ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH THĂM DÒ DỮ LIỆU DỰ ĐOÁN BỆNH ĐỘT QUỴ

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Đỗ Phạm Phúc Tính	20522020
2	Cao Đình Duy Ngọc	20521661
3	Nguyễn Trần Gia Thế	20521940

TP. HỒ CHÍ MINH – 12/2022

MỤC LỤC

1. GIỚI THIỆU

Bộ dữ liệu Patient Characteristics Survey (PCS): 2017 chứa dữ liệu của nhiều người về tình trạng sức khoẻ và các thông tin của họ. Chúng em sử dụng bộ dữ liệu này dùng để phân tích và dự đoán khả năng mắc bệnh đột quỵ của một người từ những thông tin mà họ đã ghi trong khảo sát.

Để thực hiện đề tài này, chúng em sử dụng những thư viện trong python như Sklearn, Numpy, Pandas, Matplotlib, Seaborn, Streamlit, imblearn. Các thuật toán để dự đoán khả năng bệnh mắc bệnh đột quỵ bằng các thuộc tính bao gồm Logistic Regression, SVM, Random Forest, Naïve Bayes.

Sau khi thực hiện đồ án này, chúng em đã phân tích và tìm ra được các thuộc tính có ảnh hưởng đến khả năng mắc bệnh đột quỵ bao gồm: Sex, Age Group, No Chronic Med Condition. Ngoài ra, chúng em còn cài đặt các mô hình học máy để tìm ra mô hình phù hợp nhất với bộ dữ liệu của chúng em đã thu thập, kết quả cho thấy cả 3 mô hình Logistic Regression, SVM, Random Forest, có kết quả cao nhất và tương đương nhau với 75.19% với độ đo accuracy và 77.19% với độ đo f1.

2. NỘI DUNG

2.1 Mô tả bộ dữ liệu

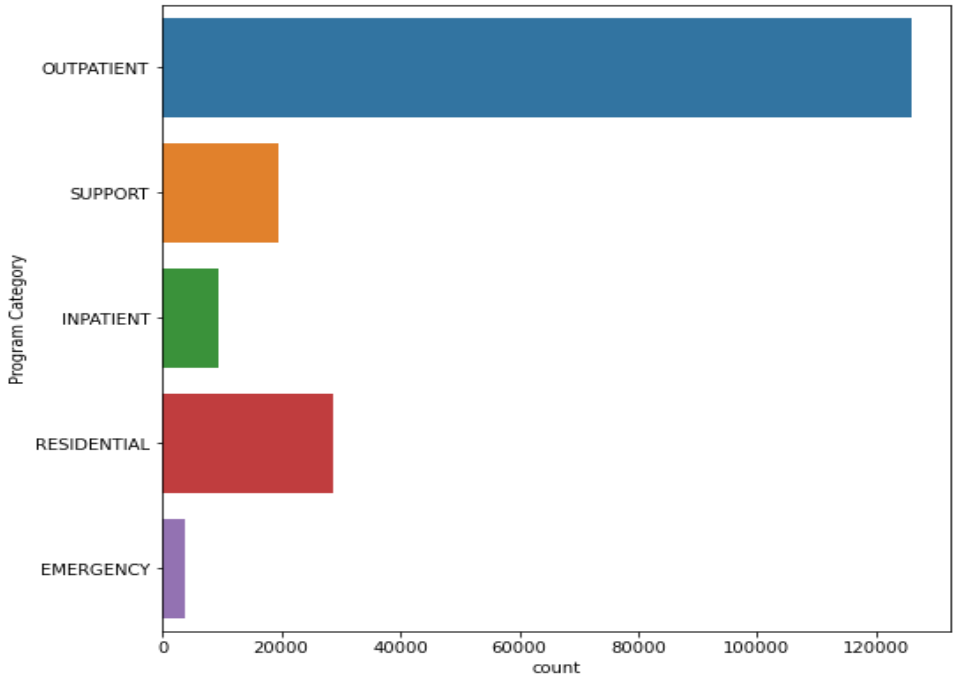
Bộ dữ liệu chúng em sử dụng là bộ dữ liệu Patient Characteristics Survey (PCS): 2017 được thu thập từ trang web data.world. Bộ dữ liệu gồm 187192 dòng và 67 thuộc tính. Trong đó có 65 thuộc tính phân loại, 2 thuộc tính kiểu số và thuộc tính mục tiêu là “Stroke”.

Bộ dữ liệu không có giá trị bị khuyết, tuy nhiên lại có những giá trị "UNKNOWN", "CLIENT DIDN'T ANSWER", "CLIENT DID NOT ANSWER”, “UNKNOWN RACE”, “UNKNOWN EMPLOYMENT STATUS", "NOT APPLICABLE", "UNKNOWN EMPLOYMENT HOUR”.

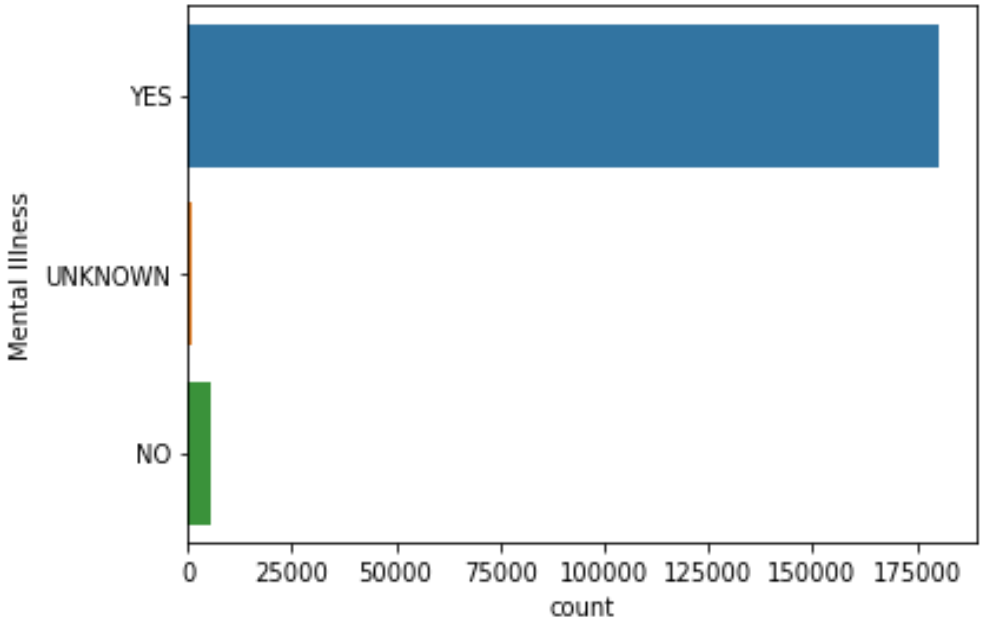
	Survey Year	Program Category	Region Served	Age Group	Sex	Transgender	Sexual Orientation	Hispanic Ethnicity	Race	Living Situation	...	Unknown Insurance Coverage	No Insurance	Mec Insu
133784	2017	OUTPATIENT	NEW YORK CITY REGION	ADULT	MALE	NO, NOT TRANSGENDER	BISEXUAL	NO, NOT HISPANIC/LATINO	UNKNOWN RACE	PRIVATE RESIDENCE	...	NO	NO	
19582	2017	EMERGENCY	LONG ISLAND REGION	ADULT	FEMALE	UNKNOWN	UNKNOWN	NO, NOT HISPANIC/LATINO	BLACK ONLY	PRIVATE RESIDENCE	...	UNKNOWN	YES	UNK
41477	2017	OUTPATIENT	LONG ISLAND REGION	ADULT	MALE	NO, NOT TRANSGENDER	STRAIGHT OR HETEROSEXUAL	NO, NOT HISPANIC/LATINO	WHITE ONLY	PRIVATE RESIDENCE	...	NO	NO	
130607	2017	OUTPATIENT	NEW YORK CITY REGION	CHILD	MALE	NO, NOT TRANSGENDER	UNKNOWN	NO, NOT HISPANIC/LATINO	BLACK ONLY	PRIVATE RESIDENCE	...	NO	NO	
130418	2017	RESIDENTIAL	NEW YORK CITY REGION	ADULT	MALE	NO, NOT TRANSGENDER	STRAIGHT OR HETEROSEXUAL	NO, NOT HISPANIC/LATINO	WHITE ONLY	PRIVATE RESIDENCE	...	NO	NO	

5 rows × 67 columns

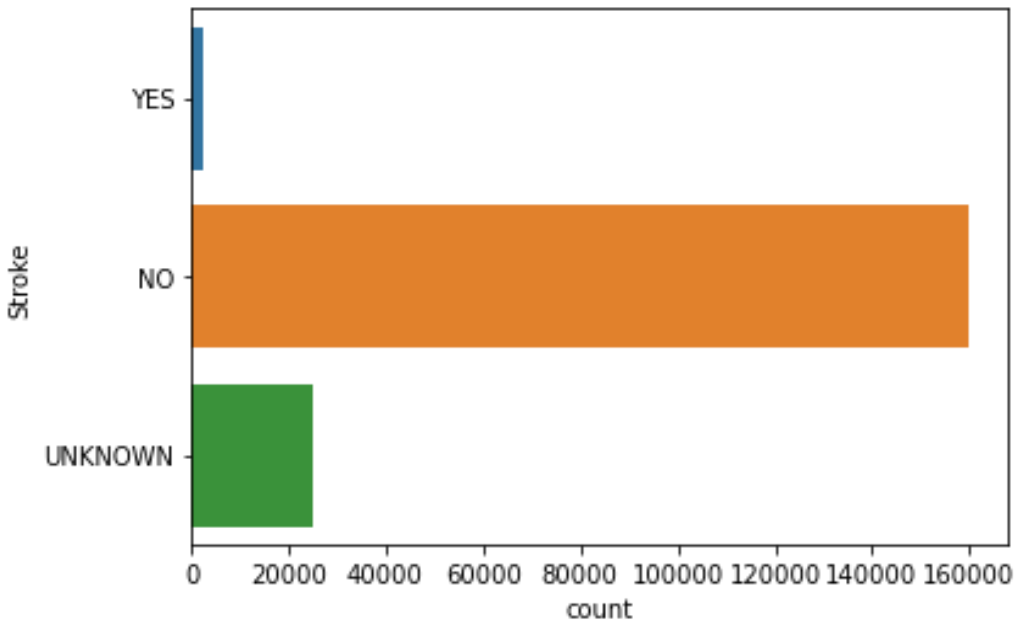
Hình 1. 5 mẫu ngẫu nhiên có trong bộ dữ liệu



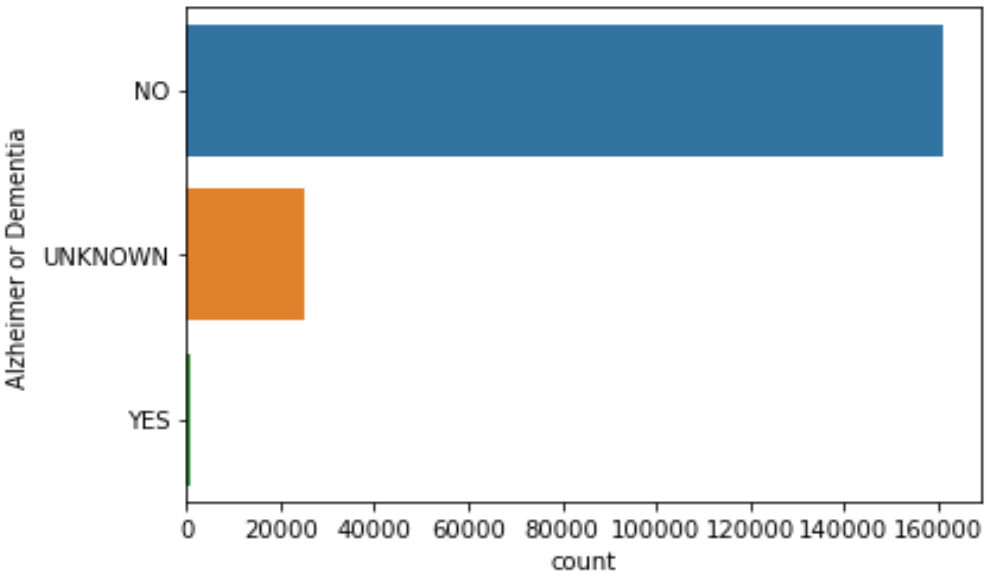
Hình 2. Biểu đồ thể hiện tần suất của các giá trị trong thuộc tính Program Category



Hình 3. Biểu đồ thể hiện tần suất của các giá trị trong thuộc tính Mental Illness.



Hình 4. Biểu đồ thể hiện tần suất của các giá trị trong thuộc tính Stroke

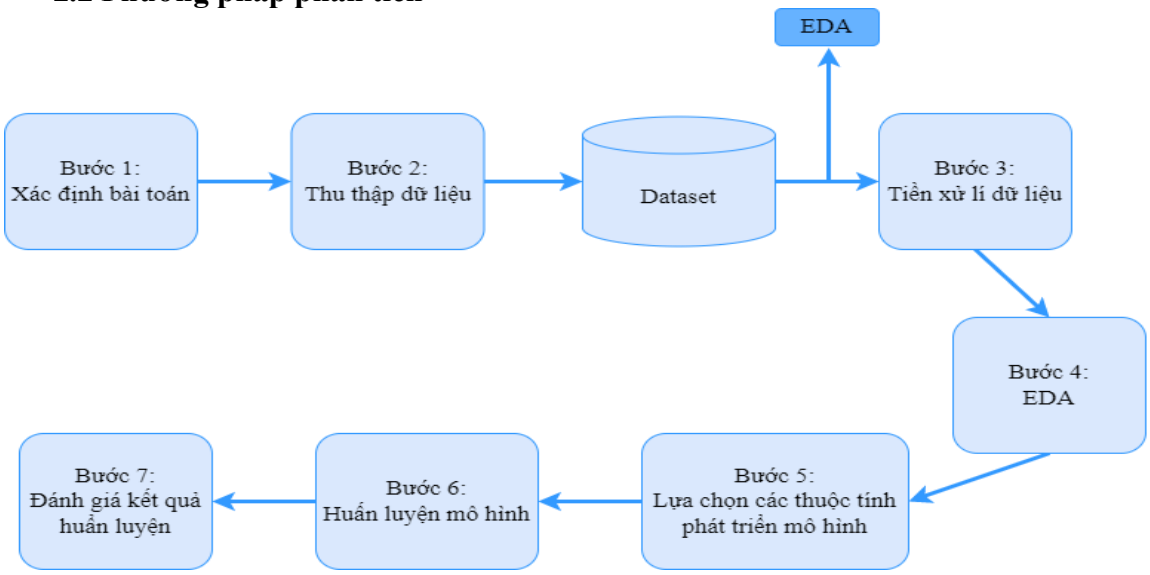


Hình 5. Biểu đồ thể hiện tần suất của các giá trị trong thuộc tính Alzheimer or Dementia

- Từ các hình 2,3,4,5 có thể các các giá trị của mỗi thuộc tính đều có sự chênh lệch rất về số lượng.
- Đối với thuộc tính "Program Category": nhãn OUTPATIENT xuất hiện nhiều nhất với 126,064 lần, trong khi 4 nhãn còn lại tần suất ít hơn rất nhiều lần: RESIDENTIAL là 28,734 lần, SUPPORT (19,411 lần), INPATIENT (9,282), EMERGENCY (3,701).
- Đối với thuộc tính "Stroke": nhãn NO chiếm đa số với 159830 lần xuất hiện, có tỉ lệ 85.39%, nhiều hơn nhiều lần so với 2 nhãn còn lại. Nhãn UNKNOWN (25,126), nhãn YES (2236).
- Với thuộc tính Mental Illness: nhãn YES chiếm đa số với tỉ lệ 18,0346/187,192 (96.34%), nhãn NO ít hơn nhiều lần với 5,655/187,192, nhãn UNKNOWN thấp nhất với tỉ lệ 1,191/187,192.

- Còn ở thuộc tính Alzheimer or Dementia: nhãn NO xuất hiện nhiều nhất, có tỉ lệ 161,026/187,192 (86.02%) tiếp theo là nhãn UNKNOWN 25,126/187,292 và thấp nhất là nhãn YES (1,040/187,192).

2.2 Phương pháp phân tích



Hình 6. Quy trình thực hiện đồ án

2.3 Xử lý bộ dữ liệu

2.3.1 Xử lý dữ liệu UNKNOW

- Các giá trị "UNKNOWN", "CLIENT DIDN'T ANSWER", "CLIENT DID NOT ANSWER", "UNKNOWN RACE", "UNKNOWN EMPLOYMENT STATUS", "NOT APPLICABLE", "UNKNOWN EMPLOYMENT HOUR" được chuyển thành các giá trị NaN.
- Xoá các thuộc tính có tỉ lệ giá trị NaN >70%. Sau khi xoá, bộ dữ liệu còn 187.192 dòng và 65 thuộc tính.
- Xoá tất cả các dòng có giá trị NaN sau đó xoá các cột chỉ có 1 giá trị. Sau bước này bộ dữ liệu còn 37.468 dòng và 59 thuộc tính.

2.3.2 Chuẩn hoá dữ liệu

- Đối với các biến phân loại kiểu object, chúng em dùng thuật toán LabelEncoder để chuyển các giá trị thành các con số.

2.4 Phân tích thăm dò

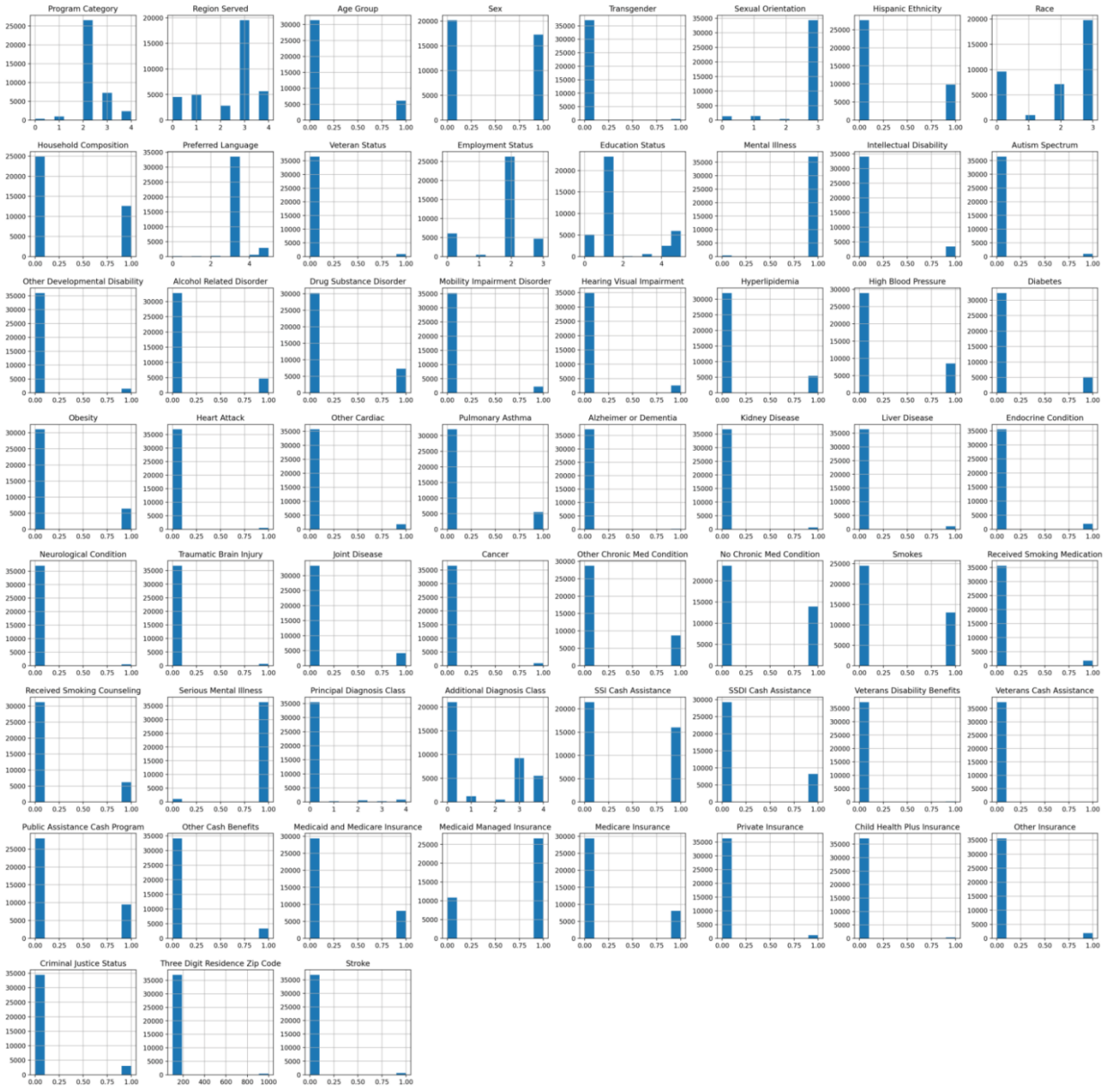
2.4.1 Thống kê mô tả

- Trong bộ dữ liệu đã xử lý, thuộc tính "Three Digit Residencce Zip Code" là biến liên tục, các thuộc tính còn lại là biến phân loại.

Bảng 1. Thống kê mô tả của thuộc tính "Three Digit Residence Zip Code"

	Three Digit Residence Zip Code
count	37468
mean	127.773

std	93.236
min	100
25%	104.750
50%	113.000
75%	130.000
max	999



Hình 7. Biểu đồ thể hiện tần suất của các giá trị trong các thuộc tính sau khi xử lý.

2.4.2 Trực quan biến phân loại

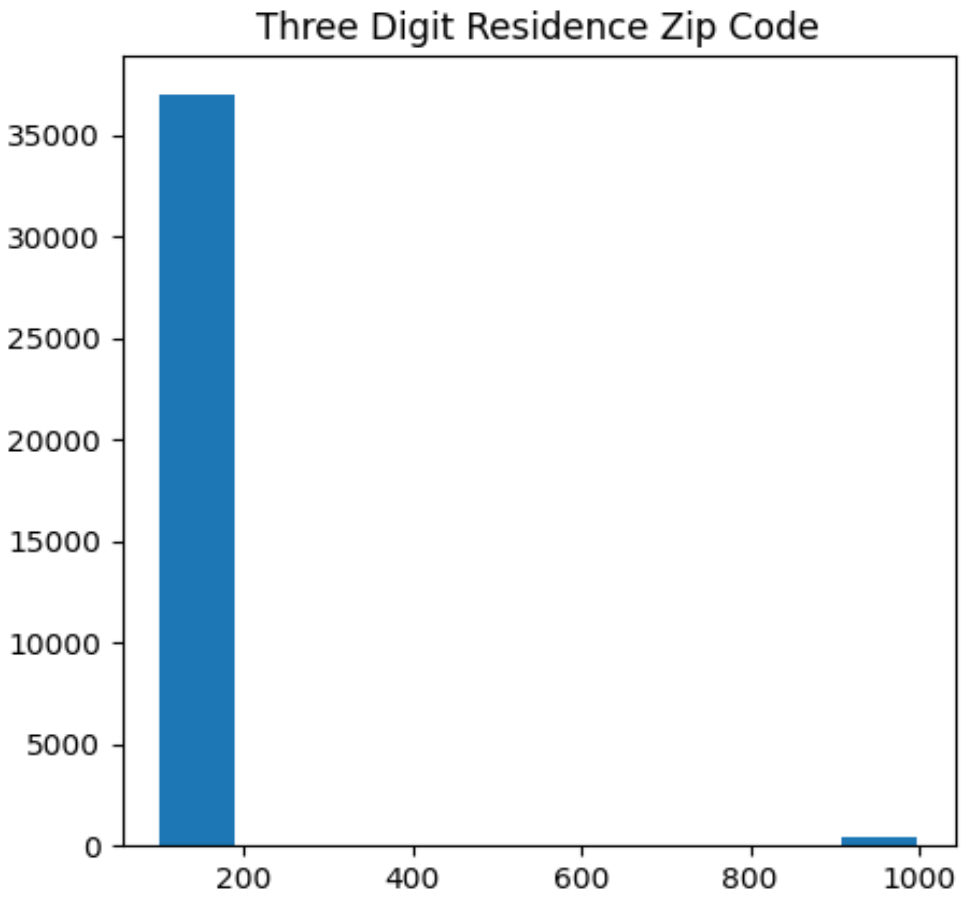
Có thể thấy với các thuộc tính chỉ có 2 giá trị 0-1 thì phần lớn có sự chênh lệch nhiều về số lượng giữa 2 giá trị, số lượng nhãn 0 lớn gấp nhiều lần nhãn 1.

Riêng thuộc tính Sex, Household Composition, Smokes, No Chronic Med Condition, SSI Cash Assistance sự chênh lệch giữa nhãn 0 và nhãn 1 tương đối, nhỏ hơn 2 lần.

Đối với thuộc tính Serious Mental Illness và Mental Illness thì ngược lại, số lượng nhãn 1 nhiều hơn gấp nhiều lần so với nhãn 0 (Nhãn YES nhiều hơn nhãn NO)

Đối với thuộc tính có nhiều nhãn, sự chênh lệch giữa các nhãn vẫn có sự chênh lệch khá lớn. Thuộc tính Program Category thì số lượng nhãn 2 (OUTPATIENT nhiều nhất, gấp nhiều lần số lượng các nhãn khác), thuộc tính Region Served, nhãn 3 (NEW YORK CITY REGION) chiếm số lượng nhiều nhất và đa số. Thuộc tính Sexual Orientation, nhãn 3 (STRAIGHT OR HETEROSEXUAL) chiếm đa số. Thuộc tính Race, nhãn 3 (WHITE ONLY) chiếm phần lớn, tuy nhiên sự chênh lệch so với nhãn 0 và 2 không nhiều bằng chênh lệch ở các thuộc tính khác. Thuộc tính Prefered Language, nhãn 3 (ENGLISH) chiếm nhiều nhất, có sự chênh lệch rất nhiều lần so với các nhãn còn lại. Thuộc tính Employed Status, nhãn 2 (NOT IN LABOR FORCE: UNEMPLOYED AND NOT LOOKING FOR WORK) chiếm đa số. Đối với các thuộc tính Education Status, Principal Diagnosis Class, Additional Diagnosis Class, các nhãn chiếm đa số lần lượt các với thuộc tính là: 1 (MIDDLE SCHOOL TO HIGH SCHOOL), 0 (MENTAL ILLNESS), 0 (MENTAL ILLNESS).

2.4.3 Trực quan biến liên tục



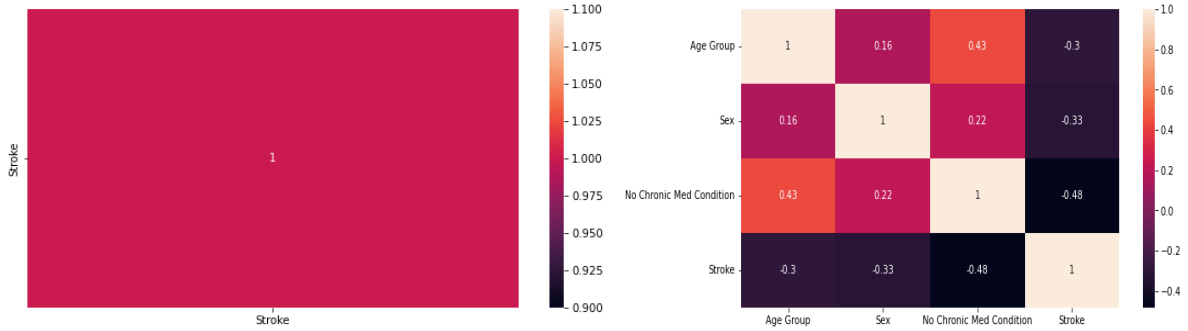
Hình 8. Biểu đồ thể hiện tần suất của các giá trị
trong thuộc tính “Three Digit Residence Zip Code” khi xử lý.

Có thể thấy giá trị tập trung nhiều nhất ở giá trị dưới 150, chiếm 98,90% trong khi giá trị 999 chiếm 1,1% tương ứng với 414 giá trị trong tổng số 37.468 giá trị có trong bộ dữ liệu sạch.

2.4.4 Độ tương quan.

Với bộ dữ liệu 37,468 dòng và 59 cột, chúng em tính độ tương quan giữa các thuộc tính so với thuộc tính "Stroke" thì nhận ra không có thuộc tính nào ảnh hưởng đến thuộc tính "Stroke" (hệ số tương quan >=0.3). Điều này được giải thích như sau: do dữ liệu bị mất cân bằng ở hầu như ở tất cả thuộc tính, ở mỗi giá trị của các thuộc tính tương ứng các giá trị 0 -1 ở biến "Stroke" cũng bị mất cân bằng, dẫn đến việc không có thuộc tính nào ảnh hưởng đến thuộc tính "Stroke". Điều này cũng sẽ làm cho không có thuộc tính nào để thực nghiệm các mô hình. Do đó chúng em thực hiện phương pháp cân bằng dữ liệu.

Phương pháp cân bằng dữ liệu được chúng em sử dụng là phương pháp SMOTE. Phương pháp này sẽ giúp tăng số lượng nhãn ít hơn. Tỉ lệ nhãn 0-1 sau khi dược chúng em tăng nhãn ít hơn là 1:1. Cụ thể, nhãn 0 (NO) – 1 (YES) là 36855:36855. Như vậy, bộ dữ liệu sau khi được cân bằng gồm 73,710 dòng và 59 thuộc tính. Chúng em thực hiện tính lại độ tương quan giữa các thuộc tính so với thuộc tính "Stroke" và lấy giá trị hệ số tương quan >=0.3.



Hình 9. Ma trận tương quan trước (trái) và sau (phải) khi cân bằng.

Từ 2 hình trên có thể thấy, sau khi thực hiện phương pháp cân bằng dữ liệu, độ tương quan giữa các thuộc tính đã được cải thiện, bằng chứng là xuất hiện 3 thuộc tính có ảnh hưởng đến thuộc tính "Stroke". 3 thuộc tính này cũng sẽ là 3 thuộc tính để chúng em thực hiện thí nghiệm các mô hình dự đoán khả năng bị đột quỵ.

2.5 Thực nghiệm mô hình

2.5.1 Các mô hình huấn luyện.

- Bài toán của chúng em thuộc dạng bài toán phân loại. Do đó, Chúng em lựa chọn các mô hình phân loại Logistic Regression, SVM, Naive Bayes và Random Forest.
- Các độ đo đánh giá được chúng em lựa chọn bao gồm accuracy_score, f1_score.

2.5.2 Bộ dữ liệu huấn luyện.

Sau khi xác định được các thuộc tính có ảnh hưởng tới thuộc tính phụ thuộc "Stroke", chúng em tạo một bảng dữ liệu gồm 3 thuộc tính độc lập cùng với thuộc tính phụ thuộc "Stroke" gồm 73,710 dòng và 4 cột. Sau đó, dùng thư viện train_test_split để chia tập dữ liệu 73,710 dòng và 4 thuộc tính thành 2 bộ: huấn luyện và kiểm thử.

Bảng 2. Kích thước các bộ dữ liệu sau khi được chia

Bộ dữ liệu	Kích thước	Tỉ lệ nhãn 0: 1 của cột Stroke
Huấn luyện	58968 dòng, 4 cột	0: 29495 1: 29473
Kiểm thử	14742 dòng, 4 cột	0: 7393 1: 7349

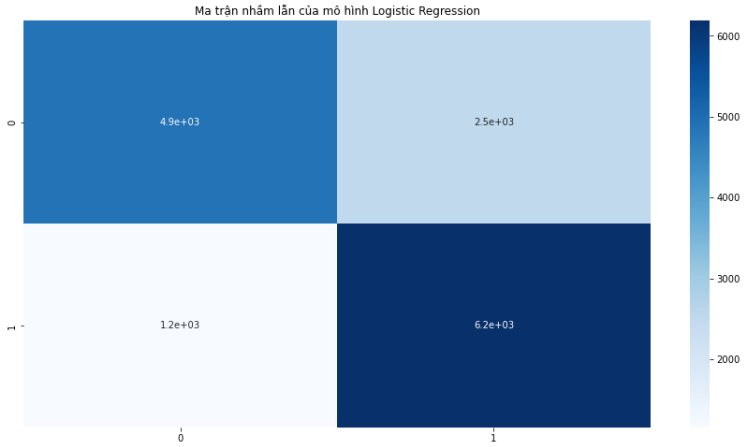
2.5.3 Huấn luyện và đánh giá kết quả

Chúng em tiến hành cài đặt các mô hình Logistic Regression, SVM, Naïve Bayes và Random Forest để huấn luyện bộ dữ liệu huấn luyện đã được chuẩn bị trước đó. Cuối cùng kiểm tra kết quả trên tập kiểm thử. Bảng 3 thể hiện kết quả của mô hình sau khi huấn luyện và kiểm thử.

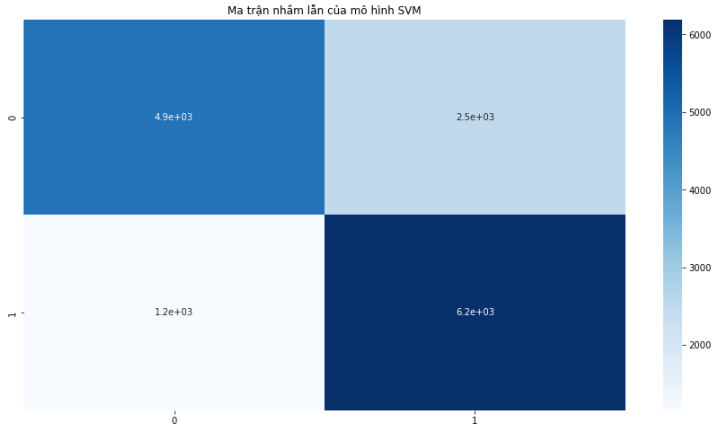
Bảng 3. Kết quả của các mô hình trên bộ dữ liệu kiểm thử

Mô hình	Accuracy_score	F1_score
Logistic Regression	75.19%	77.19%
SVM	75.19%	77.19%
Naïve Bayes	70.05%	77.19%
Random Forest	75.19%	77.19%

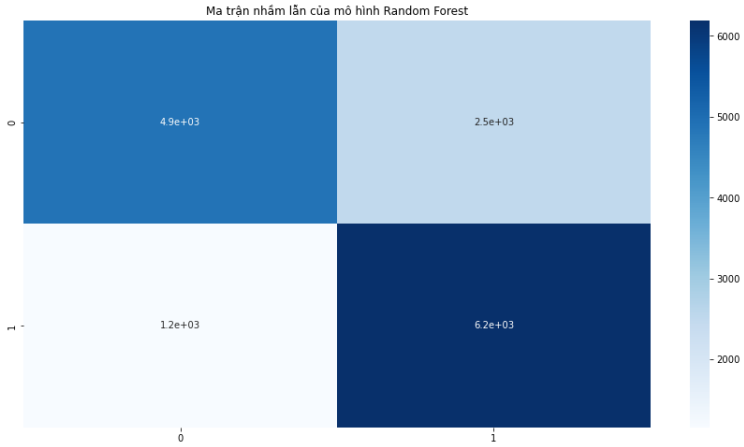
Từ bảng 3 cho thấy, Mô hình tốt nhất trong các mô hình mà chúng em thí nghiệm khi sử dụng các thuộc tính Age Group, Sex, No Chronic Med Condition để dự đoán khả năng mắc bệnh đột quỵ bao gồm Logistic Regression, SVM, Random Forest với độ đo F1 là 77.19%, accuracy là 75.19%. Tuy nhiên, các mô hình cho kết quả dự đoán có độ chính xác không cao và không có sự chênh lệch nhiều. Do các thuộc tính này có ảnh hưởng tới thuộc tính “Stroke” không cao, hệ số tương quan của thuộc tính Sex đối với thuộc tính “Stroke” là -0.33, thuộc tính No Chronic Med Condition là -0.48, thuộc tính Age Group là -0.3, đều là hệ số tương quan thấp, sự ảnh hưởng yếu. Từ đó có thể kết luận, sử dụng các thuộc tính Age Group, Sex, No Chronic Med Condition để dự đoán khả năng mắc bệnh đột quỵ đạt hiệu quả không tốt.



Hình 7. Ma trận nhầm lẫn của mô hình Logistic Regression

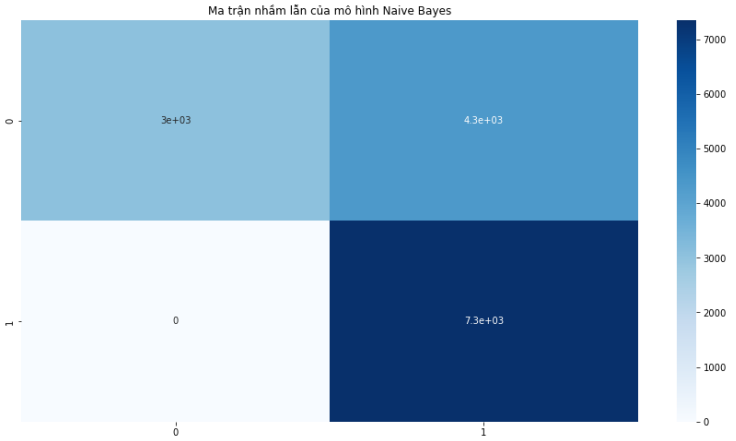


Hình 8. Ma trận nhầm lẫn của mô hình SVM



Hình 9. Ma trận nhầm lẫn của mô hình Random Forest

Đối với 3 mô hình Logistic Regression, SVM, Random Forest, các mô hình dự đoán nhãn 1 đúng nhiều hơn nhãn 0. Tỉ lệ sự đoán sai của nhãn 0 cũng cao hơn nhãn 1, tỉ lệ đoán sai của nhãn 0 là 33.82%, nhãn 1 là 15.76%.



Hình 10. Ma trận nhầm lẫn của mô hình Naïve Bayes

Mô hình dự đoán nhãn 1 đúng hoàn toàn 100% toàn bộ nhãn 1 có trong bộ dữ liệu, trong khi nhãn 0 mô hình dự đoán nhầm sang nhãn 1 rất nhiều, tỉ lệ nhãn 0 được mô hình dự đoán sang nhãn 1 là hơn 4343, chiếm hơn 58.74%. Cho thấy, mô hình dự đoán nhãn 1 rất chuẩn.

3. KẾT LUẬN

Trong quá trình làm việc, nhóm em đã tiến hành các việc như sau:

- Xử lý các giá trị "UNKNOWN", "CLIENT DIDN'T ANSWER", "CLIENT DID NOT ANSWER", "UNKNOWN RACE", "UNKNOWN EMPLOYMENT STATUS", "NOT APPLICABLE", "UNKNOWN EMPLOYMENT HOUR".
- Xoá các thuộc tính có tỉ lệ giá trị NaN >70%.
- Xoá tất cả các dòng có giá trị NaN sau đó xoá các cột chỉ có 1 giá trị.
- Dùng thuật toán LabelEncoder chuẩn hoá các giá trị phân loại trong bộ dữ liệu
- Phân tích thăm dò và trực quan các thuộc tính của bộ dữ liệu.
- Tìm ra được các thuộc tính có ảnh hưởng đến thuộc tính phụ thuộc "Stroke".
- Huấn luyện mô hình: áp dụng các mô hình Logistic Regression, SVM, Naïve Bayes, Random Forest để thực hiện huấn luyện trên bộ dữ liệu, dùng các độ đo accuracy, f1 và ma trận nhầm lẫn để đánh giá kết quả của các mô hình. Kết quả đạt được, các mô hình bao gồm Logistic Regression, SVM, Random Forest với độ đo F1 là 77.19%, accuracy là 75.19%.

TÀI LIỆU THAM KHẢO

[1] "data.world," [Trực tuyến]. Available: https://data.world/data-ny-gov/8itk-gcdy/workspace/file?filename=patient-characteristics-survey-pcs-2017-1.csv.

[2] "Scikit-learn," [Trực tuyến]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[3] "Scikit-learn," [Trực tuyến]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[4] "Scikit-learn," [Trực tuyến]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html.

[5] "Scikit-learn," [Trực tuyến]. Available: https://numpy.org/.

[6] "Scikit-learn," [Trực tuyến]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Đỗ Phạm Phúc Tính	Phân công nhiệm vụ, Thu thập dữ liệu, Code xử lý và phân tích dữ liệu, Viết báo cáo
2	Cao Đình Duy Ngọc	Code Demo, Code xử lý và phân tích dữ liệu, Sửa source code
3	Nguyễn Trần Gia Thế	Code xử lý và phân tích dữ liệu, Làm slide, Viết báo cáo,

PHỤ LỤC

Bảng 4. Danh sách các thuộc tính của bộ dữ liệu Patient Characteristics Survey (PCS): 2017

STT	Tên thuộc tính	Mô tả	Kiểu dữ liệu	Miền dữ liệu
1	Survey Year	Năm khảo sát	Numeric	[2017]
2	Program Category	Loại chương trình	Categorical	['EMERGENCY', 'INPATIENT', 'OUTPATIENT', 'RESIDENTIAL', 'SUPPORT']
3	Region Served	Khu vực phục vụ	Categorical	['CENTRAL NY REGION', 'HUDSON RIVER REGION', 'LONG ISLAND REGION', 'NEW YORK CITY REGION', 'WESTERN REGION']
4	Age Group	Nhóm tuổi	Categorical	['ADULT', 'CHILD', 'UNKNOWN']
5	Sex	Giới tính	Categorical	['FEMALE', 'MALE', 'UNKNOWN']
6	Transgender	Chuyển giới	Categorical	['CLIENT DIDN'T ANSWER', 'NO, NOT TRANSGENDER', 'UNKNOWN', 'YES, TRANSGENDER']
7	Sexual Orientation	Khuynh hướng tính dục	Categorical	['BISEXUAL', 'CLIENT DID NOT ANSWER', 'LESBIAN OR GAY', 'OTHER', 'STRAIGHT OR HETEROSEXUAL', 'UNKNOWN']
8	Hispanic Ethnicity	Dân tộc Tây Ban Nha	Categorical	['NO, NOT HISPANIC/LATINO', 'UNKNOWN', 'YES, HISPANIC/LATINO']
9	Race	Chủng tộc	Categorical	['BLACK ONLY', 'MULTI-RACIAL', 'OTHER', 'UNKNOWN RACE', 'WHITE ONLY']
10	Living Situation	Điều kiện sống	Categorical	['INSTITUTIONAL SETTING', 'OTHER LIVING SITUATION',

				'PRIVATE RESIDENCE', 'UNKNOWN']
11	Household Composition	Thành phần hộ gia đình	Categorical	['COHABITATES WITH OTHERS', 'LIVES ALONE', 'NOT APPLICABLE', 'UNKNOWN']
12	Preferred Language	Ngôn ngữ ưa thích	Categorical	['AFRO-ASIATIC', 'ALL OTHER LANGUAGES', 'ASIAN AND PACIFIC ISLAND', 'ENGLISH', 'INDO-EUROPEAN', 'SPANISH', 'UNKNOWN']
13	Veteran Status	Cựu chiến binh	Categorical	['NO' 'UNKNOWN' 'YES']
14	Employment Status	Tình trạng việc làm	Categorical	['EMPLOYED', 'NON-PAID/VOLUNTEER', 'NOT IN LABOR FORCE : UNEMPLOYED AND NOT LOOKING FOR WORK', 'UNEMPLOYED, LOOKING FOR WORK', 'UNKNOWN EMPLOYMENT STATUS']
15	Number Of Hours Worked Each Week	Số giờ làm việc mỗi tuần	Categorical	['01-14 HOURS', '15-34 HOURS', '35 HOURS OR MORE', 'NOT APPLICABLE', 'UNKNOWN EMPLOYMENT HOURS']
16	Education Status	Học vấn	Categorical	['COLLEGE OR GRADUATE DEGREE', 'MIDDLE SCHOOL TO HIGH SCHOOL', 'NO FORMAL EDUCATION', 'OTHER' 'PRE-K TO FIFTH GRADE', 'SOME COLLEGE', 'UNKNOWN']
17	Special Education Services	Giáo dục đặc biệt	Categorical	['NO', 'NOT APPLICABLE', 'UNKNOWN', 'YES']
18	Mental Illness	Bệnh tâm thần	Categorical	['NO', 'UNKNOWN', 'YES']
19	Intellectual Disability	Thiểu năng trí tuệ	Categorical	['NO', 'UNKNOWN', 'YES']
20	Autism Spectrum	Bệnh tự kỉ	Categorical	['NO', 'UNKNOWN', 'YES']
21	Other Developmental Disability	Khuyết tật phát triển khác	Categorical	['NO', 'UNKNOWN', 'YES']
22	Alcohol Related Disorder	Rối loạn liên quan đến cồn	Categorical	['NO', 'UNKNOWN', 'YES']
23	Drug Substance Disorder	Rối loạn chất gây nghiện	Categorical	['NO', 'UNKNOWN', 'YES']
24	Mobility Impairment Disorder	Rối loạn suy giảm vận động	Categorical	['NO', 'UNKNOWN', 'YES']
25	Hearing Visual Impairment	Khiếm thính, khiếm thị	Categorical	['NO', 'UNKNOWN', 'YES']
26	Hyperlipidemia	Mỡ máu cao	Categorical	['NO', 'UNKNOWN', 'YES']
27	High Blood Pressure	Huyết áp cao	Categorical	['NO', 'UNKNOWN', 'YES']
28	Diabetes	Bệnh tiểu đường	Categorical	['NO', 'UNKNOWN', 'YES']
29	Obesity	Béo phì	Categorical	['NO', 'UNKNOWN', 'YES']

30	Heart Attack	Đau tim	Categorical	['NO', 'UNKNOWN', 'YES']
31	Stroke	Đột quỵ	Categorical	['NO', 'UNKNOWN', 'YES']
32	Other Cardiac	Các bệnh tim khác	Categorical	['NO', 'UNKNOWN', 'YES']
33	Pulmonary Asthma	hen suyễn	Categorical	['NO', 'UNKNOWN', 'YES']
34	Alzheimer or Dementia	Alzheimer hoặc mất trí nhớ	Categorical	['NO', 'UNKNOWN', 'YES']
35	Kidney Disease	Suy thận	Categorical	['NO', 'UNKNOWN', 'YES']
36	Liver Disease	Bệnh gan	Categorical	['NO', 'UNKNOWN', 'YES']
37	Endocrine Condition	Nội tiết tố	Categorical	['NO', 'UNKNOWN', 'YES']
38	Neurological Condition	Tình trạng thần kinh	Categorical	['NO', 'UNKNOWN', 'YES']
39	Traumatic Brain Injury	Chấn thương sọ não	Categorical	['NO', 'UNKNOWN', 'YES']
40	Joint Disease	Bệnh khớp	Categorical	['NO', 'UNKNOWN', 'YES']
41	Cancer	Ung thư	Categorical	['NO', 'UNKNOWN', 'YES']
42	Other Chronic Med Condition	Tình trạng mãn tính khác	Categorical	['NO', 'UNKNOWN', 'YES']
43	No Chronic Med Condition	Không có bệnh mãn tính	Categorical	['NO', 'UNKNOWN', 'YES']
44	Unknown Chronic Med Condition	Không xác định bệnh mãn tính	Categorical	['NO', 'YES']
45	Smokes	Hút thuốc	Categorical	['NO', 'UNKNOWN','YES']
46	Received Smoking Medication	Nhận cai thuốc bằng thuốc	Categorical	['NO', 'UNKNOWN','YES']
47	Received Smoking Counseling	Nhận tư vấn cai thuốc lá	Categorical	['NO', 'UNKNOWN','YES']
48	Serious Mental Illness	Bệnh tâm thần nghiêm trọng	Categorical	['NO', 'UNKNOWN','YES']
49	Principal Diagnosis Class	Chuẩn đoán chính	Categorical	['MENTAL ILLNESS' 'NOT MI - DEVELOPMENTAL DISORDERS' 'NOT MI - ORGANIC MENTAL DISORDER' 'NOT MI - OTHER' 'SUBSTANCE-RELATED AND ADDICTIVE DISORDERS' 'UNKNOWN']
50	Additional Diagnosis Class	Chuẩn đoán bổ sung	Categorical	['MENTAL ILLNESS' 'NOT MI - DEVELOPMENTAL DISORDERS' 'NOT MI - ORGANIC MENTAL DISORDER' 'NOT MI - OTHER' 'SUBSTANCE-RELATED AND ADDICTIVE DISORDERS' 'UNKNOWN']
51	SSI Cash Assistance	Hỗ trợ SSI	Categorical	['NO' 'UNKNOWN' 'YES']
52	SSDI Cash Assistance	Hô trợ SSDI	Categorical	['NO' 'UNKNOWN' 'YES']
53	Veterans Disability Benefits	Lợi ích cựu chiến binh	Categorical	['NO' 'UNKNOWN' 'YES']
54	Veterans Cash Assistance	Hỗ trợ tiền cựu chiến binh	Categorical	['NO' 'UNKNOWN' 'YES']

55	Public Assistance Cash Program	Chương trình hỗ trợ tiền	Categorical	['NO' 'UNKNOWN' 'YES']
56	Other Cash Benefits	Các hỗ trợ tiền khác	Categorical	['NO' 'UNKNOWN' 'YES']
57	Medicaid and Medicare Insurance	Bảo hiểm và trợ cấp y tế	Categorical	['NO' 'UNKNOWN' 'YES']
58	Unknown Insurance Coverage	Bảo hiểm không khác định	Categorical	['NO' 'UNKNOWN' 'YES']
59	No Insurance	Không có bảo hiểm	Categorical	['NO' 'YES']
60	Medicaid Insurance	Bảo hiểm y tế trợ cấp	Categorical	['NO' 'UNKNOWN' 'YES']
61	Medicaid Managed Insurance	Bảo hiểm y tế được trợ cấp và quản lí	Categorical	['NO' 'NOT APPLICABLE' 'UNKNOWN' 'YES']
62	Medicare Insurance	Bảo hiểm y tế	Categorical	['NO' 'UNKNOWN' 'YES']
63	Private Insurance	Bảo hiểm tư nhân	Categorical	['NO' 'UNKNOWN' 'YES']
64	Child Health Plus Insurance	Bảo hiểm sức khỏe trẻ em	Categorical	['NO' 'UNKNOWN' 'YES']
65	Other Insurance	Bảo hiểm khác	Categorical	['NO' 'UNKNOWN' 'YES']
66	Criminal Justice Status	Tội phạm hình sự	Categorical	['NO' 'UNKNOWN' 'YES']
67	Three Digit Residence Zip Code	Mã nơi cư trú	Numeric	[100, 999]