KHAI PHÁ DỮ LIỆU

Decision Tree Model – Dự đoán bệnh mùa

Sinh viên thực hiện:
Nguyễn Hoài Nam
Nguyễn Huỳnh Hương

Giới thiệu về bài toán

Bài toán sử dụng mô hình Decision Tree để dự đoán bệnh mùa dựa trên dữ liệu bệnh nhân chứa thông tin nhân khẩu (năm sinh, dân tộc), địa lý (phường/xã, quận/huyện, tỉnh/thành phố), mã bệnh và thời gian điều trị. Mô hình học cách phân loại bệnh dựa trên các đặc trưng này, từ đó xác định nguy cơ mắc bệnh theo mùa của từng cá nhân.
Decision Tree cho phép xây dựng các quy tắc phân loại dễ hiểu, hỗ trợ y tế trong việc dự báo và phòng ngừa bệnh mùa một cách chính xác và kịp thời.
Kết quả huấn luyện và đánh giá cho thấy mô hình đạt độ chính xác **71%**. Các chỉ số đánh giá chi tiết như sau:

- **Precision**: dao động từ 0.66 đến 0.73 giữa các lớp bệnh.
- **Recall**: dao động từ 0.66 đến 0.73.
- **F1-score**: trung bình đạt **0.70** (macro).

Mô hình cho thấy hiệu quả phân loại tương đối đồng đều giữa các nhóm bệnh, đặc biệt với các nhóm có tần suất xuất hiện cao.

Nội dung

1. Giới thiệu về Decision Tree

Cây quyết định (*Decision Tree*) là một mô hình học máy phổ biến, được sử dụng để giải quyết các bài toán phân loại và hồi quy. Mô hình biểu diễn các quyết định dưới dạng cấu trúc cây, với các nút (*node*) đại diện cho điều kiện, nhánh (*branch*) thể hiện lựa chọn, và lá (*leaf*) biểu thị kết quả dự đoán.

Dựa trên bộ dữ liệu `SoLieuCPDieuTri2016.xlsx`, cây quyết định giúp phân tích và dự đoán mối quan hệ giữa các bệnh và mùa trong năm. Mô hình này dễ hiểu, trực quan, hỗ trợ các cơ sở y tế đưa ra quyết định nhanh chóng, tối ưu hóa phòng ngừa và điều trị, từ đó nâng cao hiệu quả chăm sóc sức khỏe cộng đồng.

1. Giới thiệu về Decision Tree

Entropy

$$\text{Entropy}(S) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

- S đại diện cho bộ dữ liệu cần tính entropy.
- m là số lượng lớp (hoặc nhãn) trong tập dữ liệu.
- p_i là tỷ lệ các điểm dữ liệu thuộc lớp i so với tổng số điểm trong tập S.
- Entropy đo mức độ tinh khiết (pure) hay hỗn loạn của tập dữ liệu.

1. Giới thiệu về Decision Tree

Information Gain

$$IG(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \cdot \text{Entropy}(S_v)$$

- A là thuộc tính đang được xét để phân chia dữ liệu.
- $\text{Values}(A)$ là tập các giá trị có thể của thuộc tính A.
- S_v là tập con của S với $A = v$.
- $|S|$ và $|S_v|$ lần lượt là số lượng phần tử trong S và S_v.
- Information Gain đo lường mức giảm entropy sau khi phân chia dữ liệu theo thuộc tính A.

2. Giới thiệu bộ dữ liệu

Đây là bộ dữ liệu Số liệu chi phí điều trị năm 2016 (SoLieuCPDieuTri2016.xlsx) có 68,762 hàng và 17 cột :

Ý nghĩa các cột dữ liệu:

- **ID:** Mã định danh của bệnh nhân.
- **NAMSINH:** Năm sinh của bệnh nhân.
- **DANTOC:** Dân tộc của bệnh nhân (ví dụ: Kinh, Mông, Jarai, v.v.).
- **TENPXA:** Tên phường/xã.
- **TENQUANHUYEN:** Tên quận/huyện.
- **TENTINHTHANH:** Tên tỉnh/thành phố.
- **MAICD:** Mã ICD – mã quốc tế phân loại bệnh.
- **CHANDOAN:** Chẩn đoán bệnh.

2. Giới thiệu bộ dữ liệu

- **NGAYVAO:** Thời gian bệnh nhân vào viện.
- **NGAYRA:** Thời gian bệnh nhân xuất viện.
- **TONGCP:** Tổng chi phí điều trị.
- **BHYT_TT:** Số tiền được bảo hiểm y tế thanh toán cho bệnh nhân.
- **Thời gian điều trị:** Số ngày điều trị, tính chính xác theo thời gian (thập phân, ví dụ: 0.00741 tương đương vài phút).
- **Unnamed 14:** Tổng số ngày điều trị (đã được làm tròn).
- **Unnamed 2:** Không xác định.
- **Unnamed 15:** Không xác định.
- **Unnamed 16:** Không xác định.

3. Hiểu dữ liệu

Kiểm tra giá trị thiếu:

- Các cột chứa toàn bộ giá trị thiếu:
  - `Unnamed: 2`, `Unnamed: 15`, `Unnamed: 16`
- Còn lại: không có giá trị thiếu

Gợi ý xử lý:

- Xoá các cột không chứa thông tin (`Unnamed: 2`, `Unnamed: 15`, `Unnamed: 16`)

3. Hiểu dữ liệu

- **Thống kê mô tả bộ dữ liệu**

	ID	NAMSINH	NGAYVAO	NGAYRA	TONGCP	BHYT_TT	Thời gian điều trị	Unnamed: 14
count	68762	68762	68762	68762	68762	68762	68762	68762
mean	1.607e+16	1976.34	2016-07-17	2016-07-25	4.41e+06	1.80e+06	7.50	7.44
min	1.513e+16	1909.0	2010-04-09	2016-01-01	1.80	-353702.8	-5.97	-6.00
25%	1.605e+16	1955.0	2016-05-12	2016-05-20	718131.99	0.0	2.28	2.0
50%	1.608e+16	1979.0	2016-08-02	2016-08-09	1.84e+06	0.0	4.86	5.0
75%	1.610e+16	1996.0	2016-10-16	2016-10-24	4.59e+06	1.24e+06	8.28	8.0
max	1.704e+16	2016.0	2016-12-31	2016-12-31	1.64e+09	2.04e+07	2142.29	2142.0
std	3.65e+13	26.14	–	–	8.25e+06	5.87e+06	17.49	17.48

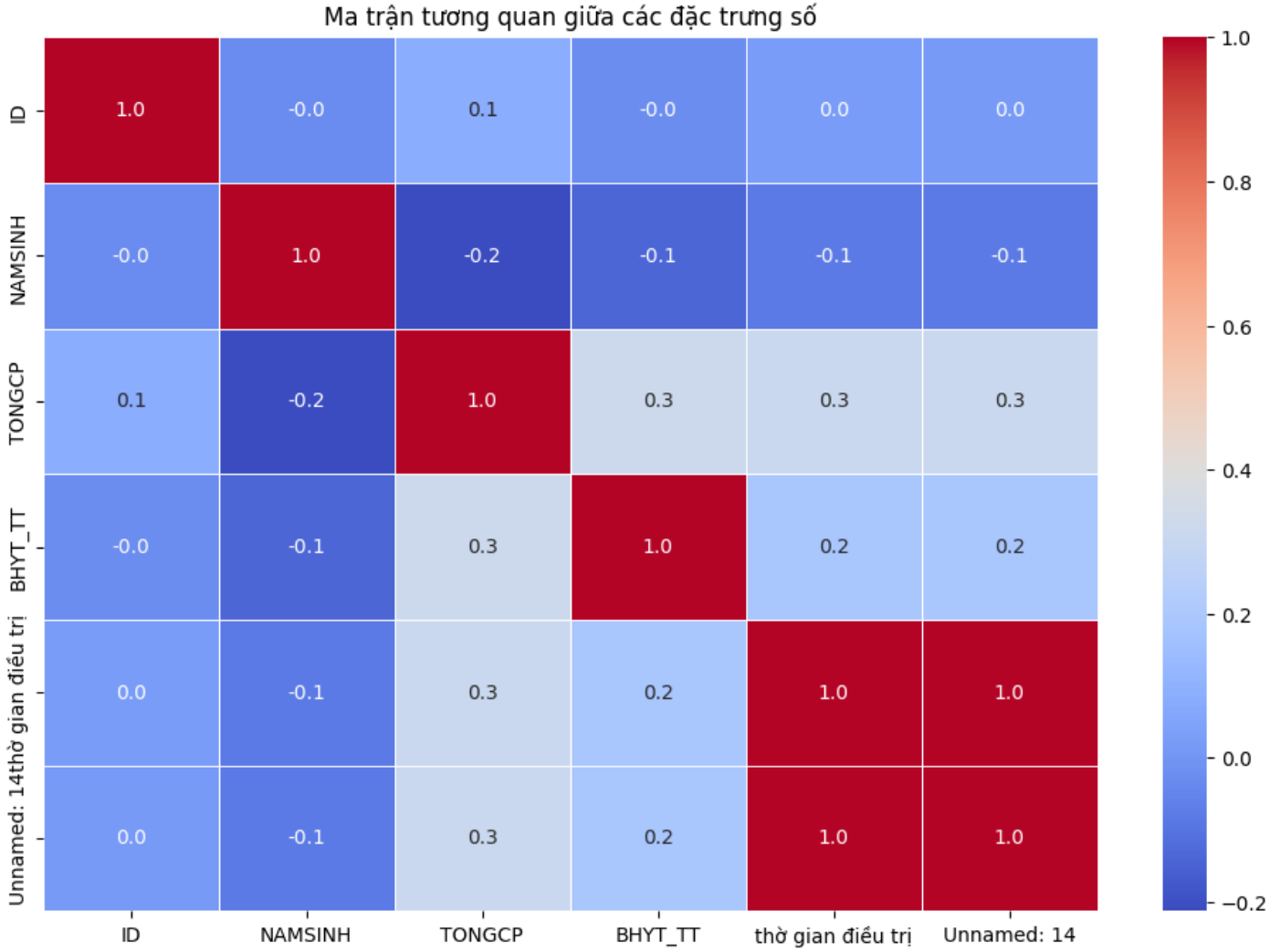
3. Hiểu dữ liệu

Phát hiện giá trị âm bất thường:

- Cột `BHYT_TT` có **1** giá trị âm.
- Cột `Unnamed: 14` có **2** giá trị âm.

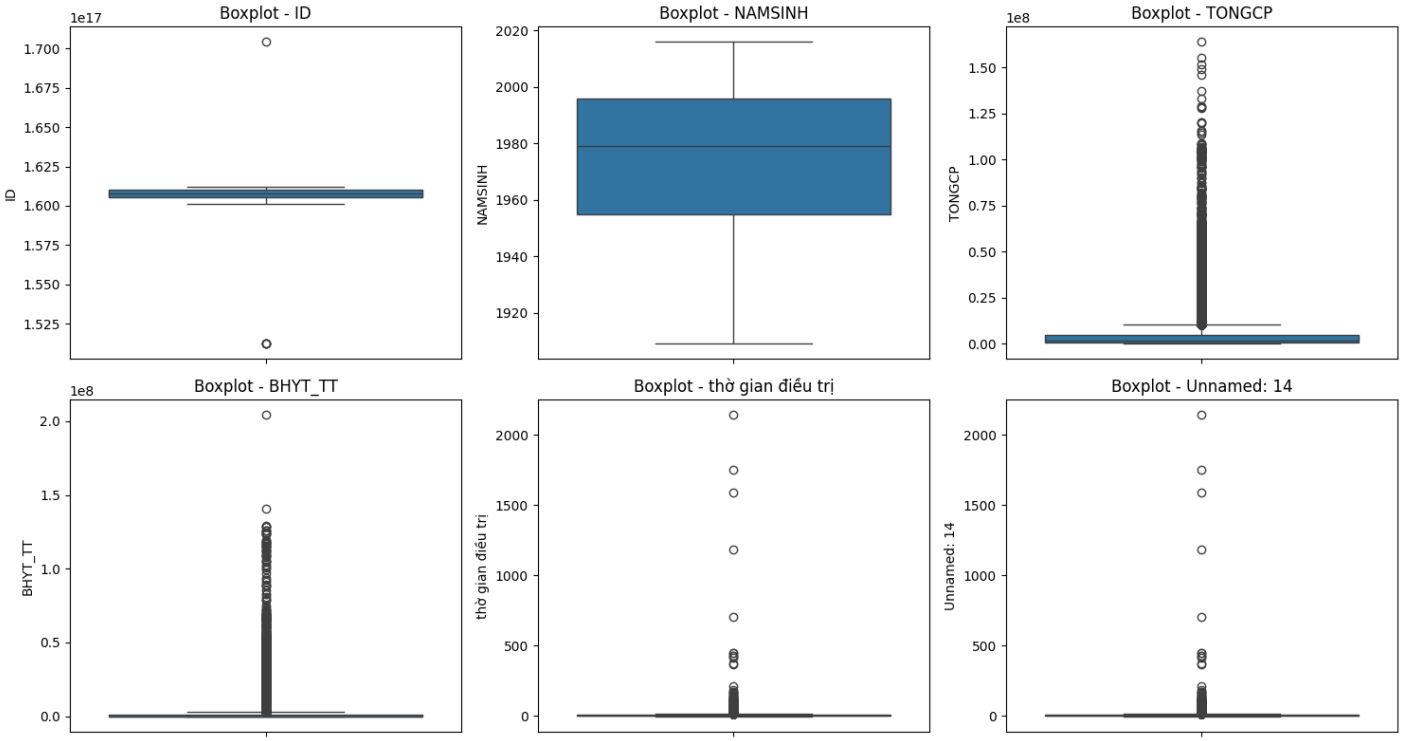
3. Hiểu dữ liệu

Biểu đồ tương quan



3. Hiểu dữ liệu

Outliers



3. Hiểu dữ liệu

- **Xử lý trùng lặp:**
  - Số lượng dòng có `ID` trùng lặp: **956**
  - Gợi ý: loại bỏ các dòng trùng lặp để tránh sai lệch thống kê và huấn luyện mô hình.

- **Xử lý giá trị không hợp lệ:**
  - Một số cột chứa giá trị `"Không xác định"`.
  - Gợi ý: thay thế bằng giá trị phù hợp (ví dụ: `NaN`) hoặc xử lý theo chiến lược riêng (bỏ dòng, gộp nhóm, gán giá trị phổ biến, v.v.)

4. Tiền xử lý dữ liệu

- Đổi tên cột `'Unnamed:14'` thành `'treatment_time'`
- Loại bỏ các dòng trùng lặp theo cột `'ID'`, giữ lại dòng đầu tiên
- Xóa các cột: `'thời gian điều trị'`, `'ID'`
- Xử lý các giá trị âm trong cột: `'BHYT_TT'` và `'treatment_time'`
- Xử lý outlier các cột: `'TONGCP'`, `'BHYT_TT'`, `'treatment_time'`, bằng IQR.
- Xóa các giá trị 'không xác định' trong các cột: `'TENPXA'`, `'TENQUANHUYEN'`, `'TENTINHTHANH'`
- Chuẩn hóa cột 'MAICD' và 'CHANDOAN':Loại bỏ dấu chấm phẩy cuối mỗi giá trị, sau đó tách các mã bệnh thành từng dòng riêng để thuận tiện cho việc phân tích.

4. Tiền xử lý dữ liệu

- Tạo đặc trưng thời gian:Trích xuất thông tin tháng–năm từ cột NGAYVAO và chuẩn hóa định dạng ngày cho các cột NGAYVAO và NGAYRA.
- Tạo đặc trưng ’Age’: bằng cách lấy năm hiện tại(2016) - ’NGAYVAO’
- Dựa vào tháng của cột ’NGAYVAO’ để tạo biến mục tiêu (season).
- Các mùa ở Việt Nam:
- Mùa Xuân: tháng 2, 3, 4
- Mùa Hè: tháng 5, 6, 7
- Mùa Thu: 8, 9, 10
- Mùa Đông: tháng 11, 12, 1

4. Tiền xử lý dữ liệu

- Xử lý cột 'DANTOC': Vì cột 'DANTOC' có giá trị 'Kinh' tức dân tộc kinh chiếm hơn 99% data nên ta có thể xử lý như sau. Kinh = Kinh, không phải kinh: other
- Đặt y = 'season' và xóa các cột 'season', 'NGAYVAO', 'NGAYRA', 'month-year', 'NAMSINH', vì không còn giá trị trong việc phân tích.

4. Tiền xử lý dữ liệu

DataFrame sau khi xử lý

	DANTOC	TENPXA	TENQUANHUYEN	TENTINHTHANH	MAICD	CHANDOAN	TONGCP	BHYT_TT	treatment_time	AGE
0	1	Phường Nhơn Bình	Thành phố Qui Nhơn	Bình Định	S01.1	Vết thương hở của mi mắt và vùng quanh mắt	15000.0000	0.0000	0	28
1	1	Phường Nhơn Phú	Thành phố Qui Nhơn	Bình Định	I20	Cơn đau thắt ngực	83346.5000	83346.5000	0	18
2	1	Xã Nhơn Lộc	Thị xã An Nhơn	Bình Định	J68.2	Viêm hô hấp trên	159.9990	0.0000	0	36
3	1	Xã Nhơn Lộc	Thị xã An Nhơn	Bình Định	J68.2	Hạ calci máu	159.9990	0.0000	0	36
4	1	Xã Nhơn Lộc	Thị xã An Nhơn	Bình Định	P71.0	Viêm hô hấp trên	159.9990	0.0000	0	36

Chia dữ liệu train/test

Tỷ lệ chia: 80% train — 20% test

- Dữ liệu được chia thành hai tập:
  - **Tập huấn luyện (Train set):** chiếm 80% tổng dữ liệu, được sử dụng để huấn luyện mô hình.
  - **Tập kiểm tra (Test set):** chiếm 20% tổng dữ liệu, được sử dụng để đánh giá hiệu quả mô hình trên dữ liệu chưa từng thấy.
- Việc chia dữ liệu giúp kiểm tra khả năng tổng quát hoá của mô hình và tránh overfitting.
- Phân chia đảm bảo tỷ lệ giữa hai nhãn được giữ nguyên (stratified splitting).

4. Tiền xử lý dữ liệu

- Mã hóa biến mục tiêu bằng `LabelEncoder`
- Mã hóa các cột chuỗi: `TENPXA`, `TENQUANHUYEN`, `TENTINHTHANH`, `MAICD`, `CHANDOAN`
- Mã hóa chuyển dữ liệu dạng chuỗi (categorical) thành dạng số để mô hình máy học có thể xử lý được.

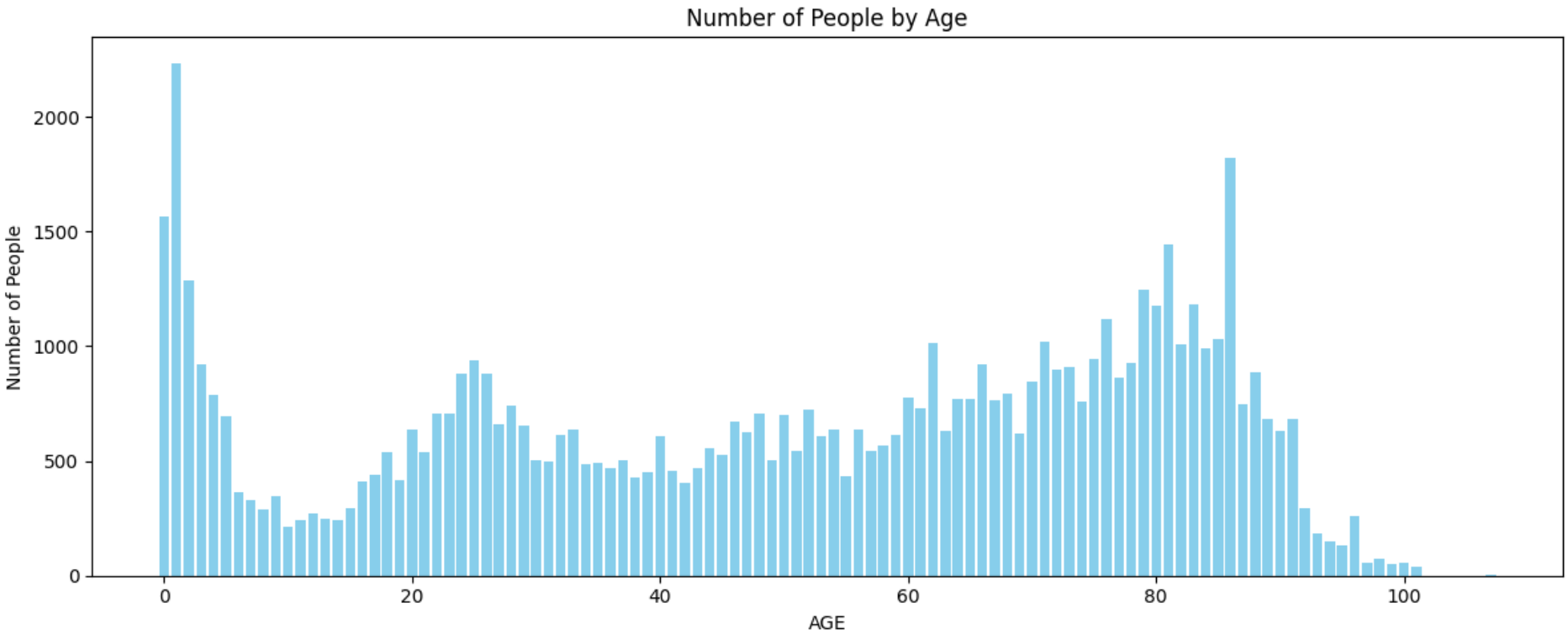
4. Tiền xử lý dữ liệu

Dữ liệu sau khi được mã hóa

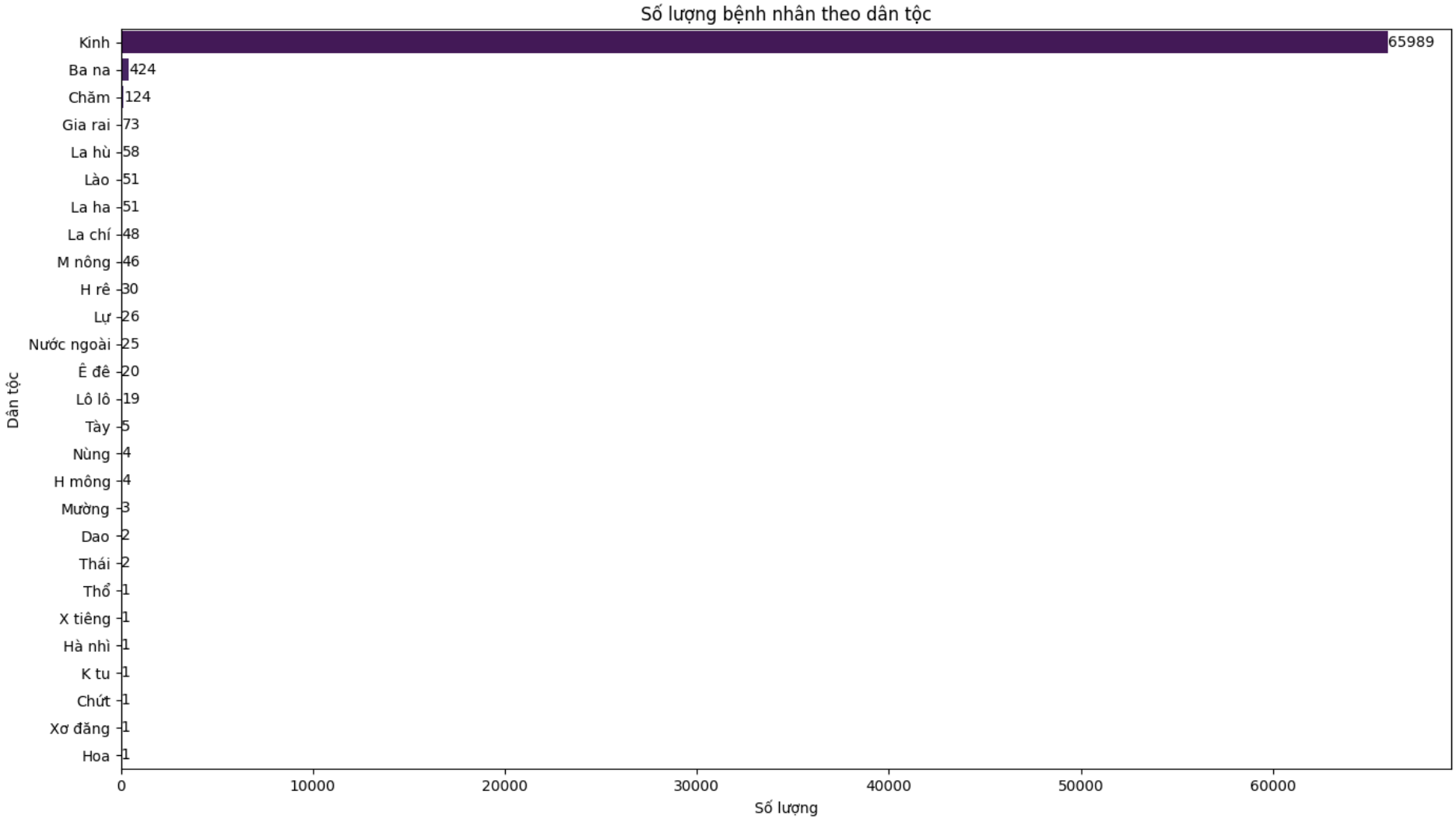
	TENPXA	TENQUANHUYEN	TENTINHTHANH	MAICD	CHANDOAN
78968	113	275	5	1709	18734
9812	188	275	5	1055	12671
133794	128	275	5	1877	10574
95678	100	275	5	2039	3248
49504	186	275	5	904	13220

5 Trực quan hóa dữ liệu

5.1 Số lượng bệnh nhân theo độ tuổi

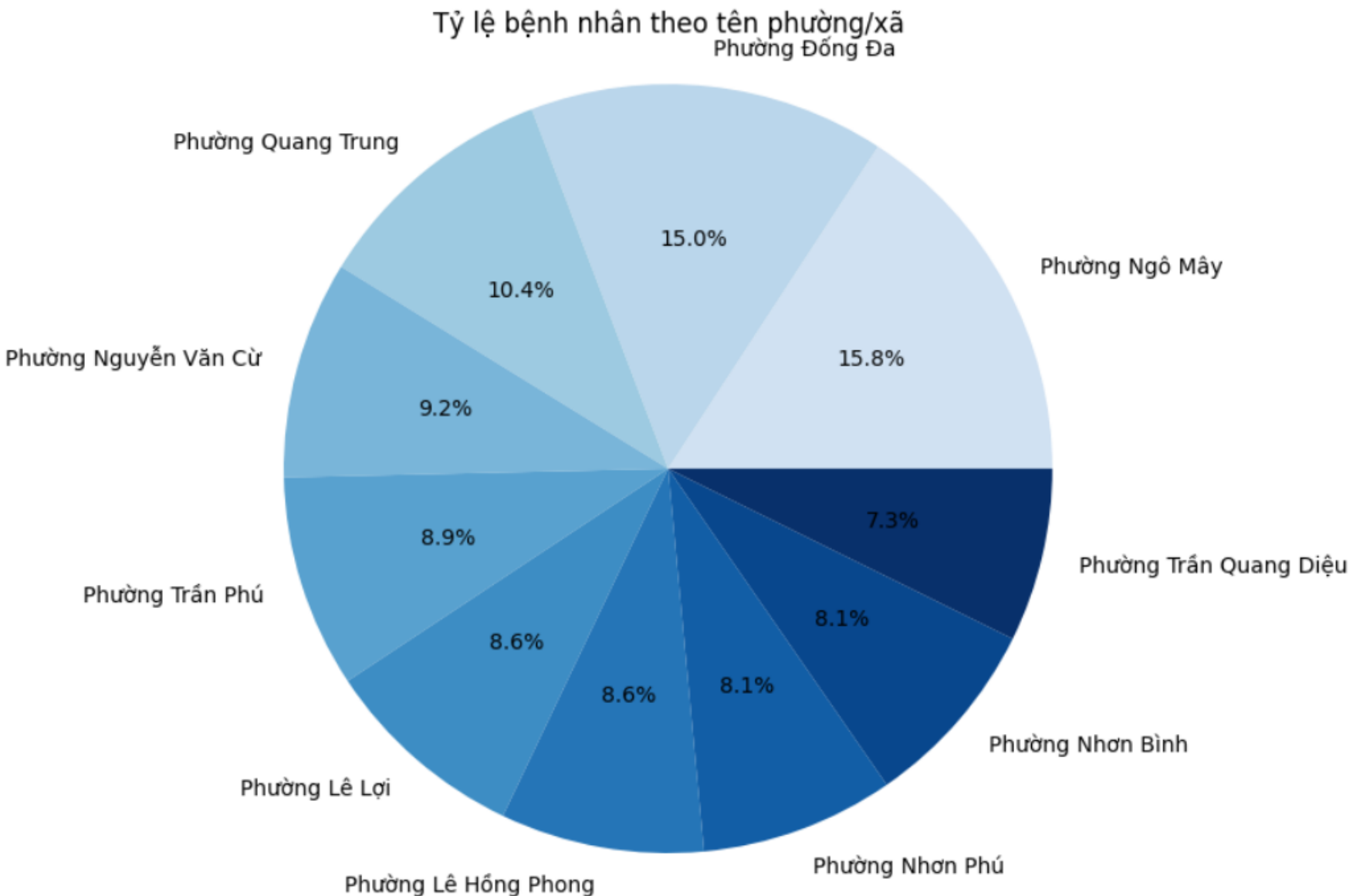


5.2 Số lượng bệnh nhân theo dân tộc



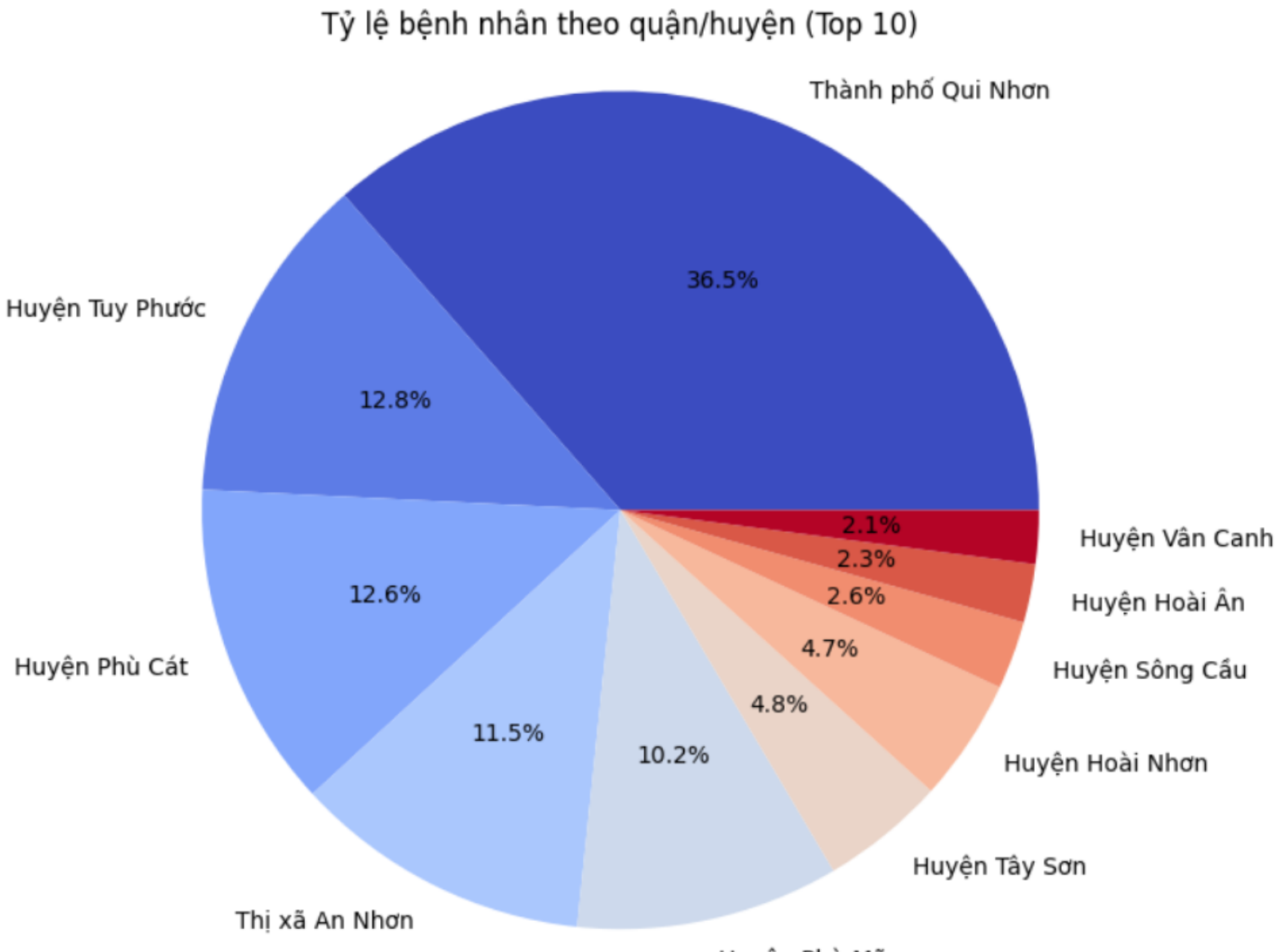
5 Trực quan hóa dữ liệu

5.3 Tỷ lệ bệnh nhân theo Phường/Xã (top 10)



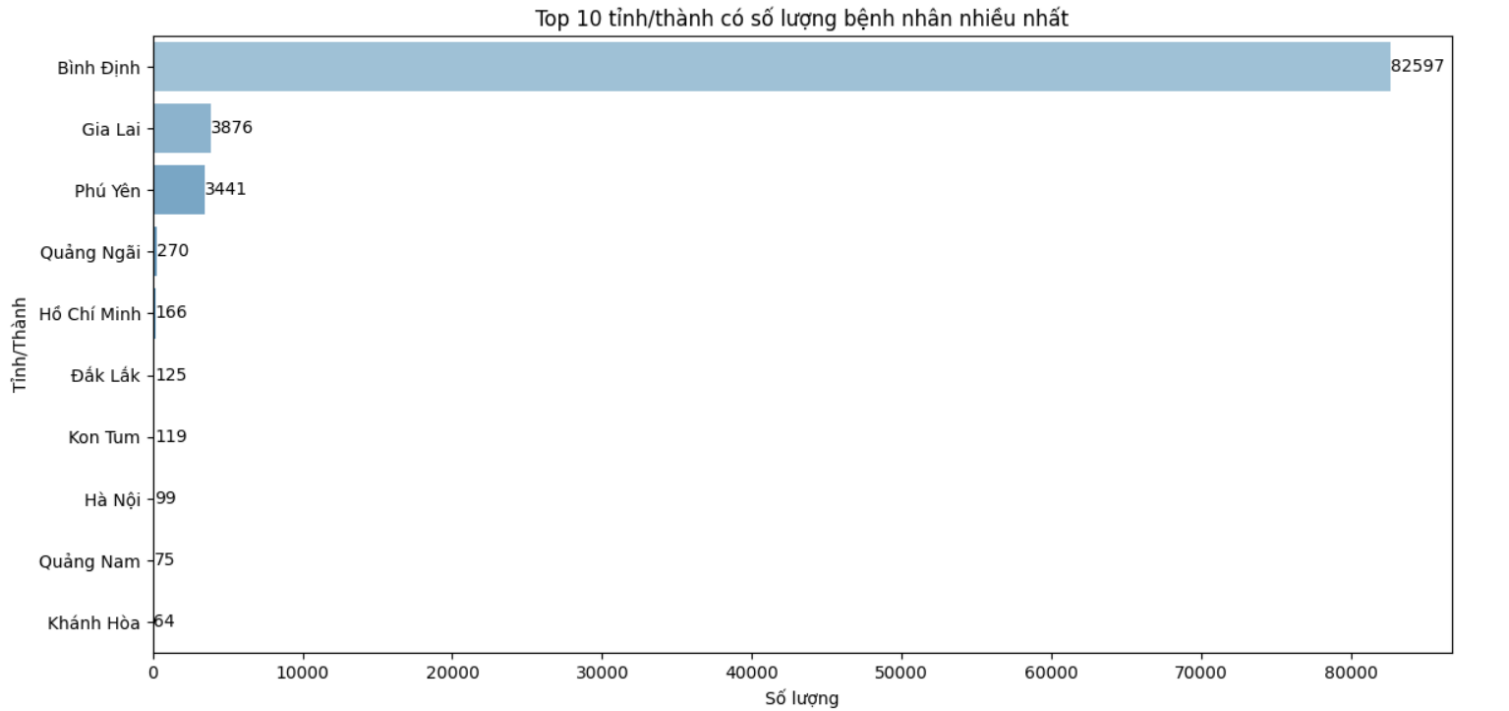
5 Trực quan hóa dữ liệu

5.4 Tỷ lệ bệnh nhân theo Quận/Huyện(top 10)



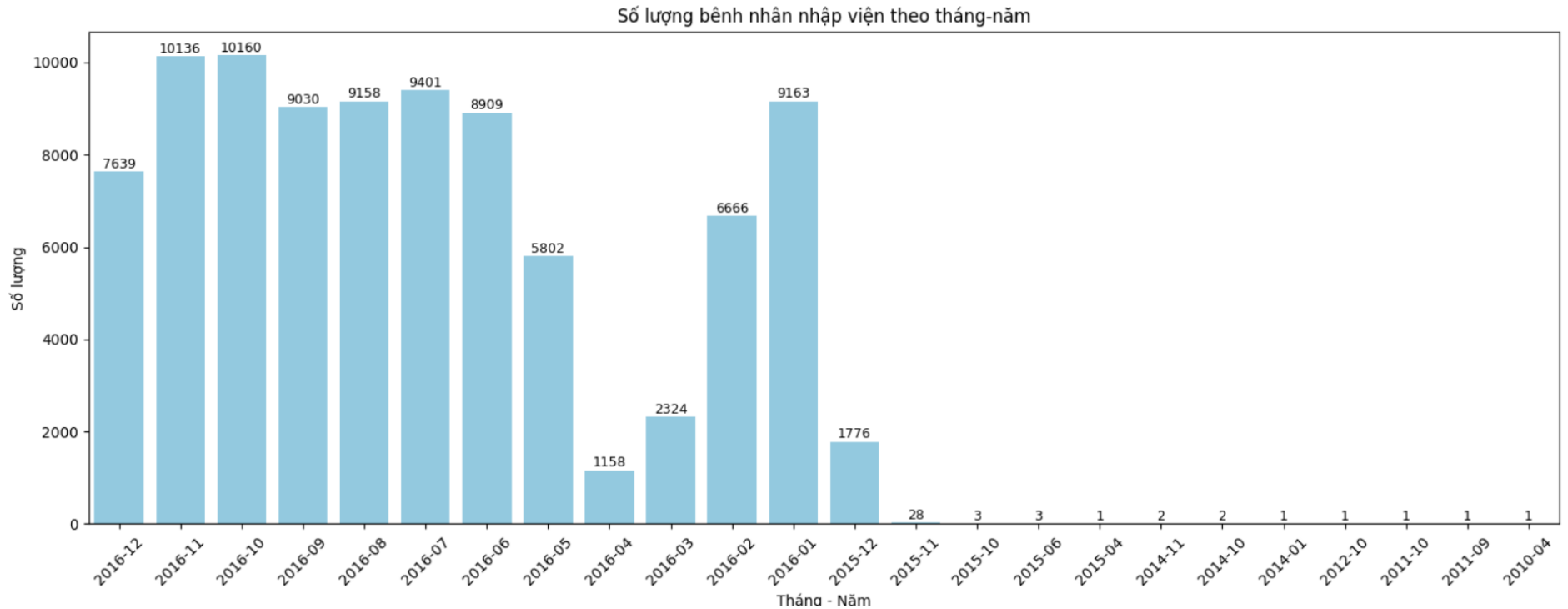
5 Trực quan hóa dữ liệu

5.5 Top 10 tỉnh Thành có số lượng bệnh nhân nhiều nhất



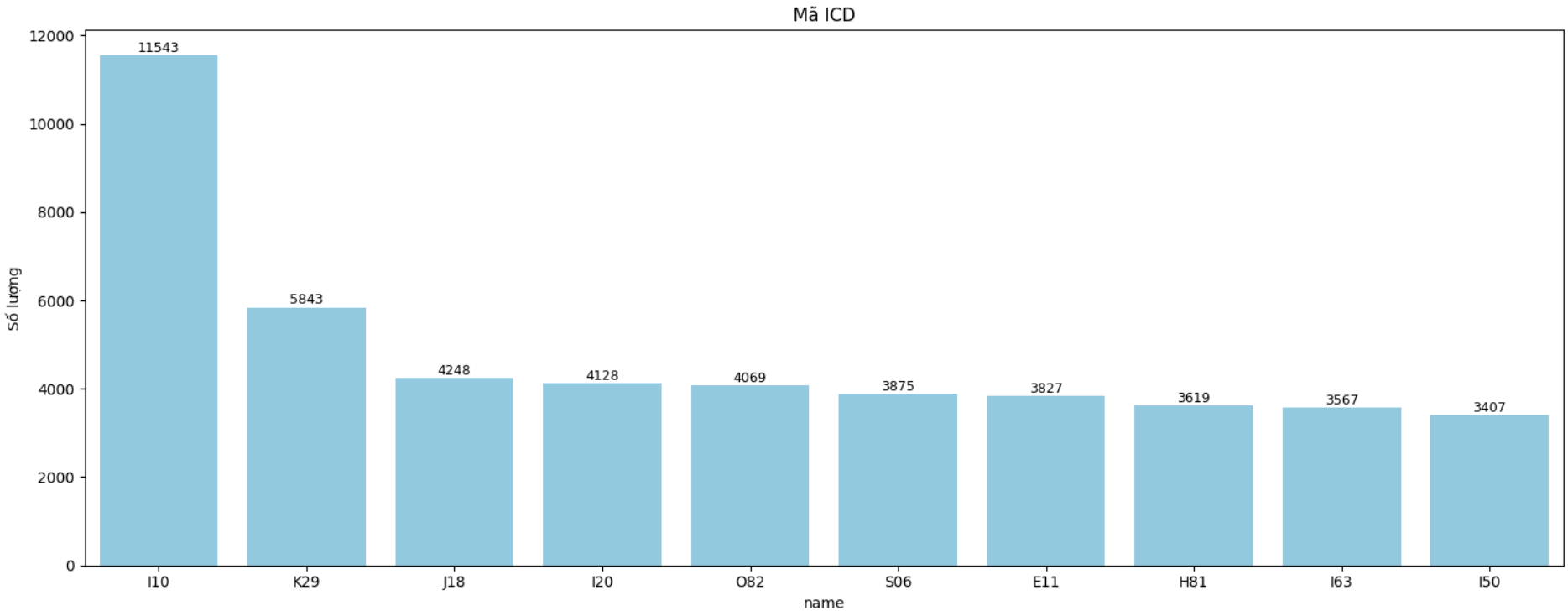
5 Trực quan hóa dữ liệu

5.6 Số lượng bệnh nhân nhập viện theo tháng-năm.



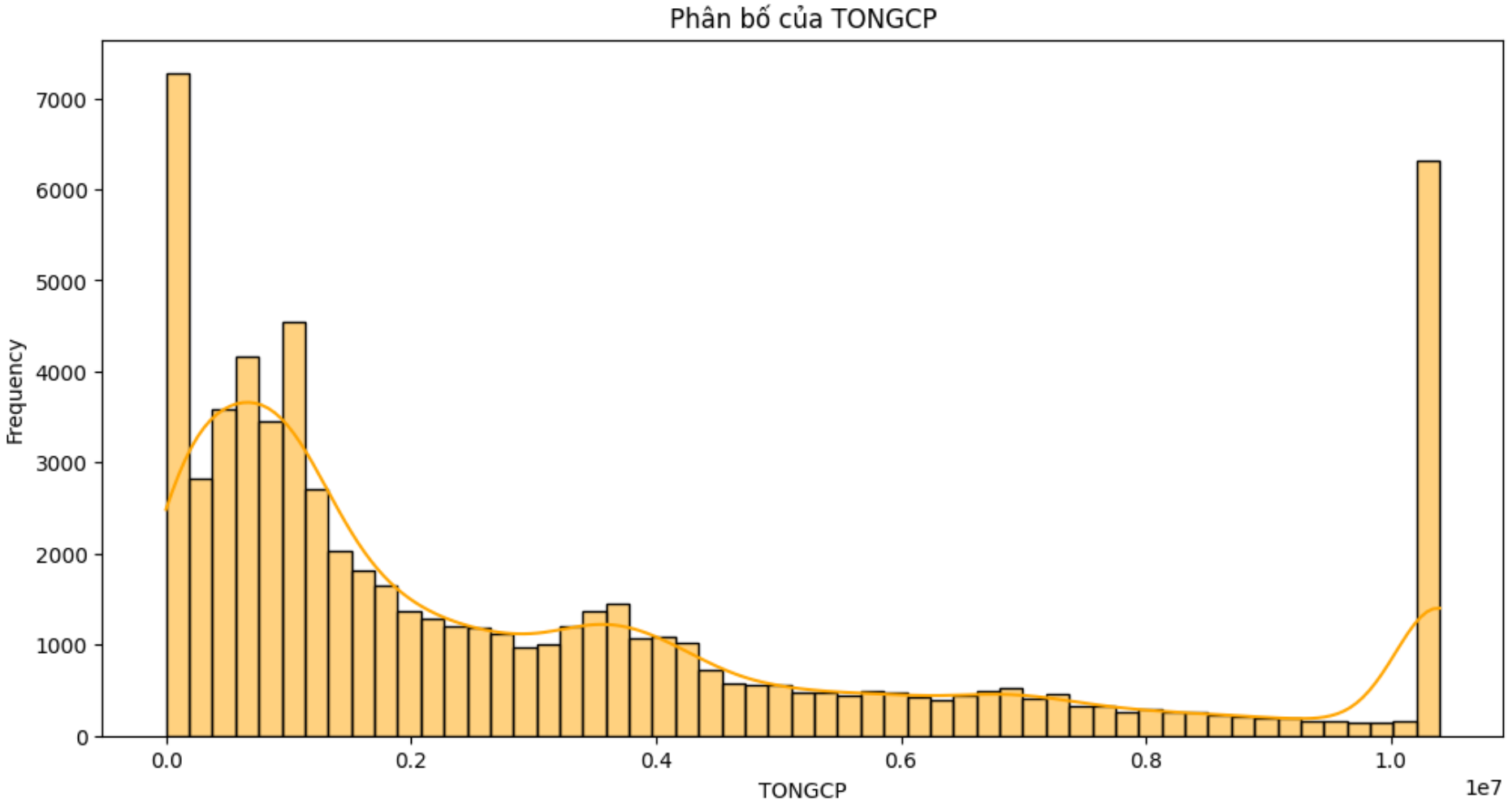
5 Trực quan hóa dữ liệu

5.7 Tần suất mã ICD



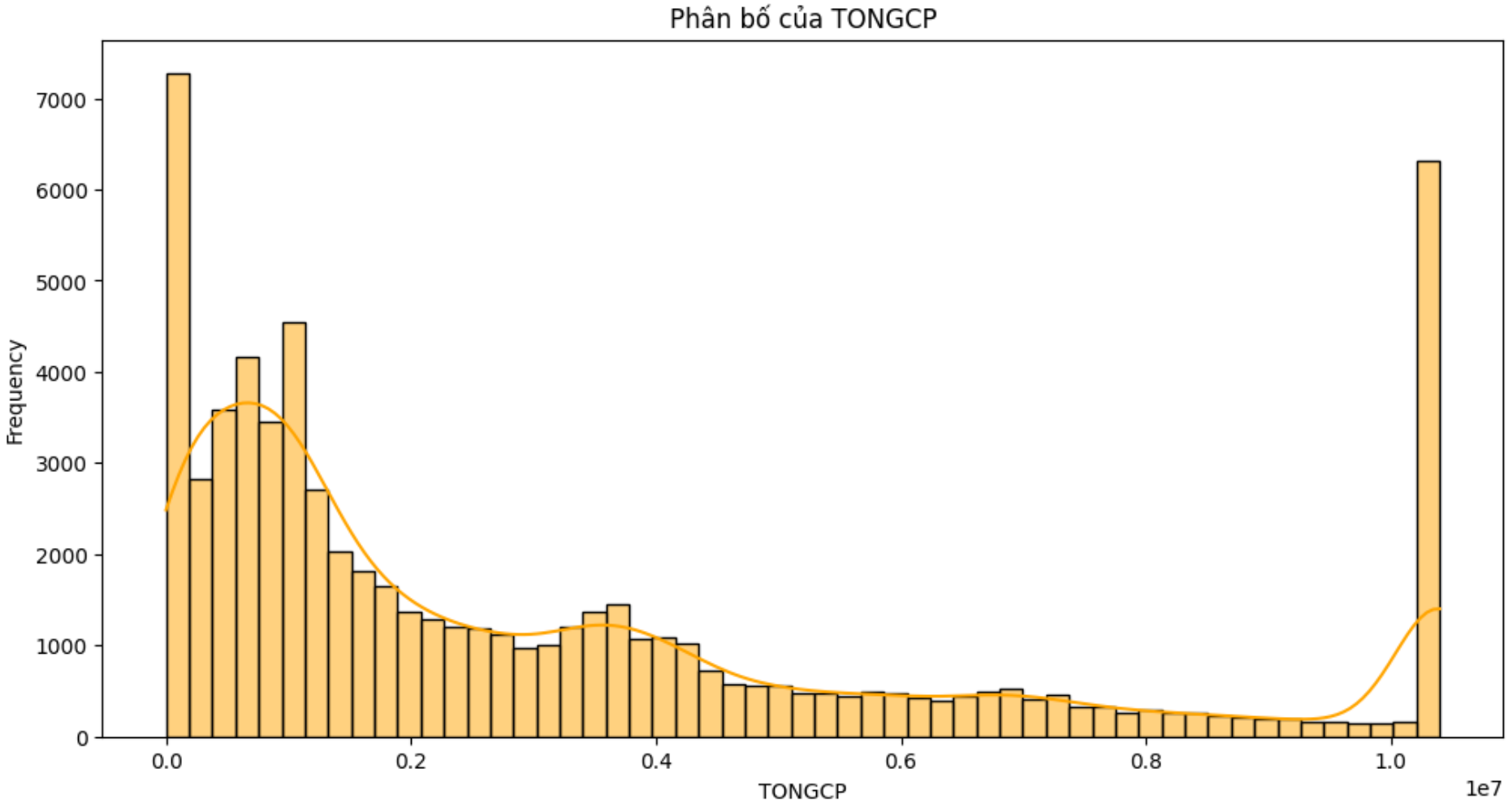
5 Trực quan hóa dữ liệu

5.8 Tần suất Chẩn đoán bệnh



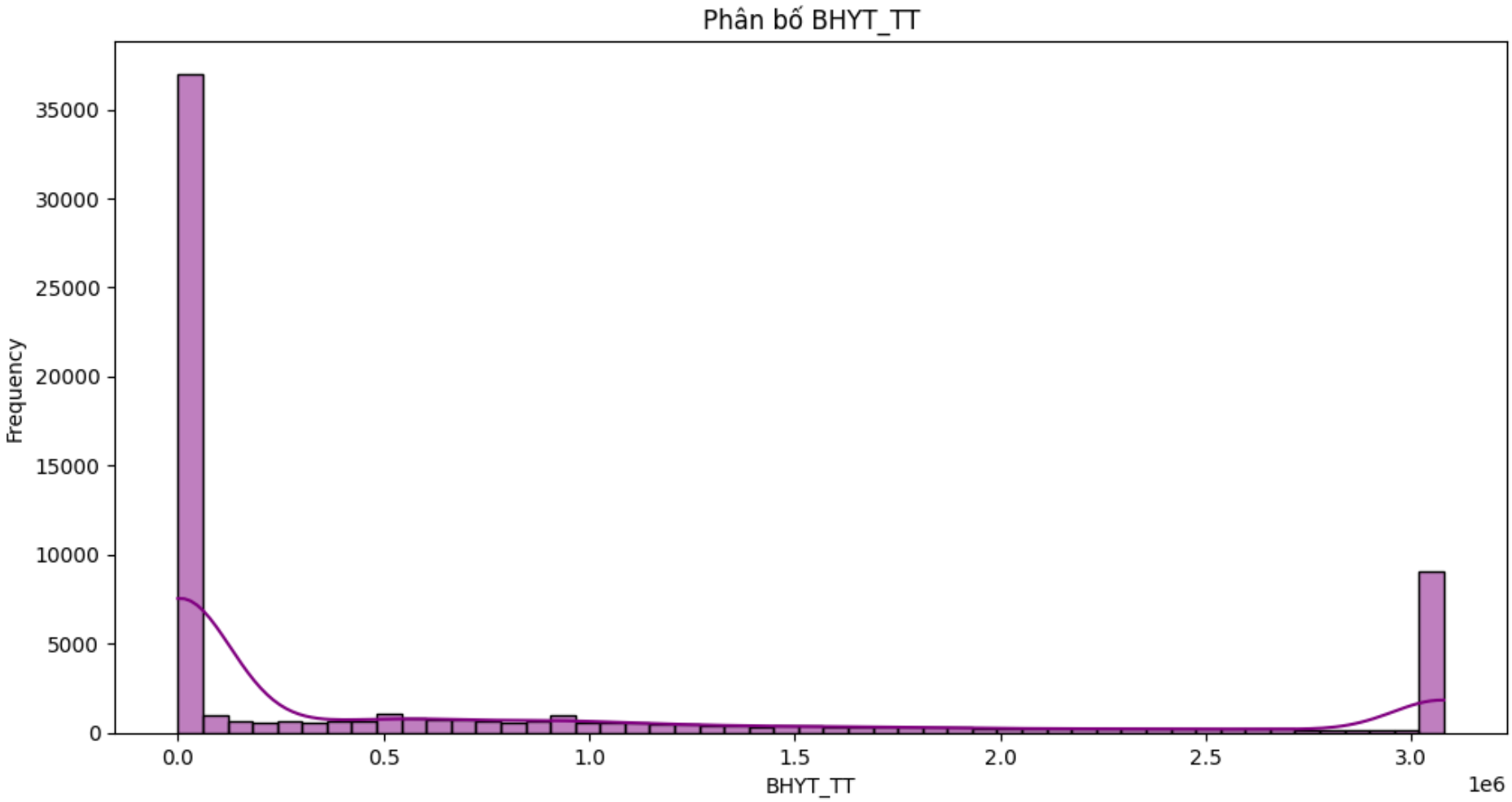
5 Trực quan hóa dữ liệu

5.9 Phân bố của Tổng chi phí



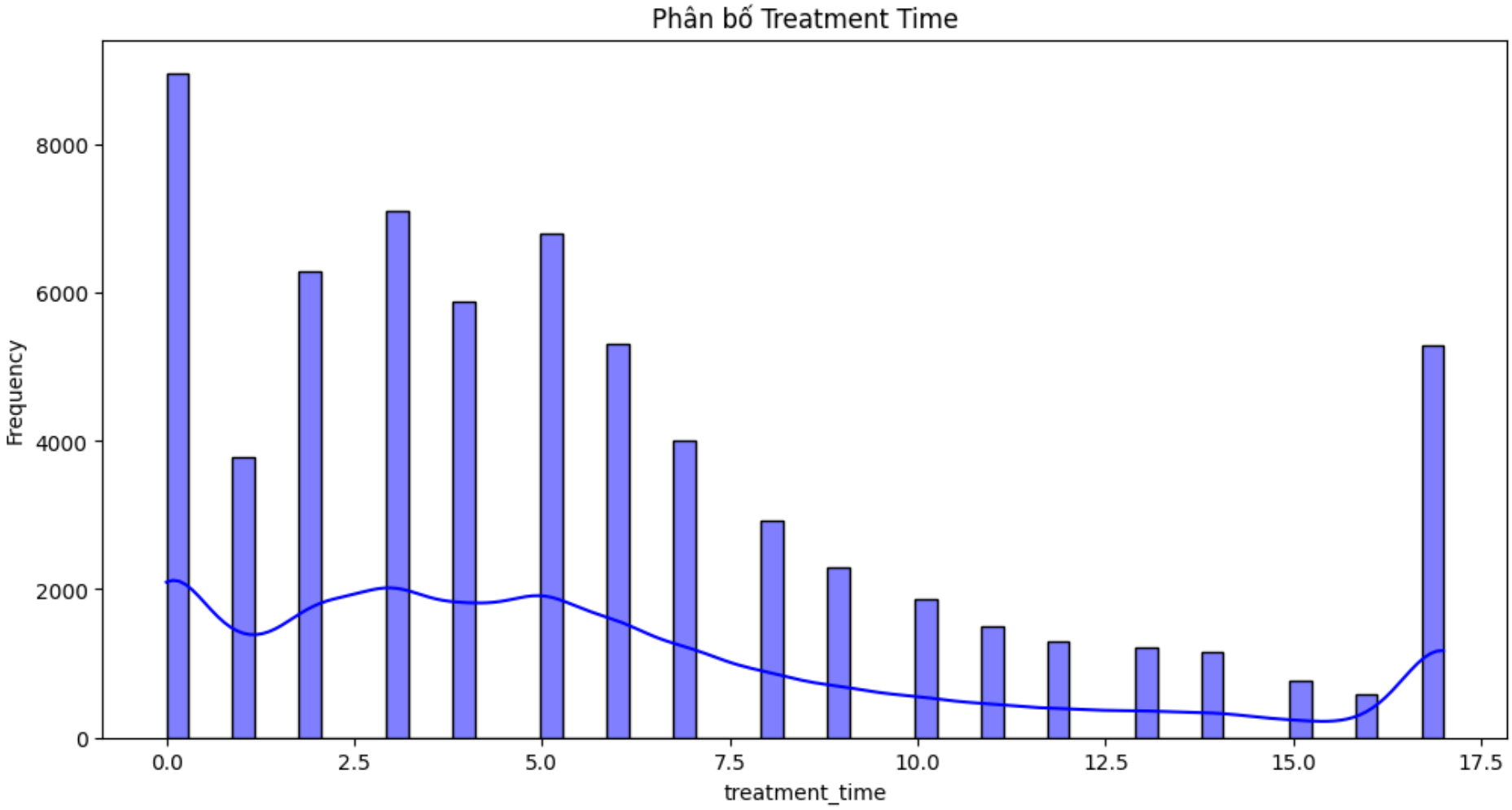
5 Trực quan hóa dữ liệu

5.10 Phân bố của BHYT_TT



5 Trực quan hóa dữ liệu

5.11 Phân bố của treatment_time



6. Lựa chọn/trích xuất đặc trưng

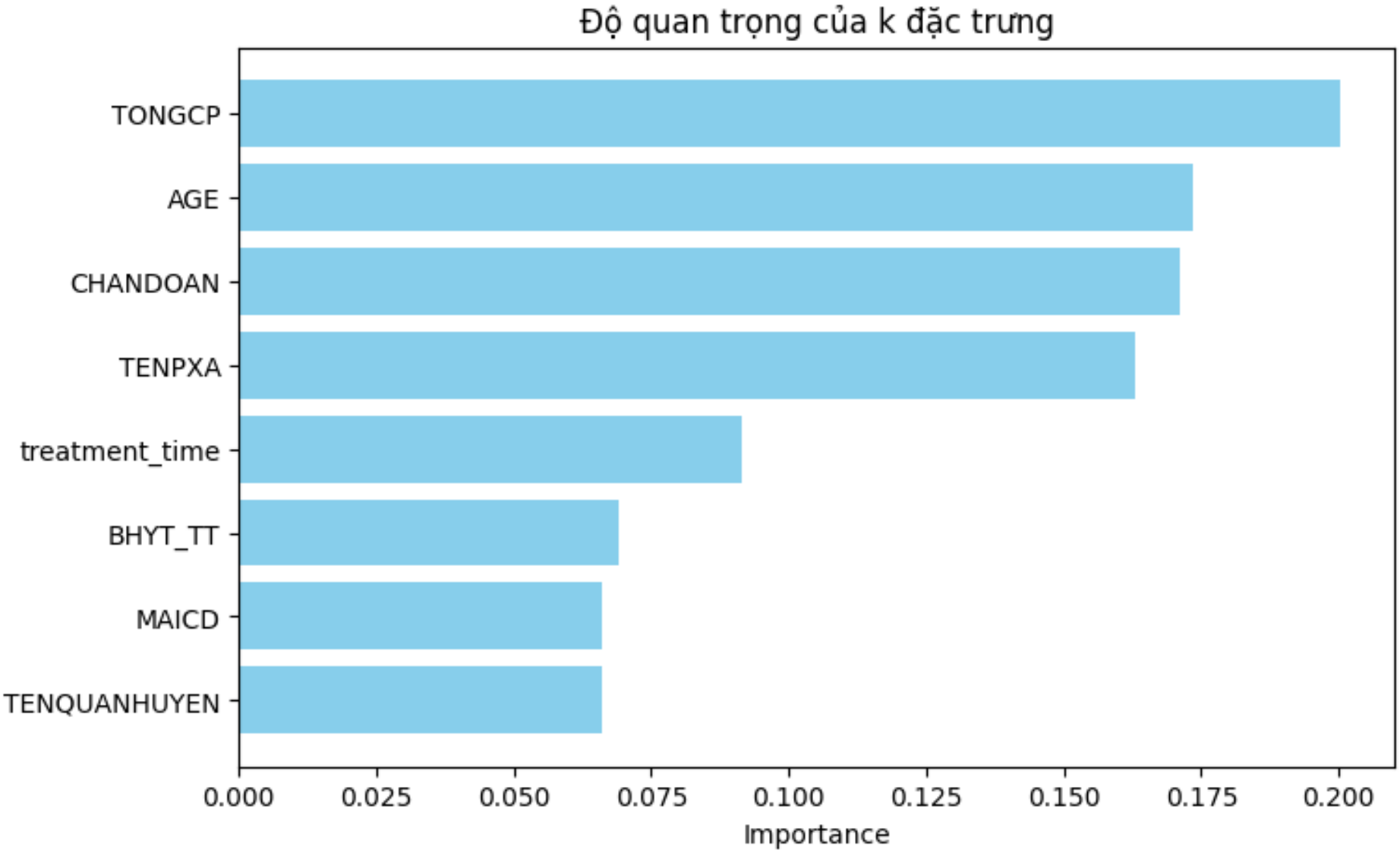
- Sử dụng mô hình `DecisionTreeClassifier` để tính độ quan trọng của từng đặc trưng.
- Chọn ra 8 đặc trưng quan trọng nhất phục vụ huấn luyện mô hình.

Top 8 đặc trưng quan trọng:

Đặc trưng	Độ quan trọng
TONGCP	0.1991
AGE	0.1714
CHANDOAN	0.1703
TENPXA	0.1596
treatment_time	0.0904
BHYT_TT	0.0679
MAICD	0.0653
TENQUANHUYEN	0.0634
TENTINHTHANH	0.0103
DANTOC	0.0023

6. Lựa chọn/trích xuất đặc trưng

Top 8 đặc trưng quan



7. Huấn luyện và Đánh giá mô hình

7.1 Khởi tạo và huấn luyện mô hình Decision Tree

- Sử dụng `DecisionTreeClassifier` với `random_state=42` để đảm bảo tính tái lập và `criterion='entropy'` nhằm tối ưu hóa việc phân chia nhánh dựa trên chỉ số entropy.
- Huấn luyện mô hình trên tập dữ liệu `x_train` (8 đặc trưng được chọn) và `y_train` (biến mục tiêu `Mùa`).
- Quá trình huấn luyện giúp mô hình học các mẫu và quy luật trong dữ liệu, tạo cơ sở cho việc dự đoán mùa xảy ra bệnh một cách chính xác.

7. Huấn luyện và Đánh giá mô hình

7.2 Đánh giá mô hình

- Sử dụng `classification_report` để tính `precision`, `recall` và `f1-score` cho từng lớp (mùa: 0, 1, 2, 3) dựa trên dự đoán (`y_pred`) và giá trị thực (`y_test`).
- **Kết quả theo mùa:**
  - Mùa xuân (0): precision 0.73, recall 0.70, f1-score 0.72 (9,838 mẫu).
  - Mùa hạ (1): precision 0.68, recall 0.73, f1-score 0.70 (8,291 mẫu).
  - Mùa thu (2): precision 0.73, recall 0.72, f1-score 0.72 (9,818 mẫu).
  - Mùa đông (3): precision 0.66, recall 0.66, f1-score 0.66 (3,525 mẫu).
- Độ chính xác tổng thể (accuracy): **71%**, tính bằng `accuracy_score`.
- Chỉ số trung bình: macro avg = 0.70, weighted avg = 0.71 $\Rightarrow$ hiệu suất ổn định giữa các lớp.

7. Huấn luyện và Đánh giá mô hình

7.3 Ma trận nhầm lẫn

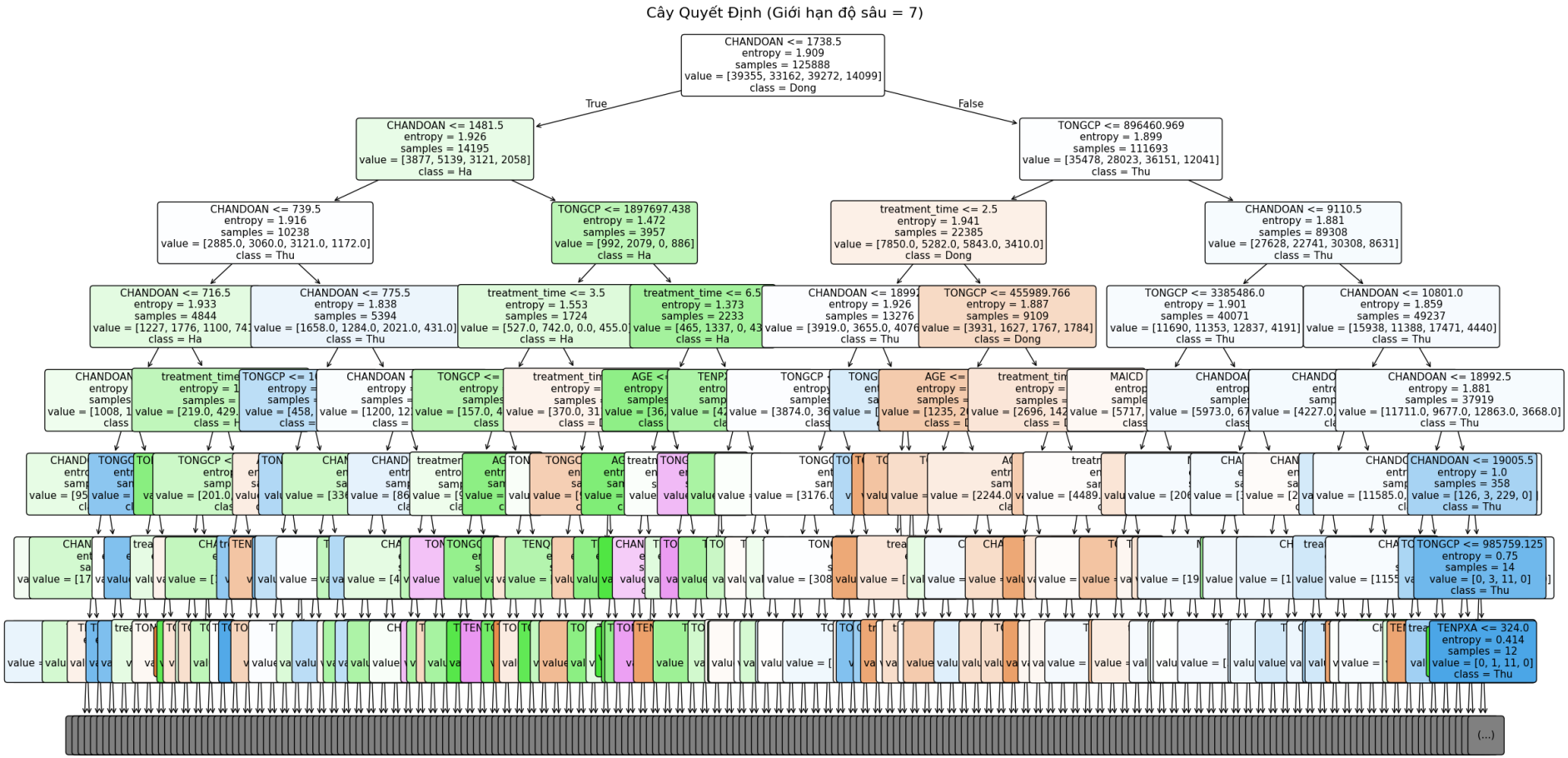
- Sử dụng ma trận nhầm lẫn để đánh giá hiệu quả mô hình phân loại bệnh theo mùa.

Ma trận nhầm lẫn:

$$\begin{bmatrix} 6860 & 1151 & 1293 & 534 \\ 792 & 6043 & 1082 & 374 \\ 1160 & 1302 & 7055 & 301 \\ 524 & 408 & 267 & 2326 \end{bmatrix}$$

- Các giá trị trên đường chéo chính (6860, 6043, 7055, 2326) là số mẫu được dự đoán đúng theo mùa.
- Các giá trị ngoài đường chéo thể hiện sự nhầm lẫn giữa các mùa, ví dụ: 1293 mẫu mùa xuân (0) bị nhầm sang mùa thu (2).
- Mô hình dự đoán tốt nhất ở mùa xuân (0) và mùa thu (2), nhưng vẫn có nhầm lẫn giữa mùa hạ(1) và mùa đông(3).

8. Phân tích cấu trúc cây quyết định



8. Phân tích cấu trúc cây quyết định

- Cây quyết định sử dụng các đặc trưng quan trọng như **CHANDOAN**, **TONGCP**, **treatment_time**, và **AGE** để phân chia dữ liệu.
- Các nút ở độ sâu thấp (ví dụ: tầng 0–4) có chỉ số **entropy cao** (trên 1.5) → dữ liệu còn hỗn loạn, phân bố chưa rõ giữa các mùa.
- Ở độ sâu lớn hơn (ví dụ: tầng 7), entropy giảm xuống còn khoảng **0.4** như ở ví dụ:
  - `TENPXA <= 324.0` với entropy = 0.414, cho thấy dữ liệu tại nút này đã gần như thuần (11/12 mẫu thuộc mùa thu).
- Điều này chứng minh: **càng đi sâu vào cây, sự phân chia càng rõ ràng và hiệu quả hơn**, tuy nhiên cây sẽ trở nên rối và khó quan sát nếu hiển thị toàn bộ khi độ sâu > 10.
- Để trực quan, cây được giới hạn ở độ sâu = 7. Trong thực tế, có thể tiếp tục mở rộng để tăng độ chính xác nếu cần và nếu chấp nhận đánh đổi khả năng diễn giải.

9. Kết luận

- Mô hình cây quyết định (`Decision Tree`) đã được xây dựng thành công để phân loại bệnh theo mùa dựa trên dữ liệu điều trị năm 2016.
- Kết quả đánh giá cho thấy:
  - Độ chính xác tổng thể (**Accuracy**) đạt **71%**.
  - Chỉ số trung bình `precision`, `recall`, `f1-score` đều đạt khoảng **0.70**, cho thấy hiệu suất khá ổn định giữa các mùa.
  - Mô hình hoạt động tốt nhất với các lớp có nhiều mẫu (xuân, hạ, thu).
- Ma trận nhầm lẫn cho thấy phần lớn dự đoán đúng tập trung ở đường chéo, tuy nhiên vẫn còn một số nhầm lẫn giữa các mùa như xuân – thu, hạ – thu.
- Mô hình có tiềm năng hỗ trợ các cơ sở y tế dự báo xu hướng bệnh theo mùa, từ đó chủ động phòng ngừa và phân bổ nguồn lực hiệu quả.
- Định hướng cải thiện:
  - Áp dụng các mô hình mạnh hơn như `Random Forest`, `Gradient Boosting` để tăng độ chính xác và khả năng tổng quát.

Xin cảm ơn Cô và các bạn đã dành thời gian lắng nghe phần trình bày của nhóm em!