



PNL-DA52

# PHÂN TÍCH VÀ XÂY DỰNG CÁC MÔ HÌNH HỌC MÁY DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG



Đoàn Ngọc Tuấn

# NỘI DUNG



TỔNG QUAN



PHÂN TÍCH



XÂY DỰNG  
MÔ HÌNH



ĐỀ XUẤT  
KẾT LUẬN



# TỔNG QUAN

# LÝ DO CHỌN ĐỀ TÀI

- Bệnh tiểu đường là một bệnh mãn tính phổ biến, ảnh hưởng hàng triệu người trên toàn thế giới.
- **Mục tiêu:** phân tích và dự đoán bệnh tiểu đường để nâng cao chất lượng cuộc sống và giảm thiểu biến chứng.
  - **Nghiên cứu:**
  - **Phòng ngừa và quản lý:**
  - **Giảm biến chứng và chi phí điều trị.**

# DATASET

- Bộ dữ liệu dự đoán bệnh tiểu đường là tập hợp dữ liệu y tế và nhân khẩu học từ bệnh nhân, cùng với tình trạng bệnh tiểu đường của họ.
- Gồm 100.000 dòng, 10 thuộc tính.

<b>id</b>	<b>gender</b>	<b>age</b>	<b>hypertension</b>	<b>heart_disease</b>	<b>smoking_history</b>	<b>bmi</b>	<b>HbA1c_level</b>	<b>blood_glucose_level</b>	<b>diabetes</b>
1	Female	80	0	1	never	25.19	6.6	140	0
2	Female	54	0	0	No Info	27.32	6.6	80	0
3	Male	28	0	0	never	27.32	5.7	158	0
4	Female	36	0	0	current	23.45	5	155	0
5	Male	76	1	1	current	20.14	4.8	155	0
6	Female	20	0	0	never	27.32	6.6	85	0
7	Female	44	0	0	never	19.31	6.5	200	1
8	Female	79	0	0	No Info	23.86	5.7	85	0
9	Male	42	0	0	never	33.64	4.8	145	0
10	Female	32	0	0	never	27.32	5	100	0
11	Female	53	0	0	never	27.32	6.1	85	0
12	Female	54	0	0	former	54.7	6	100	0
13	Female	78	0	0	former	36.05	5	130	0
14	Female	67	0	0	never	25.69	5.8	200	0
15	Female	76	0	0	No Info	27.32	5	160	0
16	Male	78	0	0	No Info	27.32	6.6	126	0
17	Male	15	0	0	never	30.36	6.1	200	0
18	Female	42	0	0	never	24.48	5.7	158	0
19	Female	42	0	0	No Info	27.32	5.7	80	0

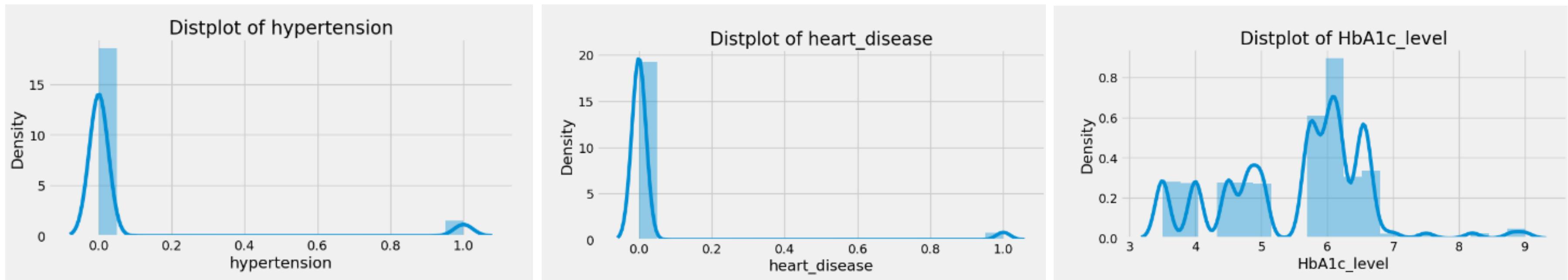
# DATASET

Thuộc tính	Ý nghĩa
id	Mã bệnh nhân
gender	Giới tính
age	Tuổi
hypertension	Bệnh nhân có bị huyết áp cao hay không
heart_disease	Bệnh nhân có bệnh tim mạch hay không
smoking_history	Tiền sử hút thuốc lá của bệnh nhân
bmi	Chỉ số bmi
HbA1c_level	Lượng đường trung bình trong máu
blood_glucose_level	Mức đường huyết
diabetes	Bệnh nhân có bị bệnh tiểu đường hay không



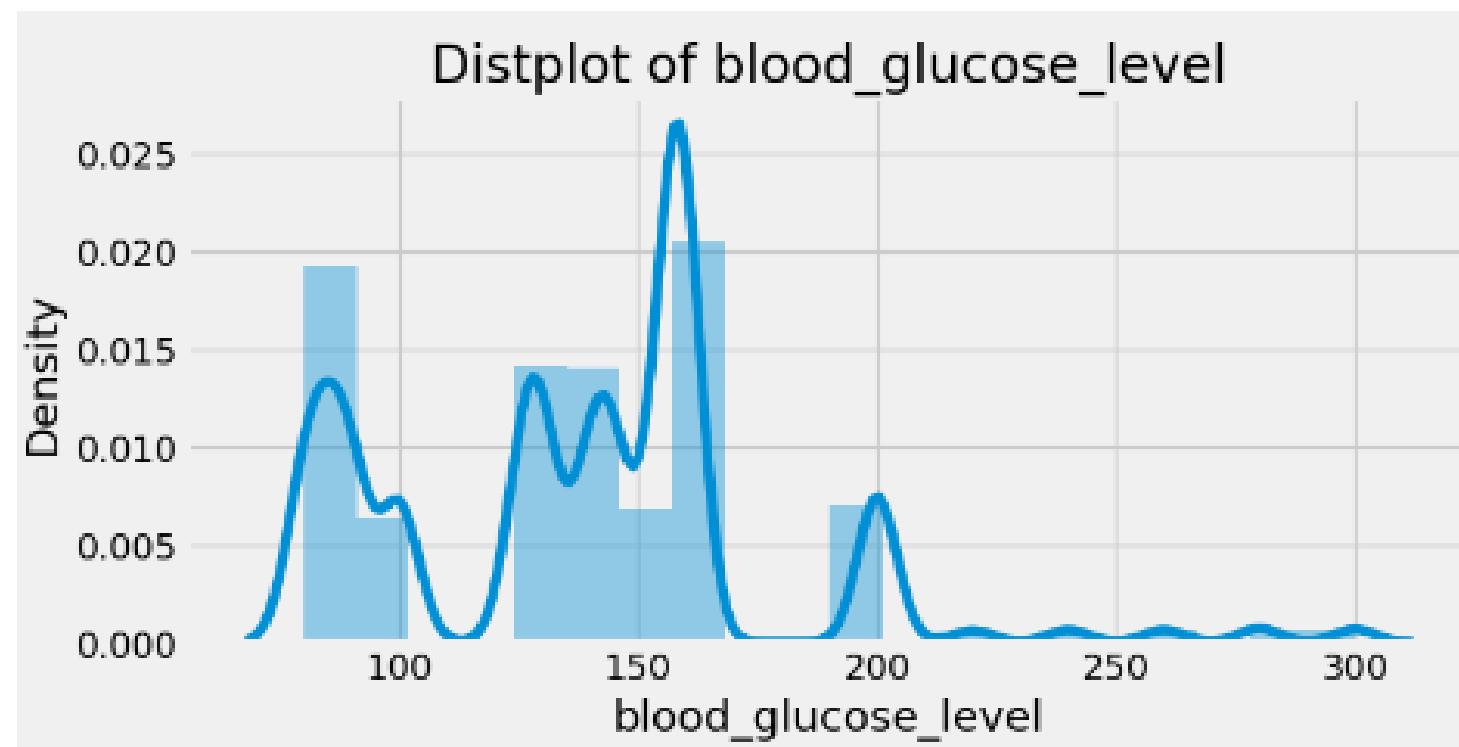
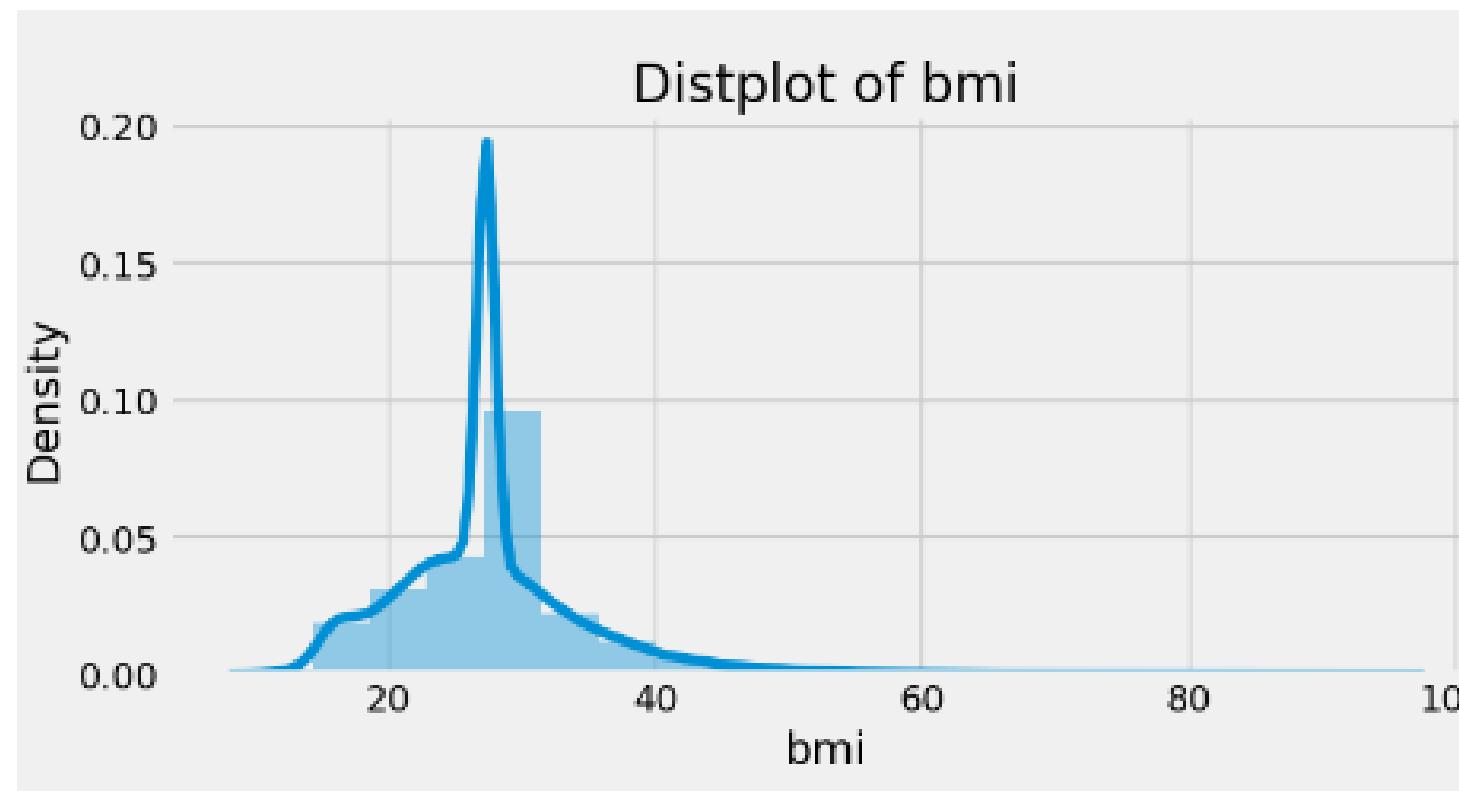
**PHÂN TÍCH**

# PHÂN TÍCH TỔNG QUAN



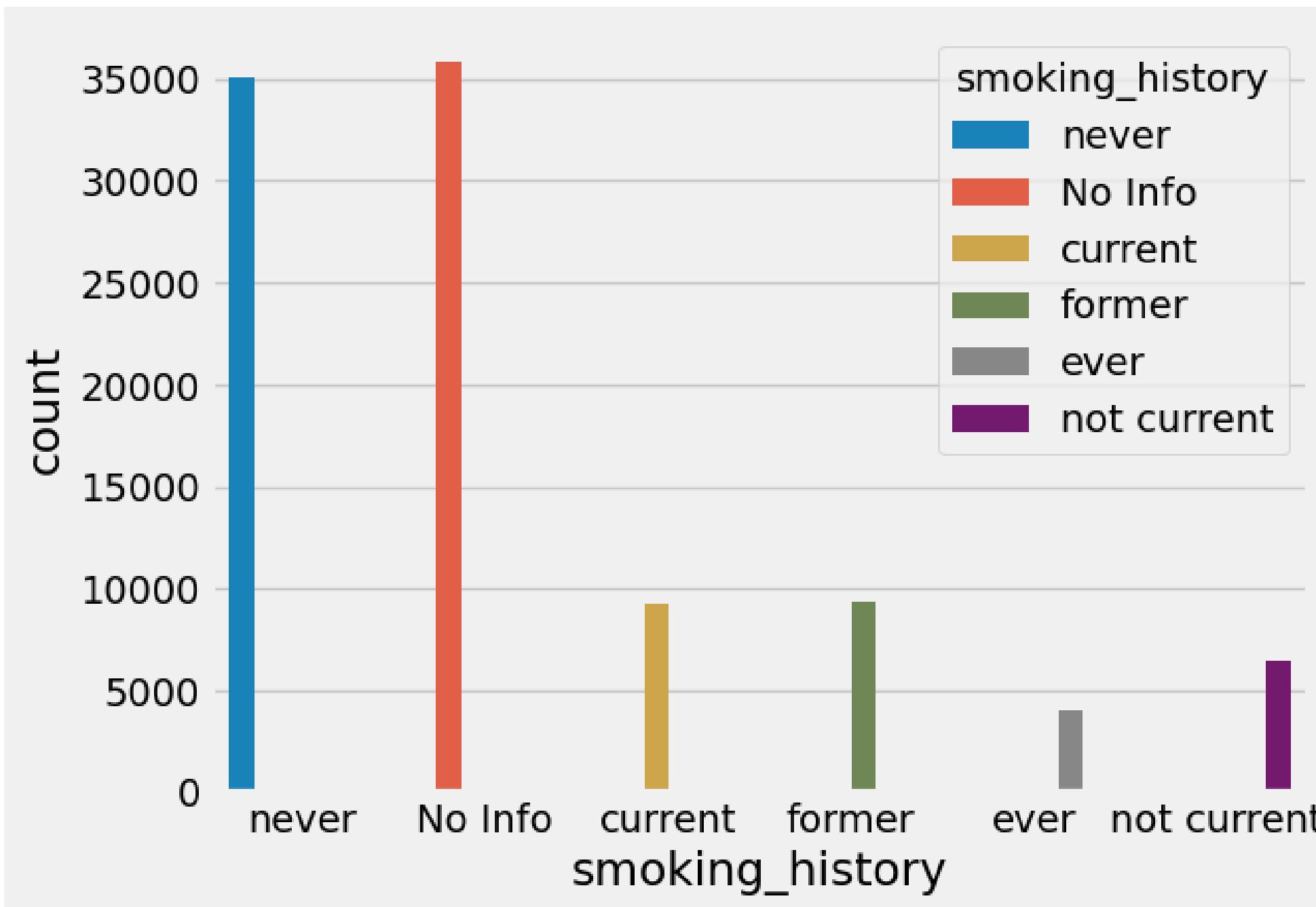
- Đa phần các bệnh nhân không bị huyết áp cao.
- Các bệnh nhân có vấn đề về tim mạch chiếm số lượng thấp.
- HbA1c có mức độ chủ yếu từ 6-7.

# PHÂN TÍCH TỔNG QUAN



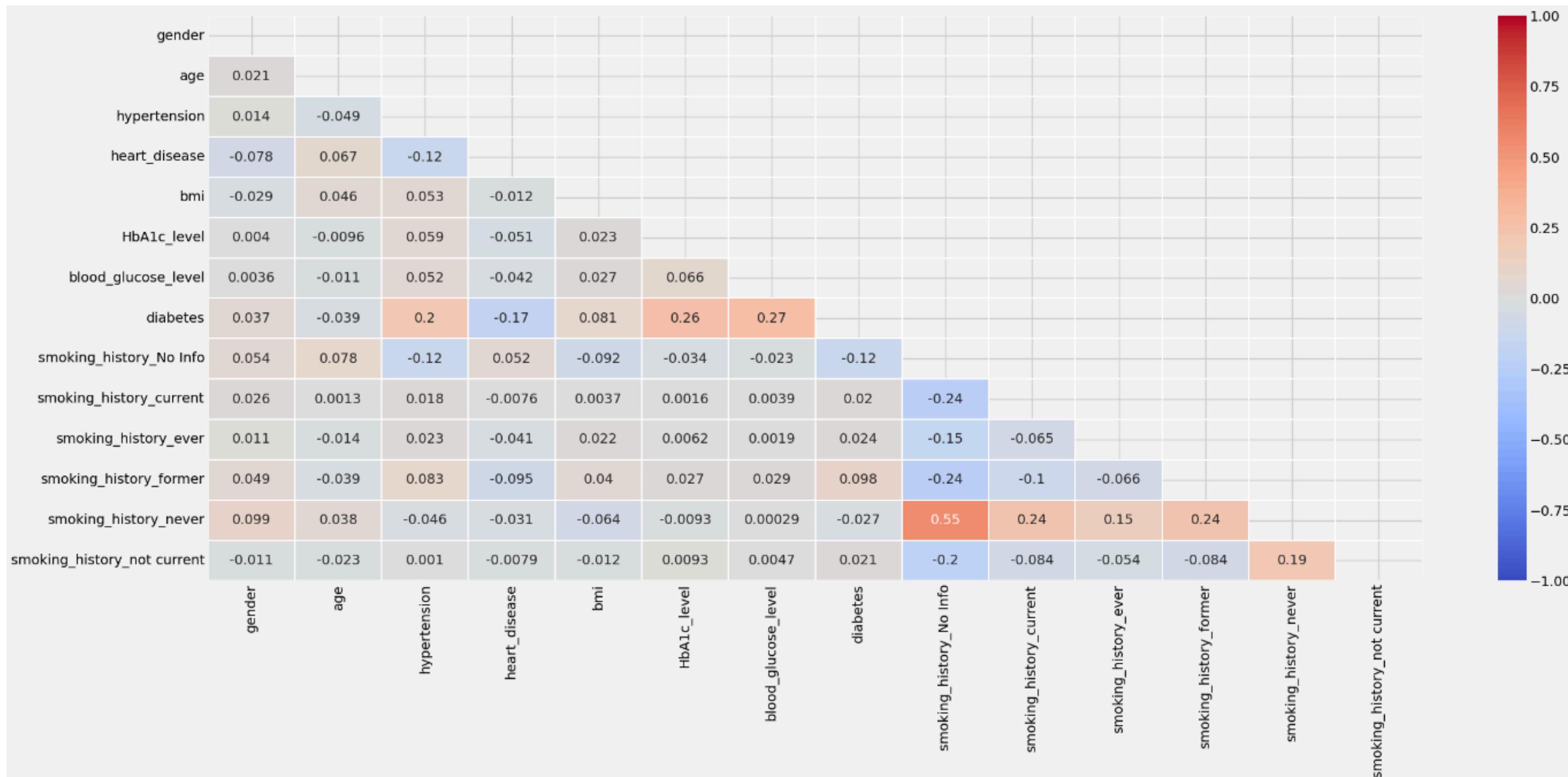
- Chỉ số BMI của bệnh nhân chủ yếu trong khoảng 20-35.
- Mức đường huyết chủ yếu trong khoảng 130-160.

# PHÂN TÍCH TỔNG QUAN



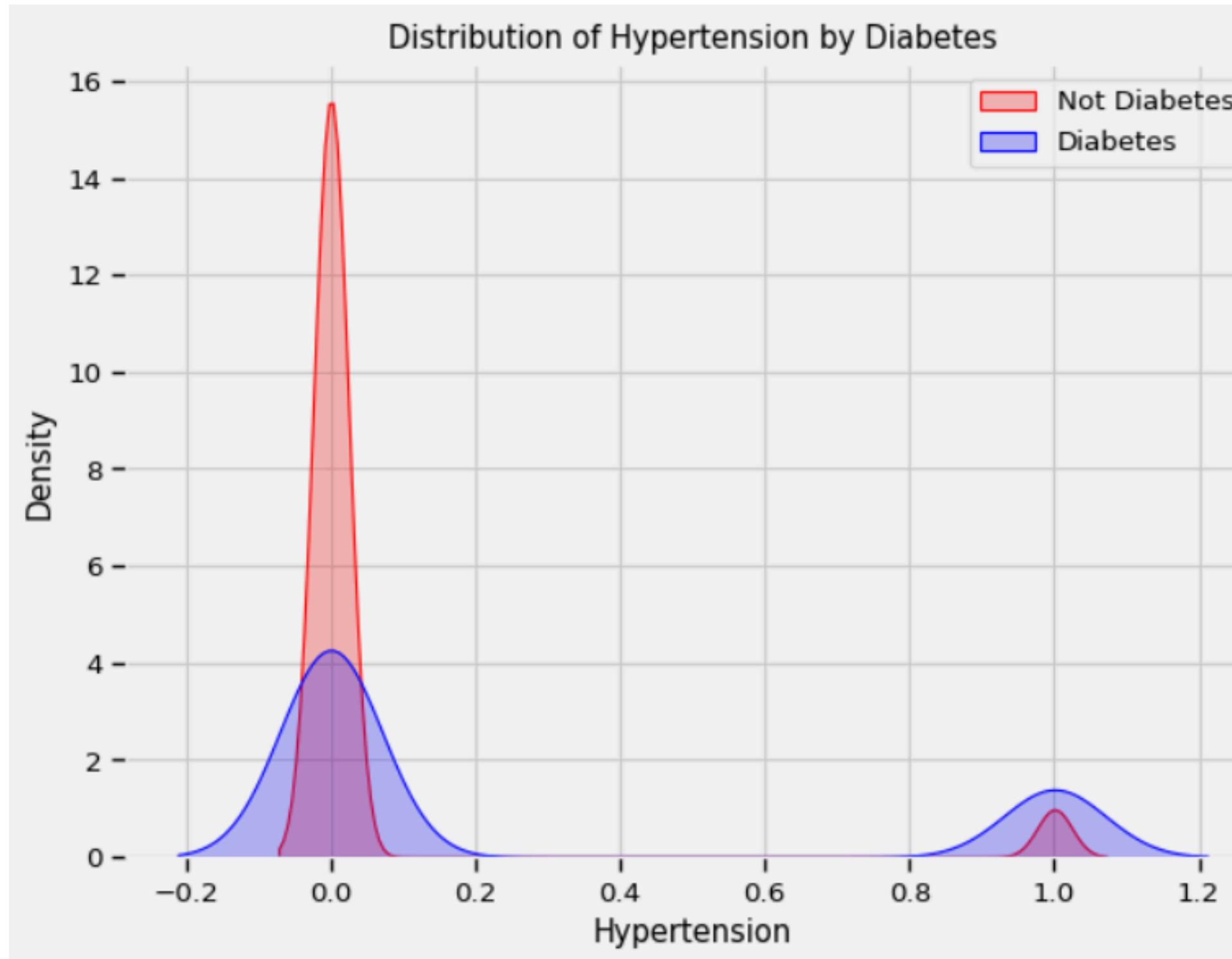
- Hầu hết các bệnh nhân đều chưa bao giờ hút thuốc hoặc chưa có thông tin về tiền sử hút thuốc lá của bệnh nhân.

# PHÂN TÍCH CHI TIẾT



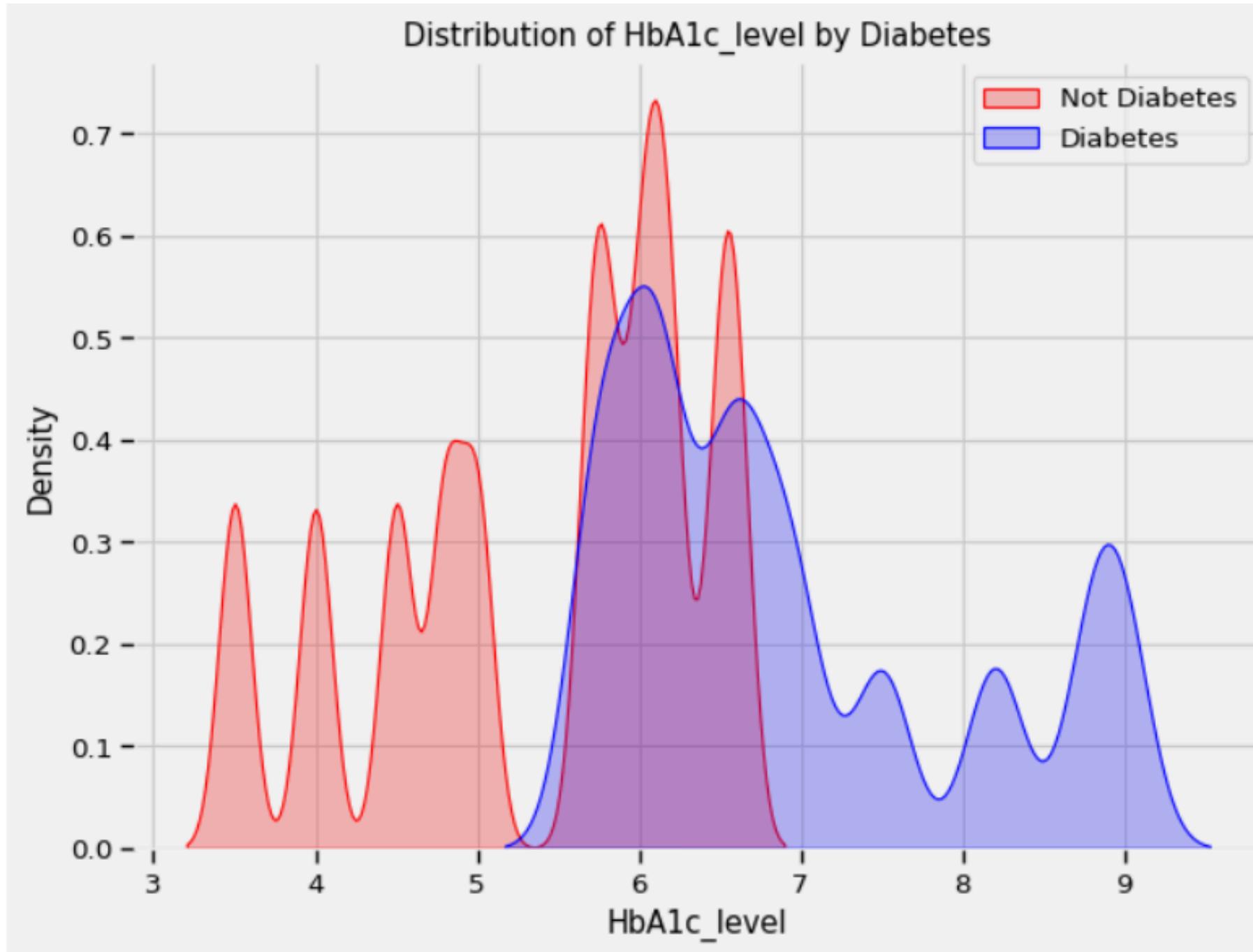
- **Diabetes** có sự tương quan tương đối mạnh đối với **heart\_disease**, **hypertension**, **HbA1c\_level**, **blood\_glucose\_level**
- Tương quan thuận đổi với **hypertension**, **HbA1c\_level**, **blood\_glucose\_level** và tương quan nghịch với **heart\_disease**

# PHÂN TÍCH CHI TIẾT



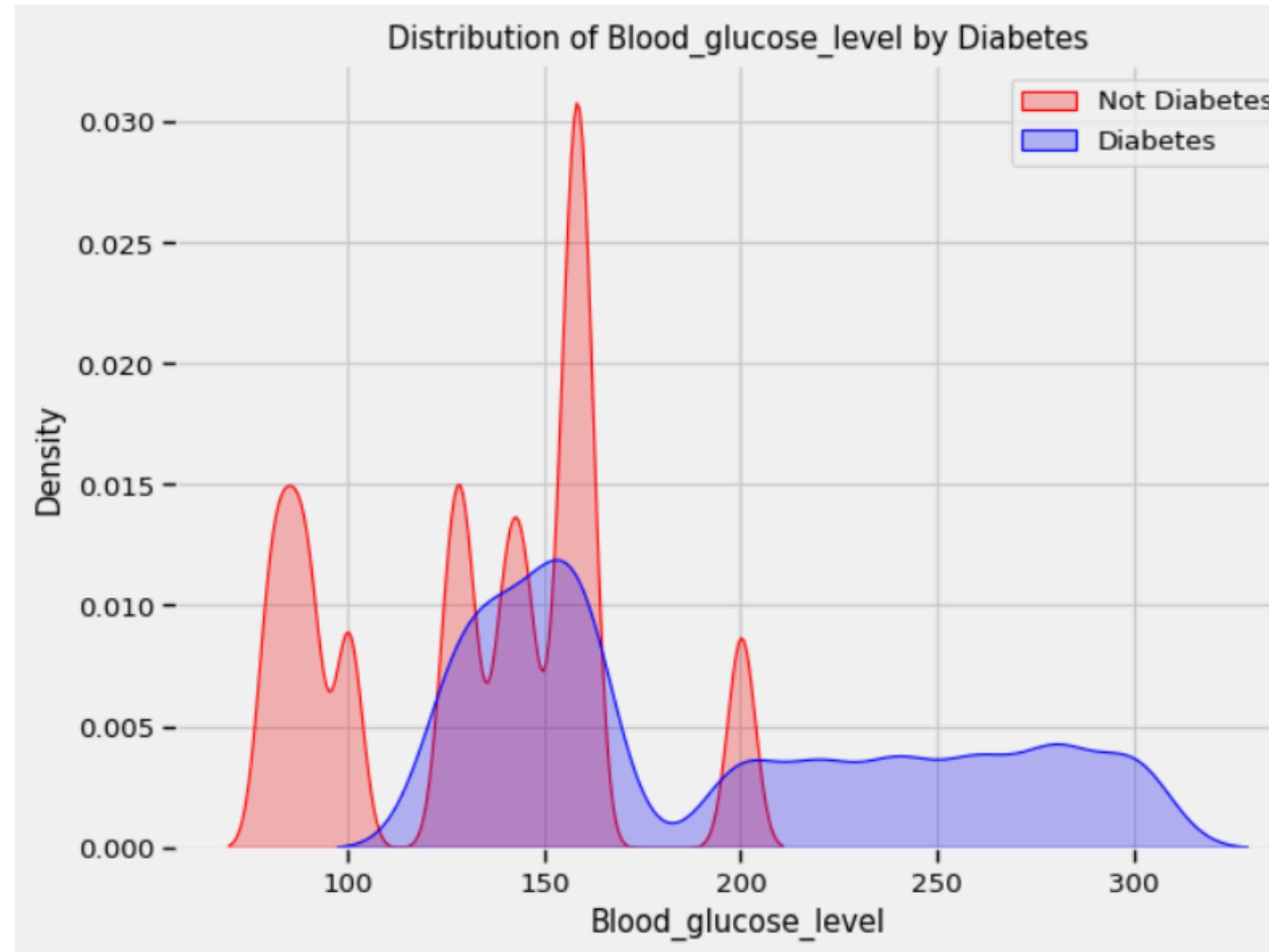
- Những bệnh nhân không bị huyết áp cao (**hypertension = 0**) thì tỉ lệ không bị bệnh tiểu đường **cao hơn gấp 4 lần** so với tỉ lệ bị bệnh.
- Những bệnh nhân **huyết áp cao** thì khả năng bị bệnh tiểu đường **khá cao**, tỉ lệ bị bệnh cao hơn tỉ lệ không bị bệnh.

# PHÂN TÍCH CHI TIẾT



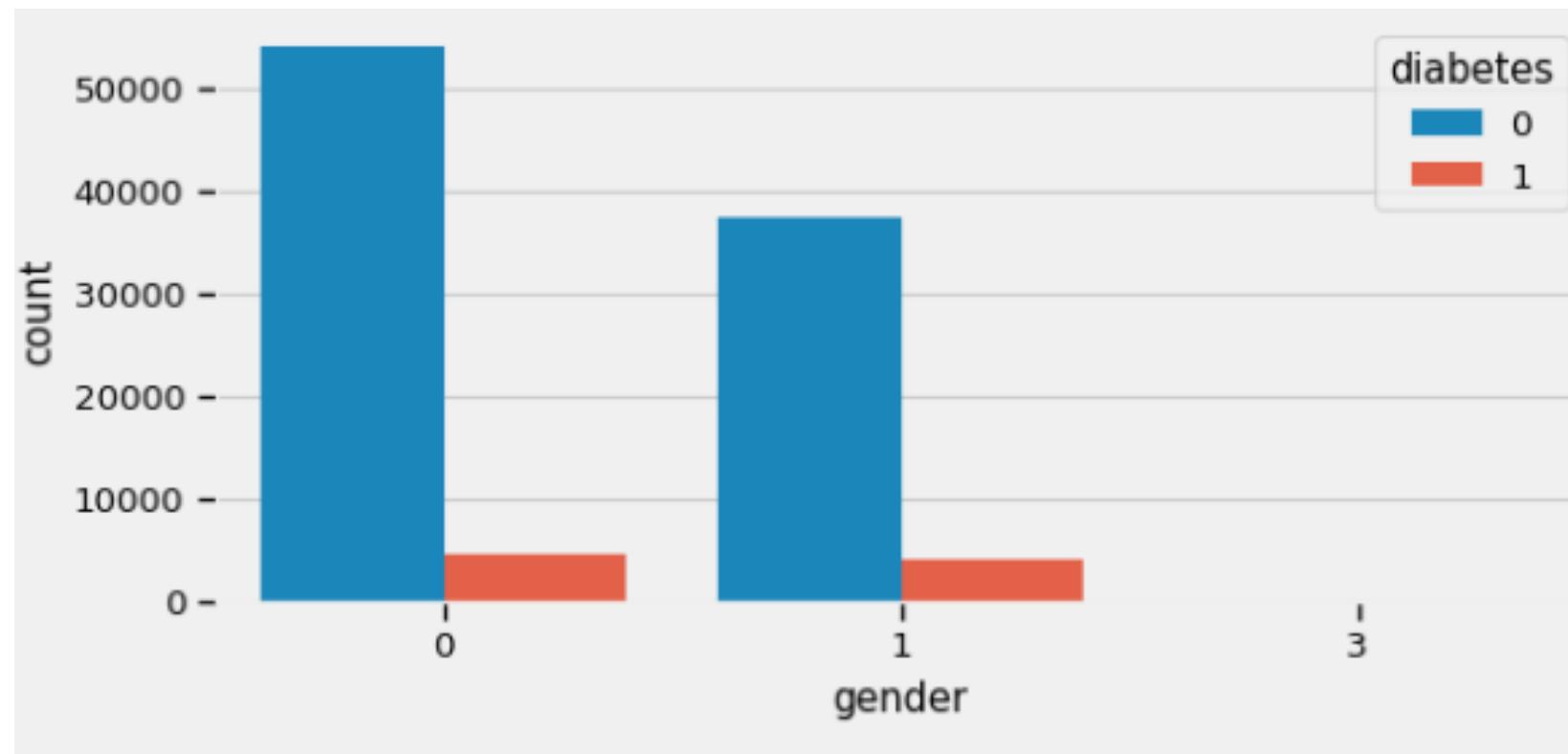
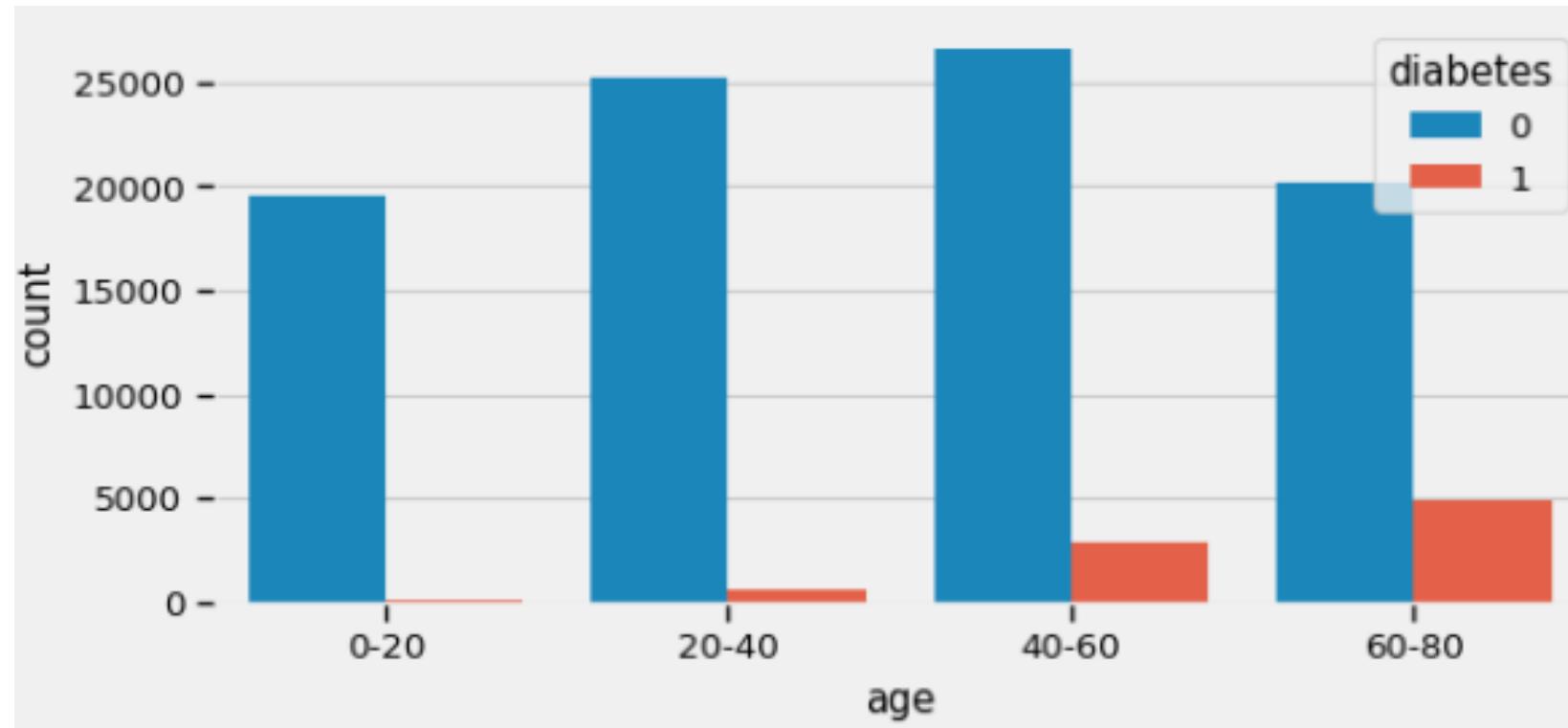
- Những bệnh nhân có chỉ số lượng đường trong máu  $< 5$  (**HbA1c\_level < 5**) thì chắc chắn sẽ **không bị** bệnh tiểu đường.
- Những bệnh nhân có chỉ số lượng đường trong máu  $> 7$  chắc chắn sẽ **bị** bệnh tiểu đường
- Đối với những bệnh nhân có HbA1c\_level từ khoảng 5.2 đến 6.8 thì vẫn sẽ có khả năng bị bệnh, tỉ lệ bệnh gần bằng tỉ lệ không bị bệnh.

# PHÂN TÍCH CHI TIẾT



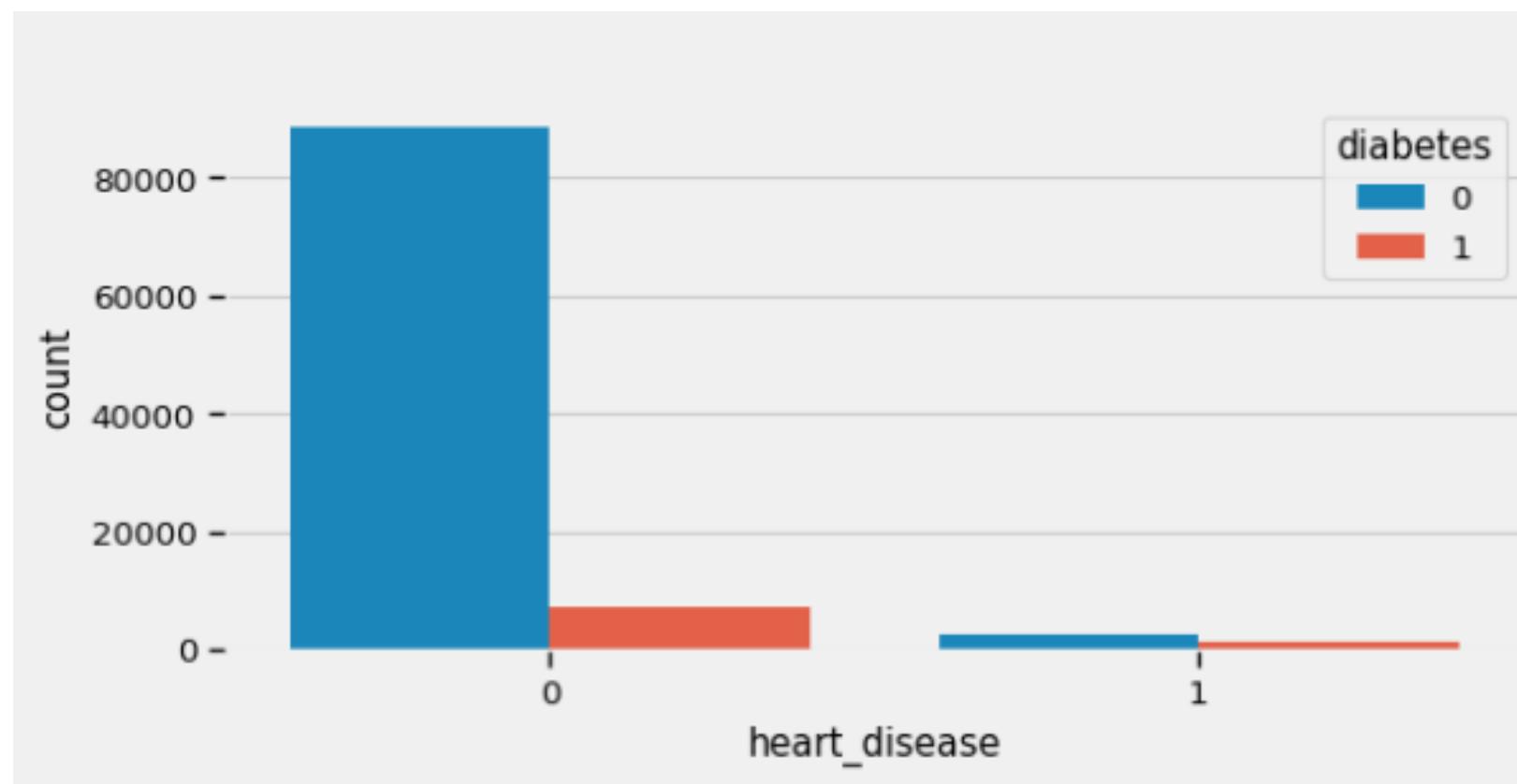
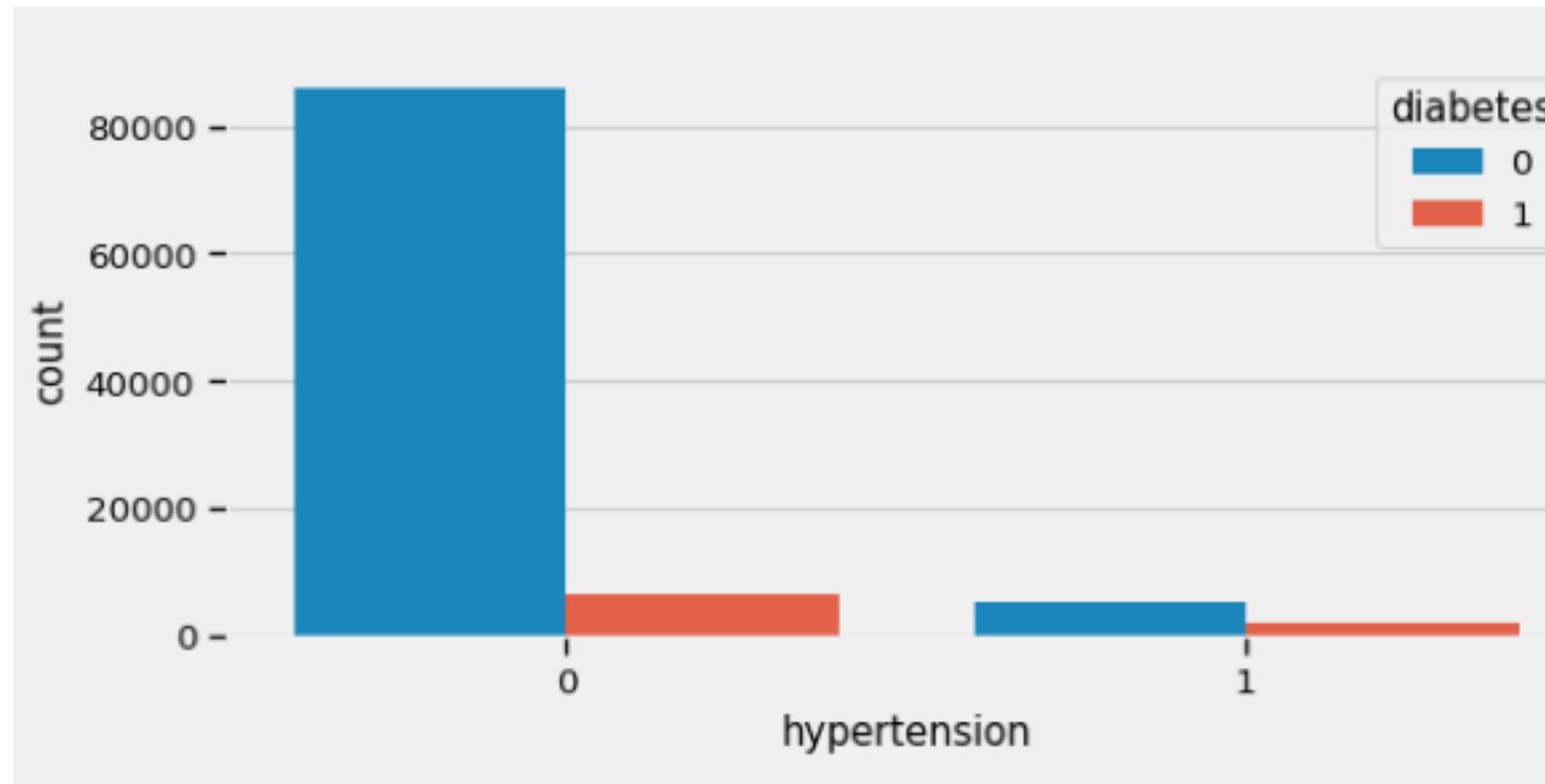
- Những bệnh nhân có mức đường huyết thấp, dưới 100 mg/dL thì chắc chắn không bị bệnh tiểu đường
- Những bệnh nhân có mức đường huyết cao,  $> 200$  mg/dL thì khả năng rất cao sẽ mắc bệnh tiểu đường
- Đối với những bệnh nhân có mức đường huyết từ 100 - 200 thì cũng có thể mắc bệnh, nhưng tỉ lệ thấp hơn.

# PHÂN TÍCH CHI TIẾT



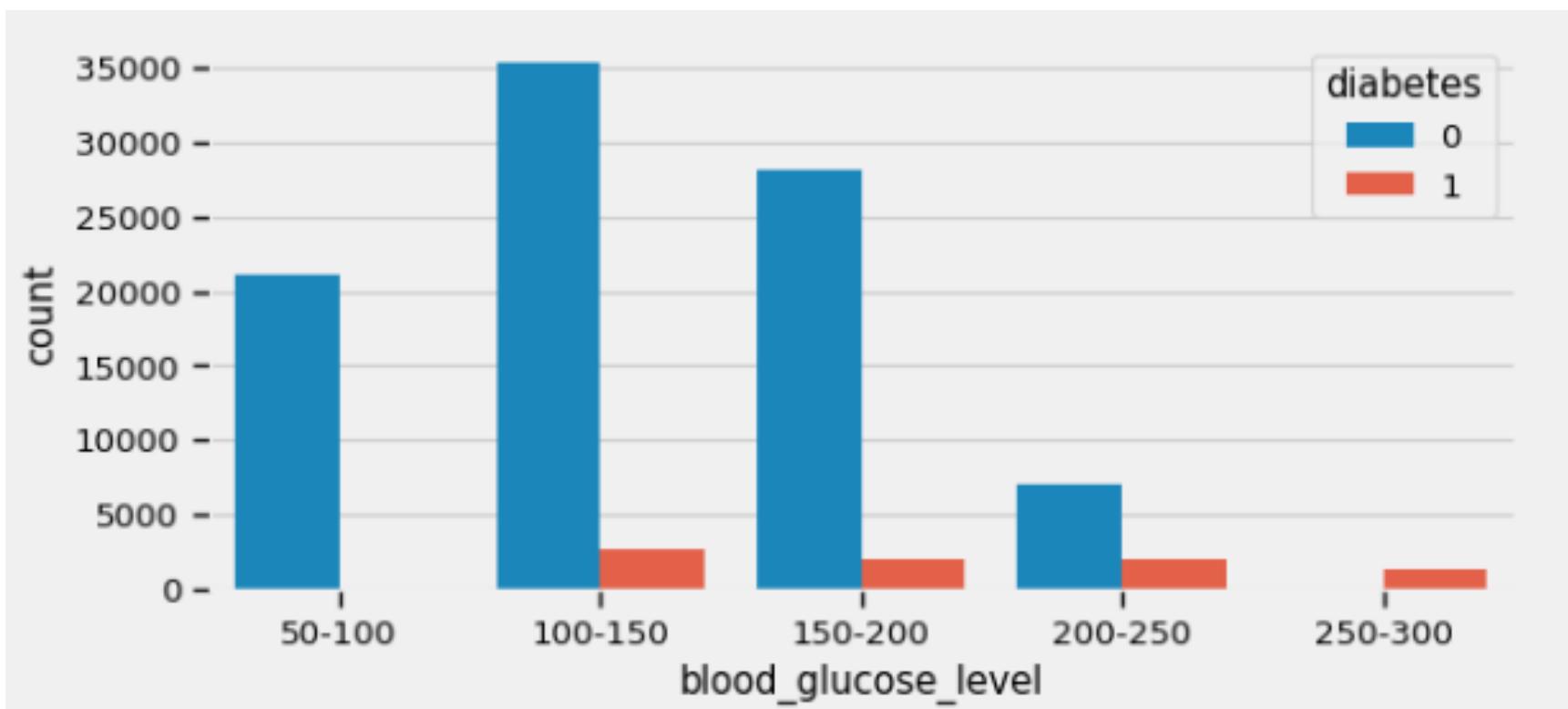
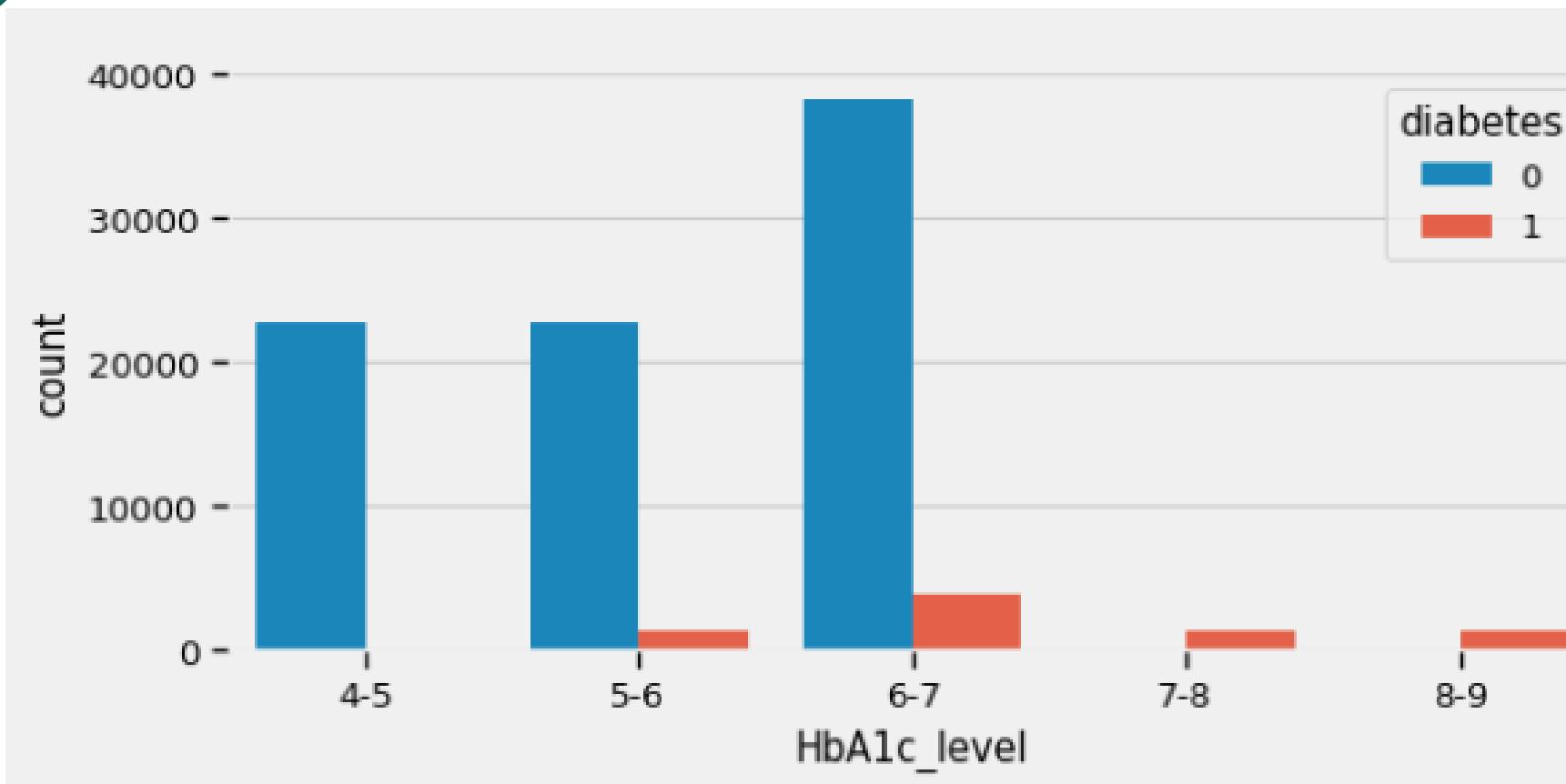
- Nhóm bệnh nhân có độ tuổi > 40 thì khả năng mắc bệnh tiểu đường của họ khá cao.
- Nhóm bệnh nhân nam có nguy cơ mắc bệnh cao hơn nhóm bệnh nhân nữ.

# PHÂN TÍCH CHI TIẾT



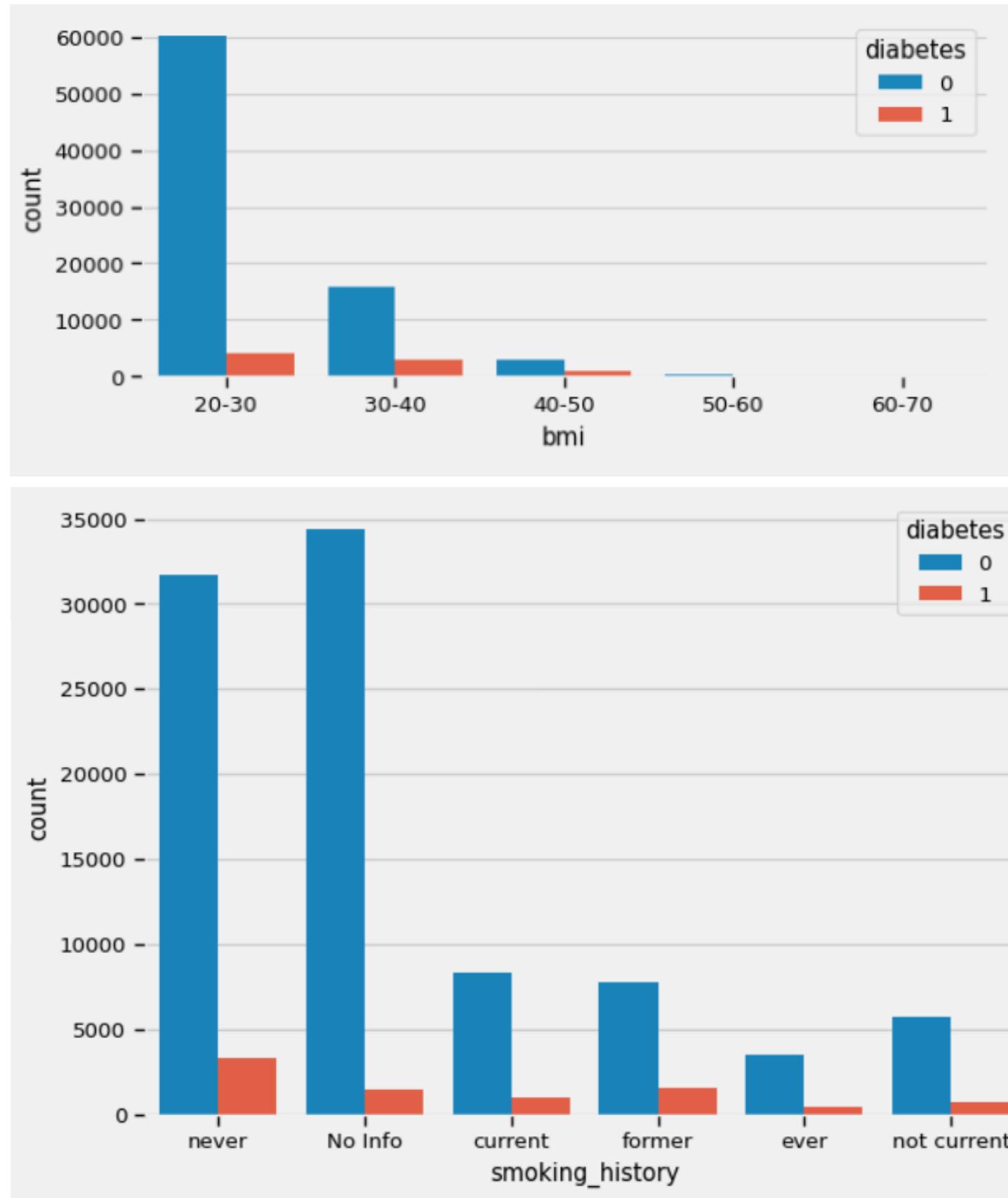
- Những bệnh nhân có vấn đề về huyết áp, tim mạch thì khả năng mắc bệnh cao hơn hẳn so với nhóm còn lại.

# PHÂN TÍCH CHI TIẾT



- Các chỉ số HbA1c\_level và blood glucose level tỉ lệ thuận với tỉ lệ mắc bệnh tiểu đường của bệnh nhân.

# PHÂN TÍCH CHI TIẾT

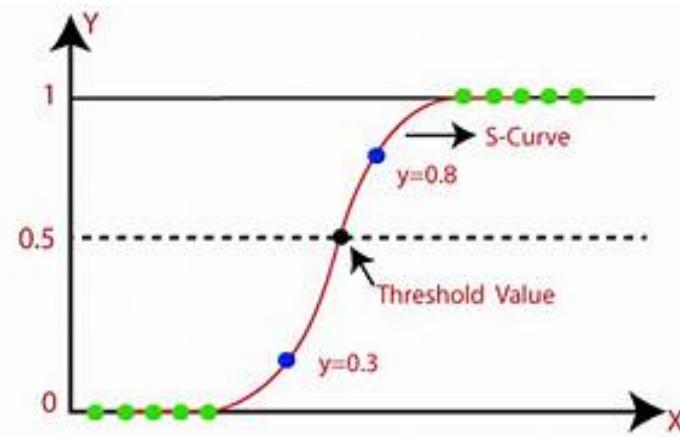


- Những bệnh nhân béo phì cấp độ 2, cấp độ 3 ( $bmi > 30$ ) sẽ làm tăng khả năng mắc bệnh tiểu đường.
- Nhóm bệnh nhân hiện tại hoặc trước đó đã từng hút thuốc lá sẽ tăng nguy cơ mắc bệnh tiểu đường.

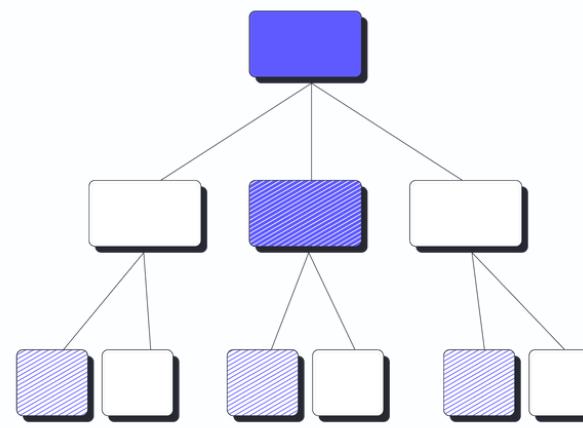


# XÂY DỰNG MÔ HÌNH

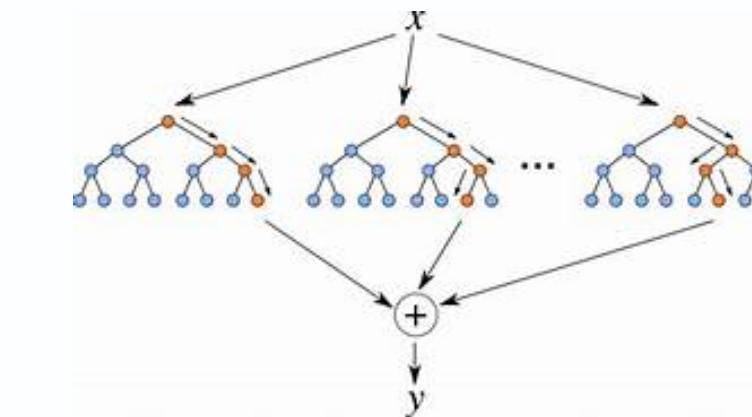
# THUẬT TOÁN SỬ DỤNG



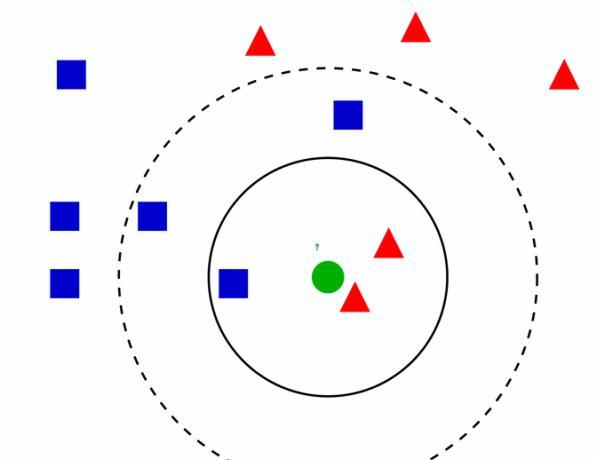
Logistic Regression



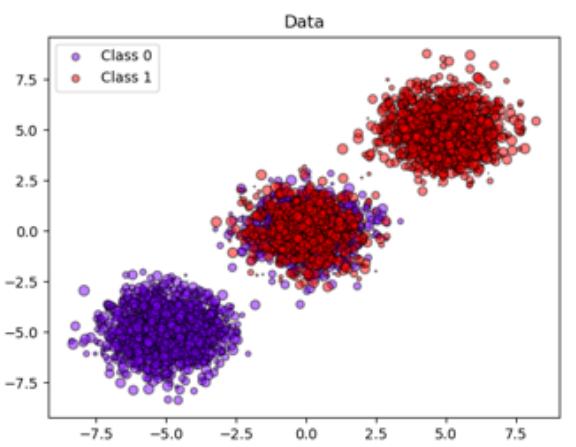
Decision Tree



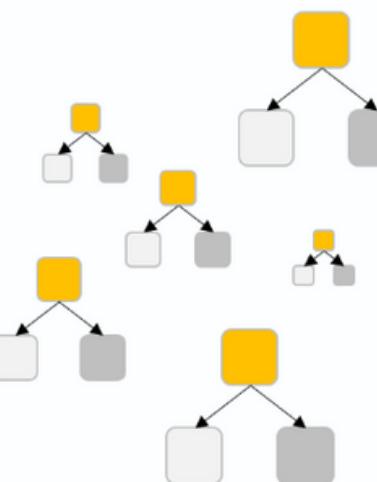
Random Forest



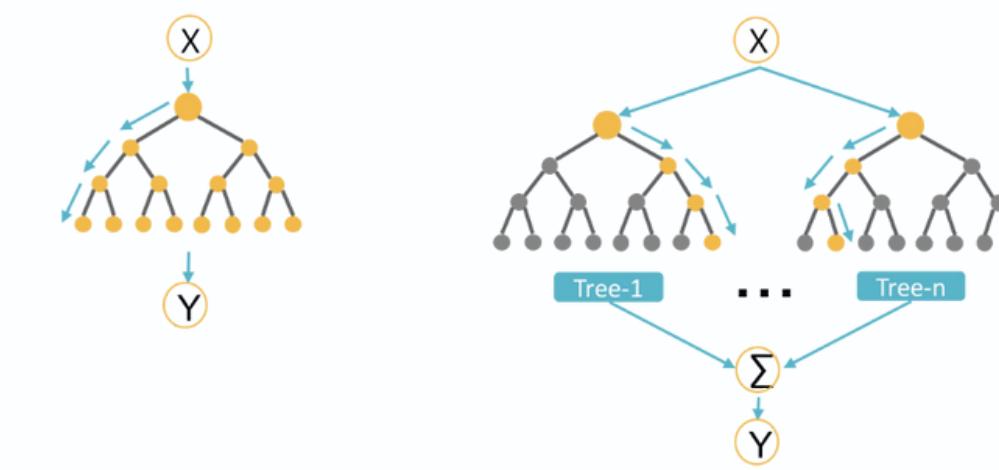
KNN



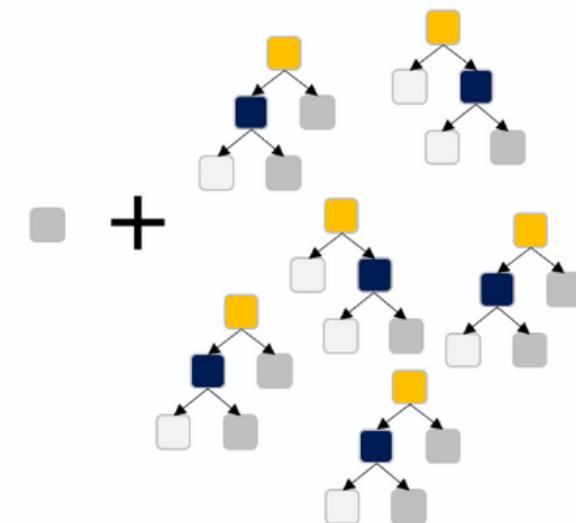
GaussianNB



Ada Boost



XGBoost



Gradient Boosting

# THUẬT TOÁN SỬ DỤNG



```
models = [
    LogisticRegression(),
    DecisionTreeClassifier(),
    GaussianNB(),
    GridSearchCV(RandomForestClassifier(), param_grid={
        'n_estimators': [10, 50, 100, 200],
    }),
    GridSearchCV(KNeighborsClassifier(), param_grid={
        'n_neighbors': [10, 50, 100, 200],
    }),
    GridSearchCV(AdaBoostClassifier(), param_grid={
        'n_estimators': [10, 50, 100, 200],
    }),
    GridSearchCV(GradientBoostingClassifier(), param_grid={
        'n_estimators': [10, 50, 100, 200],
    }),
    GridSearchCV(XGBClassifier(), param_grid={
        'n_estimators': [10, 50, 100, 200],
    }),
]
```

## Label Encoder và OneHotEncoder những cột dữ liệu dạng chuỗi.

	gender	smoking_history
0	Female	never
1	Female	No Info
2	Male	never
3	Female	current
4	Male	current



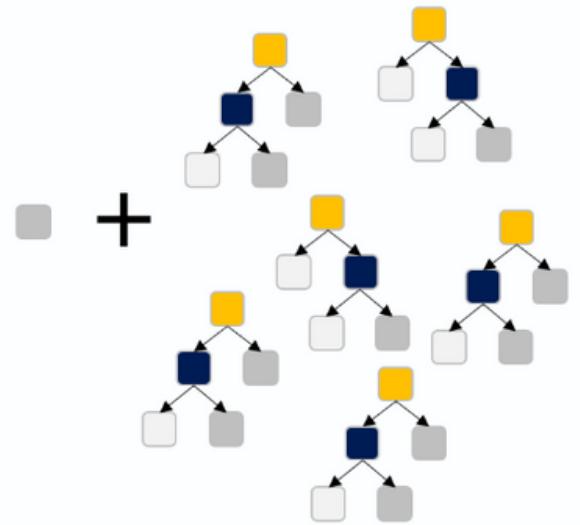
	gender	smoking_history_No Info	smoking_history_current	smoking_history_ever	smoking_history_former	smoking_history_never	smoking_history_not current
0	0	False	False	False	False	True	False
1	0	True	False	False	False	False	False
2	1	False	False	False	False	True	False
3	0	False	True	False	False	False	False
4	1	False	True	False	False	False	False

# SO SÁNH METRICS



Model	Accuracy	Precision		Recall		F1-score	
		0	1	0	1	0	1
LogisticRegression	0.95	0.96	0.87	0.99	0.98	0.98	0.67
Decision Tree	0.95	0.97	0.72	0.97	0.73	0.97	0.73
GaussianNB	0.90	0.97	0.46	0.93	0.65	0.95	0.54
Random Forest	0.97	0.97	0.95	1.00	0.69	0.98	0.80
KNN	0.95	0.95	0.96	1.00	0.47	0.97	0.63
AdaBoost	0.97	0.97	0.97	1.00	0.69	0.98	0.81
<b>GradientBoosting</b>	<b>0.97</b>	<b>0.97</b>	<b>0.99</b>	<b>1.00</b>	<b>0.69</b>	<b>0.99</b>	<b>0.81</b>
XGBoost	0.97	0.97	1.00	1.00	0.67	0.99	0.81

- Dựa vào các độ đo Recall, Precision, F1-score trên từng lớp 0, 1 và Accuracy thì Model Gradient Boosting có sự vượt trội hơn so với các mô hình khác.



→ Gradient Boosting là mô hình tốt nhất cho việc dự đoán bệnh tiểu đường của bệnh nhân.

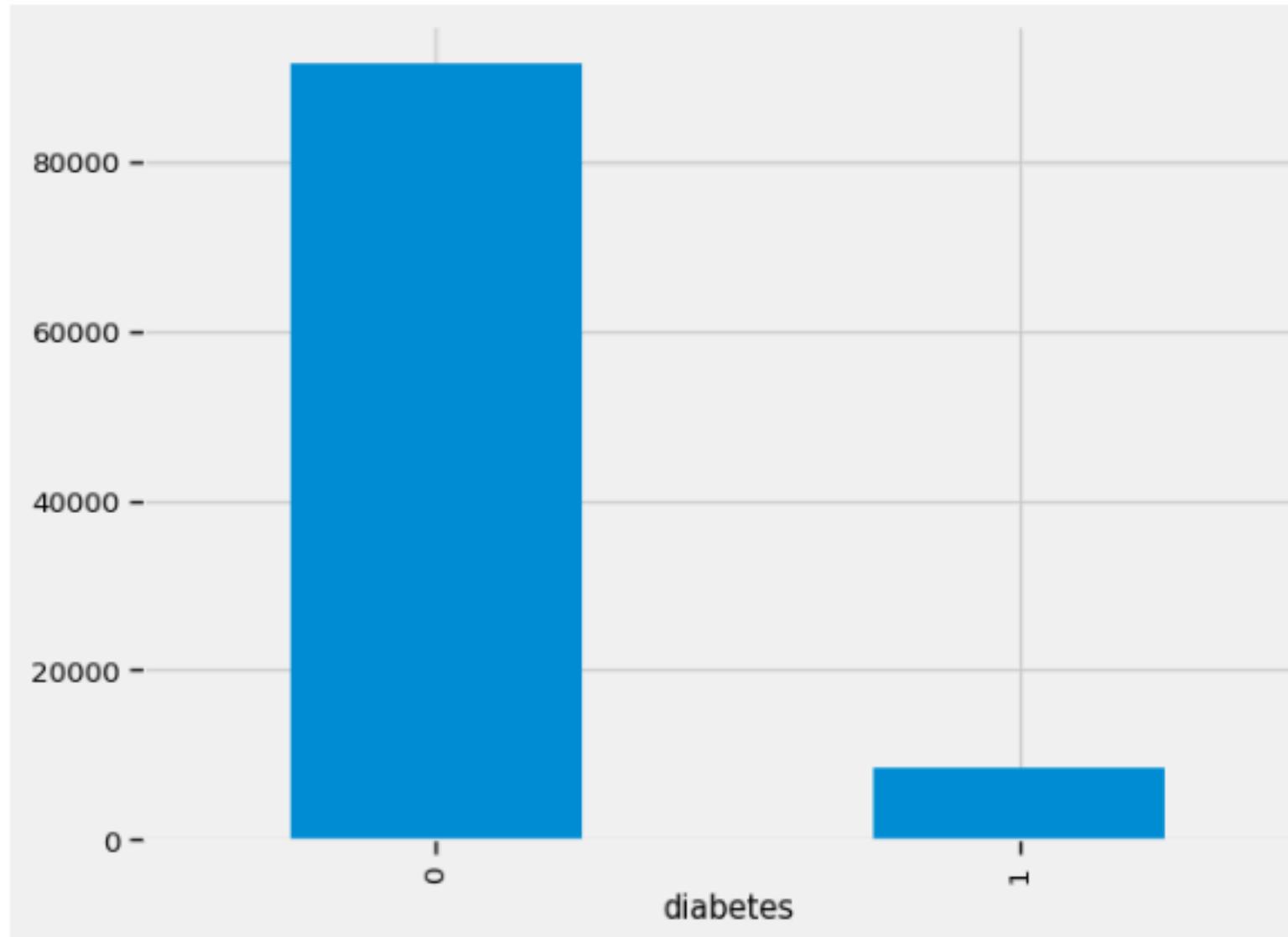
## Gradient Boosting

	precision	recall	f1-score	support
0	0.97	1.00	0.99	18292
1	0.99	0.69	0.81	1708
accuracy			0.97	20000
macro avg	0.98	0.84	0.90	20000
weighted avg	0.97	0.97	0.97	20000

# XỬ LÝ MẤT CÂN BẰNG DỮ LIỆU



## Diabetes



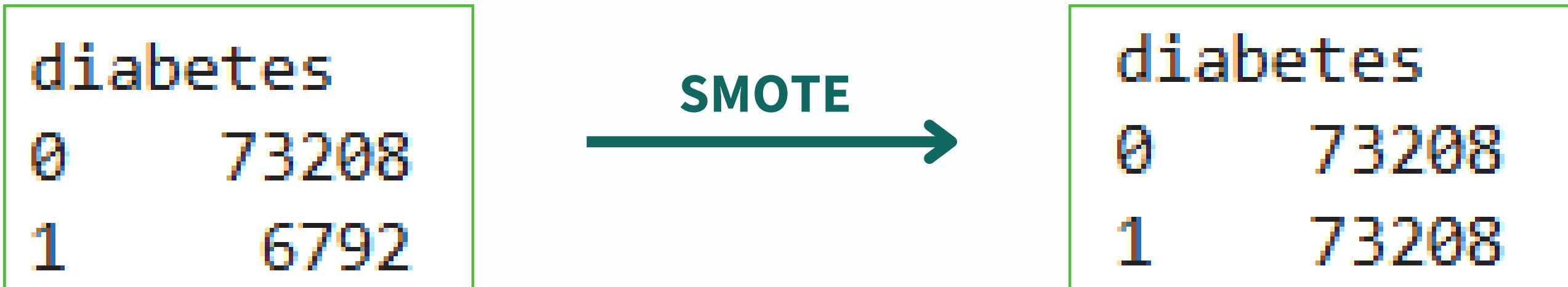
### Phương pháp sử dụng:

- OverSampling
- UnderSampling
- OverSampling + MinMaxScaler

# OVERSAMPLING



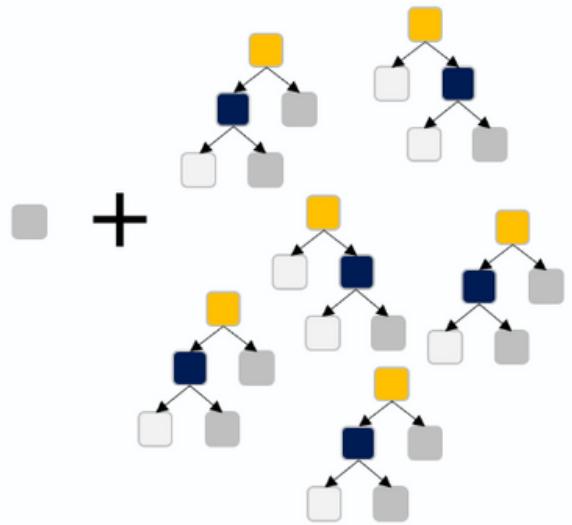
Sử dụng phương pháp **SMOTE (Synthetic Minority Over-sampling Technique)**



# SO SÁNH METRICS



Model	Accuracy	Precision		Recall		F1-score	
		0	1	0	1	0	1
LogisticRegression	0.93	0.98	0.5	0.94	0.75	0.96	0.64
Decision Tree	0.95	0.98	0.69	0.97	0.74	0.97	0.72
GaussianNB	0.84	0.98	0.32	0.84	0.82	0.91	0.46
Random Forest	0.96	0.97	0.84	0.99	0.72	0.98	0.78
KNN	0.90	0.98	0.46	0.92	0.77	0.95	0.58
AdaBoost	0.97	0.97	0.93	1.00	0.70	0.98	0.80
<b>GradientBoosting</b>	<b>0.97</b>	<b>0.97</b>	<b>0.96</b>	<b>1.00</b>	<b>0.69</b>	<b>0.98</b>	<b>0.80</b>
XGBoost	0.97	0.97	0.92	0.99	0.71	0.99	0.80



## Gradient Boosting

- Các metrics sau khi OverSampling khá tương đồng với metrics với dataset ban đầu, không có sự cải thiện trên lớp 1.
- Model Gradient Boosting có sự vượt trội hơn so với các mô hình khác.

	precision	recall	f1-score	support
0	0.97	1.00	0.98	18292
1	0.96	0.69	0.80	1708
accuracy			0.97	20000
macro avg	0.96	0.85	0.89	20000
weighted avg	0.97	0.97	0.97	20000

# UNDERSAMPLING

Giảm số dòng dữ liệu có diabetes = 0 bằng với số dòng có diabetes = 1

diabetes	
0	91500
1	8500

UnderSampling  
→

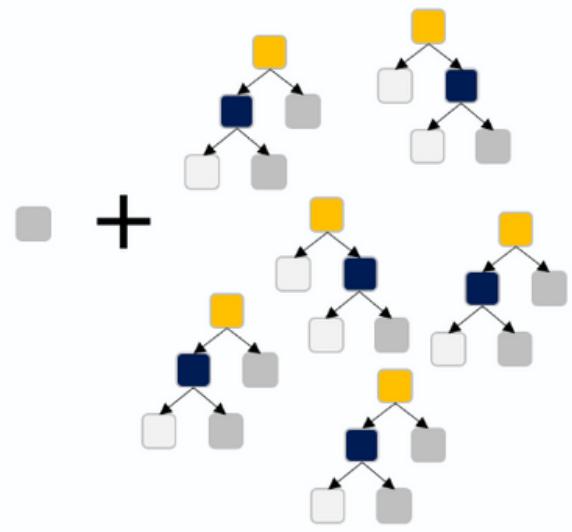
diabetes	
1	8500
0	8500

# SO SÁNH METRICS



Model	Accuracy	Precision		Recall		F1-score	
		0	1	0	1	0	1
LogisticRegression	0.87	0.87	0.87	0.87	0.87	0.87	0.87
Decision Tree	0.88	0.88	0.88	0.88	0.88	0.88	0.88
GaussianNB	0.83	0.81	0.84	0.85	0.80	0.83	0.82
Random Forest	0.91	0.91	0.90	0.90	0.91	0.90	0.91
KNN	0.86	0.85	0.87	0.88	0.85	0.86	0.86
<b>AdaBoost</b>	<b>0.91</b>	<b>0.92</b>	<b>0.90</b>	<b>0.90</b>	<b>0.92</b>	<b>0.91</b>	<b>0.91</b>
<b>GradientBoosting</b>	<b>0.91</b>	<b>0.92</b>	<b>0.90</b>	<b>0.90</b>	<b>0.92</b>	<b>0.91</b>	<b>0.91</b>
XGBoost	0.91	0.91	0.90	0.90	0.91	0.91	0.91

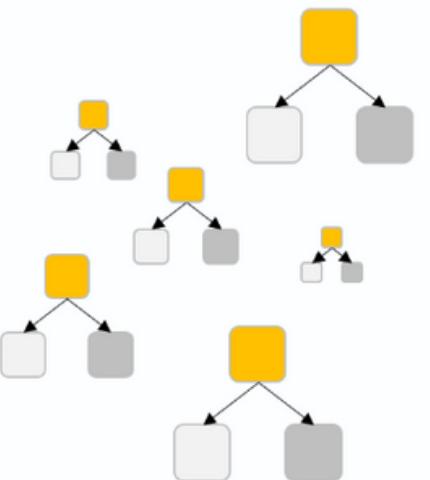
# KẾT LUẬN



Gradient Boosting

- Các metrics có sự cải thiện rõ rệt trên lớp 1, nhưng trên lớp 0 và accuracy có sự giảm nhẹ.
- Nên sử dụng model sau khi UnderSampling để có được hiệu quả cao trên cả hai lớp.

→ Gradient Boosting, AdaBoost là hai mô hình tốt nhất cho việc dự đoán bệnh tiểu đường của bệnh nhân.



Ada Boost

# OVERSAMPLING - MINMAXSCALER



	gender	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes	smoking_history_No Info	smoking_history_current	smoking_history_e
0	0.0	0.549550	0.0	0.0	0.115850	0.545455	0.545455	1.0	0.0	0.0	0.0
1	1.0	0.837337	0.0	1.0	0.217823	0.545455	0.545455	1.0	0.0	0.0	0.0
2	1.0	0.624625	1.0	0.0	0.217823	0.400000	0.818182	1.0	0.0	0.0	1.0
3	1.0	0.912412	0.0	0.0	0.199873	1.000000	0.363636	1.0	0.0	0.0	0.0
4	0.0	0.662162	0.0	0.0	0.217823	0.636364	0.359091	1.0	0.0	0.0	0.0

# SO SÁNH METRICS



Model	Accuracy	Precision		Recall		F1-score	
		0	1	0	1	0	1
LogisticRegression	0.88	0.89	0.88	0.88	0.88	0.88	0.88
Decision Tree	0.88	0.88	0.88	0.88	0.88	0.88	0.88
GaussianNB	0.83	0.81	0.84	0.85	0.80	0.83	0.82
Random Forest	0.90	0.91	0.90	0.90	0.91	0.90	0.90
KNN	0.88	0.88	0.88	0.89	0.88	0.88	0.88
<b>AdaBoost</b>	<b>0.91</b>	<b>0.92</b>	<b>0.90</b>	<b>0.90</b>	<b>0.92</b>	<b>0.91</b>	<b>0.91</b>
<b>GradientBoosting</b>	<b>0.91</b>	<b>0.92</b>	<b>0.90</b>	<b>0.90</b>	<b>0.92</b>	<b>0.91</b>	<b>0.91</b>
XGBoost	0.91	0.91	0.90	0.90	0.91	0.91	0.91



**ĐÁNH GIÁ - KẾT LUẬN**

# ĐỀ XUẤT - KẾT LUẬN

**Những bệnh nhân có khả năng mắc bệnh tiểu đường là người có những đặc điểm sau:**

- Lớn tuổi
- Bị huyết áp cao
- HbA1c\_level khoảng từ 7-9
- Từng hút thuốc lá trong quá khứ
- Mức đường huyết  $> 200 \text{ mg/dL}$
- Béo phì độ 2 trở lên

**Đề xuất đối với nhóm bệnh nhân này**

- Thay đổi thói quen ăn uống, kiểm soát khẩu phần ăn để giảm cân.
- Bỏ hút thuốc lá.
- Thường xuyên tập thể dục thể thao để nâng cao sức khỏe.
- Theo dõi sức khỏe định kỳ.

# ĐỀ XUẤT - KẾT LUẬN

## Kết luận về xây dựng mô hình:

- **Gradient Boosting, AdaBoost** là mô hình có các metrics **vượt trội** hơn so với các mô hình khác khi so sánh trên dataset ban đầu, sau khi OverSampling, UnderSampling, UnderSampling + MinMaxScaler.
- **Gradient Boosting, AdaBoost** là hai mô hình **tốt nhất** cho việc dự đoán bệnh tiểu đường của bệnh nhân. Nên huấn luyện mô hình sau khi đã UnderSampling, nó sẽ cải thiện giá trị các độ đo trên cả lớp 0 và 1, tăng độ chính xác của mô hình.



PNL-DA52

**THANK YOU FOR  
YOUR ATTENTION!**