Chương 1: Giới thiệu

Hiện nay, bệnh tiểu đường là một căn bệnh mãn tính phổ biến nhất trên thế giới. Bệnh tiểu đường là một căn bệnh mãn tính nghiêm trọng khiến mọi người mất khả năng điều chỉnh hiệu quả lượng glucose trong máu và có thể dẫn đến giảm chất lượng cuộc sống và tuổi thọ. Sau khi các loại thực phẩm khác nhau được phân hủy thành đường trong quá trình tiêu hóa, đường sẽ được giải phóng vào máu. Điều này báo hiệu tuyến tụy tiết ra insulin. Insulin giúp các tế bào trong cơ thể sử dụng lượng đường trong máu để tạo năng lượng. Bệnh tiểu đường thường được đặc trưng bởi cơ thể không sản xuất đủ insulin hoặc không thể sử dụng insulin được sản xuất hiệu quả khi cần thiết.

Các biến chứng như bệnh tim, mất thị lực, cắt cụt chi dưới và bệnh thận có liên quan đến lượng đường cao mãn tính vẫn còn trong máu đối với những người mắc bệnh tiểu đường. Mặc dù không có cách chữa khỏi bệnh tiểu đường, nhưng các chiến lược như giảm cân, ăn uống lành mạnh, vận động và điều trị y tế có thể làm giảm tác hại của căn bệnh này ở nhiều bệnh nhân. Chẩn đoán sớm có thể dẫn đến thay đổi lối sống và điều trị hiệu quả hơn, khiến các mô hình dự đoán nguy cơ mắc bệnh tiểu đường trở thành công cụ quan trọng đối với cộng đồng và các quan chức y tế công cộng.

Ở Hoa Kỳ, bệnh tiểu đường ảnh hưởng đến hàng triệu người Mỹ mỗi năm và gây gánh nặng tài chính đáng kể cho nền kinh tế. Quy mô của vấn đề này cũng rất quan trọng để nhận ra. Trung tâm Kiểm soát và Phòng ngừa Dịch bệnh đã chỉ ra rằng tính đến năm 2018, 34,2 triệu người Mỹ mắc bệnh tiểu đường và 88 triệu người mắc bệnh tiền tiểu đường. Hơn nữa, CDC ước tính rằng cứ 5 bệnh nhân tiểu đường thì có 1 người và khoảng 8 trong số 10 người tiền tiểu đường không biết về nguy cơ của mình.

Mặc dù có nhiều loại bệnh tiểu đường khác nhau, nhưng bệnh tiểu đường loại II là dạng phổ biến nhất và tỷ lệ mắc bệnh thay đổi theo độ tuổi, trình độ học vấn, thu nhập, địa điểm, chủng tộc và các yếu tố xã hội khác quyết định sức khỏe. Phần lớn gánh nặng bệnh tật cũng rơi vào những người có tình trạng kinh tế xã hội thấp hơn. Bệnh tiểu đường cũng đặt gánh nặng lớn lên nền kinh tế, với chi phí được chẩn đoán là bệnh tiểu đường khoảng 327 tỷ đô la và tổng chi phí cho bệnh tiểu đường và tiền tiểu đường không được chẩn đoán lên tới 400 tỷ đô la mỗi năm.

Dựa trên dữ liệu được thu thập năm 2015 của CDC khảo sát qua điện thoại về vấn đề liên quan đến sức khỏe (BRFSS), bài báo cáo sẽ tập trung phân tích và xử lý dữ liệu để trả lời cho các câu hỏi:

1. Dựa vào các yếu tố được khảo sát từ BRFSS có thể đưa ra dự đoán 1 cá nhân có thể bị tiểu đường hoặc tiền tiểu đường hay không ?

2. Những yếu tố nào quan trọng trong việc dự đoán, và có thể lựa chọn 1 tập hợp con các yếu tố đó để tạo 1 câu hỏi ngắn để dự đoán tình trạng tiểu đường của 1 người hay không ?

Để đạt được kết quả phân tích trên, nhóm đã đề xuất 1 quy trình phân tích và xử lý dữ liệu bao gồm:

1. Khám phá dữ liệu (EDA)
2. Làm sạch dữ liệu
3. Kiểm định các giả thuyết
4. Xây dựng, đánh giá và lựa chọn mô hình
5. Tổng hợp và kết luận

Chương 2: Phân tích

2.1. Khám phá dữ liệu (EDA)

Khám phá dữ liệu là bước quan trọng nhất trong phân tích và xử lý dữ liệu. Ở bước này, ta sẽ tìm hiểu, phát hiện các vấn đề của dữ liệu và các mối quan hệ tiềm ẩn giữa các biến với nhau từ đó tìm ra phương án xử lý hiệu quả để đạt được mục tiêu mong muốn.

2.1.1.Mục tiêu

- Kiểm tra cấu trúc dữ liệu.
- Phát hiện giá trị khuyết và dòng trùng lặp.
- Phát hiện vấn đề có trong dữ liệu.
- Phát hiện các mối quan hệ tiềm ẩn giữa các biến trong bộ dữ liệu.
- Xác định các yếu tố có thể ảnh hưởng lớn đến mô hình dự đoán.

2.1.2.Kết quả thu được:

***Cấu trúc dữ liệu:**

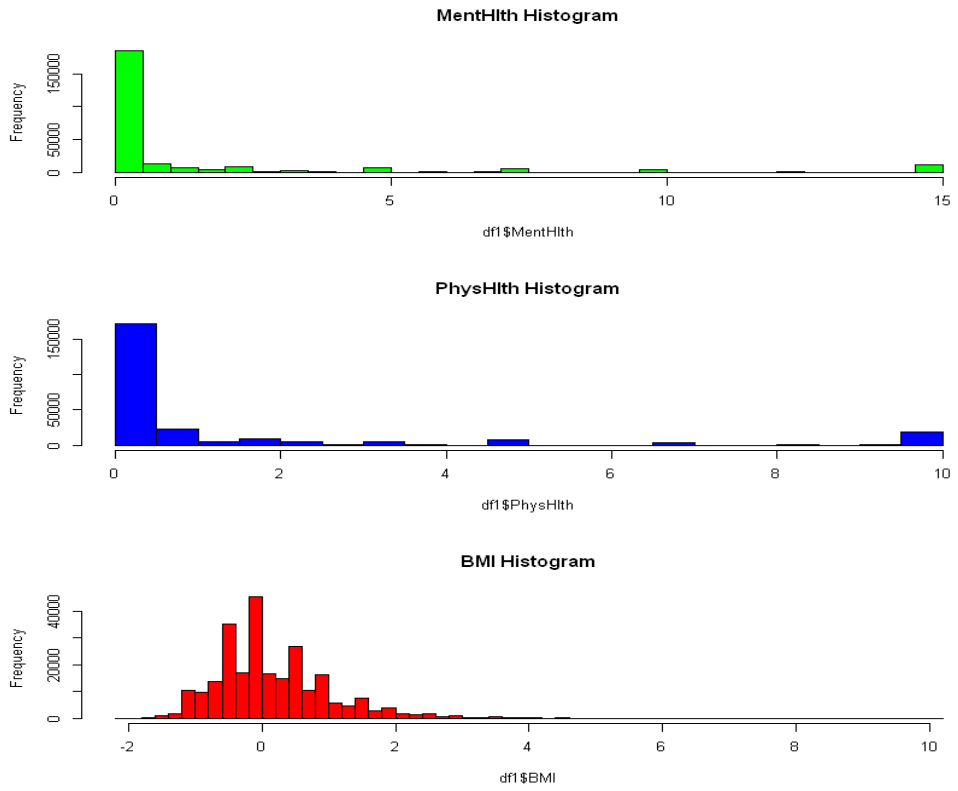
Bộ dữ liệu cung cấp gồm **253.680 dòng** và **22 cột**, trong đó có **21 biến độc lập** và **một biến phụ thuộc** là *Diabetes_012*. Biến *Diabetes_012* nhận giá trị là 0 (người không bị tiểu đường), 1 (người bị tiền tiểu đường) và 2 (người bị tiểu đường).Qua đó xác định đây là một bài toán phân loại đa danh mục (**multinomial classification**).

Phần lớn các biến trong dữ liệu là biến nhị phân. Các biến như *GenHlth*, *Age*, *Education*, và *Income* không phải biến nhị phân nhưng thuộc loại biến định tính. Trong khi đó, các biến *BMI*, *MentHlth*, và *PhysHlth* là những biến định lượng rời rạc.

***Dữ liệu khuyết và trùng lặp:**

Bộ dữ liệu đã được làm sạch và không có chứa các dữ liệu khuyết. Có khoảng 24000 dòng bị trùng nhau, chiếm khoảng 9.5% dữ liệu gốc. Những dòng trùng nhau này chủ yếu rơi vào nhóm không bị tiểu đường.

***Phân phối dữ liệu:**



***Hình 1**: Histogram của các biến BMI, MentHlth, và PhysHlth*

Biểu đồ phân phối cho các biến định lượng cho thấy phân phối của cả 3 biến đều đang bị lệch phải. Phân phối của *MentHlth* và *PhysHlth* cũng khá tương đồng, có dấu hiệu đa cộng tuyến. Cụ thể hơn như sau:

Đặc điểm của từng biến:

- **MentHlth và PhysHlth:** Hai biến này đều có phần lớn giá trị tập trung ở mức thấp, cụ thể là gần 0, cho thấy phần lớn người tham gia khảo sát báo cáo rất ít vấn đề về sức khỏe tâm lý và thể chất. Tuy nhiên, một số giá trị lớn hơn xuất hiện với tần suất giảm dần, dẫn đến phân phối lệch phải. Phân phối tương đồng của hai biến này gợi ý rằng các yếu tố sức khỏe tâm lý và thể chất có thể có mối liên hệ hoặc ảnh hưởng lẫn nhau.

- **BMI:** Phân phối của biến BMI cũng lệch phải nhưng không mạnh bằng hai biến trên. Phần lớn dữ liệu BMI nằm trong phạm vi gần giá trị trung tâm, cho thấy một phân phối gần chuẩn hơn so với MentHlth và PhysHlth.

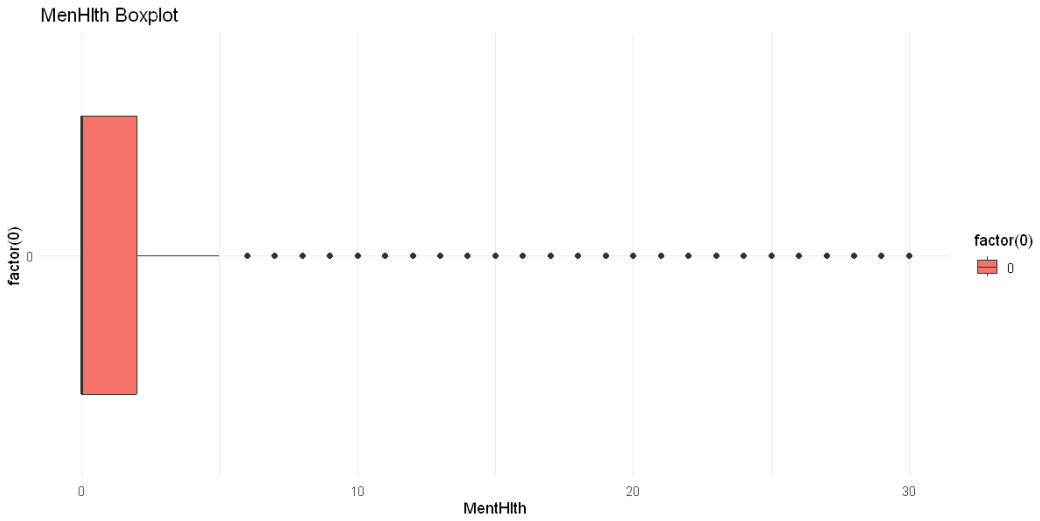
Mối quan hệ giữa các biến:

- Sự tương đồng trong phân phối của **MentHlth** và **PhysHlth** có thể là dấu hiệu của mối tương quan giữa hai biến này, tức là những người có vấn đề về sức khỏe tâm lý thường cũng có các vấn đề về sức khỏe thể chất (hoặc ngược lại).

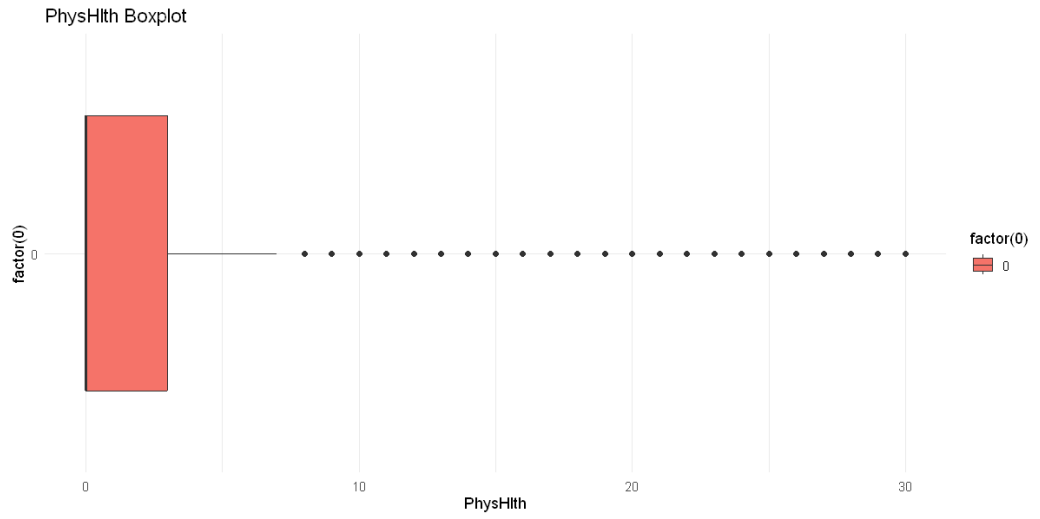
Điều này cần được kiểm tra thêm bằng phân tích hệ số tương quan hoặc các kỹ thuật thống kê khác.

- **BMI** có phân phối khác biệt hơn, nhưng sự lệch nhẹ về bên phải gợi ý rằng có một số ít cá nhân có BMI cao bất thường, có thể ảnh hưởng đến kết quả nếu không được xử lý đúng.

Tiếp theo, các Boxplot cho các biến định lượng *BMI*, *MentHlth*, và *PhysHlth* cho thấy có xuất hiện outliers.



***Hình 2**: Boxplot cho biến MentHlth*



***Hình 3**: Boxplot cho biến PhysHlth*

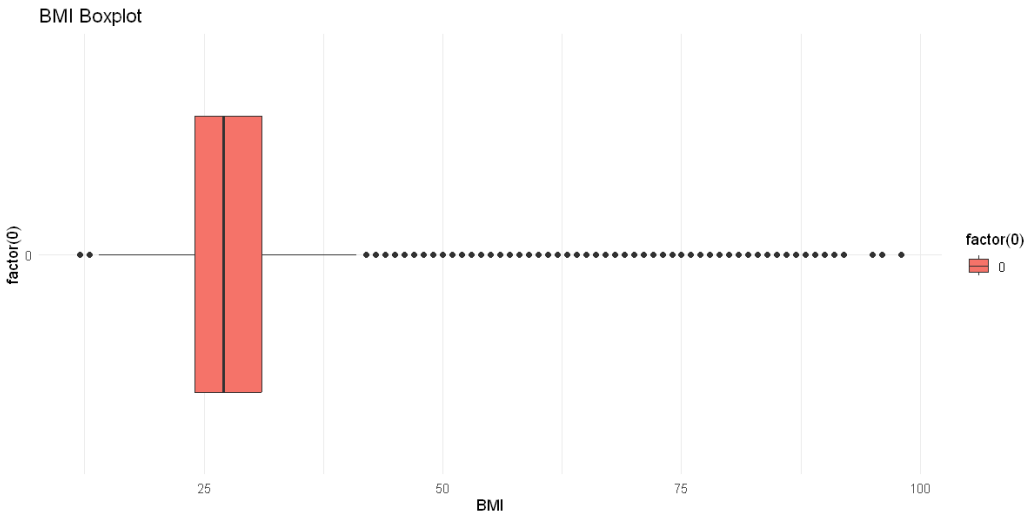
Do phân phối khả tương đồng của *MentHlth* và *PhysHlth*, nên ta có nhận xét chung cho 2 biến như sau:

Phân phối chính:

- Phần lớn dữ liệu của 2 biến tập trung ở mức thấp (gần 0 ngày), thể hiện qua hộp nằm ở phía dưới đồ thị.
- Trung vị gần sát với giá trị 0, cho thấy phần lớn mẫu khảo sát có rất ít hoặc không có vấn đề về sức khỏe tâm lý.

Ngoại lệ (outliers):

- Có nhiều điểm dữ liệu nằm ngoài phạm vi của râu, phân bố rải rác từ giá trị 10 đến hơn 20 ngày. Đây là các cá nhân báo cáo có vấn đề sức khỏe tâm lý nhiều ngày trong tháng.
- Số lượng ngoại lệ khá lớn, điều này cho thấy phân phối của 2 biến thực sự bị lệch phải mạnh.



***Hình 4**: Boxplot cho biến BMI*

Về BMI, có thể thấy rằng phân phối của biến là không đồng đều, với các giá trị ngoại lệ cho thấy có một vài cá nhân có BMI rất cao hoặc rất thấp.

Phân phối chính:

- Phần lớn dữ liệu BMI nằm trong khoảng trung bình (gần giá trị trung vị), tuy nhiên, có sự phân tán ở cả hai phía với một số giá trị bất thường (outliers) nằm xa giá trị trung tâm.
- Dữ liệu BMI tập trung chủ yếu trong khoảng từ 24 đến 31.

Ngoại lệ (Outliers):

Có nhiều điểm ngoại lệ nằm ngoài phạm vi râu, tập trung ở hai phía:

- **BMI thấp:** Một vài điểm dưới mức 20, điều này có thể đại diện cho những cá nhân có BMI dưới mức tiêu chuẩn, tức là gầy yếu.

- **BMI cao:** Rất nhiều điểm trên mức 35, thậm chí vượt qua mức 75 hoặc hơn. Điều này cho thấy có một số cá nhân trong mẫu có BMI rất cao, có khả năng thuộc nhóm "béo phì" hoặc "béo phì nghiêm trọng".
- Số lượng điểm ngoại lệ phía trên (BMI cao) nhiều hơn phía dưới (BMI thấp), cho thấy dữ liệu có phân phối lệch phải. Tức là, trong tập mẫu, có nhiều cá nhân có BMI cao bất thường so với cá nhân có BMI thấp bất thường.

Qua tính toán, ta có được tỉ lệ outliers của BMI là 3,9%. MentHlth là 14,3% và PhysHlth là 16,1%.

Về các **biến định tính,** nhóm đưa ra một số nhận nhận xét sơ lược như sau:

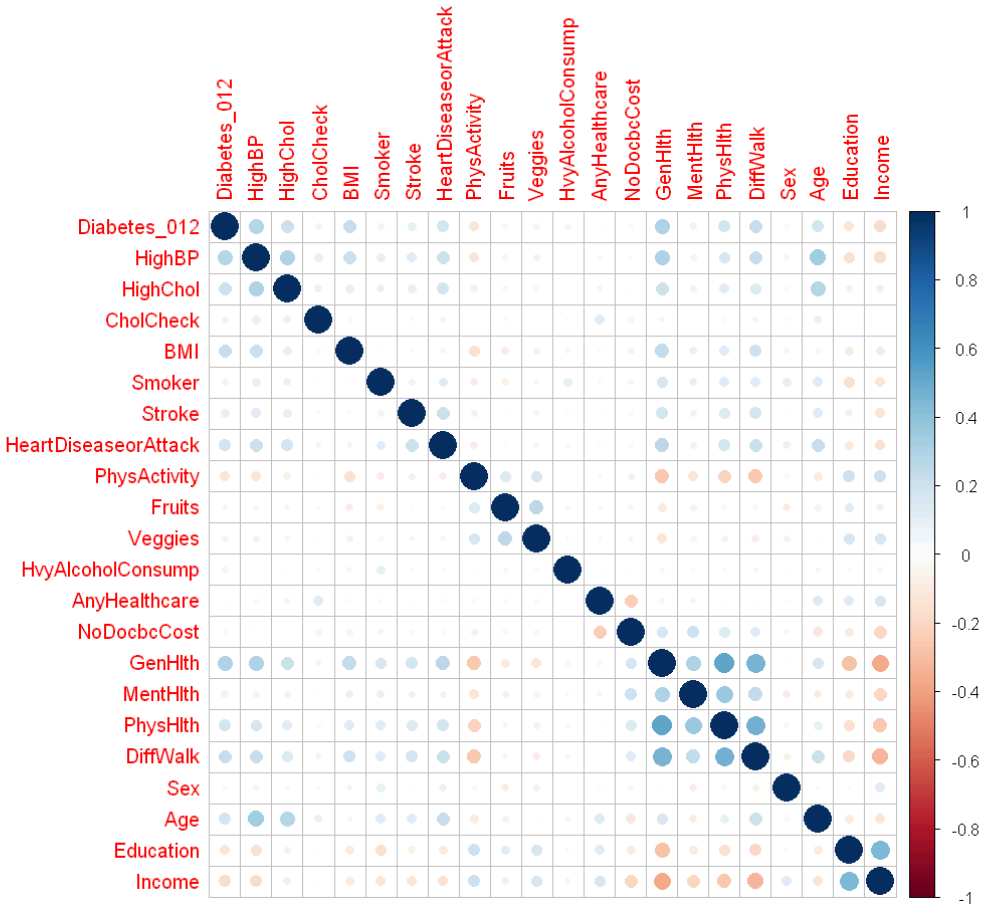
***Hình 5**: Tổng hợp các biểu đồ cột cho từng biến định tính*

- **Các yếu tố nguy cơ** như huyết áp cao, cholesterol cao, và hút thuốc có tỷ lệ khá đáng kể trong mẫu nghiên cứu.
- **Thói quen sinh hoạt lành mạnh**, như ăn trái cây, rau, và tham gia hoạt động thể chất, chiếm tỷ lệ cao.
- **Tiếp cận y tế** nhìn chung tốt, với phần lớn người tham gia sử dụng dịch vụ y tế và ít gặp khó khăn về chi phí.

- **Sức khỏe tổng quát** được đánh giá ở mức trung bình, nhưng có một số ít người gặp khó khăn nghiêm trọng (như đột quỵ, bệnh tim, khó khăn khi đi bộ).

Đối với biến dự đoán *Diabetes_012,* tỷ lệ số lượng quan trắc trong 3 nhóm lần lượt là: 0.84 - 0.018 - 0.14 (tương ứng với nhóm không bị tiểu đường - tiền tiểu đường - tiểu đường) số liệu này chứng tỏ dữ liệu đang bị **mất cân bằng nặng**.

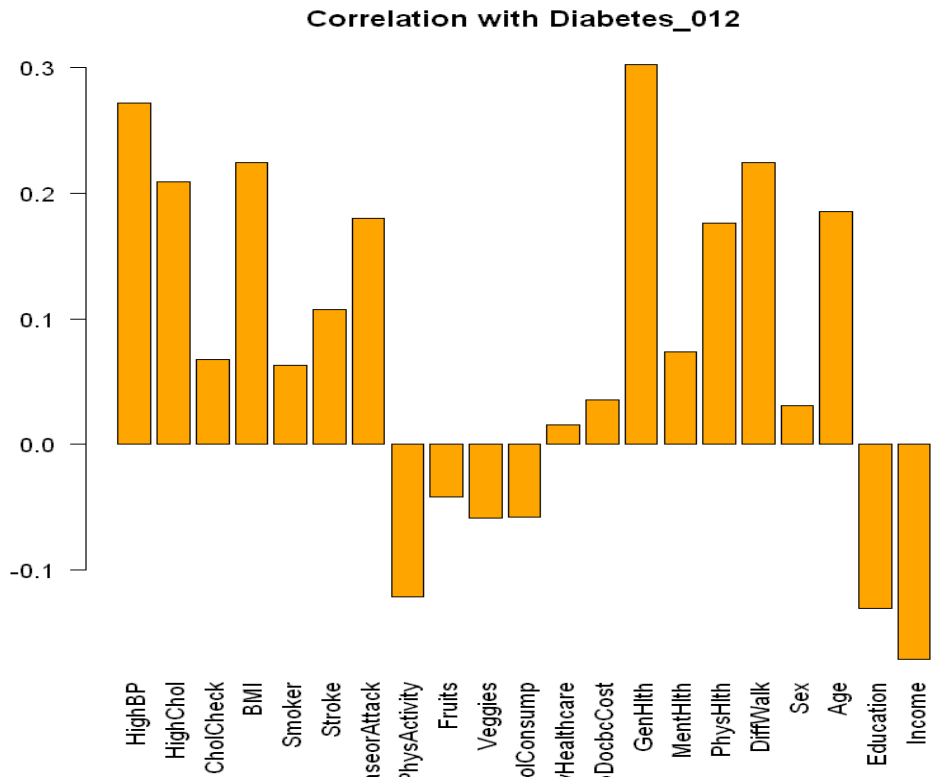
***Tương quan:**



***Hình 6**: Biểu đồ ma trận tương quan (Correlation Matrix) giữa các biến*

Một số biến trong dataset có tương quan mạnh với nhau:

- (GenHlth, PhysHlth), (PhysHlth, DiffWalk), (Education, Income), (GenHlth, DiffWalk), (MenthHlth, PhysHlth): tương quan thuận
- (GenHlth, Income), (DiffWalk, Income): tương quan nghịch



***Hình** 7: Biểu đồ tương quan giữa biến dự đoán Diabetes_012 và các biến định tính*

Các biến có khả năng tương quan mạnh với biến dự đoán: HighBP, HighChol, BMI, CholCheck, Stroke, HeartDiseaseorAttack, PhysActivity, GenHlth, PhysHlth, DiffWalk, Age, Education, Income.

2.1.3. Nhận xét tổng quan về bộ dữ liệu

Đa số các quan trắc trong dataset này nằm trong độ tuổi trung niên, có thu nhập khá tốt và cũng có lối sống lành mạnh. Bên cạnh đó cũng quan sát thấy số lượng các quan trắc không bị bệnh tiểu đường cũng chiếm đa số.

Một số vấn đề có trong bộ dữ liệu:

- Dữ liệu trùng lặp
- Outliers
- Mất cân bằng dữ liệu
- Tương quan phức tạp giữa các biến
- Chứa nhiều biến định tính, có thể không phù hợp với các mô hình yêu cầu về phân phối dữ liệu.

2.1.4. Phương án đề xuất

- **Dữ liệu trùng lặp**: do tỷ lệ dữ liệu trùng < 10% và chủ yếu rơi vào class chiếm đa số nên có thể thực hiện xóa dữ liệu trùng.
- **Outliers**: đối với biến BMI thì biến này đóng vai trò quan trọng trong dự đoán tiểu đường, đối với biến MentHlth và PhysHlth thì tỷ lệ outliers chiếm trên 10% nên không thể thực hiện xóa dữ liệu. Nhóm đề xuất 1 phương án an toàn là scale lại dữ liệu của các biến này để giảm ảnh hưởng outliers.
- **Mất cân bằng dữ liệu**: Nhóm có ít dữ liệu nhất là nhóm tiền tiểu đường vẫn có hơn 3000 quan trắc, đủ để thử tất cả các phương pháp xử lý mất cân bằng dữ liệu (Under-sampling, Over-sampling, Under-Over sampling và SMOTE).
- **Tương quan phức tạp giữa các biến**: Cần thực hiện kiểm định để xác nhận lại mối tương quan giữa các biến, cũng như cần chú ý lựa chọn biến cho mô hình và kiểm tra đa cộng tuyến.

2.2. Làm sạch dữ liệu

Trước khi tiến hành phân tích và xây dựng mô hình, việc làm sạch dữ liệu là một bước quan trọng để đảm bảo chất lượng và độ tin cậy của kết quả. Trong phần này, nhóm thực hiện các thao tác nhằm xử lý các vấn đề phổ biến trong tập dữ liệu như giá trị thiếu (missing values), dữ liệu trùng lặp (duplicate rows), và dữ liệu ngoại lai (outliers).

Ngoài ra, các biến trong dữ liệu được chuẩn hóa và chuyển đổi nếu cần thiết để phù hợp với yêu cầu của mô hình. Các bước làm sạch không chỉ giúp tăng hiệu suất mô hình mà còn đảm bảo rằng các kết luận rút ra từ dữ liệu là chính xác và đáng tin cậy.

Để giải quyết các vấn đề đã phát hiện trong bộ dữ liệu bao gồm xử lý dữ liệu trùng lặp, outlier và mất cân bằng dữ liệu, Nhóm chia phần làm sạch dữ liệu làm 2 phần, làm sạch dữ liệu cho kiểm định giả thuyết và làm sạch dữ liệu cho xây dựng mô hình.

2.2.1. Làm sạch dữ liệu cho kiểm định giả thuyết

Do kiểm định giả thuyết nhạy cảm với các outliers nên cần xử lý trước khi bước vào kiểm định giả thuyết:

- Thực hiện gom nhóm những nhãn hiếm cho biến Education và đánh nhãn lại: 1 (Less than high school), 2 (High School), 3 (College), 4 (College graduate).

- Đối với outliers của biến định lượng, nhóm chọn sử dụng Robust Scaling để xử lý. Phương pháp này hoạt động bằng cách chuẩn hóa dữ liệu dựa trên median (trung vị) và IQR thay vì mean và standard deviation như trong Standard Scaling.

Lý do lựa chọn:

- **Bảo toàn thông tin:** Phương pháp này không loại bỏ ngoại lệ mà giảm tác động của chúng, đảm bảo dữ liệu vẫn đầy đủ và phù hợp cho các phân tích sâu hơn.
- **Phù hợp với dữ liệu lệch:** Các biến như **MentHlth** và **PhysHlth** có phân phối lệch phải và tỷ lệ outlier cao, khiến Robust Scaling trở thành một lựa chọn phù hợp để giảm sự ảnh hưởng của các giá trị bất thường mà không làm thay đổi tính chất tổng quát của dữ liệu.
- **Tăng hiệu quả mô hình:** Chuẩn hóa bằng Robust Scaling có thể giúp các thuật toán phân tích và mô hình học máy, đặc biệt là những thuật toán nhạy cảm với tầm giá trị của dữ liệu, hoạt động hiệu quả hơn.

Bên cạnh Robust Scaler còn có phương pháp phổ biến là Log Transform để giảm ảnh hưởng của outliers nhưng trong quá trình EDA phát hiện ra các biến MentHlth và PhysHlth có số lượng lớn các giá trị 0 nên phương pháp này không phù hợp.

2.2.2. Làm sạch dữ liệu cho xây dựng mô hình

Bước này cũng sẽ thực hiện gom nhóm những nhãn hiếm cho biến Education và sử dụng Robust Scaler để giảm ảnh hưởng của outlier. Các bước thực hiện quá trình này như sau:

- Thực hiện gom nhóm nhãn hiếm cho biến Education giống phần trước
- Loại bỏ tất cả duplicate rows để giảm bớt các thông tin gây nhiễu cho mô hình.
- Để tránh bị rò rỉ thông tin khi xử lý dữ liệu cho mô hình, dữ liệu sau khi xử lý được chia thành hai phần theo tỷ lệ 7:3 cho tập train và test.
- Tính toán IQR và median của các đặc trưng trên tập train, sau đó lưu lại các giá trị này để sử dụng cho các quan trắc mới trong tương lai hoặc tập test.
- Scale dữ liệu trên cả tập train và test bằng các thông số IQR và median đã được tính toán và lưu trữ.
- Cuối cùng, xử lý vấn đề mất cân bằng dữ liệu trên tập train bằng cách áp dụng các phương pháp đã được học, chẳng hạn như undersampling, oversampling, under-over sampling và SMOTE.

Quá trình này đảm bảo dữ liệu được chuẩn bị tốt nhất cho việc huấn luyện và đánh giá mô hình.

2.3. Kiểm định các giả thuyết

Kiểm định giả thuyết nhằm xác nhận các mối quan hệ hoặc khác biệt có ý nghĩa thống kê giữa các biến trong tập dữ liệu. Kết quả kiểm định sẽ cung cấp nền tảng để đưa ra các quyết định, củng cố các phát hiện quan trọng, và đảm bảo rằng các kết luận dựa trên dữ liệu không phải do ngẫu nhiên.

2.3.1. Mục tiêu

- Xác nhận mối quan hệ của các biến giải thích với biến dự đoán
- Xác nhận mối quan hệ tương quan giữa các biến giải thích

2.3.2. Thực hiện

Sử dụng bộ dữ liệu được làm sạch ở bước *2.2.1*
Đặt các giả thuyết:

1. Giả thuyết (H01): không có sự khác biệt về số lượng/ tỷ lệ quan trắc của biến X giữa các nhóm không tiểu đường, tiền tiểu đường và tiểu đường.
2. Giả thuyết (H02): không có sự khác biệt về số lượng/ tỷ lệ quan trắc giữa biến X và biến Y trong mỗi nhóm không tiểu đường, tiền tiểu đường và tiểu đường.

Các kiểm định sử dụng:

- Kiểm định **Chi-square kết hợp với Monte Carlo Simulation** giữa các biến định tính. Do dữ liệu bị mất cân bằng nên kết hợp kiểm định với Monte Carlo Simulation để đảm bảo tính chính xác cho kiểm định, phương pháp này cũng ít hao tốn tài nguyên hơn bootstrap.
- Kiểm định **ANOVA/ Kruskal-Wallis** giữa biến định lượng và định tính/ định lượng. Kiểm định ANOVA có yêu cầu về phân phối chuẩn cho dữ liệu nên nhóm em sử dụng thêm kiểm định phi tham số là kruskal-wallis để đảm bảo được tính chính xác cho thống kê.

2.3.3. Kết quả

Với mức ý nghĩa alpha = 0.05.

- Kết quả kiểm định cho giả thuyết H01 cho các p-value đều tiệm cận 0, vậy nên ta bác bỏ giả thuyết H01 cho tất cả các biến giải thích. Điều này có nghĩa là tất cả các biến giải thích đều có tương quan với biến dự đoán.

- Kết quả kiểm định cho giả thuyết H02: ta sẽ bác bỏ giả thuyết cho 1 cặp biến nếu có p-value <= 0.05 trong ít nhất 1 nhóm. Nghĩa là sự khác biệt giữa X và Y phụ thuộc vào nhóm tình trạng tiểu đường. Điều này gợi ý rằng mối quan hệ giữa X và Y có thể thay đổi tùy theo tình trạng tiểu đường của từng nhóm. Kết quả sau kiểm định cho thấy phần lớn là bác bỏ H02, cho thấy có nhiều sự tương quan giữa các biến giải thích.

2.3.4. Kết luận

Sau quá trình kiểm định quan hệ giữa các biến, ta có được các nhận xét:

- Tất cả các biến giải thích đều có đóng góp vào mô hình.

- Đa số các biến giải thích có mối quan hệ tương quan, điều này làm mô hình khả năng rất cao bị đa cộng tuyến. Tuy nhiên nếu loại bỏ hết đa cộng tuyến thì giảm hiệu suất mô hình do mâu thuẫn với nhận xét 1.

2.3.5. Phương án đề xuất

Sử dụng các phương pháp co hệ số cho mô hình Logistic Regression hoặc các mô hình học máy để xử lý vấn đề đa cộng tuyến.

2.4. Xây dựng, đánh giá và lựa chọn mô hình

Quá trình xây dựng mô hình bao gồm các bước: lựa chọn đặc trưng, tối ưu hóa siêu tham số (hyperparameter tuning) và đánh giá hiệu suất dựa trên các thước đo như độ chính xác (accuracy), độ chính xác dự đoán (precision), độ bao phủ (recall), F1-score và Kappa.

Sau khi phân tích và làm sạch dữ liệu, nhóm quyết định sử dụng hai thuật toán Logistic Regression và Random Forests để giải quyết bài toán.

Các thuật toán LDA và QDA không được lựa chọn vì yêu cầu các biến giải thích phải tuân theo phân phối chuẩn, trong khi dữ liệu của chúng ta bao gồm nhiều biến nhị phân và định tính. Thuật toán Naive Bayes cũng bị loại trừ do đòi hỏi tính độc lập giữa các biến, điều mà tập dữ liệu này không đảm bảo.

Vì tập dữ liệu bị mất cân bằng, nhóm đặc biệt chú trọng đến các chỉ số Recall, F1-score và Kappa, trong đó Recall cho nhóm 1 và nhóm 2 (tiền tiểu đường và tiểu đường) được ưu tiên hơn cả.

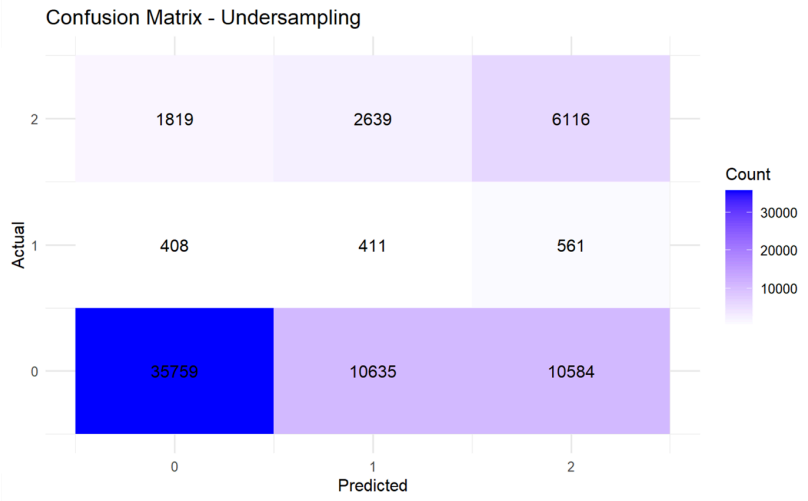
2.4.1. Mô hình Logistic Regression

Đây là bài toán phân loại đa danh mục nên sẽ sử dụng multinomial logistic của thư viện nnet trong R.

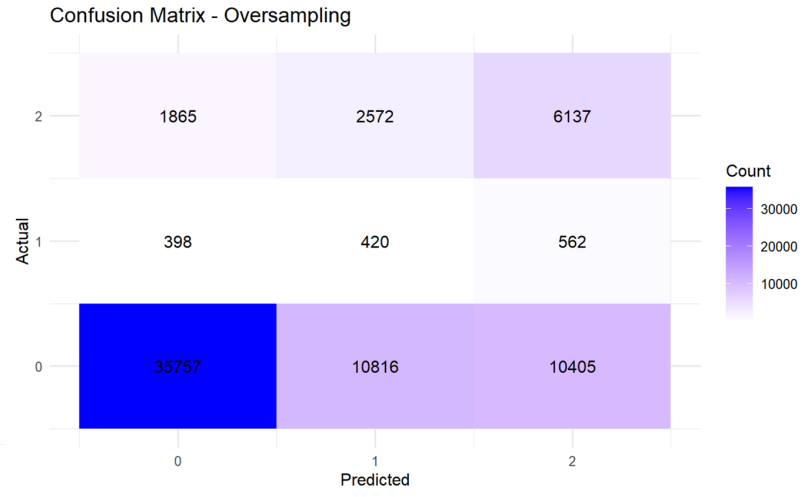
Huấn luyện mô hình với tất cả các tập dữ liệu train được xử lý mất cân bằng (ở bước *2.2.2*) và đánh giá trên tập test để lựa xem tập dữ liệu nào tốt nhất.

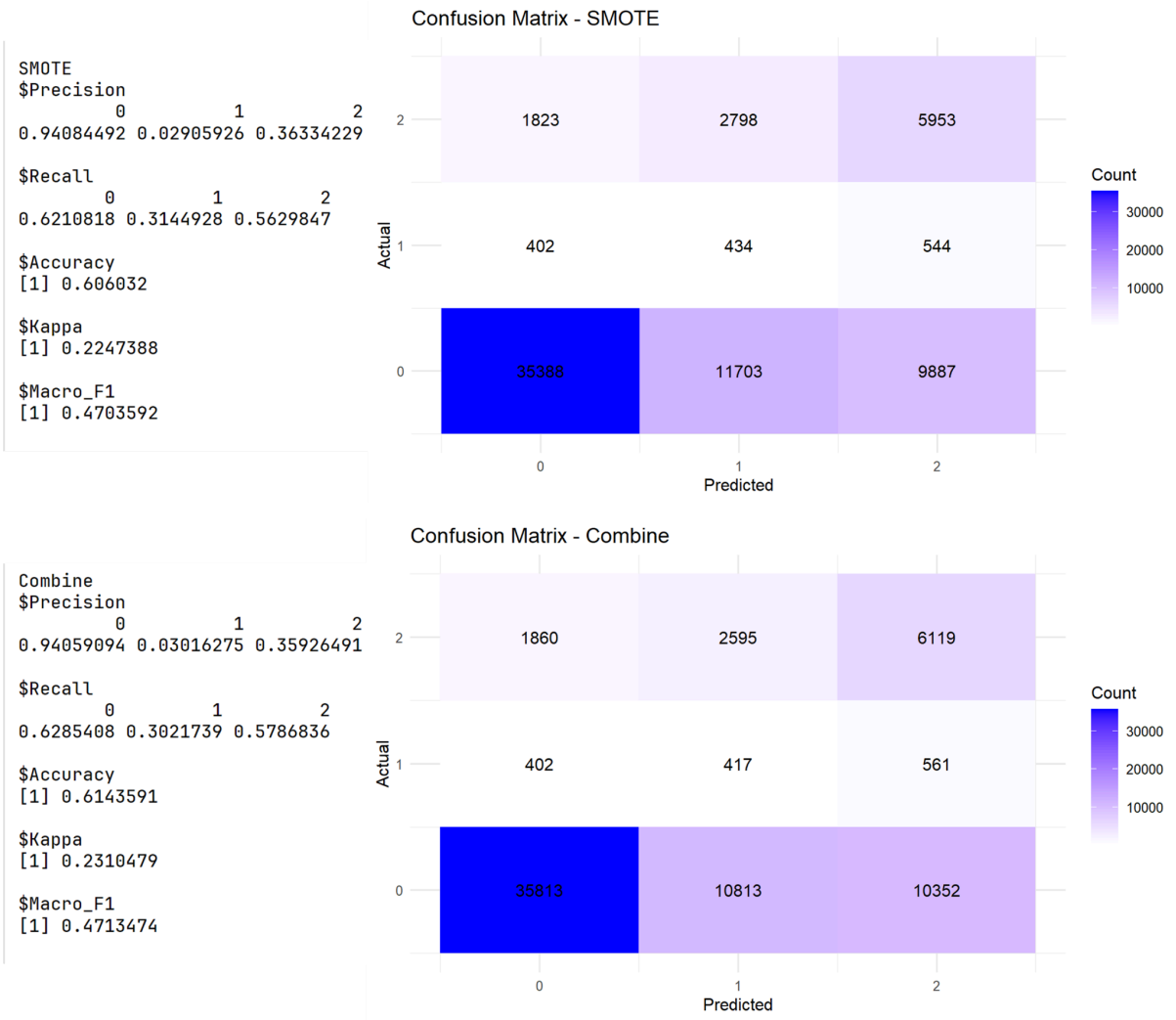
Trước tiên ta sẽ sử dụng tất cả các biến để huấn luyện mô hình (do tất cả các biến đều có tương quan với biến dự đoán).

```
Undersampling
$Precision
         0          1          2
0.94137314 0.03003288 0.35432478

$Recall
         0         1         2
0.6275931 0.2978261 0.5783998

$Accuracy
[1] 0.6134451

$Kappa
[1] 0.2301439

$Macro_F1
[1] 0.4697235
```




```
Oversampling
$Precision
         0          1          2
0.94047870 0.03041715 0.35880496

$Recall
         0         1         2
0.6275580 0.3043478 0.5803859

$Accuracy
[1] 0.6138513

$Kappa
[1] 0.2308088

$Macro_F1
[1] 0.4717102
```





***Hình 8**: Các kết quả Evaluation và Confusion Matrix lần lượt của 4 mô hình*

Dựa vào kết quả chạy thử nghiệm cho thấy model được huấn luyện tốt nhất với tập train_combine. Tiếp theo ta sẽ thử sử dụng công cụ lựa chọn biến tự động để xem có thể trích ra 1 tập con yếu tố làm cải thiện mô hình không với hàm stepAIC.

[Kết quả thí nghiệm]

Kết quả cho thấy tập các yếu tố tốt nhất cho mô hình vẫn là tập ban đầu với tất cả các biến giải thích. Tuy nhiên thì mô hình này đang gặp vấn đề đa cộng tuyến. Nên ta sẽ sử dụng các phương pháp co hệ số để xử lý.

2.4.3. Logistic Regression và phương pháp co hệ số

Sử dụng thư viện glmnet để huấn luyện mô hình multinomial logistic với co hệ số. Thực hiện huấn luyện với các phương pháp co hệ số khác nhau (lasso, ridge, elastic net)

Kết quả thu được cho thấy phương pháp co hệ số Ridge cho kết quả tốt nhất. Tuy nhiên các chỉ số đánh giá của mô hình không quá tốt, cụ thể:

[Kết quả thí nghiệm]

Điều này có thể nghi ngờ có mối quan hệ phi tuyến trong bộ dữ liệu, hoặc các vấn đề khác về chẩn đoán mô hình. Ta sẽ thử mô hình học máy để xem kết quả có được cải thiện hay không.

2.4.4. Mô hình học máy Random Forests

Random Forests là một mô hình phổ biến trong lĩnh vực học máy nhờ khả năng xử lý cả dữ liệu tuyến tính và phi tuyến. Một trong những ưu điểm nổi bật của mô hình này là việc sử dụng phương pháp bootstrap kết hợp với kỹ thuật bagging. Điều này giúp Random Forests giảm thiểu hiện tượng overfitting, đồng thời cải thiện độ chính xác và tính ổn định của dự đoán.

Cụ thể, bootstrap cho phép tạo ra nhiều tập dữ liệu con bằng cách lấy mẫu ngẫu nhiên từ tập dữ liệu gốc, trong đó một mẫu có thể được chọn nhiều lần. Mỗi tập dữ liệu con được sử dụng để huấn luyện một cây quyết định trong rừng, và kết quả cuối cùng được tổng hợp bằng cách trung bình (đối với bài toán hồi quy) hoặc bỏ phiếu đa số (đối với bài toán phân loại).

Ngoài ra, Random Forests còn có khả năng đánh giá tầm quan trọng của các đặc trưng (feature importance), giúp chúng ta hiểu rõ hơn về ảnh hưởng của từng biến đến kết quả mô hình. Đây là một lợi thế quan trọng khi làm việc với các bài toán phức tạp hoặc dữ liệu có nhiều biến đầu vào.

Kết quả:

[Kết quả thí nghiệm]

Mặc dù model Random Forest cho recall của class 2 khá cao, tuy nhiên recall cho class 1 lại gần như bằng 0. Có thể tuning các tham số cho mô hình bằng phương pháp k-fold cross-validation. Kết quả sau khi tuning cho thấy recall cho 2 class này có xu hướng tương quan nghịch với nhau chứ không thể đồng thời cũng cao.

Hơn nữa, macro_F1 cũng chỉ đạt 0.45, cũng tương đương với model logistic đã xây dựng ở trên. **Điều này cho thấy cả 2 model tuyến tính và phi tuyến đều không hoạt động tốt trên tập dữ liệu.**

2.5. Tổng hợp phân tích

[tổng hợp các nhiệm vụ và kết quả đã làm, nhận xét]

Chương 3: Kết luận

Nhận xét: Kết quả sau khi xây dựng các mô hình phân loại để dự đoán tình trạng bệnh tiểu đường của 1 người dựa trên bộ dữ liệu được cung cấp không đủ mạnh để đưa vào ứng dụng thực tế. Có thể vấn đề nằm ở bộ dữ liệu được cung cấp không có đủ thông tin để dự đoán biến mục tiêu hoặc có chứa thông tin nhiễu. Nhất là đối với class 1 (tiền tiểu đường), khi dữ liệu trong nhóm này rất ít so với 2 class còn lại nên việc dự đoán nhóm này rất khó dù đã có các biện pháp xử lý mất cân bằng.

Kết luận: Trả lời cho 2 câu hỏi ở đầu bài:

1. ***Dựa vào các yếu tố được khảo sát từ BRFSS có thể đưa ra dự đoán 1 cá nhân có thể bị tiểu đường hoặc tiền tiểu đường hay không ?***

Không, vì mô hình xây dựng dựa trên các yếu tố và dữ liệu được khảo sát phân loại chưa tốt.

2. ***Những yếu tố nào quan trọng trong việc dự đoán, và có thể lựa chọn 1 tập hợp con các yếu tố đó để tạo 1 câu hỏi ngắn để dự đoán tình trạng tiểu đường của 1 người hay không ?***

Tất cả các yếu tố được khảo sát đều có đóng góp vào việc dự đoán, có 1 vài yếu tố quan trọng như là BMI, PhysHlth, HighBP. Tuy nhiên vì không thể xây dựng được mô hình đủ tốt dựa trên các yếu tố ban đầu của dữ liệu nên cũng không thể lựa chọn được 1 tập hợp con các yếu tố cho dự đoán.

Giải pháp:

1. Thu thập thêm dữ liệu những người bị tiền tiểu đường (class 1).
2. Chuyển qua phân loại nhị phân, dự đoán 1 người bị tiểu đường hoặc không. (loại bỏ hết các quan trắc thuộc class 1 hoặc gom nhãn class 1 và 2).

Do tính chất về mức độ bệnh và phương pháp điều trị cho người tiểu đường và tiền tiểu đường là khác nhau nên việc chuyển hướng sang phân loại nhị phân có thể không giải thích được hết vấn đề đặt ra.

Theo khảo sát dân cư ở Mỹ thì số lượng người tiền tiểu đường nhiều hơn số lượng người tiểu đường nên việc dự đoán tiền tiểu đường cũng rất cần thiết. Vậy nên tùy theo tình huống và mục tiêu cụ thể có thể lựa chọn 1 trong 2 phương án trên để tiếp tục phân tích.