc độ xử lý nhanh, hiệu quả với dữ liệu lớn và phức tạp.

**d. LightGBM**

* Một biến thể của thuật toán boosting, tương tự như XGBoost nhưng thường nhanh hơn và tiết kiệm bộ nhớ hơn.
* Hiệu suất tương đương hoặc đôi khi tốt hơn XGBoost trên một số bộ dữ liệu.
* Thường được chọn khi cần tăng tốc độ huấn luyện hoặc hạn chế bộ nhớ sử dụng.

**e. Explainable AI (SHAP, LIME)**

* Các phương pháp SHAP và LIME được sử dụng để giải thích các mô hình phức tạp như Random Forest, XGBoost, LightGBM.
* SHAP cung cấp cách giải thích toàn cục và cục bộ rõ ràng về đóng góp của từng biến.
* LIME cho phép giải thích cục bộ cho từng dự đoán cụ thể.

**7.2. Phương án tối ưu đề xuất**

Phương án đề xuất để triển khai thực tế là:

**XGBoost kết hợp SHAP**

**7.3. Lý do lựa chọn XGBoost + SHAP**

* XGBoost cho hiệu suất dự đoán cao, phù hợp với dữ liệu lớn và phức tạp.
* Có khả năng xử lý tốt dữ liệu mất cân bằng.
* Kết hợp SHAP giúp cung cấp khả năng giải thích minh bạch về các yếu tố tác động đến dự đoán, đáp ứng yêu cầu về giải thích trong lĩnh vực y tế.

**7.4. So sánh khách quan các mô hình**

* **Logistic Regression:** Phù hợp với mục tiêu đơn giản, dễ giải thích nhưng độ chính xác thường thấp hơn trên dữ liệu phức tạp.
* **Random Forest:** Hiệu suất tốt, ít bị overfitting nhưng khả năng giải thích quyết định chi tiết hạn chế hơn.
* **XGBoost:** Thường cho hiệu suất vượt trội, xử lý tốt dữ liệu lớn và mất cân bằng, nhưng cần bổ sung công cụ giải thích (SHAP) để tăng khả năng minh bạch.
* **LightGBM:** Nhanh hơn, tiết kiệm bộ nhớ, nhưng mức độ hỗ trợ cộng đồng và công cụ giải thích đôi khi ít hơn XGBoost.

**7.5. Khuyến nghị triển khai thực tế**

* Thực hiện đánh giá và so sánh hiệu suất chi tiết các mô hình trên dữ liệu validation.
* Sử dụng SHAP để cung cấp thông tin giải thích rõ ràng về kết quả của mô hình, tạo điều kiện thuận lợi cho việc ứng dụng kết quả vào thực tế lâm sàng.

**VIII. Đề xuất hướng phát triển sâu hơn:**

* Xây dựng hệ thống dashboard để đánh giá rủi ro thời gian thực khi bệnh nhân nhập viện.
* Phân tích sâu hơn về các nhóm nguy cơ theo bệnh lý nền hoặc tuổi tác.
* Nghiên cứu sử dụng mô hình học sâu (deep learning) với dữ liệu lớn.