**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT**

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue circle with white text

Description automatically generated

**BÁO CÁO**

**BÁO CÁO BÀI TẬP GIỮA KÌ**

**MÔN HỌC: TRÍ TUỆ NHÂN TẠO**

**ĐỀ TÀI**

**DỰ BÁO CHI PHÍ BẢO HIỂM SỨC KHỎE DỰA TRÊN CÁC YẾU TỐ ẢNH HƯỞNG BẰNG MÔ HÌNH HỒI QUY TUYẾN TÍNH**

**CÁN BỘ GIẢNG DẠY**

**TS. Nguyễn Hoàng Long**

**SINH VIÊN THỰC HIỆN**

Nguyễn Thị Bảo Ngọc 2221050172

Lớp DCCTCLC67A

**Hà Nội, 2024**

**MỤC LỤC**

[**MỤC LỤC** 1](#_Toc186375822)

[**DANH MỤC HÌNH ẢNH** 3](#_Toc186375823)

[I. TÓM TẮT ĐỀ TÀI 5](#_Toc186375824)

[II. MỤC TIÊU 6](#_Toc186375825)

[III. PHƯƠNG PHÁP NGHIÊN CỨU 6](#_Toc186375826)

[1. Sử dụng mô hình Logistic Regression 6](#_Toc186375827)

[2. Sử dụng mô hình Naïve Bayes 6](#_Toc186375828)

[3. Sử dụng mô hình Decision Tree 7](#_Toc186375829)

[4. Sử dụng mô hình Random Forest 7](#_Toc186375830)

[IV. CÁC BƯỚC THỰC HIỆN 8](#_Toc186375831)

[1. Tìm kiếm dữ liệu 8](#_Toc186375832)

[2. Import các thư viện cần thiết 8](#_Toc186375833)

[3. Đọc tệp dữ liệu 8](#_Toc186375834)

[4. Xử lý dữ liệu 8](#_Toc186375835)

[5. Chọn biến độc lập tốt 8](#_Toc186375836)

[6. Chia tệp dữ liệu 8](#_Toc186375837)

[7. Xây dựng mô hình 8](#_Toc186375838)

[**7.1 Mô hình Logistic Regression** 8](#_Toc186375839)

[**7.2 Mô hình Naïve Bayes** 8](#_Toc186375840)

[**7.3 Mô hình Decision Tree** 8](#_Toc186375841)

[**7.4 Mô hình Random Forest** 8](#_Toc186375842)

[V. KẾT LUẬN 8](#_Toc186375843)

# **DANH MỤC HÌNH ẢNH**

[Hình 1 - Import các thư viện 8](#_Toc181317630)

[Hình 2 - Hiển thị dữ liệu 9](#_Toc181317631)

[Hình 3 - Kiểm tra giá trị trùng lặp 9](#_Toc181317632)

[Hình 4 - Tìm giá trị trùng lặp 9](#_Toc181317633)

[Hình 5 - Xóa giá trị trùng lặp 10](#_Toc181317634)

[Hình 6 - Thông tin dữ liệu 10](#_Toc181317635)

[Hình 7 - Đếm và kiểm tra dữ liệu có đồng nhất hay không 11](#_Toc181317636)

[Hình 8 - Kiểm tra khoảng trắng trong cột dữ liệu object 11](#_Toc181317637)

[Hình 9 - Xử lý các dữ liệu bị thiếu 12](#_Toc181317638)

[Hình 10 - Chỉ số của các cột dữ liệu 12](#_Toc181317639)

[Hình 11 - Tính chỉ số z\_score để tìm giá trị ngoại lai 13](#_Toc181317640)

[Hình 12 - Xóa các dữ liệu ngoại lai 13](#_Toc181317641)

[Hình 13 - Biểu đồ pairplot thể hiện mối tương quan giữa các biến 14](#_Toc181317642)

[Hình 14 Biểu đồ barplot tương quan giữa Age và Charges 15](#_Toc181317643)

[Hình 15 Biểu đồ barplot tương quan giữa Age, Sex với Charges 16](#_Toc181317644)

[Hình 16 Biểu đồ barplot tương quan giữa smoker với charges 17](#_Toc181317645)

[Hình 17 Biểu đồ barplot tương quan giữa Region với Charges 18](#_Toc181317646)

[Hình 18 - Biểu đồ scatter thể hiện mối tương quan 3 biến 20](#_Toc181317647)

[Hình 19 Biểu đồ scatter tương quan giữa bmi, sex với charges 21](#_Toc181317648)

[Hình 21 Kiểm tra tính đa cộng tuyến 22](#_Toc181317649)

[Hình 20 - Hàm Forward Selection 23](#_Toc181317650)

[Hình 22 - Chia tệp dữ liệu 24](#_Toc181317651)

[Hình 23 - Kết quả mô hình 25](#_Toc181317652)

[Hình 24 - Kiểm tra tính chuẩn tắc 26](#_Toc181317653)

[Hình 25 - KIểm tra tính đồng nhất 26](#_Toc181317654)

[Hình 27 - Kiểm tra hiện tượng fitting 27](#_Toc181317655)

1. **TÓM TẮT ĐỀ TÀI**

Bệnh tiểu đường, một trong những bệnh mãn tính phổ biến nhất tại Hoa Kỳ, không chỉ ảnh hưởng đến hàng triệu người mỗi năm mà còn tạo ra gánh nặng tài chính đáng kể cho nền kinh tế. Đây là một bệnh mãn tính nghiêm trọng, khi cơ thể mất khả năng điều chỉnh hiệu quả mức đường trong máu, dẫn đến giảm chất lượng cuộc sống và tuổi thọ.

Các biến chứng như bệnh tim, mất thị lực, cắt cụt chi dưới và bệnh thận thường liên quan đến mức đường máu cao kéo dài ở người mắc bệnh tiểu đường. Mặc dù không có phương pháp chữa trị, việc giảm cân, ăn uống lành mạnh, hoạt động thể chất và nhận điều trị y tế có thể giảm thiểu tác động của bệnh này.

Quy mô của vấn đề này rất đáng chú ý, với hàng triệu người Mỹ mắc bệnh tiểu đường và tiền tiểu đường. Mặc dù có nhiều loại tiểu đường, tiểu đường loại II là phổ biến nhất, và tỷ lệ mắc bệnh này thay đổi theo nhiều yếu tố. Bệnh tiểu đường không chỉ ảnh hưởng đến sức khỏe cá nhân mà còn tạo ra gánh nặng lớn cho nền kinh tế với chi phí khổng lồ hàng năm.

Đề tài này tập trung vào việc dự báo bệnh tiểu đường, mục tiêu là phát triển mô hình dự đoán để xác định nguy cơ mắc bệnh và từ đó áp dụng các biện pháp phòng ngừa và quản lý hiệu quả. Đây là một lĩnh vực nghiên cứu quan trọng để giúp cải thiện sức khỏe cộng đồng và giảm thiểu tác động của bệnh tiểu đường trong xã hội.

1. **MỤC TIÊU**

Dự đoán bệnh tiểu đường từ câu hỏi khảo sát BRFSS: Mục tiêu chính của đề tài là khám phá khả năng sử dụng các câu hỏi khảo sát từ Hệ thống giám sát các yếu tố rủi ro hành vi (BRFSS) để dự đoán xác định liệu một cá nhân có mắc bệnh tiểu đường hay không.

Xác định yếu tốc rủi ro quan trọng trong dự đoán tiểu đường: đề tài tập trung vào việc xác định những yếu tố rủi ro quan trọng nhất có khả năng dự đoán nguy cơ mắc bệnh tiểu đường một cách chính xác. Bằng cách phân tích dữ liệu từ BRFSS, chúng ta sẽ tìm ra những yếu tố ảnh hưởng nhiều nhất đến việc dự đoán bệnh tiểu đường.

Sử dụng yếu tố rủi ro để dự đoán bệnh tiểu đường: đề xuất tiếp cận sử dụng một phần nhỏ các yếu tố rủi ro để xây dựng một mô hình dự đoán chính xác liệu một cá nhân có mắc bệnh tiểu đường hay không. Mục tiêu là tạo ra một ứng dụng web tích hợp mô hình tốt nhất, giúp dự đoán nguy cơ mắc bệnh tiểu đường dựa trên các yếu tố rủi ro quan trọng đã xác định. Điều này sẽ hỗ trợ trong việc chuẩn đoán sớm, phòng ngừa và quản lý bệnh tiểu đường một cách hiệu quả và tiện lợi.

1. **PHƯƠNG PHÁP NGHIÊN CỨU**
2. **Sử dụng mô hình Logistic Regression**

**Ưu điểm:** Đầu tiên, mô hình Logistic Regression rất dễ hiểu và dễ áp dụng. Với tính đơn giản, nó thích hợp cho cả người mới bắt đầu trong lĩnh vực Machine learning. Thứ hai, tính tường minh của Logistic Regression là một ưu điểm lớn. Kết quả từ mô hình này có thể được giải thích một cách rõ ràng, giúp hiểu rõ hơn về tác động của các biến đầu vào đến biến phụ thuộc. Ngoài ra, khả năng xử lý dữ liệu đầu vào của Logistic Regression cũng là một điểm mạnh. Mô hình này có thể xử lý cả biến đầu vào liên tục và rời rạc, đồng thời có khả năng làm việc tốt với dữ liệu nhiễu.

**Nhược điểm:** Tuy nhiên, mô hình Logistic Regression cũng có nhược điểm của mình. Đầu tiên, giả định về mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc có thể là hạn chế, đặc biệt đối với dữ liệu có mối quan hệ phi tuyến. Thứ hai, mô hình này dễ bị ảnh hưởng bởi dữ liệu nhiễu. Nếu có nhiễu hoặc outliers trong dữ liệu, dự đoán từ Logistic Regression có thể không chính xác. Cuối cùng, Logistic Regression không thể xử lý được các mối quan hệ phức tạp giữa biến đầu vào và biến phụ thuộc, cũng như không hiệu quả khi xử lý multicollinearity - sự tương quan mạnh giữa các biến đầu vào.

1. **Sử dụng mô hình Naïve Bayes**

**Ưu điểm:** Mô hình Naïve Bayes được đánh giá cao về tính dễ triển khai và hiệu quả, đặc biệt là trên các tập dữ liệu lớn. Độ phức tạp thấp cũng là điểm mạnh của mô hình này, giúp giảm thời gian huấn luyện và dự đoán. Với khả năng xử lý tốt trên dữ liệu lớn và dữ liệu thưa, Naïve Bayes thường cho hiệu suất tốt và không gặp vấn đề overfitting. Mô hình này cũng linh hoạt khi xử lý cả biến đầu vào rời rạc và liên tục.

**Nhược điểm:** Tuy nhiên, mô hình Naïve Bayes có giả định về sự độc lập giữa các biến, đôi khi không phản ánh đúng mối quan hệ thực tế giữa chúng. Điều này có thể ảnh hưởng đến kết quả dự đoán. Một nhược điểm khác là Naïve Bayes yêu cầu lượng dữ liệu đầu vào đủ lớn để hoạt động hiệu quả. Ngoài ra, mô hình không thể học được các mối quan hệ phức tạp giữa các biến và không xử lý tốt với missing data.

1. **Sử dụng mô hình Decision Tree**

**Ưu điểm**: Một trong những ưu điểm lớn nhất của Decision Tree là tính dễ hiểu và dễ giải thích. Cấu trúc cây quyết định dễ dàng được biểu diễn dưới dạng đồ thị, giúp người dùng hiểu rõ quá trình ra quyết định của mô hình. Decision Tree có khả năng xử lý cả dữ liệu phân loại và dữ liệu số, không yêu cầu chuẩn hóa dữ liệu và có khả năng xử lý missing values một cách tương đối tốt. Mô hình Decision Tree cũng linh hoạt trong việc xử lý các mối quan hệ phi tuyến, có khả năng tạo ra các quy tắc quyết định phức tạp và phù hợp với dữ liệu không phải tuyến tính.

**Nhược điểm:** Tuy nhiên, mô hình Decision Tree dễ bị overfitting, đặc biệt khi cây quyết định quá sâu và phức tạp. Điều này có thể dẫn đến việc mô hình không tổng quát hóa tốt trên dữ liệu mới. Một nhược điểm khác của Decision Tree là sự nhạy cảm đối với sự biến đổi nhỏ trong dữ liệu đầu vào, có thể dẫn đến sự không ổn định trong dự đoán của mô hình. Decision Tree cũng có thể dẫn đến việc tạo ra quá nhiều lá (leaves) nhỏ, dẫn đến việc mô hình trở nên quá phức tạp và khó diễn giải.

1. **Sử dụng mô hình Random Forest**

**Ưu điểm:** Một trong những ưu điểm lớn nhất của Random Forest là khả năng xử lý overfitting, một vấn đề thường gặp trong các cây quyết định đơn lẻ. Bằng cách kết hợp nhiều cây quyết định, Random Forest giảm nguy cơ overfitting và tạo ra một mô hình tổng quát hóa tốt. Random Forest có khả năng làm việc hiệu quả trên dữ liệu có số lượng biến lớn và không yêu cầu chuẩn hóa dữ liệu. Nó cũng có khả năng xử lý cả biến dạng số và biến dạng phân loại một cách hiệu quả. Mô hình Random Forest cung cấp một ước lượng đáng tin cậy về tầm quan trọng của các biến đầu vào đối với mục tiêu dự đoán, giúp hiểu rõ hơn về quá trình ra quyết định của mô hình.

**Nhược điểm:** Mặc dù Random Forest giảm nguy cơ overfitting so với cây quyết định đơn lẻ, nhưng vẫn có khả năng overfitting trên các bộ dữ liệu nhỏ hoặc dữ liệu noise. Một nhược điểm khác của Random Forest là đòi hỏi nhiều tài nguyên tính toán hơn so với một cây quyết định đơn lẻ, đặc biệt khi số lượng cây lớn và có nhiều biến đầu vào. Random Forest có thể trở nên khó giải thích do sự phức tạp của mô hình, đặc biệt khi số lượng cây lớn và có nhiều biến đầu vào, làm giảm khả năng diễn giải của mô hình.

1. **CÁC BƯỚC THỰC HIỆN**
   1. **Tìm kiếm dữ liệu**

Hệ thống Giám sát Các Yếu tố Rủi ro Hành vi (BRFSS) là một cuộc khảo sát hàng năm về sức khỏe được CDC tiến hành thông qua điện thoại, thu thập câu trả lời từ hơn 400.000 người Mỹ. Cuộc khảo sát này tập trung vào việc thu thập thông tin về các hành vi rủi ro liên quan đến sức khỏe, tình trạng sức khỏe mãn tính và việc sử dụng dịch vụ phòng ngừa. BRFSS đã được thực hiện hàng năm từ năm 1984. Tập dữ liệu cho năm 2015, có sẵn trên Kaggle, bao gồm câu trả lời từ 441.455 cá nhân với 330 đặc tính. Các đặc tính này bao gồm cả các câu hỏi trực tiếp hỏi người tham gia cũng như các biến tính toán dựa trên câu trả lời của mỗi cá nhân.

Bộ dữ liệu **"diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.csv"** là một bộ dữ liệu sạch chứa 70.692 câu trả lời từ BRFSS 2015 của CDC. Bộ dữ liệu này được chia đều 50-50 giữa những người không mắc bệnh tiểu đường và những người bị tiền tiểu đường hoặc mắc tiểu đường. Biến mục tiêu **"Diabetes\_binary"** có 2 giá trị: 0 đại diện cho không mắc tiểu đường, và 1 đại diện cho tiền tiểu đường hoặc mắc tiểu đường. Bộ dữ liệu này bao gồm 21 biến đặc tính và đã được cân bằng để phù hợp với mục tiêu nghiên cứu hoặc phân tích dữ liệu.

**Thông tin tệp dữ liệu:**

* **Diabetes\_binary:** Biến nhị phân cho biết một người có mắc bệnh tiểu đường hay không (0 = không, 1 = có).
* **HighBP:** Cho biết người đó có bị cao huyết áp hay không (0 = không, 1 = có).
* **HighChol:** Cho biết người đó có bị cholesterol cao hay không (0 = không, 1 = có).
* **CholCheck:** Cho biết người đó có kiểm tra cholesterol trong vòng 5 năm gần đây (0 = không, 1 = có).
* **BMI:** Chỉ số khối cơ thể (Body Mass Index), đo lường mỡ cơ thể dựa trên chiều cao và cân nặng.
* **Smoker:** Đo lường việc một người đã hút ít nhất 100 điếu thuốc trong suốt cuộc đời của họ (0 = chưa hút, 1 = đã hút).
* **Stroke:** Cho biết người đó đã từng bị đột quỵ hay chưa (0 = chưa, 1 = rồi).
* **HeartDiseaseorAttack:** Cho biết người đó đã từng bị bệnh tim hoặc cơn đau tim hay chưa (0 = chưa, 1 = rồi).
* **PhysActivity:** Cho biết người đó có tham gia hoạt động thể chất trong 30 ngày gần đây ngoài công việc không (0 = không, 1 = có).
* **Fruits:** Cho biết người đó có ăn trái cây 1 lần trở lên mỗi ngày không (0 = không, 1 = có).
* **Veggies:** Cho biết người đó có ăn rau củ 1 lần trở lên mỗi ngày không (0 = không, 1 = có).
* **HvyAlcoholConsump:** (đàn ông trưởng thành >=14 ly mỗi tuần và phụ nữ trưởng thành >=7 ly mỗi tuần) (0 = không 1 = có).
* **AnyHealthcare:** Cho biết người đó có bất kỳ loại bảo hiểm chăm sóc sức khỏe nào, bao gồm bảo hiểm y tế, các chương trình trả trước như HMO..(0 = không, 1 = có)
* **NoDocbcCost:** Biến này đo lường việc trong vòng 12 tháng gần đây có một thời điểm bạn cần phải thăm bác sĩ nhưng không thể vì lý do chi phí hay không (0 = không, 1 = có).
* **GenHlth:** Biến này đo lường cảm nhận tổng quan về sức khỏe của một cá nhân trên một thang điểm từ 1 đến 5 (1 = Xuất sắc, 2 = Rất tốt, 3 = Tốt, 4 = Trung bình, 5 = Kém).
* **MentHlth:** Ngày sức khỏe tinh thần kém thang điểm 1-30 ngày.
* **PhysHlth:** Số ngày bệnh tật hoặc thương tích trong 30 ngày qua, thang điểm 1-30
* **DiffWalk:** Cho biết người đó có gặp khó khăn khi đi bộ không.( 0 = không gặp khó khăn, 1 = gặp khó khăn)
* **Sex:** Giới tính của người đó (1 = nam, 0 = nữ, hoặc mã hóa tương tự).
* **Age:** Biến "Age" (\_AGEG5YR) được chia thành 13 cấp độ tuổi, bao gồm các khoảng tuổi như sau: 1 = 18-24 tuổi, 2 = 25-29 tuổi, 3 = 30-34 tuổi, 4 = 35-39 tuổi, 5 = 40-44 tuổi, 6 = 45-49 tuổi, 7 = 50-54 tuổi, 8 = 55-59 tuổi, 9 = 60-64 tuổi, 10 = 65-69 tuổi, 11 = 70-74 tuổi, 12 = 75-79 tuổi, 13 = 80 tuổi trở lên.
* **Education:** Đo lường mức độ giáo dục của một cá nhân trên một thang điểm từ 1 đến 6 (1 = Chưa học/ Chỉ học mẫu giáo), 2 = Tiểu học, 3 = Trung học, 4 = Tốt nghiệp trung học, 5 = Đại học, 6 = Tốt nghiệp đại học rồi).
* **Income:** Mức thu nhập thang (1-8), bao gồm các thang: 1 = dưới $10.000; 2 = Từ $10,000 đến dưới $15,000; 3 = Từ $15,000 đến dưới $20,000; 4 = Từ $20,000 đến dưới $25,000; 5 = Từ $25,000 đến dưới $35,000; 6 = Từ $35,000 đến dưới $50,000; 7 = Từ $50,000 đến dưới $75,000; 8 = $75.000 trở lên.
  1. **Import các thư viện cần thiết**

A screen shot of a computer code

Description automatically generated

Hình 1- Import thư viện

* 1. **Đọc tệp dữ liệu**

A screenshot of a computer

Description automatically generated

Hình 2- Đọc tệp dữ liệu

* 1. **Xử lý dữ liệu**
  2. **Quan sát dữ liệu cân bằng**

|  |
| --- |
| Hình 3 - Quan sát dữ liệu  **Nhận xét:** Tập dữ liệu cân bằng giữa 2 lớp đầu ra:   * Lớp 0: Không bị bệnh ~ 50%. * Lớp 1: Bị bệnh ~ 50%.   Vì vậy, tập dữ liệu này không cần xử lý cân bằng. |

* 1. **Loại bỏ các giá trị trùng lặp**
  2. **Xử lý dữ liệu thiếu**
  3. **Xử lý giá trị ngoại lai**

|  |
| --- |
|  |

**Nhận xét:** Tập dữ liệu cân bằng giữa 2 lớp đầu ra:

* Lớp 0: Không bị bệnh ~ 50%.
* Lớp 1: Bị bệnh ~ 50 %.

Vì vậy, tập dữ liệu này không cần xử lý cân bằng.

* 1. **Feature Engineering**

### **5.1 Feature Creation**

* 1. **Chọn biến độc lập tốt.**
  2. **Chia tệp dữ liệu**
  3. **Xây dựng mô hình**

**7.1 Mô hình Logistic Regression**

**7.2 Mô hình Naïve Bayes**

**7.3 Mô hình Decision Tree**

**7.4 Mô hình Random Forest**

1. **KẾT LUẬN**