CHẨN ĐOÁN ĐÁI THÁO ĐƯỜNG TRÊN NHIỀU MÔ HÌNH

Lê Tuấn Khôi

TABLE OF CONTENTS

01

DATASET OVERVIEW

TỔNG QUAN

Đái tháo đường là một trong những bệnh mạn tính phổ biến nhất ở Hoa Kỳ nói riêng và toàn cầu nói chung. Bệnh đem lại gánh nặng tài chính đáng kể cho nền kinh tế do những tác động xấu, biến chứng lên sức khỏe, chất lượng cuộc sống cũng như tuổi thọ. Đái tháo đường mặc dù hiện tại không thể chữa trị tận gốc nhưng có thể phòng ngừa và chẩn đoán sớm, do đó các các mô hình dự đoán trở thành công cụ quan trọng đối với Y Tế và cộng đồng.

Mục tiêu xây dựng project này là xây dựng những mô hình Machine Learning để dự đoán nguy cơ mắc bệnh Đái tháo đường dựa trên các yếu tố nguy cơ của người dân.

- Những yếu tố nguy cơ nào ảnh hưởng đến việc mắc bệnh?
- Mô hình máy học nào có thể dự đoán tốt cho bệnh đái tháo đường?

TỔNG QUAN DỮ LIỆU

The Behavioral Risk Factor Surveillance System (BRFSS): hằng năm, CDC Hoa Kỳ thu thập phản hồi khảo sát hơn 400.000 người dân Mỹ về những vấn đề liên quan sức khỏe, hành vi, tình trạng bệnh mạn tính, phương pháp phòng ngừa.

Đối với project này, bộ dữ liệu được lấy từ dữ liệu mở của CDC Hoa Kỳ năm 2022.
Tham chiếu với hai bộ dữ liệu Kaggle là:

- BRFSS (2015)
- Diabetes Health Indicators (2015)

Bộ dữ liệu gốc chứa 445.132 phản hồi với 328 features dùng để đánh giá nhiều bệnh khác nhau, trong đó bao gồm cả bệnh Đái tháo đường.

EDA

Dữ liệu thô ban đầu gồm **445.132 dòng, 328 cột,** không có duplicate, rất nhiều giá trị null.

Dựa vào BRFSS 2022 Codebook tiến hành trích lọc các features gợi ý liên quan đến bệnh Đái tháo đường

1	DIABETE4	Diabetes	Có ai nói rằng bạn mắc bệnh đái tháo đường không?	1 = yes, 0 = no
2	_BMI5	BMI	Chỉ số khối cơ thể	
3	SMOKE100	Smoker	Từ đó đến giờ bạn từng hút nhiều hơn 100 điếu thuốc chưa? (100 điếu thuốc = 5 gói)	1 = yes, 0 = no
4	CVDSTRK3	Stroke	Từng bị đột quỵ hay chưa?	1 = yes, 0 = no
5	_MICHD	HeartDisease	Có bệnh tim mạch như bệnh mạch vành hay nhồi máu cơ tim hay không?	1 = yes, 0 = no
6	_TOTINDA	PhysicalActivity	Số ngày hoạt động thể chất trong vòng 1 tháng qua? (không bao gồm lao động trong công việc)	1 = yes, 0 = no
7	_RFDRHV8	Drinker	Nghiện rượu hay không? (Nam >=14, nữ >=7 lon mỗi tuần)	1 = yes, 0 = no
8	_HLTHPLN	Healthcare	Có bảo hiểm Y tế không?	1 = yes, 0 = no
9	MEDCOST1	CannotaffordDoctor	Trong vòng 12 tháng gần nhất, bạn không thể tiếp cận dịch vụ Y tế đúng hay không?	1 = yes, 0 = no
10	_RFHLTH	GeneralHealth	Tình trạng sức khỏe tổng quát	1,2,3 = tốt 4,5 = tệ
11	MENTHLTH	MentalHealth	Tình trạng sức khỏe tinh thần: Số ngày có vấn đề về stress, trầm cảm, rối loạn cảm xúc trong vòng 1 tháng qua?	1-30
12	PHYSHLTH	PhysicalHealth	Số ngày có vấn đề về thể chất trong vòng 1 tháng qua?	1-30
13	DIFFWALK	DiffWalk	Có đi lại khó khăn không?	1 = yes, 0 = no
14	_SEX	Sex	Giới tính	1 = nam, 0 = nữ

15	_AGEG5YR	Age	Độ tuổi	1 - 13 từ 18 đến trên 80, mỗi mốc ứng với 5 năm
16	EDUCA	Education	Trình độ giáo dục cao nhất?	1 - 6 chưa từng, tiểu học, trung học, tốt nghiệp trung học, đại học cao đẳng, sau đại học
17	INCOME3	Income	Thu nhập hàng năm?	1 - 7 dưới 15k$, dưới 25k$, dưới 35k$, dưới 50k$ dưới 100k$, dưới 200k$, trên 200

PREPROCESSING DATA

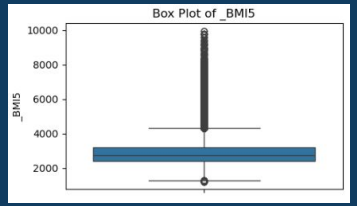
Tiến hành:

- Xử lý các giá trị null, “refuse to answer (code 7, 9)”
- ReLabel lại các trường hợp tiểu đường giả do thai kỳ, tiền tiểu đường (code 2, code 4)
- Loại bỏ outlier BMI
- Đổi tên các cột lại cho dễ tiếp cận và xây dựng mô hình

Sau khi tiến hành, bộ dữ liệu mới gồm **223.287 dòng, 17 cột,** không giá trị null và duplicated
Hai kiểu dữ liệu của biến: binary và numeric-non-binary

```
df_selected.isnull().sum()

DIABETE4       3
_BMI5      48806
SMOKE100   31777
CVDSTRK3       2
_MICHD      5021
_TOTINDA       0
_RFDRHV8       0
_HLTHPLN       0
MEDCOST1       4
_RFHLTH        0
MENTHLTH       3
PHYSHLTH       5
DIFFWALK   22155
_SEX           0
_AGEG5YR       0
EDUCA          5
INCOME3    12932
```



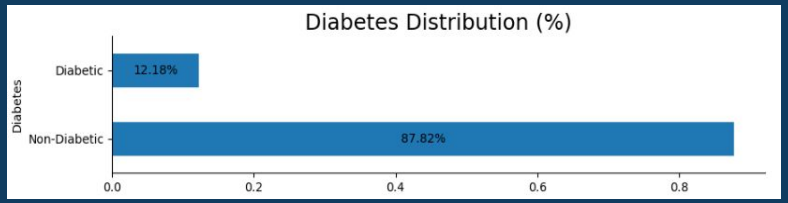

```
 0   Diabetes            223287 non-null  float64
 1   BMI                 223287 non-null  float64
 2   Smoker              223287 non-null  float64
 3   Stroke              223287 non-null  float64
 4   HeartDisease        223287 non-null  float64
 5   PhysicalActivity    223287 non-null  float64
 6   Drinker             223287 non-null  float64
 7   Healthcare          223287 non-null  float64
 8   CannotaffordDoctor  223287 non-null  float64
 9   GeneralHealth       223287 non-null  float64
 10  MentalHealth        223287 non-null  float64
 11  PhysicalHealth      223287 non-null  float64
 12  DiffWalk            223287 non-null  float64
 13  Sex                 223287 non-null  float64
 14  Age                 223287 non-null  float64
 15  Education           223287 non-null  float64
 16  Income              223287 non-null  float64
```

	count	mean	std	min	25%	50%	75%	max
Diabetes	223,287.00	0.12	0.33	0.00	0.00	0.00	0.00	1.00
BMI	223,287.00	25.82	3.37	12.00	23.00	26.00	29.00	31.00
Smoker	223,287.00	0.40	0.49	0.00	0.00	0.00	1.00	1.00
Stroke	223,287.00	0.04	0.19	0.00	0.00	0.00	0.00	1.00
HeartDisease	223,287.00	0.08	0.27	0.00	0.00	0.00	0.00	1.00
PhysicalActivity	223,287.00	0.81	0.39	0.00	1.00	1.00	1.00	1.00
Drinker	223,287.00	0.08	0.26	0.00	0.00	0.00	0.00	1.00
Healthcare	223,287.00	0.95	0.22	0.00	1.00	1.00	1.00	1.00
CannotaffordDoctor	223,287.00	0.08	0.27	0.00	0.00	0.00	0.00	1.00
GeneralHealth	223,287.00	1.14	0.34	1.00	1.00	1.00	1.00	2.00
MentalHealth	223,287.00	4.05	7.95	0.00	0.00	0.00	4.00	30.00
PhysicalHealth	223,287.00	3.75	8.02	0.00	0.00	0.00	3.00	30.00
DiffWalk	223,287.00	0.12	0.32	0.00	0.00	0.00	0.00	1.00
Sex	223,287.00	0.49	0.50	0.00	0.00	0.00	1.00	1.00
Age	223,287.00	7.59	3.58	1.00	5.00	8.00	11.00	13.00
Education	223,287.00	5.15	0.97	1.00	4.00	5.00	6.00	6.00
Income	223,287.00	7.03	2.42	1.00	5.00	7.00	9.00	11.00

02

ANALYZE

ANALYZE

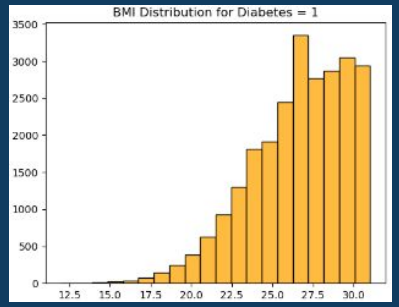


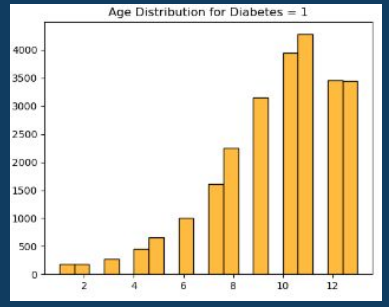
Bộ dữ liệu không cân bằng:

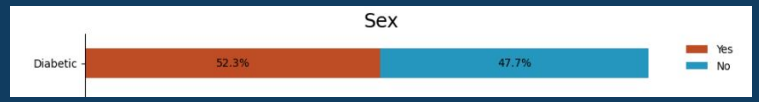
- 196.087 người không mắc bệnh Đái tháo đường, chiếm 87.82%
- 27.200 người mắc bệnh, chiếm 12.18%

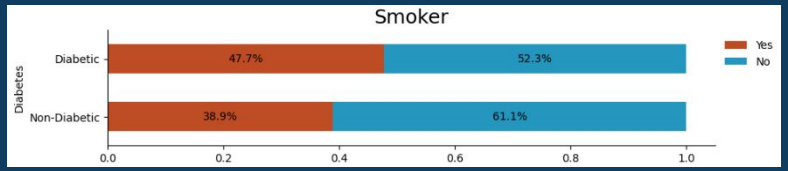
<u>Yếu tố nguy cơ:</u>

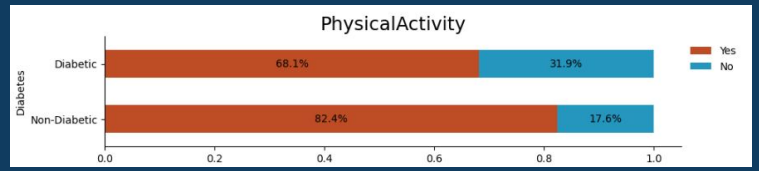
- Số người mắc bệnh tăng theo: BMI, nhóm tuổi
- Nam giới có xu hướng mắc bệnh nhiều hơn
- Tỉ lệ người hút thuốc cao hơn ở người mắc bệnh
- Người mắc bệnh ít hoạt động thể chất hơn



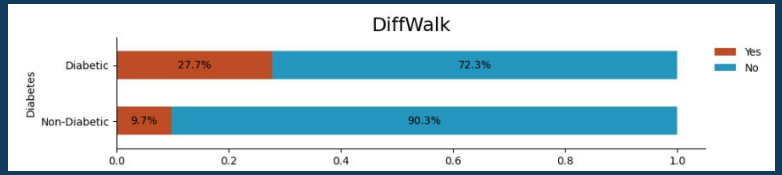


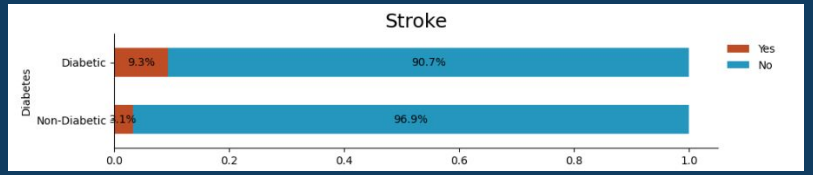


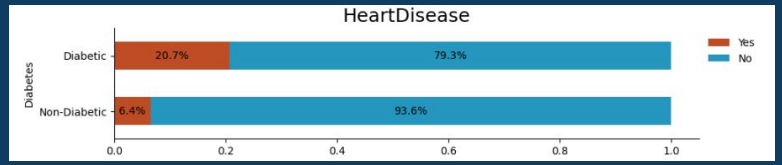




ANALYZE



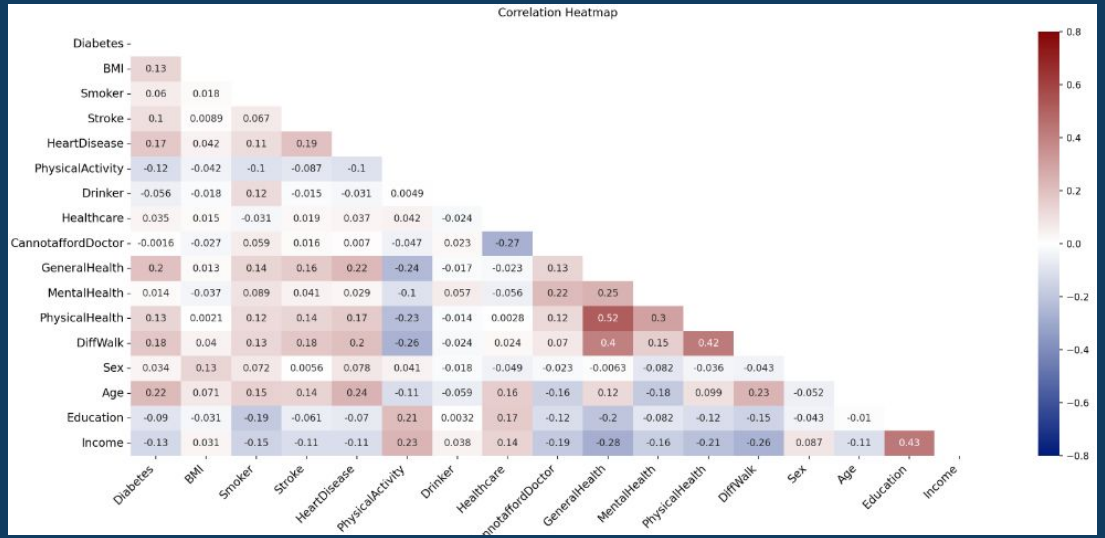




Biến chứng: người mắc bệnh

- Tỉ lệ đi lại khó khăn hơn do có thể do biến chứng mạch máu ở chân
- Tỉ lệ măc bệnh đột quỵ đi cùng cao hơn
- Tỉ lệ mắc bệnh tim mạch đi cùng cao hơn

ANALYZE



Sử dụng correlation: Đái tháo đường có

- Tương quan dương với BMI, tuổi, đi lại khó khăn, sức khỏe tổng quát và sức khỏe thể chất, bệnh tim mạch
- Tương quan âm với Vận động thể chất, thu nhập

Ngoài ra:

- Sức khỏe tinh thần và thể chất có tương quan khá cao với nhau
- Không có khả năng chi trả cho bác sĩ có tương quan âm với thu nhập
- Giáo dục có tương quan dương với thu nhập

03

MODELS PREDICTION

CHUẨN BỊ

	BMI	Smoker	Stroke	HeartDisease	PhysicalActivity	Drinker	Healthcare	CannotaffordDoctor	GeneralHealth	MentalHealth	PhysicalHealth	DiffWalk	Sex
1	27.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00
2	26.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	3.00	2.00	0.00	0.00
4	22.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	2.00	0.00	2.00	0.00	0.00
6	23.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00
7	28.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...
445122	22.00	1.00	1.00	0.00	0.00	0.00	1.00	0.00	2.00	1.00	30.00	0.00	1.00
445123	24.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	2.00	7.00	0.00	0.00	0.00
445124	30.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	15.00	0.00	0.00	1.00
445126	31.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00
445128	29.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00	2.00	2.00	0.00	0.00

223287 rows × 16 columns

Chuẩn hóa dữ liệu:

- SMOTE
- StandardScaler

Splitting train_test 70:30

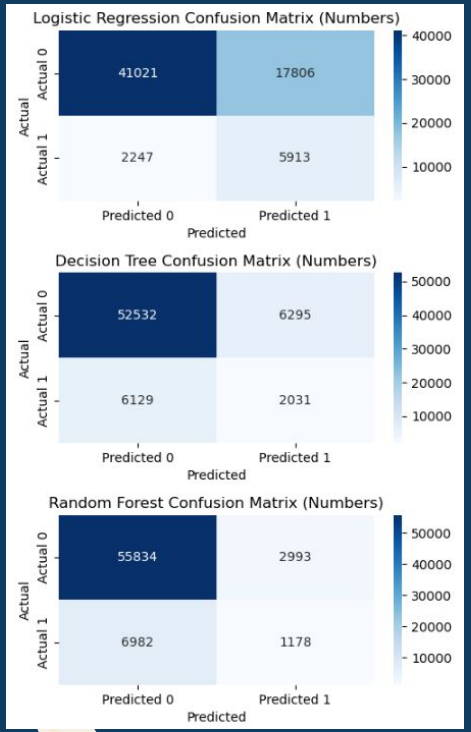
```
X_train_sm.shape, y_train_sm.shape, X_test.shape, y_test.shape
```

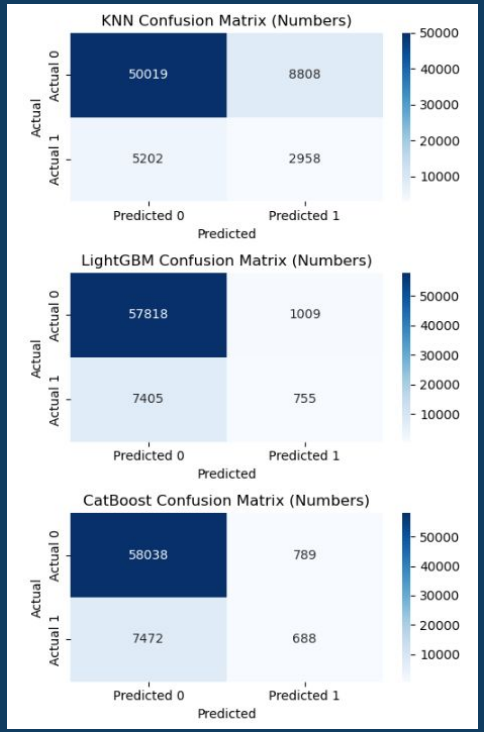
```
((274520, 16), (274520,), (66987, 16), (66987,))
```

Lựa chọn các mô hình để xây dựng:

- Logistic Regression
- Decision Tree
- Random Forest
- KNN
- LightGBM
- CatBoost

KẾT QUẢ





	Model	ROC AUC	Precision	Recall	F1-Score
0	Logistic Regression	0.78	0.60	0.71	0.59
1	Decision Tree	0.58	0.57	0.57	0.57
2	Random Forest	0.72	0.59	0.55	0.55
3	KNN	0.67	0.58	0.61	0.59
4	LightGBM	0.77	0.66	0.54	0.54
5	CatBoost	0.77	0.68	0.54	0.54

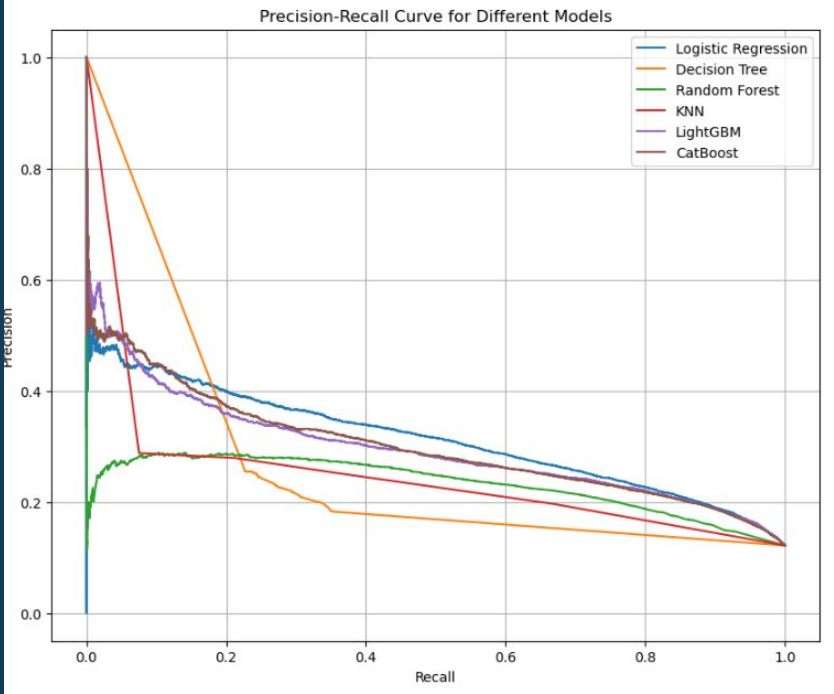
Logistic Regression: có độ chính xác tổng thể cao nhất. Precision và Recall của mô hình này cũng tương đối tốt, cân bằng tốt giữa việc dự đoán đúng các trường hợp dương tính và âm tính.

Decision Tree: có ROC AUC và Precision thấp nhất, không phải là lựa chọn tốt.

KNN và Random Forest: có hiệu suất trung bình, KNN có Recall và F1-Score cao hơn có thể điều chỉnh thêm.

CatBoost và LightGBM: có ROC AUC và Precision cao, nhưng Recall và F1-Score thấp hơn do khả năng phát hiện trường hợp dương tính không tốt.

KẾT QUẢ



- Logistic Regression: Bắt đầu với Precision cao ở các giá trị Recall thấp.
  Giảm dần Precision khi Recall tăng, cho thấy sự cân bằng tốt giữa Precision và Recall.
- LightGBM: cũng giữ được cân bằng tốt, tương đối ổn định.
- KNN và Decision Tree: nó sự sụt giảm nhanh về Precision khi Recall tăng, cho thấy chúng không phải là lựa chọn tốt.
- Random Forest và CatBoost: nằm ở mức trung bình, với sự cân bằng tốt hơn so với Decision Tree và KNN nhưng không tốt bằng Logistic Regression và LightGBM.

TUNING

```
Best Params for Logistic Regression: {'C': 0.01, 'solver': 'liblinear'}
Best Params for Decision Tree: {'max_depth': None, 'min_samples_leaf': 5, 'min_samples_split': 20}
Best Params for KNN: {'n_neighbors': 5, 'weights': 'distance'}
Best Params for Random Forest: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}
Best Params for LightGBM: {'learning_rate': 0.1, 'n_estimators': 100, 'num_leaves': 100}
Best Params for CatBoost: {'depth': 10, 'iterations': 100, 'learning_rate': 0.2}
```

Model	ROC AUC	Precision (macro avg)	Recall (macro avg)	F1-Score (macro avg)
Logistic Regression	0.78	0.60	0.71	0.59
Decision Tree	0.67	0.58	0.56	0.57
KNN	0.66	0.58	0.60	0.59
Random Forest	0.72	0.59	0.55	0.55
LightGBM	0.77	0.66	0.54	0.54
CatBoost	0.76	0.66	0.54	0.55

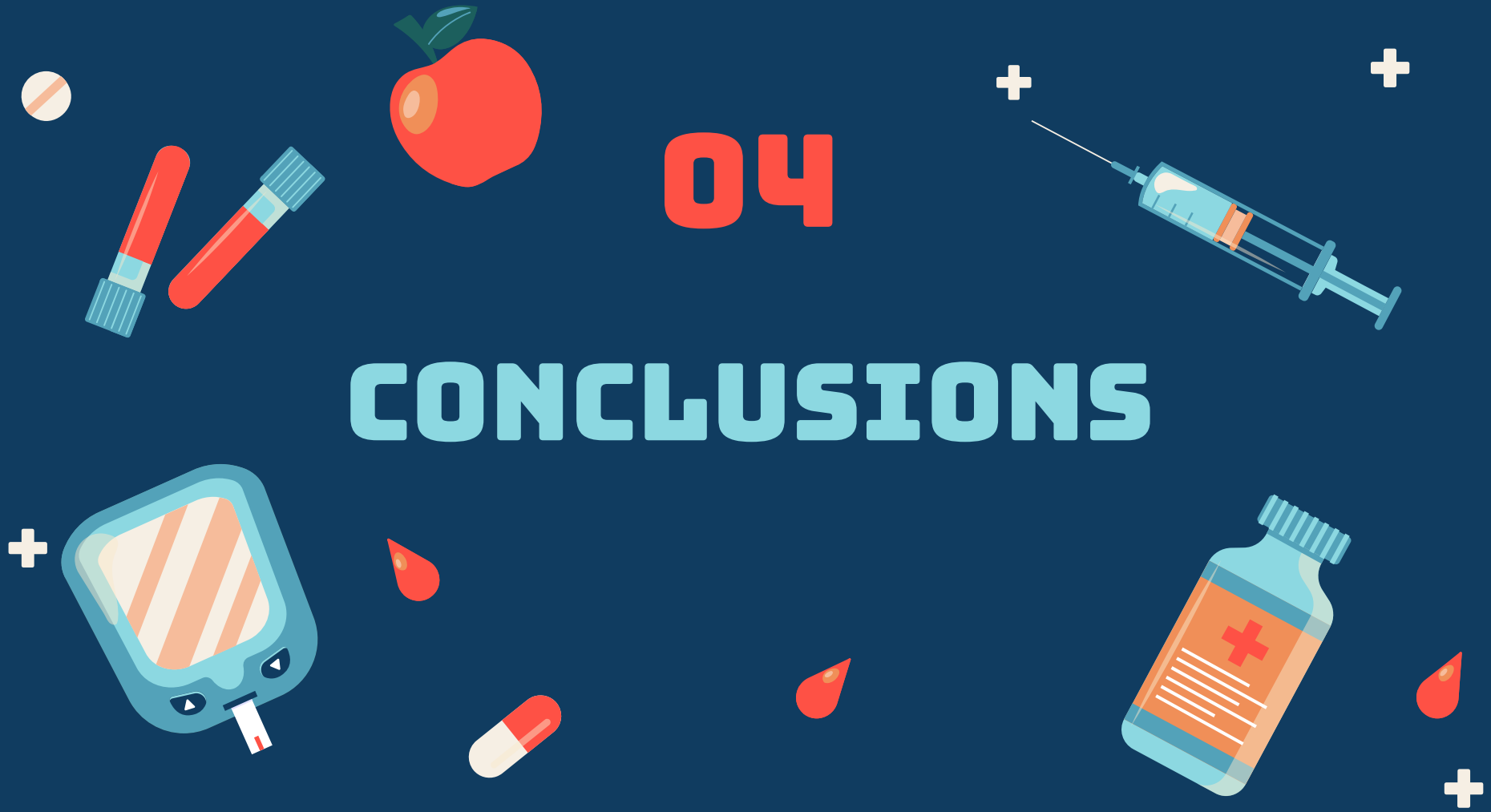
- Decision Tree cải thiện rõ rệt nhất sau khi tuning, đặc biệt là ROC AUC
- Logistic Regression và Random Forest: đã được tối ưu hóa tốt ngay từ đầu, không có sự thay đổi đáng kể sau tuning.
- LightGBM và CatBoost: giữ nguyên hiệu suất hoặc chỉ có sự thay đổi nhỏ, cho thấy đã hoạt động tốt trước khi tuning.
- KNN: không có sự thay đổi lớn và thậm chí có sự giảm nhẹ về ROC AUC sau tuning, có thể do tuning không tìm ra được siêu tham số tối ưu hơn.

Logistic Regression là lựa chọn tốt nhất

	Feature	Coefficient	Absolute Coefficient
13	Age	0.71	0.71
0	BMI	0.44	0.44
8	GeneralHealth	0.34	0.34
15	Income	-0.20	0.20
5	Drinker	-0.20	0.20
3	HeartDisease	0.14	0.14
12	Sex	0.12	0.12
6	Healthcare	0.09	0.09
11	DiffWalk	0.08	0.08
14	Education	-0.07	0.07
4	PhysicalActivity	-0.07	0.07
2	Stroke	0.03	0.03
9	MentalHealth	0.03	0.03
7	CannotaffordDoctor	-0.02	0.02
1	Smoker	-0.02	0.02
10	PhysicalHealth	0.00	0.00

04

CONCLUSIONS



CONCLUSIONS

Đái tháo đường có

- Liên quan đến **BMI**, **tuổi tác**, **sức khỏe tổng quát và thể chất**, cũng như bệnh tim mạch. Điều này cho thấy những yếu tố này có thể gia tăng nguy cơ mắc bệnh đái tháo đường.
- Tương quan âm với thu nhập cho thấy thu nhập cao hơn có thể giúp tiếp cận dịch vụ y tế và chăm sóc sức khỏe tốt hơn. Nghiện rượu cũng có thể do yếu tố liên quan đến đời sống và ăn uống.
- Giới tính có ảnh hưởng nhưng không quá mạnh đến khả năng mắc bệnh

Logistic Regression là lựa chọn tốt nhất cho ứng dụng y tế vì:

- ROC AUC cao nhất, cho thấy khả năng phân biệt tốt giữa các lớp.
- Recall cao nhất, đảm bảo phát hiện được nhiều trường hợp bệnh nhất, điều rất quan trọng trong y tế.
- Việc chọn Logistic Regression sẽ giúp đảm bảo rằng ít trường hợp bệnh bị bỏ sót, đồng thời giữ được độ chính xác trong các dự đoán.

Có thể cần thêm các thử nghiệm và điều chỉnh để đạt được hiệu suất tốt nhất cho LR nói riêng và các mô hình khác hoặc thêm mô hình mới nói riêng.

THANK YOU

FOR YOUR ATTENTION