Big Data Course

Capstone Project   
Final Report

For students (instructor review required)

ⓒ2023 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

**PREDICTING THE INCIDENCE OF DIABETES**

**DỰ ĐOÁN TỶ LỆ MẮC BỆNH TIỂU ĐƯỜNG**

15/08/2024

TEAM Y2K3

Nguyễn Thị Ngọc Mai

Lâm Thị Diễm My

Trần Thanh Trúc

Nguon Gia Linh

Lê Khánh Linh

Đào Ngọc Quý

Nội dung

1. Giới thiệu (Introduction)

1.1. Thông tin nền tảng (Background Information)

1.2. Động lực và mục tiêu (Motivation and Objective)

1.3. Thành viên và vai trò công việc (Members and Role Assignments)

1.4. Lịch trình và các mốc thời gian (Schedule and Milestones)

2. Thực hiện dự án (Project Execution)

2.1. Mô tả kịch bản giả định (Simulated Scenario Description)

2.2. Lựa chọn và mô tả tập dữ liệu (Datasets Selection and Description)

2.3. Đường ống thu thập dữ liệu (Data Ingestion Pipeline)

2.4. Xử lý biến đổi dữ liệu (Data Transformation Processing)

2.5. Truy vấn dữ liệu và tạo thông tin (Data Query and Insight)

3. Kết quả (Results)

3.1. Các đoạn mã nguồn và mã nguồn thu thập dữ liệu (Data Ingestion Scripts and Code)

3.2. Các đoạn mã nguồn và mã nguồn biến đổi dữ liệu (Data Transformation Scripts and Code)

3.3. Mô tả các tập dữ liệu đã biến đổi (Description and Sample of Transformed Datasets)

3.4. Trình bày kết quả truy vấn dữ liệu (Data Visualization of Query Results)

4. Projected Impact

4.1. Thành tựu và những lợi ích (Accomplishments and Benefits)

4.2. Những cải tiến trong tương lai (Future Improvements)

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Giới thiệu (Introduction)

**1.1. Thông tin nền tảng (Background Information)**

Bệnh tiểu đường là một trong những bệnh lý mạn tính phổ biến nhất trên thế giới, ảnh hưởng đến sức khỏe của hàng trăm triệu người. Bệnh tiểu đường xảy ra khi cơ thể không thể sử dụng insulin một cách hiệu quả, dẫn đến tăng lượng đường trong máu. Nếu không được quản lý và điều trị kịp thời, bệnh tiểu đường có thể gây ra các biến chứng nghiêm trọng như bệnh tim mạch, suy thận, mất thị lực và thậm chí tử vong.

Trong những năm gần đây, tỷ lệ mắc bệnh tiểu đường đã tăng lên một cách đáng báo động. Theo ước tính của Tổ chức Y tế Thế giới (WHO), vào năm 2019, có khoảng 463 triệu người trưởng thành mắc bệnh tiểu đường trên toàn cầu. Dự báo tỷ lệ này sẽ tiếp tục gia tăng trong tương lai nếu không có những biện pháp can thiệp kịp thời.

Việc dự đoán chính xác tỷ lệ mắc bệnh tiểu đường trong cộng đồng rất quan trọng để các cơ sở y tế và chính phủ có thể đưa ra các chiến lược phòng ngừa, quản lý bệnh hiệu quả. Điều này sẽ giúp giảm gánh nặng bệnh tật, tăng cường sức khỏe cộng đồng và tiết kiệm chi phí chăm sóc y tế.

**1.2. Động lực và mục tiêu (Motivation and Objective)**

Việc dự đoán sớm và chính xác tỷ lệ mắc bệnh tiểu đường là một yếu tố then chốt trong việc hỗ trợ các chuyên gia y tế đưa ra những biện pháp can thiệp kịp thời. Điều này không chỉ giúp giảm thiểu các biến chứng nghiêm trọng mà còn giảm bớt gánh nặng bệnh tật cho cộng đồng. Trong bối cảnh sự phát triển của công nghệ và khoa học dữ liệu, việc áp dụng các phương pháp phân tích dữ liệu và học máy (machine learning) mở ra những cơ hội mới trong việc dự đoán khả năng mắc bệnh tiểu đường dựa trên các yếu tố nguy cơ như độ tuổi, giới tính, chỉ số BMI, tiền sử bệnh lý, và các chỉ số y tế khác.

* Với những lý do trên, mục tiêu của dự án này là:

- Xây dựng các mô hình dự đoán chính xác tỷ lệ mắc bệnh tiểu đường dựa trên các nguồn dữ liệu khác nhau.

- Cung cấp các công cụ và thông tin hữu ích để hỗ trợ các cơ sở y tế và chính phủ trong việc quản lý và phòng ngừa bệnh tiểu đường hiệu quả.

- Góp phần nâng cao sức khỏe cộng đồng thông qua việc dự báo và quản lý tốt hơn bệnh tiểu đường.

**1.3. Thành viên và vai trò công việc (Members and Role Assignments)**

Dự án này được thực hiện bởi một nhóm gồm 6 thành viên với các vai trò như sau:

* Nguyễn Thị Ngọc Mai : Nhóm trưởng

– Thu thập dữ liệu

– Xử lý dữ liệun lệch

– Xây dựng mô hình học máy

– Mất cân bằng dữ liệu

– Tổng hợp kết quả, viết báo cáo

* Lâm Thị Diễm My :

– Thu thập dữ liệu

– Xử lý dữ liệu lệch

– Xây dựng mô hình học máy

– Mất cân bằng dữ liệu

– Tổng hợp kết quả, viết báo cáo

* Trần Thanh Trúc :

– Trực quan hóa dữ liệu

– Xử lý dữ liệu lệch

– Xây dựng mô hình học máy

– Đánh giá mô hình

* Nguon Gia Linh :

– Thu thập dữ liệu

– Xây dựng mô hình học máy

– Đánh giá mô hình

– Trực quan hóa dữ liệu

– Tổng hợp kết quả, viết báo cáo

* Lê Khánh Linh :

– Thu thập dữ liệu

– Xây dựng mô hình học máy

– Đánh giá mô hình

– Trực quan hóa dữ liệu

– Tổng hợp kết quả, viết báo cáo

* Đào Ngọc Quý :

– Trực quan hóa dữ liệu

– Xây dựng mô hình học máy

– Đánh giá mô hình

**1.4. Lịch trình và các mốc thời gian (Schedule and Milestones)**

1. Thu thập dữ liệu và chuẩn bị dữ liệu (1/6/2014 -> 20/6/2024)

2. Tiền xử lý dữ liệu (21/6/2024 -> 7/7/2024)

3. Lựa chọn mô hình & Huấn luyện mô hình (8/7/2024->28/7/2024)

4. Đánh giá mô hình (29/7/2024 -> 2/8/2024)

5. Kiểm thử Mô hình với Dữ liệu Thực tế (3/8/2024 -> 5/8/2024)

6. Tổng hợp kết quả và hoàn thiện báo cáo (6/8/2024 -> 10/8/2024)

2. Thực hiện dự án (Project Execution)

**2.1. Mô tả kịch bản giả định (Simulated Scenario Description)**

Một nhóm gồm 10.000 bệnh nhân từ 30 đến 60 tuổi, cả nam và nữ, đã đăng ký tham gia chương trình sàng lọc. Mỗi bệnh nhân được yêu cầu cung cấp các thông tin sức khỏe ban đầu, bao gồm:

Độ tuổi: Độ tuổi hiện tại của bệnh nhân.

Giới tính: Nam hoặc Nữ.

Chỉ số BMI: Được tính toán dựa trên chiều cao và cân nặng của bệnh nhân.

Tiền sử bệnh lý: Thông tin về việc bệnh nhân có từng bị bệnh tim mạch, cao huyết áp, hoặc tiền sử gia đình có người mắc tiểu đường hay không.

Hút thuốc: Tiền sử hút thuốc lá của bệnh nhân.

Mức đường huyết: Kết quả kiểm tra đường huyết hiện tại.

Chỉ số HbA1c: Kết quả xét nghiệm cho thấy mức đường huyết trung bình của bệnh nhân trong 2-3 tháng gần nhất.

Các thông tin này sau đó được nhập vào hệ thống, nơi mô hình học máy của chúng ta sẽ phân tích và đưa ra dự đoán về nguy cơ mắc bệnh tiểu đường cho từng bệnh nhân.

Mục tiêu:

Đối với bệnh nhân có nguy cơ cao: Họ sẽ được thông báo và được khuyến nghị thực hiện các xét nghiệm chi tiết hơn, đồng thời có thể bắt đầu các biện pháp phòng ngừa hoặc điều trị sớm.

Đối với bệnh nhân có nguy cơ thấp: Họ sẽ nhận được các khuyến nghị về việc duy trì lối sống lành mạnh và theo dõi sức khỏe định kỳ.

Bằng cách áp dụng mô hình dự đoán này, bệnh viện không chỉ giúp bệnh nhân hiểu rõ hơn về tình trạng sức khỏe của mình mà còn tối ưu hóa quá trình sử dụng nguồn lực y tế, tập trung vào những trường hợp cần thiết nhất.

**2.2. Lựa chọn và mô tả tập dữ liệu (Datasets Selection and Description)**

Bộ dữ liệu gồm 99 977 quan sát, 9 thuộc tính

Các thuộc tính của bộ dữ liệu

* Gender: Giới tính đề cập đến giới tính sinh học của cá nhân, có thể có tác động đến khả năng mắc bệnh tiểu đường của họ. Có hai loại trong đó (male , female)
* Age: Tuổi tác là một yếu tố quan trọng vì bệnh tiểu đường thường được chẩn đoán ở người lớn tuổi. Độ tuổi nằm trong khoảng từ (0-80) trong tập dữ liệu của chúng tôi.
* Hypertension: Tăng huyết áp là một tình trạng bệnh lý trong đó huyết áp trong động mạch tăng cao liên tục. Nó có giá trị (0 hoặc 1) trong đó 0 cho biết họ không bị tăng huyết áp và với 1 có nghĩa là họ bị tăng huyết áp.
* Heart disease: Bệnh tim là một tình trạng y tế khác có liên quan đến việc tăng nguy cơ phát triển bệnh tiểu đường. Nó có giá trị (0 hoặc 1) trong đó 0 cho biết họ không mắc bệnh tim và với 1 có nghĩa là họ mắc bệnh tim.
* Smoking history: Tiền sử hút thuốc cũng được coi là một yếu tố nguy cơ của bệnh tiểu đường và có thể làm trầm trọng thêm các biến chứng liên quan đến bệnh tiểu đường.Nó có giá trị (0 hoặc 1) trong đó 0 cho biết họ không thuốc và với 1 có nghĩa là họ có hút thuốc
* BMI (Body Mass Index): là thước đo lượng mỡ trong cơ thể dựa trên cân nặng và chiều cao. Giá trị BMI cao hơn có liên quan đến nguy cơ mắc bệnh tiểu đường cao hơn. Phạm vi chỉ số BMI trong tập dữ liệu là từ (10,16 đến 71,55). BMI dưới 18,5 là thiếu cân, 18,5-24,9 là bình thường, 25-29,9 là thừa cân và từ 30 trở lên là béo phì.
* HbA1c level(Hemoglobin A1c): là thước đo lượng đường trong máu trung bình của một người trong 2-3 tháng qua. Mức độ cao hơn cho thấy nguy cơ phát triển bệnh tiểu đường cao hơn. Hầu hết mức HbA1c trên 6,5% cho thấy bệnh tiểu đường.
* Blood glucose level: Mức đường huyết đề cập đến lượng glucose trong máu tại một thời điểm nhất định. Mức đường huyết cao là dấu hiệu chính của bệnh tiểu đường.
* Diabetes: Là biến mục tiêu được dự đoán giá trị (0 hoặc 1) với giá trị 1 biểu thị sự hiện diện của bệnh tiểu đường và 0 biểu thị nguy cơ mắc bệnh.

**2.3. Đường ống thu thập dữ liệu (Data Ingestion Pipeline)**

**Nhập thư viện và thiết lập môi trường:** Bước đầu tiên trong quá trình xử lý dữ liệu bệnh tiểu đường là thiết lập môi trường và nhập các thư viện cần thiết như Pandas, NumPy, Matplotlib, Seaborn, và Scikit-learn. Các thư viện này cung cấp công cụ để xử lý dữ liệu, trực quan hóa, và xây dựng các mô hình học máy, hỗ trợ việc phân tích các thuộc tính như Giới tính, Tuổi, Tăng huyết áp, Bệnh tim, Tiền sử hút thuốc, BMI, HbA1c, Mức đường huyết, và Tình trạng tiểu đường.

**Đọc dữ liệu:** Bước tiếp theo là nạp bộ dữ liệu tiểu đường từ tệp CSV vào một cấu trúc có tổ chức, thường là DataFrame, sử dụng Python và thư viện Pandas. Mục tiêu chính của bước này là chuyển dữ liệu từ nguồn lưu trữ vào môi trường làm việc, nhằm chuẩn bị cho các bước phân tích và mô hình hóa tiếp theo liên quan đến việc dự đoán tỷ lệ mắc bệnh tiểu đường dựa trên các thuộc tính đã được cung cấp.

**2.4. Xử lý biến đổi dữ liệu (Data Transformation Processing)**

**2.4.1. Chuyển đổi dữ liệu**

**- Chuyển đổi thuộc tính giới tính (gender)**: Để các thuật toán học máy có thể xử lý dữ liệu dạng chuỗi, cần chuyển đổi giá trị giới tính từ "female" và "male" thành các giá trị số tương ứng. Cụ thể, "female" được chuyển thành 0 và "male" thành 1.

- Sử dụng phương thức **fit\_transform** của đối tượng **LabelEncoder** trong thư viện scikit-learn để chuẩn hóa dữ liệu dạng chuỗi (categorical data) thành số (numerical data).Việc chuyển đổi các giá trị chuỗi thành số giúp các thuật toán này có thể sử dụng được dữ liệu.

A screenshot of a computer

Description automatically generated

**2.4.2. Xử lý dữ liệu trùng, xử lý dữ liệu rỗng**

- Loại bỏ dữ liệu trùng lặp :Dữ liệu trùng lặp có thể gây ra bias trong mô hình học máy, làm sai lệch kết quả và đánh giá không chính xác về hiệu suất của mô hình. Việc loại bỏ dữ liệu trùng lặp đảm bảo rằng mỗi quan sát đều là duy nhất và không ảnh hưởng tiêu cực đến mô hình.

- Loại bỏ dữ liệu bị null : Dữ liệu thiếu (null) có thể gây ra vấn đề khi huấn luyện mô hình, làm cho các thuật toán học máy không thể xử lý được hoặc dẫn đến kết quả không chính xác. Loại bỏ hoặc xử lý dữ liệu thiếu giúp duy trì tính toàn vẹn của tập dữ liệu và đảm bảo rằng các mô hình học máy được huấn luyện trên dữ liệu đầy đủ và nhất quán.

**2.4.3. Xử lý các giá trị ngoại biên**

- Các giá trị ngoại biên có thể làm thay đổi trung bình và độ lệch chuẩn của tập dữ liệu, làm giảm độ chính xác của các phân tích thống kê.Loại bỏ các giá trị ngoại biên giúp cải thiện chất lượng dữ liệu, đảm bảo rằng mô hình học máy được huấn luyện trên dữ liệu đại diện và chính xác.

- Sử dụng biểu đồ boxplot để phát hiện các giá trị ngoại biên. Biểu đồ này hiển thị phạm vi giữa các phần tư (IQR) và các giá trị nằm ngoài phạm vi này được coi là ngoại biên.

A graph with a number of text

Description automatically generated with medium confidence

Từ biểu đồ trên ta có thể thấy:

* HbA1c level và blood glucose level: Hai biến này có nhiều giá trị ngoại biên nhất, thể hiện qua các chấm nằm ngoài "ria" của boxplot. Điều này cho thấy có những bệnh nhân có mức HbA1c và đường huyết bất thường so với phần lớn các bệnh nhân khác.
* Age: Biến tuổi có một số ít giá trị ngoại biên, có thể là những bệnh nhân lớn tuổi hơn so với đa số.
* Các biến còn lại: Các biến gender, hypertension, heart disease, smoking history, bmi và diabetes không có hoặc có rất ít giá trị ngoại biên, cho thấy các biến này phân bố khá tập trung.

- Những bệnh nhân có giá trị ngoại biên có thể có những đặc điểm riêng biệt, ảnh hưởng đến nguy cơ và tiến triển của bệnh tiểu đường. Việc phân tích sâu hơn về các trường hợp này có thể cung cấp thông tin hữu ích cho việc điều trị và chăm sóc bệnh nhân. Và Giá trị ngoại biên có thể ảnh hưởng đến kết quả của các mô hình dự đoán. Nên chúng tôi cân nhắc việc loại bỏ hoặc xử lý các giá trị ngoại biên này để đảm bảo tính ổn định và độ chính xác của mô hình.

- Sau khi cân nhắc chúng tôi đã thực hiện loại bỏ các giá trị ngoại lai bằng cách sử dụng phương pháp IQR (Interquartile Range - khoảng tứ phân vị). Đầu tiên, chúng tôi tính toán các phân vị Q1 (25%) và Q3 (75%) của tập dữ liệu. Sau đó, xác định khoảng tứ phân vị (IQR) bằng cách lấy hiệu của Q3 và Q1.

- Công thức tính IQR:

**IQR=Q3−Q1**

- Dựa trên IQR, chúng tôi tính toán các ngưỡng trên và dưới để xác định giá trị ngoại lai. Cuối cùng, những giá trị nằm ngoài khoảng này được loại bỏ khỏi tập dữ liệu, tạo ra một tập dữ liệu mới sạch hơn và đáng tin cậy hơn cho các phân tích tiếp theo.

- Dữ liệu chúng tôi có phân bố đặc biệt, không tuân theo phân phối chuẩn. Trong trường hợp này, việc sử dụng quy tắc 1.5 \* IQR có thể không phù hợp để xác định giá trị ngoại lai. Vì vậy chúng tôi đã cộng thêm 1 vào IQR có thể giúp điều chỉnh ngưỡng để phù hợp hơn với đặc điểm phân bố của dữ liệu.  Mục tiêu của chúng tôi là không phải là loại bỏ tất cả các giá trị ngoại lai, mà chỉ là loại bỏ những giá trị ngoại lai cực đoan nhất. Việc cộng thêm 1 vào IQR có thể giúp bạn đạt được mục tiêu này bằng cách nới rộng ngưỡng xác định giá trị ngoại lai.

**2.4.4. Xử lý độ lệch dữ liệu**

- Trong bài toán dự đoán bệnh tiểu đường, tập dữ liệu về bệnh tiểu đường của chúng tôi có sự mất cân bằng lớp, với số lượng mẫu của lớp "không mắc bệnh" lớn hơn nhiều so với lớp "mắc bệnh". Điều này có thể dẫn đến việc các mô hình học máy thiên về dự đoán lớp đa số và bỏ qua lớp thiểu số, gây ra hiệu suất kém trong việc phát hiện các trường hợp mắc bệnh. Do đó chúng tôi đã sử dụng **kỹ thuật Hybrid Methods** (kết hợp thủ công giữa Oversampling và Undersampling) giúp cân bằng lại tỷ lệ giữa các lớp bằng cách tăng số lượng mẫu của lớp thiểu số (oversampling) và giảm số lượng mẫu của lớp đa số (undersampling). Điều này giúp mô hình học tập được các đặc trưng của cả hai lớp một cách hiệu quả hơn.

- Vì kỹ thuật Hybrid Methods tận dụng ưu điểm của cả Oversampling và Undersampling, trong đó:

* Oversampling giúp tăng cường sự hiện diện của lớp thiểu số, cung cấp thêm thông tin cho mô hình học tập. Và chúng tôi sẽ Oversampling bằng SMOTE (Synthetic Minority Over-sampling Technique). Áp dụng SMOTE lên tập huấn luyện, tạo ra tập dữ liệu mới X\_smote và y\_smote với số lượng mẫu của hai lớp được cân bằng hơn. SMOTE tạo ra các mẫu tổng hợp mới cho lớp thiểu số bằng cách nội suy giữa các mẫu hiện có của lớp đó.
* Undersampling giúp giảm bớt sự áp đảo của lớp đa số. Và chúng tôi sẽ Undersampling bằng  Random Undersampling. Áp dụng undersampling lên tập dữ liệu đã được oversampling, tạo ra tập dữ liệu cuối cùng X\_resampled và y\_resampled với số lượng mẫu của hai lớp cân bằng. Random Undersampling loại bỏ ngẫu nhiên một số mẫu từ lớp đa số.

- Hybrid Methods kết hợp cả hai kỹ thuật này, tận dụng ưu điểm của từng phương pháp và giảm thiểu nhược điểm của chúng.

**2.4.5 Xây dựng mô hình**

Sau khi sử dụng các kỹ thuật, những thuật toán để xử lý và chuyển đổi dữ liệu thì chúng tôi đã có một tập dữ liệu mới chuẩn hóa và cân bằng. Tiếp theo chúng tôi phân tích và xây dựng mô hình dự đoán dựa trên tập dữ liệu này.

Nhóm chúng tôi đã chọn ra 5 mô hình có những đặc tính mạnh mẽ và phù hợp với việc phân tích một bộ dữ liệu lớn có gần 100.000 quan sát. Sau đây chúng tôi sẽ đi phân tích từng mô hình, tiến hành đánh giá và lựa chọn 1 mô hình phù hợp nhất cho dự án này.

**Mô hình Logistic Regression**

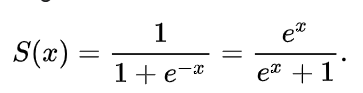
**Lý do chọn mô hình:**

* **Tính chất phân loại của bài toán:** Mô hình Logistic Regression phù hợp với các bài toán phân loại nhị phân, nơi biến mục tiêu chỉ có hai giá trị có thể xảy ra (ở đây là 0 hoặc 1 cho biến Diabetes). Điều này làm cho Logistic Regression trở thành một lựa chọn tự nhiên cho việc dự đoán khả năng mắc bệnh tiểu đường dựa trên các yếu tố sức khỏe như Gender, Age, Hypertension, Heart Disease, Smoking History, BMI, HbA1c Level, và Blood Glucose Level.
* **Hiệu quả với dữ liệu y tế:** Các thuộc tính trong tập dữ liệu của chúng tôi như: Age, BMI, và Blood Glucose Level là các biến liên tục, trong khi các thuộc tính khác như Gender, Hypertension, và Heart Disease là biến nhị phân. Logistic Regression có thể xử lý tốt cả hai loại biến này, giúp tận dụng tối đa thông tin từ dữ liệu đã cung cấp.
* **Tính đơn giản và hiệu quả:** Logistic Regression là một mô hình đơn giản, không đòi hỏi quá nhiều tính toán phức tạp, nhưng vẫn mang lại hiệu suất cao trong việc dự đoán. Điều này rất phù hợp với chúng tôi vì chúng tôi muốn có một mô hình dễ triển khai và giải thích trong bối cảnh nghiên cứu y tế.

**Nguyên lý hoạt động:**

Phân loại nhị phân: Logistic Regression hoạt động bằng cách ước lượng xác suất một đối tượng thuộc vào một trong hai lớp (ví dụ: có hoặc không mắc bệnh tiểu đường). Để làm điều này, mô hình áp dụng một hàm logistic (hay còn gọi là sigmoid) lên một tổ hợp tuyến tính của các biến đầu vào.

Hàm sigmoid: Hàm sigmoid có dạng

​

trong đó

S(x): là xác suất của sự kiện y bằng 1, tức là xác suất mắc bệnh tiểu đường.

x là một tổ hợp tuyến tính của các biến đầu vào (như tuổi, BMI, mức đường huyết …)

e là cơ số của logarit tự nhiên (khoảng 2.71828)

Ở mô hình này chúng tôi có sử dụng tham số **random\_state=42** nhằm để thiết lập hạt giống cho bộ sinh số ngẫu nhiên trong quá trình huấn luyện mô hình. Điều này giúp đảm bảo rằng kết quả của mô hình có thể được tái lập mỗi khi chạy lại quá trình huấn luyện với cùng một dữ liệu và tham số. Đây là một thực hành quan trọng để đảm bảo tính nhất quán trong nghiên cứu, giúp chúng tôi có thể so sánh kết quả một cách công bằng giữa các lần chạy khác nhau.

**Mô hình Random Forest**

**Lý do chọn mô hình Random Forest**

Random Forest là một lựa chọn tốt cho bài toán dự đoán bệnh tiểu đường do hiệu suất tốt, khả năng xử lý dữ liệu mất cân bằng và tính ổn định. Ưu điểm của Random Forest:

* **Hiệu suất tốt:** Random Forest thường cho kết quả dự đoán tốt trên nhiều loại dữ liệu khác nhau, bao gồm cả dữ liệu y tế. Nó có khả năng xử lý tốt các mối quan hệ phi tuyến tính và tương tác giữa các biến, điều này thường xảy ra trong dữ liệu thực tế.
* **Khả năng xử lý dữ liệu mất cân bằng:** Random Forest có thể tự điều chỉnh để xử lý các lớp dữ liệu không cân bằng, như trong trường hợp dữ liệu bệnh tiểu đường, nơi số lượng người không mắc bệnh thường lớn hơn nhiều so với số lượng người mắc bệnh.
* **Tính ổn định:** Random Forest ít bị ảnh hưởng bởi nhiễu trong dữ liệu và có khả năng chống overfitting (quá khớp) tốt hơn so với một số mô hình khác như cây quyết định đơn lẻ.
* **Quan trọng của đặc trưng:** Random Forest có thể cung cấp thông tin về tầm quan trọng của từng đặc trưng trong việc dự đoán, giúp hiểu rõ hơn về các yếu tố ảnh hưởng đến bệnh tiểu đường.

**Nguyên lý hoạt động**

Random Forest hoạt động bằng cách xây dựng nhiều cây quyết định trên các tập con dữ liệu khác nhau và kết hợp dự đoán của chúng để đưa ra dự đoán cuối cùng. Sự kết hợp này giúp giảm thiểu overfitting và tăng tính tổng quát của mô hình, đồng thời tận dụng sức mạnh của nhiều cây quyết định để đưa ra dự đoán chính xác hơn.

Ở mô hình này chúng tôi có sử dụng:

**tham số random\_state=42** nhằm để thiết lập hạt giống cho bộ sinh số ngẫu nhiên trong quá trình huấn luyện mô hình.

**n\_estimators** là số lượng cây quyết định trong rừng ngẫu nhiên. Mặc định là 100. Tăng số lượng cây thường cải thiện hiệu suất nhưng cũng làm tăng thời gian huấn luyện

**bootstrap:** Kiểm soát việc sử dụng bootstrap sampling (lấy mẫu có hoàn lại) khi xây dựng các cây. Mặc định là True. Nếu đặt là False, sẽ sử dụng toàn bộ tập dữ liệu để xây dựng mỗi cây.

**max\_depth:** Độ sâu tối đa của mỗi cây quyết định. Mặc định là None, tức là cây sẽ được phát triển đến khi tất cả các lá là thuần nhất hoặc đạt đến các điều kiện dừng khác. Giới hạn max\_depth có thể giúp ngăn ngừa overfitting.

**Mô hình K-Nearest Neighbors**

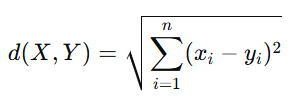
* **Lý do chọn mô hình KNN :**
* **Bản chất không tuyến tính của dữ liệu:** KNN là một mô hình phi tham số (non-parametric), nghĩa là nó không đưa ra giả định về mối quan hệ tuyến tính giữa các đặc trưng và nhãn đầu ra. Trong bộ dữ liệu, mối quan hệ giữa các yếu tố như tuổi tác, chỉ số BMI, mức HbA1c, và tình trạng bệnh tiểu đường có thể không phải là tuyến tính. KNN có thể linh hoạt trong việc tìm kiếm các mẫu phức tạp và không tuyến tính trong dữ liệu.
* **Tính đơn giản và dễ hiểu:** KNN là một trong những thuật toán đơn giản nhất, dễ dàng giải thích và triển khai. Giúp dễ hiểu cách thức mà thuật toán đưa ra quyết định, đặc biệt là khi giải thích kết quả dự đoán.
* **Khả năng xử lý dữ liệu hỗn hợp:** Bộ dữ liệu bao gồm cả biến số (như tuổi, chỉ số BMI,...) và biến phân loại (như giới tính, tiền sử hút thuốc,...). KNN có thể xử lý tốt cả hai loại biến này mà không cần nhiều tiền xử lý phức tạp.
* **Tính linh hoạt với các lựa chọn khác nhau của “k”:**Có thể điều chỉnh giá trị “k” để đạt được độ chính xác cao nhất cho bài toán của mình. Điều này giúp tối ưu hóa mô hình dựa trên đặc điểm cụ thể của bộ dữ liệu.
* **Có tính chính xác cao:** Thuật toán KNN có thể đạt độ chính xác cao, đặc biệt khi giá trị của k được chọn cẩn thận. Nó thường được sử dụng làm thuật toán chuẩn cho nhiều tác vụ học máy.

**Nguyên lý hoạt động**

Khi áp dụng thuật toán K-Nearest Neighbors (KNN) vào bài toán hồi quy, chúng ta sử dụng KNN để dự đoán giá trị liên tục thay vì dự đoán lớp như trong bài toán phân loại. Đây là cách KNN hoạt động trong bài toán hồi quy:

Chọn giá trị K: Đầu tiên, bạn phải chọn giá trị K, tức là số lượng láng giềng gần nhất mà bạn muốn sử dụng để đưa ra dự đoán. Giá trị K càng lớn, mô hình sẽ càng trở nên mượt mà hơn, nhưng cũng có thể làm giảm độ chính xác.

Xác định khoảng cách: KNN sử dụng một phương pháp đo khoảng cách (thường là khoảng cách Euclidean) để xác định mức độ tương đồng giữa các điểm dữ liệu. Càng gần, khoảng cách càng nhỏ. Sử dụng công thức tính khoảng cách Euclid để tính khoảng cách 2 điểm X , Y trong không gian n chiều, công thức khoảng cách Euclid giữa chúng là:



xi: tọa độ của điểm X tại chiều thứ i.

yi: tọa độ của điểm Y tại chiều thứ i.

n: số chiều của không gian đặc trưng

* Tìm K láng giềng gần nhất: Đối với mỗi điểm dữ liệu trong tập kiểm tra, KNN tìm K láng giềng gần nhất trong tập huấn luyện dựa trên khoảng cách được xác định. Ở bài toán của chúng tôi, giá trị k đã được thử nghiệm để tìm ra giá trị tối ưu bằng cách chạy thử và vẽ biểu đồ cuối cùng chọn được k = 13 là giá trị k tối ưu.
* Đưa ra dự đoán: Khi có K láng giềng, giá trị hồi quy có thể được dự đoán bằng cách lấy trung bình (hoặc trọng số trung bình, tùy thuộc vào cách bạn cấu hình mô hình) của giá trị của chúng.

**Mô hình Naive Bayes**

**Lý do chọn mô hình:**

* **Hiệu suất tốt với các đặc tính độc lập:** Mô hình Naive Bayes hoạt động dựa trên giả định rằng các thuộc tính của tập dữ liệu là độc lập với nhau, điều này thường không hoàn toàn đúng trong thực tế, nhưng mô hình này vẫn cho hiệu suất dự đoán khá tốt trong nhiều trường hợp. Với dữ liệu về tiểu đường, giả định này có thể chấp nhận được vì mỗi thuộc tính như tuổi, giới tính, chỉ số BMI, và tiền sử bệnh có thể được xem như là độc lập tương đối với nhau.
* **Xử lý tốt với dữ liệu dạng phân loại và liên tục:** Naive Bayes có khả năng xử lý tốt cả dữ liệu phân loại (như giới tính, tiền sử hút thuốc) và dữ liệu liên tục (như BMI, mức đường huyết), điều này rất phù hợp với tập dữ liệu về dự đoán tiểu đường của bạn.
* **Hiệu suất cao trên dữ liệu thiếu:** Naive Bayes thường cho kết quả tốt ngay cả khi dữ liệu thiếu hoặc không hoàn toàn chính xác, điều này rất hữu ích nếu tập dữ liệu của bạn có chứa các trường hợp dữ liệu không đầy đủ.

**Nguyên lý hoạt động:**

Thuật toán phân loại có giám sát, dựa trên Định lý Bayes. Ý tưởng chính của nó là dự đoán xác suất của một phần tử dữ liệu thuộc vào một lớp được sử dụng để giải quyết bài toán phân loại:

* Chọn giả định:Đầu tiên, chọn giả định về độc lập. Naive Bayes giả định rằng tất cả các thuộc tính đầu vào đều độc lập với nhau.
* Áp dụng Định lý Bayes: Naive Bayes sử dụng Định lý Bayes để tính xác suất hậu nghiệm, tức là xác suất của một sự kiện A xảy ra khi biết sự kiện B đã xảy ra. Trong bài toán phân loại, A là một phần tử dữ liệu, B là một giả thiết về lớp mà A thuộc về.
* Tính xác suất: Đối với mỗi lớp, Naive Bayes tính xác suất của mỗi thuộc tính dữ liệu thuộc về lớp đó. Xác suất này được tính dựa trên tần suất xuất hiện của thuộc tính trong lớp.
* Đưa ra dự đoán: Khi có xác suất cho mỗi lớp, Naive Bayes dự đoán lớp của một điểm dữ liệu bằng cách chọn lớp có xác suất hậu nghiệm cao nhất.

**Mô hình XGBoost**

XGBoost (Extreme Gradient Boosting) là một giải thuật được base trên gradient boosting, tuy nhiên kèm theo đó là những cải tiến to lớn về mặt tối ưu thuật toán, về sự kết hợp hoàn hảo giữa sức mạnh phần mềm và phần cứng, giúp đạt được những kết quả vượt trội cả về thời gian training cũng như bộ nhớ sử dụng

Lý do nên sử dụng mô hình XGBoost

* **Hiệu suất cao:**

- XGBoost thường đạt được độ chính xác và hiệu suất vượt trội so với các thuật toán boosting khác như GBM và LGBM, đặc biệt trên các bộ dữ liệu y tế lớn và phức tạp

- Độ chính xác cao của mô hình là rất quan trọng trong lĩnh vực y tế, vì nó ảnh hưởng trực tiếp đến chẩn đoán và điều trị bệnh

* **Khả năng xử lý dữ liệu tốt:**

- Dữ liệu y tế thường có nhiều đặc trưng khác nhau, bị khuyết, và kết hợp cả dữ liệu phân loại và hồi quy

- XGBoost có khả năng xử lý tốt các dạng dữ liệu này, giúp tận dụng tối đa thông tin từ nguồn dữ liệu mà không cần phải tiền xử lý quá nhiều

* **Tốc độ nhanh:**

- XGBoost nhanh hơn nhiều so với GBM và LGBM trong quá trình huấn luyện và dự đoán

- Tốc độ nhanh là yếu tố quan trọng trong các ứng dụng y tế, vì cần đưa ra kết quả nhanh chóng để hỗ trợ chẩn đoán và điều trị

* **Tính linh hoạt và tùy biến:**

- XGBoost cung cấp nhiều tùy chọn và tham số để điều chỉnh, giúp người dùng có thể tối ưu hóa mô hình phù hợp với từng bài toán y học cụ thể

- Điều này cho phép người dùng dễ dàng thích ứng mô hình với các yêu cầu và ràng buộc khác nhau trong lĩnh vực y tế

* **Hỗ trợ tính toán song song:**

- XGBoost có khả năng tận dụng tốt các tài nguyên phần cứng như GPU và CPU, giúp tăng tốc quá trình huấn luyện mô hình

- Điều này rất hữu ích khi làm việc với bộ dữ liệu y tế lớn

XGBoost sử dụng một công thức tính toán cụ thể để huấn luyện và cập nhật các cây quyết định trong quá trình Gradient Boosting. Công thức chính được sử dụng trong XGBoost là:

**L(φ) = Σ l(yi, ŷi) + Σ Ω(fk)**

Trong đó:

L(φ): Hàm mất mát tổng thể (overall loss function) cần được tối thiểu hóa

l(yi, ŷi): Hàm mất mát của mỗi mẫu dữ liệu, với `yi` là giá trị thực và `ŷi` là giá trị dự đoán

Ω(fk): Hàm regularization, giúp ngăn chặn hiện tượng quá khớp (overfitting)

Cụ thể, hàm regularization Ω(fk) được định nghĩa như sau:

**Ω(fk) = γT + (1/2) \* λ ||w||2**

Trong đó:

T: Số lượng lá (leaf) của cây quyết định

w: Các trọng số của các lá

γ và λ: Các siêu tham số cần được điều chỉnh để cân bằng giữa độ chính xác và độ phức tạp của mô hình

Quá trình tối ưu hóa XGBoost sẽ lặp lại các bước sau:

1. Huấn luyện một cây quyết định mới để giảm thiểu hàm mất mát L(φ)

2. Cập nhật các trọng số của mô hình bằng cách sử dụng gradient descent

3. Lặp lại quá trình huấn luyện và cập nhật cho đến khi đạt được mô hình tối ưu

Công thức này giúp XGBoost đạt được hiệu suất cao và khả năng xử lý dữ liệu phức tạp, đồng thời ngăn chặn hiện tượng quá khớp

**2.5. Truy vấn dữ liệu và tạo thông tin (Data Query and Insight)**

**2.5.1 Xây dựng mô hình**

Sau khi thực hiện các mô hình khác nhau, chúng tôi đã thu được những kết quả như sau:

\* **Mô hình Logistic Regression**

A graph with numbers and a number in blue squares

Description automatically generated

**Ý nghĩa các giá trị:**

* **True Negatives (TN) = 10643:** Số lượng mẫu thực tế không mắc bệnh tiểu đường (0) và được dự đoán đúng là không mắc bệnh (0).
* **False Positives (FP) = 1934:** Số lượng mẫu thực tế không mắc bệnh tiểu đường (0) nhưng bị dự đoán nhầm là mắc bệnh (1).
* **False Negatives (FN) = 61:** Số lượng mẫu thực tế mắc bệnh tiểu đường (1) nhưng bị dự đoán nhầm là không mắc bệnh (0).
* **True Positives (TP) = 541:** Số lượng mẫu thực tế mắc bệnh tiểu đường (1) và được dự đoán đúng là mắc bệnh (1).

**Tính toán các chỉ số đánh giá:**

* **Accuracy (Độ chính xác):** (TN + TP) / (TN + FP + FN + TP) 10643 + 541) / (10643+1934+61+541) ≈ 0.8486
* **Precision (Độ chính xác dương tính):** TP / (TP + FP) = 541/ (541+1934) ≈ 0.2186
* **Recall (Độ nhạy):** TP / (TP + FN) = 541/ (541 + 61) ≈ 0.8987
* **F1-score:** 2 \* (Precision \* Recall) / (Precision + Recall) ≈ 2 \* (0.2186 \* 0.8987) / (0.2186+0.8987) ≈ 0.3516

**- Nhận xét:**

* Mô hình có khả năng dự đoán rất tốt các trường hợp **không mắc bệnh tiểu đường**, thể hiện qua số lượng TN rất lớn so với FP.
* Tuy nhiên, mô hình có thể cần cải thiện trong việc dự đoán các trường hợp **mắc bệnh tiểu đường**, vì số lượng FN vẫn còn khá cao so với TP.
* **Accuracy cao (0.8486)** cho thấy mô hình có độ chính xác tổng thể khá tốt.
* **Precision (0.2186)** tương đối thấp, cho thấy mô hình có thể tạo ra một số lượng đáng kể các dự đoán dương tính giả (False Positives).
* **Recall (0.8987)** ở mức khá, cho thấy mô hình có thể phát hiện được phần lớn các trường hợp thực sự mắc bệnh tiểu đường.
* **F1-score (0.3516 )** là trung bình điều hòa giữa Precision và Recall, cung cấp một đánh giá cân bằng hơn về hiệu suất của mô hình.

- Ta có thể thấy, mô hình có độ chính xác tổng thể tốt, nhưng khả năng dự đoán Lớp 1(mắc bệnh tiểu đường) thấp, nhìn chung chúng tôi nghĩ là do sự mất cân bằng về số lượng mẫu giữa hai lớp (support) có thể là một nguyên nhân dẫn đến sự chênh lệch về hiệu suất giữa các lớp mặc dù đã sử dụng class\_weight='balanced' đã giúp cải thiện recall cho lớp 1, nhưng precision vẫn còn thấp nên kéo F1 - score xuống.

**\* Mô hình Random Forest**

A graph with numbers and a square

Description automatically generated

**- Ý nghĩa các giá trị:**

* **True Negatives (TN) = 12425:** Số lượng mẫu thực tế không mắc bệnh tiểu đường (0) và được dự đoán đúng là không mắc bệnh (0).
* **False Positives (FP) = 152:** Số lượng mẫu thực tế không mắc bệnh tiểu đường (0) nhưng bị dự đoán nhầm là mắc bệnh (1).
* **False Negatives (FN) = 268:** Số lượng mẫu thực tế mắc bệnh tiểu đường (1) nhưng bị dự đoán nhầm là không mắc bệnh (0).
* **True Positives (TP) = 334:** Số lượng mẫu thực tế mắc bệnh tiểu đường (1) và được dự đoán đúng là mắc bệnh (1).

**- Tính toán các chỉ số đánh giá:**

* **Accuracy (Độ chính xác):** (TN + TP) / (TN + FP + FN + TP) = **12425**+ **334**) / (**12425** + **152**+ **268**+ **334**) ≈ 0.9681
* **Precision (Độ chính xác dương tính):** TP / (TP + FP) = **334**/ (**334** + **152**) ≈ 0.6872
* **Recall (Độ nhạy):** TP / (TP + FN) = **334**/ (**334** + **268**) ≈ 0.5548
* **F1-score:** 2 \* (Precision \* Recall) / (Precision + Recall) ≈ 2 \* (0.6872 \* 0.5548) / (0.6872 + 0.5548) ≈ 0.6140

**- Nhận xét:**

* Mô hình có khả năng dự đoán rất tốt các trường hợp **không mắc bệnh tiểu đường**, thể hiện qua số lượng TN rất lớn so với FP.
* Tuy nhiên, mô hình có thể cần cải thiện trong việc dự đoán các trường hợp **mắc bệnh tiểu đường**, vì số lượng FN vẫn còn khá cao so với TP.
* **Accuracy cao (0.9681)** cho thấy mô hình có độ chính xác tổng thể rất tốt.
* **Precision (0.6872)** tương đối cao,cho thấy rằng trong các trường hợp được dự đoán là mắc bệnh tiểu đường, phần lớn là đúng. Điều này cũng có nghĩa là mô hình ít tạo ra dự đoán dương tính giả (False Positives).
* **Recall (0.5548)** ở mức trung bình, cho thấy mô hình có khả năng phát hiện được hơn một nửa các trường hợp thực sự mắc bệnh tiểu đường.Tuy nhiên, vẫn còn một số lượng khá lớn các trường hợp mắc bệnh (False Negatives - FN) bị bỏ sót.
* **F1-score (0.6140)** là trung bình điều hòa giữa Precision và Recall, cung cấp một đánh giá cân bằng về hiệu suất của mô hình.

- Qua phân tích trên, mô hình Random Forest có hiệu suất tốt trong việc dự đoán bệnh tiểu đường, đặc biệt là trong việc xác định các trường hợp không mắc bệnh. Tuy nhiên, vẫn còn tiềm năng để cải thiện khả năng phát hiện các trường hợp mắc bệnh (Recall), nhằm giảm số lượng dự đoán âm tính sai (False Negatives). Điều này sẽ giúp mô hình trở nên hiệu quả hơn trong việc dự đoán bệnh tiểu đường, đặc biệt là trong các trường hợp nhạy cảm liên quan đến sức khỏe.

**\* Mô hình K-Nearest Neighbors**

**A blue squares with numbers and a white background

Description automatically generated**

**- Ý nghĩa các giá trị:**

* True Negatives (TN) = 12569: Số lượng mẫu thực tế không mắc bệnh tiểu đường (0) và được dự đoán đúng là không mắc bệnh (0).
* False Positives (FP) = 8: Số lượng mẫu thực tế không mắc bệnh tiểu đường (0) nhưng bị dự đoán nhầm là mắc bệnh (1).
* False Negatives (FN) = 458: Số lượng mẫu thực tế mắc bệnh tiểu đường (1) nhưng bị dự đoán nhầm là không mắc bệnh (0).
* True Positives (TP) = 144: Số lượng mẫu thực tế mắc bệnh tiểu đường (1) và được dự đoán đúng là mắc bệnh (1).

**- Tính toán các chỉ số đánh giá:**

* Accuracy (Độ chính xác): (TN + TP) / (TN + FP + FN + TP) = 12569+ 144) / (12569 + 8+ 458+ 144) ≈ 0.9646
* Precision (Độ chính xác dương tính): TP / (TP + FP) = 144/ (144 + 8) ≈ 0.9474
* Recall (Độ nhạy): TP / (TP + FN) = 144/ (144 + 458) ≈ 0.2392
* F1-score: 2 \* (Precision \* Recall) / (Precision + Recall) ≈ 2 \* (0.9474 \* 0.2392) / (0.9474 + 0.2392) ≈ 0.3817

**- Nhận xét:**

* Mô hình KNN có khả năng dự đoán tốt các trường hợp không mắc bệnh tiểu đường với số lượng TN rất lớn và độ chính xác dương tính cao.
* Tuy nhiên, mô hình gặp khó khăn trong việc phát hiện các trường hợp mắc bệnh tiểu đường, thể hiện qua số lượng FN cao và độ nhạy thấp.
* Accuracy rất cao (0.9646): Mô hình có độ chính xác tổng thể rất tốt, khi dự đoán đúng nhãn cho phần lớn các mẫu trong tập kiểm tra. Điều này cho thấy mô hình có khả năng phân loại chính xác phần lớn các trường hợp, đặc biệt là các trường hợp không mắc bệnh tiểu đường.
* Precision cao (0.9474): Trong số các trường hợp được mô hình KNN dự đoán là mắc bệnh tiểu đường, có tới 94.7% là dự đoán chính xác. Điều này cho thấy mô hình có khả năng giảm thiểu đáng kể các dự đoán dương tính giả (False Positives), tức là mô hình ít nhầm lẫn giữa các trường hợp không mắc bệnh thành mắc bệnh.
* Recall thấp (0.2392): Mô hình chỉ phát hiện được 24% các trường hợp thực sự mắc bệnh tiểu đường. Điều này có nghĩa là mô hình KNN bỏ sót một số lượng lớn các trường hợp dương tính thật (False Negatives). Mặc dù mô hình rất chính xác trong việc tránh các dự đoán sai dương tính, nhưng lại không hiệu quả trong việc phát hiện tất cả các trường hợp mắc bệnh.
* F1-score (0.3817): F1-score thấp phản ánh sự mất cân bằng giữa Precision và Recall. Mặc dù mô hình có độ chính xác cao khi dự đoán lớp dương tính, nhưng do khả năng phát hiện các trường hợp thực sự mắc bệnh (Recall) khá thấp, nên F1-score cũng bị ảnh hưởng.

- Qua phân tích trên thì mô hình KNN hoạt động rất tốt trong việc giảm thiểu các dự đoán dương tính giả, giúp giảm bớt nhầm lẫn giữa các trường hợp không mắc bệnh thành mắc bệnh. Tuy nhiên, mô hình gặp khó khăn lớn trong việc phát hiện tất cả các trường hợp thực sự mắc bệnh tiểu đường, điều này có thể dẫn đến việc bỏ sót các bệnh nhân cần điều trị. Do đó, mô hình cần được cải thiện để tăng khả năng phát hiện bệnh, hoặc có thể kết hợp với các phương pháp khác nhằm đạt được hiệu suất tốt hơn trong việc chẩn đoán bệnh tiểu đường.

**\* Mô hình Naive Bayes**

**A blue squares with numbers and a number

Description automatically generated**

**- Ý nghĩa các giá trị**

* True Negatives (TN) = 11782: Số lượng mẫu thực tế không mắc bệnh tiểu đường (0) và được dự đoán đúng là không mắc bệnh (0).
* False Positives (FP) = 795: Số lượng mẫu thực tế không mắc bệnh tiểu đường (0) nhưng bị dự đoán nhầm là mắc bệnh (1).
* False Negatives (FN) = 384: Số lượng mẫu thực tế mắc bệnh tiểu đường (1) nhưng bị dự đoán nhầm là không mắc bệnh (0).
* True Positives (TP) = 218: Số lượng mẫu thực tế mắc bệnh tiểu đường (1) và được dự đoán đúng là mắc bệnh (1).

**- Chỉ số đánh giá:**

* Accuracy (Độ chính xác): (TN + TP) / (TN + FP + FN + TP) = (11782 + 218) / (11782 + 795 + 384 + 218) ≈ 0.9105
* Precision (Độ chính xác dương tính): TP / (TP + FP) = 218 / (218 + 795) ≈ 0.2152
* Recall (Độ nhạy): TP / (TP + FN) = 218 / (218 + 384) ≈ 0.3621
* F1-score: 2 \* (Precision \* Recall) / (Precision + Recall) ≈ 2 \* (0.215 \* 0.362) / (0.215 + 0.362) ≈ 0.270

**- Nhận xét:**

* Mô hình Naive Bayes có khả năng dự đoán tốt các trường hợp không mắc bệnh tiểu đường với số lượng TN lớn.
* Tuy nhiên, mô hình gặp khó khăn trong việc phát hiện các trường hợp mắc bệnh tiểu đường, thể hiện qua số lượng FN cao và độ nhạy thấp.
* Accuracy (0.9105): Mô hình có độ chính xác tổng thể tốt, dự đoán đúng nhãn cho 91.1% số mẫu trong tập kiểm tra.
* Precision (0.2152): Thấp, cho thấy mô hình có xu hướng tạo ra nhiều dự đoán dương tính giả (False Positives).
* Recall (0.3621): Cũng thấp, chỉ ra rằng mô hình bỏ sót một số lượng đáng kể các trường hợp thực sự mắc bệnh tiểu đường (False Negatives).
* F1-score (0.270): Thấp, phản ánh sự mất cân bằng giữa Precision và Recall. Mô hình cần cải thiện cả khả năng dự đoán chính xác các trường hợp mắc bệnh và giảm số lượng dự đoán dương tính giả.

- Ta có thể thấy, Naive Bayes có độ chính xác tổng thể khá, nhưng hiệu suất dự đoán cho lớp thiểu số (lớp 1 - mắc bệnh tiểu đường) là kém nhất trong số các mô hình từ đầu đến giờ. Mô hình này có xu hướng tạo ra nhiều dự đoán dương tính giả và bỏ sót một số trường hợp thực sự mắc bệnh.

**\* Mô hình XGBoost**

A blue and white graph

Description automatically generated

**- Ý nghĩa các giá trị :**

* **True Negatives (TN) = 12524**: Số lượng mẫu thực tế không mắc bệnh tiểu đường (0) và được dự đoán đúng là không mắc bệnh (0).
* **False Positives (FP) = 53**: Số lượng mẫu thực tế không mắc bệnh tiểu đường (0) nhưng bị dự đoán nhầm là mắc bệnh (1).
* **False Negatives (FN) = 271**: Số lượng mẫu thực tế mắc bệnh tiểu đường (1) nhưng bị dự đoán nhầm là không mắc bệnh (0).
* **True Positives (TP) = 331**: Số lượng mẫu thực tế mắc bệnh tiểu đường (1) và được dự đoán đúng là mắc bệnh (1).

**- Chỉ số đánh giá:**

* **Accuracy (Độ chính xác):** (TN + TP) / (TN + FP + FN + TP) = (12524 + 331) / (12524 + 53+ 271+ 331) ≈ 0.9754
* **Precision (Độ chính xác dương tính):** TP / (TP + FP) = 331/ (331 + 53) ≈ 0.861
* **Recall (Độ nhạy):** TP / (TP + FN) = 331/ (331 + 271) ≈ 0.5498
* **F1-score:** 2 \* (Precision \* Recall) / (Precision + Recall) ≈ 2 \* (0.861 \* 0.5498) / (0.861 + 0.5498) ≈ 0.6701

**- Nhận xét:**

* Mô hình XGBoost có khả năng dự đoán rất tốt các trường hợp **không mắc bệnh tiểu đường** với số lượng TN lớn và độ chính xác dương tính cao.
* Mô hình cũng có độ nhạy và F1-score tốt, nhưng vẫn có thể cải thiện trong việc phát hiện các trường hợp **mắc bệnh tiểu đường** để nâng cao độ nhạy.
* **Accuracy rất cao (0.9754)**: Mô hình XGBoost có độ chính xác tổng thể rất tốt, khi dự đoán đúng nhãn cho phần lớn các mẫu trong tập kiểm tra.
* **Precision cao (0.861):** Trong số các trường hợp được dự đoán là mắc bệnh tiểu đường, có tới 86.1% là dự đoán chính xác. Điều này cho thấy mô hình có khả năng hạn chế việc nhầm lẫn giữa các trường hợp không mắc bệnh thành mắc bệnh, giúp giảm thiểu các dự đoán dương tính giả (False Positives).
* **Recall trung bình (0.5498)**: Mô hình phát hiện được khoảng 55% các trường hợp thực sự mắc bệnh tiểu đường. Mặc dù mô hình có độ chính xác cao trong việc tránh các dự đoán sai dương tính, nhưng lại không hiệu quả hoàn toàn trong việc phát hiện tất cả các trường hợp mắc bệnh.
* **F1-score (0.6701)**: F1-score cho thấy một sự cân bằng giữa Precision và Recall, phản ánh rằng mặc dù mô hình có độ chính xác cao trong dự đoán lớp dương tính, nhưng vẫn còn nhiều trường hợp dương tính thật bị bỏ sót.

- Qua phân tích trên ta có thể thấy, mô hình XGBoost hoạt động rất tốt trong việc giảm thiểu các dự đoán dương tính giả, với độ chính xác và Precision cao. Tuy nhiên, mô hình cần được cải thiện để tăng khả năng phát hiện tất cả các trường hợp mắc bệnh tiểu đường (tăng Recall), hoặc kết hợp với các phương pháp khác để đạt được hiệu suất tốt hơn trong việc chẩn đoán bệnh tiểu đường.

**2.5.2. Đánh giá mô hình**

**A screenshot of a computer

Description automatically generated**

Sau khi chạy trên các mô hình như trên thì chúng tôi đã vẽ biểu đồ ROC (Receiver Operating Characteristic) và nó là một công cụ trực quan mạnh mẽ để đánh giá và so sánh hiệu suất của các mô hình phân loại, đặc biệt là trong các tình huống mà dữ liệu có thể không cân bằng giữa các lớp như tập dữ liệu của chúng tôi

**A graph of a line graph

Description automatically generated with medium confidence**

**- Dựa vào biểu đồ ROC (Receiver Operating Characteristic)  chúng tôi đưa ra những nhận xét về các mô hình như sau:**

* **XGBoost:** Mô hình XGBoost thể hiện hiệu suất tốt nhất với AUC (Area Under the Curve) là 0.96. Điều này cho thấy XGBoost có khả năng phân biệt tốt nhất giữa các lớp dương tính và âm tính trong tập dữ liệu của chúng tôi.
* **Logistic Regression:** Với AUC = 0.95, Logistic Regression cũng cho thấy hiệu suất rất tốt, gần tương đương với XGBoost.
* **Random Forest:** AUC = 0.94, hiệu suất của Random Forest cũng rất tốt, chỉ kém hơn một chút so với hai mô hình trên.
* **Naive Bayes:** AUC = 0.90, mô hình Naive Bayes có hiệu suất khá tốt, nhưng không bằng ba mô hình trên.
* **K-Nearest Neighbors:** Với AUC = 0.87, K-Nearest Neighbors có hiệu suất thấp nhất trong số các mô hình được đánh giá.

**- Qua đó có thể thấy:**

* **XGBoost là mô hình có hiệu suất tốt nhất dựa trên AUC.**
* **Logistic Regression và Random Forest cũng là những lựa chọn tốt, đặc biệt nếu bạn cần một mô hình có khả năng diễn giải tốt hơn (Logistic Regression) hoặc có khả năng xử lý tốt dữ liệu phức tạp và không tuyến tính (Random Forest).**
* **Naive Bayes có thể là một lựa chọn nếu bạn cần một mô hình đơn giản và nhanh chóng, nhưng hiệu suất của nó không bằng các mô hình khác.**
* **K-Nearest Neighbors có vẻ kém hiệu quả hơn so với các mô hình khác trong trường hợp này**

**- Kết luận:**

Trong ngữ cảnh y tế, không thể tránh khỏi việc bỏ sót một số trường hợp mắc bệnh tiểu đường, điều có thể dẫn đến những hậu quả nghiêm trọng. Do đó, chúng tôi đã chọn mô hình XGBoost, được xem là lựa chọn tối ưu cho dữ liệu của chúng tôi. Bên cạnh đó, dựa trên cả ma trận nhầm lẫn và biểu đồ ROC, chúng tôi khẳng định rằng XGBoost là mô hình phù hợp nhất cho bài toán dự đoán tỷ lệ mắc bệnh tiểu đường. Với mục tiêu phân loại chính xác các trường hợp có nguy cơ mắc bệnh, XGBoost không chỉ đạt độ chính xác cao mà còn thể hiện sự cân bằng giữa độ chính xác và độ phủ. Điều này được phản ánh qua F1-score vượt trội hơn so với các mô hình khác. Hơn nữa, chỉ số AUC cao nhất (0.96) từ biểu đồ ROC chứng tỏ XGBoost có khả năng phân biệt giữa các trường hợp mắc bệnh và không mắc bệnh, giúp tối ưu hóa khả năng dự đoán và giảm thiểu tỷ lệ sai sót. Nhờ những ưu điểm vượt trội về hiệu suất phân loại và khả năng phân biệt, XGBoost là mô hình phù hợp nhất để giải quyết bài toán dự đoán tỷ lệ mắc bệnh tiểu đường của chúng tôi.

**2.5.2. Thử nghiệm mô hình XGBoost**

- Để đánh giá khả năng dự đoán của mô hình XGBoost sau khi đã huấn luyện . Trong trường hợp này, chúng tôi đã sử dụng một file Excel chứa 43 mẫu dữ liệu để kiểm tra hiệu suất của mô hình trên các dữ liệu chưa từng gặp trước đây. Mục tiêu của việc này là xác định xem mô hình có thể áp dụng tốt cho những trường hợp mới, thực tế hay không.

A screenshot of a computer

Description automatically generated

-Kết quả chúng tôi thu được cho thấy trong 43 mẫu dữ liệu, có 34 trường hợp mà mô hình dự đoán chính xác (giá trị được cho trong file excel là 'diabetes' và giá trị dự đoán của mô hình 'Predicted' bằng nhau), và có 9 trường hợp mà mô hình dự đoán sai ((giá trị được cho trong file excel là 'diabetes' và giá trị dự đoán của mô hình 'Predicted' khác nhau). Điều này có nghĩa là mô hình đã đạt được tỷ lệ chính xác là khoảng 79% trên tập dữ liệu này.

- Mặc dù mô hình XGBoost vẫn có 9 trường hợp dự đoán sai trong tổng số 43 dữ liệu, nhưng với 34 dự đoán chính xác, mô hình đã chứng tỏ độ chính xác đáng kể và đủ khả năng đáp ứng yêu cầu của dự án. Sau khi cân nhắc các yếu tố, chúng tôi quyết định rằng hiệu suất hiện tại của mô hình là đủ để sử dụng trong dự đoán tỷ lệ mắc bệnh tiểu đường. Với kết quả đạt được, mô hình này sẽ được áp dụng mà không cần thêm các bước tối ưu hóa, vì nó đã đáp ứng tốt các tiêu chí về độ tin cậy và độ chính xác mà chúng tôi đặt ra.

3. Kết quả (Results)

**3.1. Các đoạn mã nguồn và mã nguồn thu thập dữ liệu (Data Ingestion Scripts and Code)**

A black background with white text

Description automatically generated

* drive là một mô-đun trong thư viện google.colab cung cấp các phương thức để tương tác với Google Drive từ Google Colab.
* drive.mount('/content/drive') sẽ yêu cầu bạn cấp quyền truy cập Google Drive và sau đó gắn Google Drive vào thư mục /content/drive trong môi trường Google Colab.
* Sau khi gắn, bạn có thể truy cập các tệp và thư mục trong Google Drive của mình thông qua đường dẫn /content/drive/My Drive/.

A black background with white text

Description automatically generated

* pd.read\_csv() là một hàm từ thư viện pandas dùng để đọc dữ liệu từ tệp CSV vào bộ dữ liệu.
* Đường dẫn '/content/drive/MyDrive/SIC/diabetes\_prediction.csv' chỉ đến tệp CSV trong Google Drive.

**3.2. Các đoạn mã nguồn và mã nguồn biến đổi dữ liệu (Data Transformation Scripts and Code)**

A screen shot of a computer code

Description automatically generated

* Nhóm thực hiện loại bỏ những dữ liệu bị trùng lặp bằng hàm drop\_duplicates() để tránh bị sai lệch kết quả
* Dữ liệu null có thể làm giảm chất lượng và độ chính xác của mô hình dự đoán nên cần phải loại ra. Nên phải dùng hàm dropna() để loại những dữ liệu có giá trị null

A screen shot of a computer program

Description automatically generated

* Sử dụng công cụ LabelEncoder thư viện sklearn giúp chuyển đổi các giá trị thuộc tính phân loại (categorical) thành các giá trị số nguyên (integer), giúp dễ dàng xử lý chúng trong các mô hình học máy.
* Khởi tạo đối tượng label\_encoder

A black screen with white text

Description automatically generated

* data\_string\_cols chứa danh sách tên các cột dạng chuỗi trong DataFrame df, đã được xác định trước đó.
* Với mỗi cột, phương thức fit\_transform của label\_encoder sẽ:
* fit: Tìm tất cả các giá trị duy nhất trong cột đó và gán một số nguyên cho mỗi giá trị.
* transform: Chuyển đổi tất cả các giá trị chuỗi trong cột thành các số nguyên tương ứng.
* Kết quả là cột ban đầu trong DataFrame df sẽ được thay thế bằng các giá trị số.

Xem xét điều chỉnh IQR và loại bỏ các giá trị ngoại biên.

A screenshot of a computer

Description automatically generated

* q1 (Phân vị thứ nhất) là giá trị ở vị trí thứ 25% của mẫu dữ liệu, tức là điểm mà 25% dữ liệu nằm dưới giá trị này. df.quantile(0.25) tính toán giá trị phân vị thứ nhất cho tất cả các cột trong bộ dữ liệu.
* q3 (Phân vị thứ ba) là giá trị ở vị trí thứ 75% của mẫu dữ liệu, tức là điểm mà 75% dữ liệu nằm dưới giá trị này. df.quantile(0.75) tính toán giá trị phân vị thứ ba cho tất cả các cột trong bộ dữ liệu.
* Thực hiện tính toán iqr bằng cách lấy q3 trừ q1.

A screenshot of a medical report

Description automatically generated

* bmi: Chỉ số BMI có sự phân tán khá lớn cho thấy có sự biến động khá lớn trong giá trị BMI.
* HbA1c\_level:  Mức HbA1c có sự phân tán nhỏ hơn so với BMI. Một IQR nhỏ hơn có thể cho thấy rằng mức HbA1c trong tập dữ liệu có sự phân phối khá đồng đều và ít biến động hơn.
* blood\_glucose\_level: Mức đường huyết có sự phân tán rất lớn.Cho thấy sự biến động lớn trong mức đường huyết của các đối tượng trong tập dữ liệu. Điều này có thể chỉ ra rằng có nhiều giá trị khác nhau về mức đường huyết, bao gồm cả những giá trị rất cao hoặc rất thấp.

A screen shot of a computer program

Description automatically generated

* Lấy các giá trị duy nhất của 4 giá trị hypertension, heart\_disease, smoking\_history, diabetes. Việc kiểm tra các giá trị duy nhất unique() giúp xác nhận rằng các cột này chỉ chứa các giá trị 0 và 1.
* Thêm 1 vào IQR của các biến này để ngưỡng trên và ngưỡng dưới không bị ảnh hưởng khi tính toán. Điều này giúp tránh việc loại bỏ tất cả các dòng có giá trị 1, vì nếu IQR là 0, ngưỡng trên và ngưỡng dưới sẽ trùng nhau và có thể dẫn đến việc các giá trị 1 bị coi là ngoại biên và bị loại bỏ.

A computer screen with text on it

Description automatically generated

* bien\_duoi được tính bằng phân vị thứ nhất q1 trừ đi 1.5 lần khoảng tứ phân vị iqr.
* bien\_tren được tính bằng phân vị thứ ba q3 cộng thêm 1.5 lần khoảng tứ phân vị iqr.
* Loại bỏ các giá trị ngoại biên. Tạo một mặt nạ boolean, trong đó:
* df < bien\_duoi: Xác định các giá trị nhỏ hơn biên dưới.
* df > bien\_tren: Xác định các giá trị lớn hơn biên trên.
* any(axis=1): Kiểm tra từng hàng trong DataFrame để xác định xem có bất kỳ giá trị nào trong hàng đó nằm ngoài phạm vi hợp lệ.
* Cuối cùng sẽ có được f\_no\_outlier chỉ chứa các hàng không có giá trị ngoại biên.

**3.3. Mô tả các tập dữ liệu đã biến đổi (Description and Sample of Transformed Datasets)**

A screenshot of a computer

Description automatically generated

Bộ dữ liệu ban đầu khi chưa qua các bước xử lý dữ liệu có **99 977 quan sát,** 9 thuộc tính. Thuộc tính ‘gender’ có giá trị chữ là ‘female’ và ‘male’

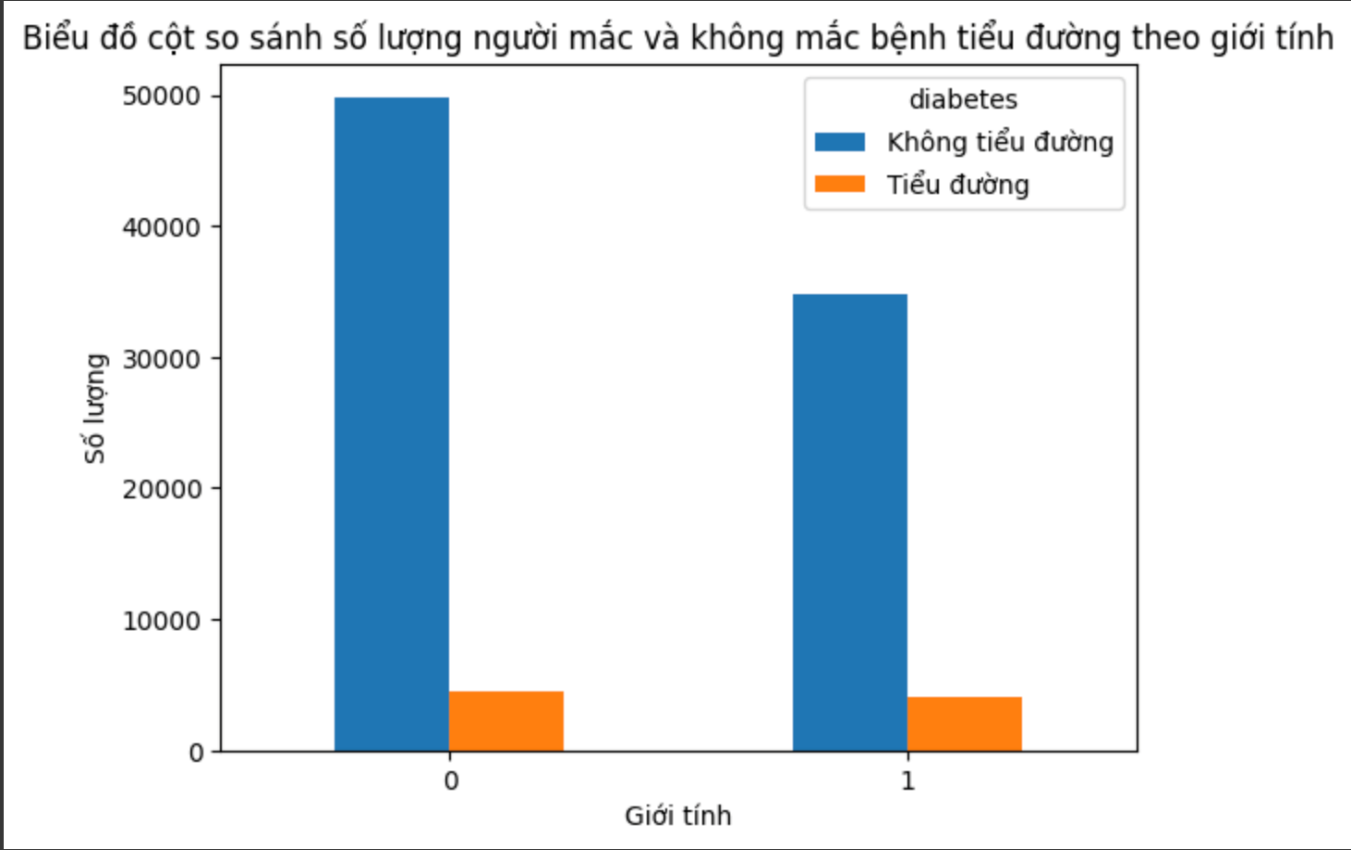
Sau khi thực hiện các bước xử lý dữ liệu thì bộ dữ liệu còn lại 65891 quan sát, 9 thuộc tính và thuộc tính ‘gender’ đã được chuyển đổi từ kiểu dữ liệu chữ sang kiểu dữ liệu số với ‘female’ là 0 và ‘male’ là 1

**3.4. Trình bày kết quả truy vấn dữ liệu (Data Visualization of Query Results)**

- Trong nghiên cứu về bệnh tiểu đường, việc phân tích ảnh hưởng của độ tuổi và giới tính đến kết quả mắc bệnh là một bước quan trọng để hiểu rõ hơn về đặc điểm của bệnh. Biểu đồ thể hiện mối quan hệ giữa độ tuổi, giới tính và kết quả mắc bệnh tiểu đường có thể mang lại nhiều mục đích quan trọng như cho phép chúng ta xác định các nhóm đối tượng có nguy cơ mắc bệnh tiểu đường cao hơn dựa trên độ tuổi và giới tính.Chúng ta cũng có thể so sánh tỷ lệ mắc bệnh giữa nam và nữ ở các độ tuổi khác nhau giúp chúng ta đánh giá tác động của yếu tố giới tính đến nguy cơ mắc bệnh tiểu đường. Ngoài ra biểu đồ này cũng có thể giúp chúng ta hiểu rõ hơn về tiến trình phát triển của bệnh tiểu đường theo thời gian và theo giới tính và những yếu tố này có thể góp phần để xây dựng các mô hình dự đoán nguy cơ mắc bệnh tiểu đường.

- Để xây dựng 2 biểu đồ dưới đây chúng tôi đã dùng những thuộc tính:

* **Gender**: Giới tính đề cập đến giới tính sinh học của cá nhân, có thể có tác động đến khả năng mắc bệnh tiểu đường của họ. Có hai loại trong đó (**male , female**)
* **Age**: Tuổi tác là một yếu tố quan trọng vì bệnh tiểu đường thường được chẩn đoán ở người lớn tuổi. Độ tuổi nằm trong khoảng từ (**0-80**) trong tập dữ liệu của chúng tôi.
* **Diabetes**: Là biến mục tiêu được dự đoán giá trị (**0 hoặc 1**) với giá trị 1 biểu thị sự hiện diện của bệnh tiểu đường và 0 biểu thị nguy cơ mắc bệnh.



*Hình 1: Biểu đồ cột so sánh số lượng người mắc bệnh và không mắc bệnh tiểu đường theo giới tính.*

- Biểu đồ cột trên so sánh số lượng người mắc và không mắc bệnh tiểu đường theo giới tính, với trục hoành thể hiện giới tính (0 là nữ, 1 là nam) và trục tung thể hiện số lượng người. Qua đó ra rút ra được:

* Nhìn chung, tỷ lệ mắc bệnh tiểu đường ở cả hai giới đều thấp so với số lượng người không mắc bệnh. Điều này thể hiện qua việc cột màu cam (tiểu đường) thấp hơn rất nhiều so với cột màu xanh (không tiểu đường) ở cả hai giới tính.
* Giới tính có thể không phải là yếu tố quyết định. Mặc dù có sự khác biệt về số lượng người mắc bệnh giữa hai giới (biểu đồ cho thấy số lượng nam giới mắc bệnh tiểu đường cao hơn nữ giới) nhưng tỷ lệ mắc bệnh ở cả hai giới đều thấp. Giới tính có thể là một yếu tố ảnh hưởng đến nguy cơ mắc bệnh tiểu đường, nhưng không phải là yếu tố quyết định duy nhất.

A graph with blue and orange bars

Description automatically generated

*Hình 2: Biểu đồ cột so sánh số lượng người mắc và không mắc bệnh tiểu đường theo nhóm tuổi*

- Biểu đồ cột trên so sánh số lượng người mắc và không mắc bệnh tiểu đường theo các nhóm tuổi được chúng tôi chia ra như sau: 0-20, 21-40, 41-60 và 61+. Qua đó có thể thấy:

* Nhìn chung, tỷ lệ mắc bệnh tiểu đường tăng lên theo độ tuổi. Nhóm tuổi 61+ có số lượng người mắc bệnh cao nhất, trong khi nhóm 0-20 tuổi có số lượng người mắc bệnh thấp nhất. Điều này phù hợp với thực tế là bệnh tiểu đường thường gặp ở người lớn tuổi hơn.
* Ngược lại, số lượng người không mắc bệnh tiểu đường giảm dần theo độ tuổi. Nhóm 0-20 tuổi có số lượng người không mắc bệnh cao nhất, trong khi nhóm 61+ có số lượng người không mắc bệnh thấp nhất.
* **N**hóm tuổi 41-60 có số lượng người mắc bệnh tiểu đường tăng đáng kể so với nhóm 21-40 tuổi, cho thấy đây là giai đoạn có nguy cơ mắc bệnh cao.

- Từ đó ta dễ dàng thấy mối liên hệ rõ ràng giữa tuổi tác và nguy cơ mắc bệnh tiểu đường. Người lớn tuổi có nguy cơ mắc bệnh cao hơn so với người trẻ tuổi. Đặc biệt, nhóm tuổi 41-60 là giai đoạn chuyển tiếp quan trọng, khi nguy cơ mắc bệnh tăng lên đáng kể.

- Biểu đồ thể hiện lượng đường trong máu trung bình, cụ thể là mức đường huyết và HbA1c (glycated hemoglobin), có vai trò quan trọng trong việc đánh giá và quản lý bệnh tiểu đường. HbA1c Level và Blood Glucose Level cung cấp thông tin quan trọng về mức độ kiểm soát tiểu đường của bệnh nhân trong thời gian dài. Sự kiểm soát không tốt có thể dẫn đến các biến chứng tiểu đường như tổn thương thần kinh, mạch máu, và các vấn đề sức khỏe khác. Mức đường huyết và HbA1c cao thường đi kèm với nguy cơ cao hơn của các biến chứng liên quan đến tiểu đường. Việc theo dõi và duy trì mức đường huyết và HbA1c trong khoảng mục tiêu giúp giảm nguy cơ các vấn đề sức khỏe nghiêm trọng. Dựa vào biểu đồ về mức độ của HbA1c và mức đường huyết, bác sĩ có thể điều chỉnh kế hoạch điều trị, bao gồm chế độ ăn uống, tập luyện, và sử dụng insulin hoặc thuốc giảm đường huyết để kiểm soát tiểu đường. Vì vậy, biểu đồ lượng đường trong máu trung bình đóng vai trò quan trọng trong việc dự đoán và đánh giá nguy cơ mắc bệnh. Nó cung cấp thông tin hữu ích cho các chuyên gia y tế trong việc chẩn đoán, theo dõi và quản lý sức khỏe.

**Biểu đồ 1: Phân bố HbA1c Level theo tình trạng mắc bệnh tiểu đường**

* HbA1c (Hemoglobin A1c) là một chỉ số đo lường lượng đường bám vào hồng cầu trong máu, phản ánh mức đường huyết trung bình trong khoảng 2-3 tháng trước đó.
* Biểu đồ này thể hiện sự phân bố của HbA1c level theo tình trạng mắc bệnh tiểu đường (diabetes = 0 là không mắc, diabetes = 1 là có mắc).
* Đối với những người không mắc bệnh tiểu đường (diabetes = 0):
* HbA1c level tập trung chủ yếu trong khoảng 5-6%, với một số giá trị lên đến 7%.
* Điều này cho thấy nhóm này có mức đường huyết ổn định, nằm trong giới hạn bình thường.
* Đối với những người mắc bệnh tiểu đường (diabetes = 1):
* HbA1c level phân bố rộng hơn, từ khoảng 6% đến 9%, tập trung chủ yếu ở 7-8%.
* Điều này cho thấy nhóm này có mức đường huyết cao hơn so với nhóm không mắc bệnh.

**A graph of lines and curves

Description automatically generated with medium confidence**

**Biểu đồ 2: Phân bố Blood Glucose Level theo tình trạng mắc bệnh tiểu đường**

* Biểu đồ này thể hiện sự phân bố của mức đường huyết (blood glucose level) theo tình trạng mắc bệnh tiểu đường.
* Đối với những người không mắc bệnh tiểu đường (diabetes = 0):
* Mức đường huyết chủ yếu nằm dưới 150 mg/dL, với một số giá trị cao hơn nhưng không nhiều.
* Điều này cho thấy nhóm này có mức đường huyết ổn định, nằm trong giới hạn bình thường.
* Đối với những người mắc bệnh tiểu đường (diabetes = 1):
* Mức đường huyết phân bố rộng hơn, với một số giá trị rất cao lên đến khoảng 250-300 mg/dL.
* Điều này cho thấy nhóm này có mức đường huyết thường xuyên cao hơn so với nhóm không mắc bệnh.

**A graph of blood glucose level

Description automatically generated**

**A blue and orange bar graph

Description automatically generated**

*Hình 5: Biểu đồ cột chồng so sánh số lượng người mắc và không mắc bệnh tiểu đường theo tiền sử hút thuốc*

- Biểu đồ cột chồng trên so sánh số lượng người mắc và không mắc bệnh tiểu đường theo tiền sử hút thuốc, với trục hoành thể hiện (0 là không có tiền sử hút thuốc, 1 là có tiền sử hút thuốc) và trục tung thể hiện tổng số lượng người (xanh là không tiểu đường, cam là tiểu đường). Qua đó ra rút ra được:

* Nhóm không có tiền sử hút thuốc (0): phần lớn không mắc bệnh tiểu đường. Điều này thể hiện qua việc cột màu xanh có số lượng người không mắc tiểu đường rất cao. Có một phần nhỏ mắc bệnh tiểu đường và số lượng người mắc vẫn khá đáng kể. Điều này nhấn mạnh rằng bệnh tiểu đường không chỉ liên quan đến hút thuốc mà còn chịu ảnh hưởng bởi nhiều yếu tố khác
* Nhóm có tiền sử hút thuốc (1): số lượng người không mắc bệnh tiểu đường chiếm đa số , tuy nhiên người mắc bệnh tiểu đường cũng chiếm một phần đáng kể. Điều này cho thấy rằng tỷ lệ mắc bệnh tiểu đường của nhóm hút thuốc cao hơn đáng kể so với nhóm không hút thuốc.
* Tóm lại, ta có thể thấy mối liên quan giữa tiền sử hút thuốc và nguy cơ mắc bệnh tiểu đường. Người có tiền sử hút thuốc có nguy cơ mắc bệnh tiểu đường cao hơn so với người không hút thuốc. Biểu đồ cho thấy vẫn có một tỷ lệ đáng kể người hút thuốc không mắc bệnh. Điều này cho thấy hút thuốc không phải là yếu tố duy nhất quyết định việc mắc bệnh tiểu đường.

**A colorful circle with text

Description automatically generated**

*Hình 6: Biểu đồ tròn thể hiện sự phân bố bệnh nhân mắc bệnh tiểu đường theo BMI*

- BMI được chia thành các nhóm: Thiếu cân, Bình thường, Thừa cân, và Béo phì. Biểu đồ trên thể hiện sự phân bố bệnh nhân mắc bệnh tiểu đường theo chỉ số khối cơ thể BMI. Biểu đồ thể hiện các tỷ lệ của các nhóm như sau:

* Béo phì: Chiếm tỷ lệ cao nhất với 50,5%. Điều này cho thấy béo phì là một yếu tố chính làm tăng nguy cơ mắc bệnh tiểu đường.
* Thừa cân: Chiếm 38,8%, là một tỷ lệ đáng kể. Người thừa cân có nguy cơ mắc bệnh tiểu đường cao hơn so với người có cân nặng bình thường.
* Bình thường: Chiếm 10%, cho thấy những người có cân nặng bình thường cũng có thể mắc bệnh tiểu đường nhưng tỷ lệ thấp hơn so với người thừa cân và béo phì.
* Thiếu cân: Chiếm tỷ lệ thấp nhất, chỉ 0,8%. Dữ liệu chỉ ra rằng thiếu cân ít liên quan đến bệnh tiểu đường hơn so với các nhóm BMI khác.

Biểu đồ cho thấy mối liên hệ chặt chẽ giữa cân nặng và nguy cơ mắc bệnh tiểu đường. Người béo phì và thừa cân có nguy cơ mắc bệnh cao hơn đáng kể so với người có cân nặng bình thường và thiếu cân. Những người có BMI bình thường hoặc thiếu cân ít có khả năng mắc bệnh tiểu đường hơn, nhưng vẫn có một tỷ lệ nhỏ mắc bệnh. Thông qua biểu đồ, ta rút ra được tầm quan trọng của việc duy trì cân nặng hợp lý để phòng ngừa bệnh tiểu đường.

A graph with blue and orange bars

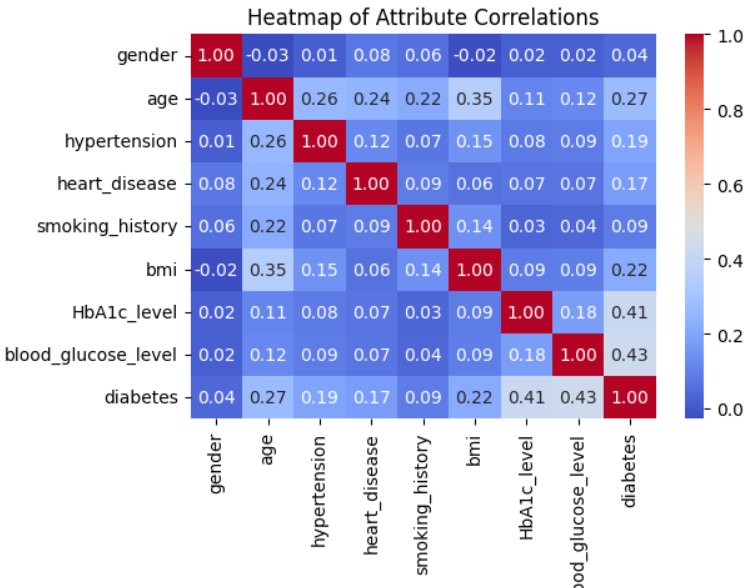
Description automatically generated

* Biểu đồ về tình trạng tăng huyết áp giúp trực quan hóa dữ liệu, nhận diện mẫu và xu hướng, và so sánh tỷ lệ mắc bệnh tiểu đường giữa các nhóm. Biểu đồ hỗ trợ phân tích và dự đoán, cho thấy rõ tăng huyết áp có thể là một yếu tố nguy cơ quan trọng đối với bệnh tiểu đường. Ngoài ra, biểu đồ còn giúp trình bày kết quả nghiên cứu một cách dễ hiểu và hỗ trợ việc ra quyết định dựa trên dữ liệu cụ thể.
* Hypertension: Nhóm tăng huyết áp (1) cho thấy tỷ lệ mắc bệnh tim cao hơn so với nhóm không tăng huyết áp (0). Điều này cho thấy tăng huyết áp có thể là một yếu tố quan trọng trong việc dự đoán bệnh tiểu đường.

A graph with blue and orange bars

Description automatically generated

* Mục đích của biểu đồ là trực quan hóa mối liên hệ giữa bệnh tiểu đường và bệnh tim, cho phép so sánh nhanh chóng số lượng người có và không có bệnh tim mắc bệnh tiểu đường. Việc hiểu rõ mối liên hệ này hỗ trợ trong việc lập kế hoạch điều trị và quản lý bệnh hiệu quả hơn. Biểu đồ cung cấp cái nhìn quan trọng về mối liên hệ giữa hai tình trạng, giúp hỗ trợ nghiên cứu và quản lý sức khỏe.
* Số lượng người mắc bệnh tiểu đường cao hơn đáng kể ở nhóm người mắc bệnh tim so với nhóm người không mắc bệnh tim. Bệnh tim là một yếu tố nguy cơ chính cho bệnh tiểu đường, có nghĩa là những người mắc bệnh tim có nguy cơ cao hơn mắc bệnh tiểu đường. Ngược lại, những người mắc bệnh tiểu đường cũng có nguy cơ cao hơn phát triển bệnh tim.
* Mức độ ảnh hưởng cao với biến mục tiêu: hypertension, heart disease



Dựa trên biểu đồ nhiệt của mối tương quan giữa các thuộc tính, từ đó có thể nhìn thấy được:

**Tương quan mạnh nhất:**

* blood glucose level (mức đường huyết) có tương quan cao nhất với diabetes với hệ số tương quan là 0.43, điều này cho thấy mức đường huyết có mối quan hệ đáng kể với bệnh tiểu đường.
* HbA1c level (mức HbA1c) cũng có tương quan tương đối cao với diabetes, với hệ số là 0.41. Điều này hợp lý, vì mức HbA1c thường được dùng để đánh giá tình trạng tiểu đường.

**Tương quan giữa các yếu tố sức khỏe khác:**

* age (tuổi) có tương quan dương với nhiều yếu tố khác như hypertension (tăng huyết áp) (0.26), heart\_disease (bệnh tim) (0.24), và bmi (0.35). Điều này gợi ý rằng tuổi cao có mối liên quan đến việc mắc các bệnh mãn tính và tình trạng sức khỏe khác như huyết áp cao, bệnh tim và chỉ số BMI cao.
* BMI có mối tương quan trung bình với age (0.35) và tương quan thấp với diabetes (0.22).

**Yếu tố ít tương quan:**

* Các yếu tố như gender (giới tính) và smoking\_history (tiền sử hút thuốc) có tương quan khá thấp với hầu hết các thuộc tính khác. Ví dụ, gender hầu như không có tương quan đáng kể với diabetes (-0.04).

Tóm lại, biểu đồ này cho thấy những thuộc tính có mối liên hệ rõ ràng nhất với bệnh tiểu đường là mức đường huyết và mức HbA1c, trong khi các yếu tố khác như chỉ số BMI và tuổi cũng đóng vai trò nhưng không quá mạnh mẽ. Những kết quả này có thể gợi ý rằng trong việc phân tích hoặc xây dựng mô hình dự đoán bệnh tiểu đường, các yếu tố như mức đường huyết và HbA1c nên được chú ý đặc biệt.

4. Projected Impact

**4.1. Thành tựu và những lợi ích (Accomplishments and Benefits)**

Dự án đã đạt được một số thành tựu đáng kể trong lĩnh vực phân tích dữ liệu y tế, đặc biệt là trong việc dự đoán nguy cơ mắc bệnh tiểu đường. Các thành tựu chính bao gồm:

* **Phát triển mô hình XGBoost có độ chính xác cao**: Mô hình XGBoost đã được xây dựng và tối ưu hóa, đạt độ chính xác lên đến 97.54% trên tập dữ liệu kiểm tra. Đây là một kết quả rất tích cực, thể hiện khả năng của mô hình trong việc hỗ trợ dự đoán chính xác tỷ lệ mắc bệnh tiểu đường, từ đó giúp cải thiện hiệu quả trong việc phát hiện sớm và quản lý bệnh.
* **Xử lý dữ liệu hiệu quả:** Dự án đã thành công trong việc làm sạch và xử lý các tập dữ liệu y tế, bao gồm việc loại bỏ các giá trị thiếu và điều chỉnh dữ liệu để phù hợp với mô hình hóa. Điều này đã giúp đảm bảo chất lượng dữ liệu cao, một yếu tố quan trọng trong việc đạt được kết quả dự đoán chính xác.
* **Cải thiện quy trình dự đoán bệnh tiểu đường:** Dự án đã góp phần cải thiện quy trình dự đoán bệnh tiểu đường bằng cách áp dụng các kỹ thuật học máy tiên tiến, từ đó giúp giảm thời gian chẩn đoán và tăng cường độ chính xác trong việc xác định các trường hợp có nguy cơ cao.

- Lợi ích từ những thành tựu này bao gồm:

* **Hỗ trợ quyết định trong y tế:** Kết quả từ mô hình dự đoán có thể hỗ trợ các bác sĩ và chuyên gia y tế trong việc ra quyết định, đặc biệt là trong việc xác định và quản lý các bệnh nhân có nguy cơ cao mắc bệnh tiểu đường. Điều này góp phần cải thiện chất lượng chăm sóc sức khỏe và tăng cường hiệu quả điều trị.
* **Tối ưu hóa quy trình chăm sóc sức khỏe:** Nhờ mô hình dự đoán có độ chính xác cao, dự án giúp giảm thiểu sai sót trong chẩn đoán, đồng thời tối ưu hóa việc sử dụng nguồn lực y tế, giảm chi phí liên quan đến các biến chứng của bệnh tiểu đường.
* **Đóng góp cho nghiên cứu khoa học:** Dự án mang lại những đóng góp quan trọng cho lĩnh vực học máy ứng dụng trong y tế, mở ra cơ hội phát triển các phương pháp mới trong việc dự đoán và phòng ngừa bệnh tiểu đường trong tương lai.

**4.2. Những cải tiến trong tương lai (Future Improvements)**

 - Mặc dù dự án đã đạt được nhiều thành tựu, vẫn còn những lĩnh vực có thể cải thiện để nâng cao hiệu quả và độ chính xác của mô hình:

* **Mở rộng dữ liệu huấn luyện:** Thu thập thêm dữ liệu từ các nguồn khác nhau sẽ giúp mô hình trở nên đa dạng và chính xác hơn. Điều này đặc biệt quan trọng trong việc đảm bảo rằng mô hình hoạt động tốt trên các tập dữ liệu khác nhau và trong các tình huống thực tế khác nhau.
* **Nâng cấp quy trình xử lý dữ liệu:** Áp dụng các kỹ thuật xử lý dữ liệu nâng cao, như sử dụng các phương pháp học sâu (deep learning) hoặc kỹ thuật tăng cường dữ liệu (data augmentation), có thể cải thiện chất lượng đầu vào cho mô hình và tăng cường khả năng dự đoán.
* **Phân tích thêm các yếu tố ảnh hưởng:** Nghiên cứu thêm về các yếu tố nhân khẩu học, lối sống, và môi trường có thể giúp xây dựng các mô hình toàn diện hơn, từ đó cung cấp các dự đoán chi tiết và chính xác hơn về nguy cơ mắc bệnh tiểu đường.

- Những cải tiến này sẽ không chỉ giúp nâng cao hiệu suất của mô hình mà còn đóng góp vào việc phát triển các phương pháp tiếp cận mới trong việc phân tích và dự đoán các bệnh lý khác.

5. Team Member Review and Comment

|  |
| --- |
| <ATTACH A TEAM PICTURE HERE> |

|  |  |
| --- | --- |
| NAME | REVIEW and COMMENT |
| Nguyễn Thị Ngọc Mai | Trong suốt quá trình học chương trình SIC của samsung cùng với sự hướng dẫn tận tình từ các thầy cô tôi đã tích lũy được nhiều kiến thức, kỹ năng, kinh nghiệm và nâng cao hiểu biết về lĩnh vực Big Data. Để rồi từ đó tôi có thể áp dụng những thức mình học được vào trong dự án cuối khóa này. Tuy về mặt nội dung vẫn còn vài thiếu sót vì nguyên nhân chưa đủ kinh nghiệm và kiến thức nhưng đó là thành quả mà tôi và nhóm của mình đã dành khá nhiều tâm huyết của mình cho nó. Quan trọng hơn, khoảng thời gian làm việc nhóm đã là một trải nghiệm đáng nhớ. Chúng tôi đã học hỏi rất nhiều từ nhau, cùng nhau nỗ lực hết mình, làm việc với tinh thần ăn ý và trách nhiệm cao. Mỗi thành viên đều đóng góp hết mình vào dự án này. Đây chính là một trong những bài học thực tiễn quý báu giúp chúng tôi chuẩn bị tốt hơn những kiến thức, kỹ năng và kinh nghiệm để nâng cao chất lượng cho những sản phẩm, dự án sau này. |
| Lâm Thị Diễm My | Thông qua khóa học SIC, tôi đã được biết thêm nhiều kiến thức về big data và những kỹ năng mềm khác. Trong quá trình làm việc nhóm, chúng tôi làm việc nhóm khá ăn ý với nhau. Tuy bài báo cáo của chúng tôi có thể không được hoàn hảo nhưng mọi người đã cùng nhau nỗ lực xây dựng và hoàn chỉnh bài báo cáo một cách tốt nhất. Tôi thật sự rất trân quý điều đó. Cảm ơn chương trình SIC đã mang đến cho tôi nhiều trải nghiệm quý báu. |
| Trần Thanh Trúc | Qua thời gian được học khóa học SIC của SamSung, tôi đã học được rất nhiều kiến thức về lĩnh vực Big Data này. Khóa học đã giúp tôi trau dồi được rất nhiều kỹ năng về tiền xử lý, trực quan hóa dữ liệu, kiến thức về sử dụng mô hình được mở rộng hơn. Khi hoàn thành bài tập lớn của khóa học, tôi đã rút ra nhiều kinh nghiệm làm việc nhóm và kỹ năng làm bài theo từng dự án cụ thể, đặc biệt là về một lĩnh vực dữ liệu y tế. Qua khóa học trên, đã giúp tôi tạo cho mình một nền tảng vững chắc cho kế hoạch thực tập sắp tới cũng như là về mục tiêu công việc tương lai của mình. |
| Nguon Gia Linh | Lúc bắt đầu quá trình học, tôi nhận thấy mình còn nhiều thiếu sót cả về kiến thức lẫn kinh nghiệm về lĩnh vực Big Data. Sau 1 khoảng thời gian tiếp xúc và nghiên cứu, cùng với sự hỗ trợ của các giáo viên và thành viên trong nhóm, tôi giờ đây cũng đã tích lũy được nhiều kiến thức hơn và hoàn thành bài báo cáo. Dù có thể bài chưa được hoàn hảo nhưng mọi người cũng đã nỗ lực rất nhiều. Cuối cùng, tôi muốn cảm ơn đến tất cả mọi người đã cùng đồng hành và giúp đỡ tôi trong thời gian qua. Khóa học đã mang đến cho tôi trải nghiệm quý báu mà nhờ đó tôi sẽ vững bước trong hành trang sau này |
| Lê Khánh Linh | Bài báo cáo của nhóm chúng tôi mặc dù vẫn còn một vài hạn chế về kiến thức, kinh nghiệm phát triển một hệ thống và tính ứng dụng thực tế tuy nhiên chúng tôi đã thể hiện những kỹ năng, kiến thức và kinh nghiệm trong suốt 3 năm học cũng như gần 3 tháng tiếp thu kiến thức từ những thầy cô khi học chương trình này. Bên cạnh đó trong quá trình làm việc nhóm tôi cũng được học hỏi từ những thành viên của mình rất nhiều, chúng tôi làm việc nhóm khá ăn ý với nhau, mọi người đều nỗ lực và làm việc rất có trách nhiệm. Đây sẽ là một trong những trải nghiệm thực tế đáng học hỏi mà tôi sẽ ghi nhớ cho sự nghiệp tương lai sau này |
| Đào Ngọc Quý | Tham gia khóa học này tôi nhận thức được rằng mình còn nhiều hạn chế và kiến thức. Đến bây giờ, khi hoàn thành sản phẩm, tôi đã tích lũy được cho mình một lượng kiến thức rất lớn. Khoảng thời gian, qua tôi được học hỏi rất nhiều từ thầy cô và các thành viên trong nhóm, mọi người đã hỗ trợ tôi rất nhiều trong suốt gần 3 tháng. Tuy trong quá trình làm có đôi chút chưa hiểu ý nhưng thời gian đã cho chúng tôi dần sửa lỗi, ăn ý nhau hơn và cuối cùng là hoàn thành sản phẩm. Dù chưa thể đóng góp quá nhiều vào sản phẩm này, tôi lại được học hỏi rất nhiều từ kiến thức, kĩ năng, kinh nghiệm của mình ở lĩnh vực Big Data. Những thực nghiệm này sẽ giúp tôi có bước đệm vững chắc để nâng cao chất lượng làm việc và hoàn thành tốt hơn các sản phẩm sau này. |

6. Instructor Review and Comment

|  |  |  |
| --- | --- | --- |
| CATEGORY | SCORE | REVIEW and COMMENT |
| IDEA | \_\_/10 |  |
| APPLICATION | \_\_/30 |  |
| RESULT | \_\_/30 |  |
| PROJECT MANAGEMENT | \_\_/10 |  |
| PRESENTATION & REPORT | \_\_/20 |  |
| TOTAL | \_\_/100 |  |

[[1]](#footnote-0)

1. [↑](#footnote-ref-0)