|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **TIỂU LUẬN MÔN HỌC**  **CÔNG NGHỆ KHOA HỌC DỮ LIỆU**  **DỰ ĐOÁN VỀ TÌNH TRẠNG BỆNH TIỂU ĐƯỜNG**  Giảng viên giảng dạy: ThS SỬ NHẬT HẠ  Sinh viên thực hiện: LÊ XUÂN TRỌNG  MSSV : 2100009483  Chuyên ngành : KHOA HỌC DỮ LIỆU  Môn học : CÔNG NGHỆ KHOA HỌC DỮ LIỆU  Khóa : 2021    **Tp.HCM, tháng 09 Năm 2024** |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **TIỂU LUẬN MÔN HỌC**  **CÔNG NGHỆ KHOA HỌC DỮ LIỆU**  **DỰ ĐOÁN VỀ TÌNH TRẠNG BỆNH TIỂU ĐƯỜNG**  Giảng viên giảng dạy:ThS SỬ NHẬT HẠ  Sinh viên thực hiện: LÊ XUÂN TRỌNG  MSSV : 2100009483  Chuyên ngành : KHOA HỌC DỮ LIỆU  Môn học : CÔNG NGHỆ KHOA HỌC DỮ LIỆU  Khóa : 2021    **Tp.HCM, tháng 09 Năm 2024** |

|  |  |
| --- | --- |
| Trường Đại học Nguyễn Tất Thành  **Khoa Công Nghệ Thông Tin**  🙜  🙜  🙝  🙝 | CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM  **Độc lập – Tự do – Hạnh phúc**  🙜 🙜 🙝 🙝 |

**NHIỆM VỤ ĐỒ ÁN**

Họ và tên: **LÊ XUÂN TRỌNG** MSSV: **2100009483**

Chuyên ngành: **Khoa học dữ liệu** Lớp: **21DTH3A** Email: **xuantrongle987@gmail.com** SĐT: **0366394982** Tên đề tài: Dự đoán về tình trạng bệnh tiểu đường

Giảng viên hướng dẫn: **ThS.Sử Nhật Hạ**

Thời gian thực hiện:  **24 /06 /2024 đến 05/ 09 /2024**

Nhiệm vụ/nội dung (mô tả chi tiết nội dung, yêu cầu, phương pháp):

**Nội dung và yêu cầu đã được thông qua Bộ môn.**

*TP.HCM, ngày 04 tháng 9 năm 2024*

|  |  |
| --- | --- |
| **Q. TRƯỞNG BỘ MÔN**  *(Ký và ghi rõ họ tên)* | **GIÁO VIÊN HƯỚNG DẪN**  *(Ký và ghi rõ họ tên)* |

**LỜI MỞ ĐẦU**

Bệnh tiểu đường hiện nay đã trở thành một trong những vấn đề sức khỏe nghiêm trọng trên toàn cầu. Với sự gia tăng của các yếu tố nguy cơ như chế độ ăn uống không lành mạnh, lối sống ít vận động, tình trạng béo phì, và yếu tố di truyền, việc dự đoán nguy cơ mắc bệnh tiểu đường đã trở thành một phần quan trọng trong việc xác định những người có nguy cơ cao và áp dụng các biện pháp phòng ngừa hiệu quả.

Nghiên cứu này tập trung vào việc áp dụng các kỹ thuật học máy để xây dựng mô hình dự đoán nguy cơ mắc bệnh tiểu đường dựa trên các yếu tố nguy cơ y tế. Sử dụng dữ liệu lâm sàng và các biến đầu vào như chỉ số khối cơ thể (BMI), mức độ đường huyết, tình trạng huyết áp, tiền sử gia đình, và các thói quen sinh hoạt hàng ngày, chúng tôi nhằm mục tiêu tạo ra một công cụ dự đoán chính xác và đáng tin cậy để xác định những người có nguy cơ cao mắc bệnh tiểu đường.

Nghiên cứu này không chỉ nhằm mục tiêu cải thiện khả năng dự đoán mà còn hướng tới ứng dụng thực tiễn trong việc đưa ra các biện pháp can thiệp sớm, từ đó giảm thiểu nguy cơ mắc bệnh tiểu đường và cải thiện chất lượng cuộc sống của cộng đồng. Mục tiêu cuối cùng của nghiên cứu là tạo ra một công cụ hữu ích để hỗ trợ quyết định lâm sàng và chăm sóc sức khỏe cá nhân.

**LỜI CẢM ƠN**

Nếu như phải gửi một lời cảm ơn dành cho các thầy cô đã trợ giúp cho mình thì cho em xin phép mượn câu nói của ngài William a Warrrd, ông từng nói rằng:

“Một người thầy trung bình chỉ biết nói

Một người thầy giỏi biết giải thích

Một người thầy chúng biết minh họa

Một người thầy vĩ đại biết truyền cảm hứng”

Nếu phải dành lời cảm ơn em muốn mượn câu nói này để dành cho các thầy cô phụ trách, cũng như là thầy Sử Nhật Hạ giảng viên phụ trách môn khai thác dữ liệu và ứng dụng đã giúp cũng như hỗ trợ chúng em hoàn thành được đồ án về chủ đề dự đoán về vấn đề suy tim. Cũng từ đồ án này em nhận ra mình cần phải có được tội lỗi phàm ăn để khơi lên cơn đói khát về kiến thức của mình từ đó ăn bớt những hạn chế của mình trong con sông kiến thức cũng như tránh xa các tội lỗi kiêu ngạo, lười biếng và đố kị để làm tròn nên những kinh nghiệm mà mình cần phải học hỏi từ mọi người để bước gần hơn tới thành công. Em cũng mong nhận được những lời gớp ý giúp em cải thiện cũng hoàn thiện đồ án hơn.

Cuối cùng em xin kính chúc quí thầy cô có một ngày tốt lạnh.

Em chân thành cảm ơn

Sinh viên thực hiện

LÊ XUÂN TRỌNG

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ II NĂM HỌC 2023 - 2024** |

**PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN**

BM-ChT-11

Môn thi: Công Nghệ Khoa Học Dữ Liệu Lớp học phần:21DTH2D

Nhóm sinh viên thực hiện:

1. Lê Xuân Trọng Tham gia đóng góp:100%

2. Lê Hoài Ân Tham gia đóng góp:100%

3. Tăng Quốc Trung Tham gia đóng góp:100%

4. Tham gia đóng góp:

5. Tham gia đóng góp:

6. Tham gia đóng góp:

7. Tham gia đóng góp:

8. Tham gia đóng góp:

Ngày thi: 04/09/2024 Phòng thi: L.607

Đề tài tiểu luận/báo cáo của sinh viên : Dự án về dự đoán tình trạng bệnh tiểu đường

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo | Gồm 3 chương | 1 |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần | Chương 2 và 3 | 5 |  |
| * Lập luận | Chương 1 | 2 |  |
| * Kết luận |  | 1 |  |
| Trình bày | Theo chuẩn format luận văn font chữ 13, canh trái, phải,… | 1 |  |
| **TỔNG ĐIỂM** |  | **10** |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)* |

**NHẬN XÉT CỦA GIẢNG VIÊN GIẢNG DẠY**

*Tp.HCM, Ngày 28 tháng 12 năm 2022*

**Giảng viên giảng dạy**

(Ký tên và ghi rõ họ tên)

**MỤC LỤC**

[LỜI MỞ ĐẦU iii](#_Toc31777)

[LỜI CẢM ƠN v](#_Toc30099)

[CHƯƠNG 1: GIỚI THIỆU 1](#_Toc6506)

[1.1 Giới thiệu đề tài. 1](#_Toc914)

[1.2.Lý do chọn đề tài 2](#_Toc23070)

[1.3.Mục tiêu đề tài. 3](#_Toc8684)

[1.4.Phương pháp đề tài. 4](#_Toc15071)

[1.5.Đối tượng và phạm vi nghiên cứu. 5](#_Toc32608)

[1.5.1. Đối Tượng Nghiên Cứu 5](#_Toc8836)

[1.5.2. Phạm Vi Nghiên Cứu 6](#_Toc19214)

[CHƯƠNG 2 ỨNG DỤNG VÀ THUẬT TOÁN 7](#_Toc12220)

[2.1 Giới thiệu bài toán. 7](#_Toc27204)

[2.2 Mô tả thuật toán. 7](#_Toc13990)

[2.2.1. Thuật toán Logistic Regression 8](#_Toc31872)

[2.2.2. Thuật toán Random Forest 8](#_Toc3365)

[2.2.3. Thuật toán K-Nearest Neighbors 9](#_Toc19212)

[2.3. Xây dựng bộ dữ liệu. 9](#_Toc16000)

[2.3.1. Bộ dữ liệu gồm các cột  9](#_Toc9880)

[2.3.2. Tiền xử lý dữ liệu  10](#_Toc2343)

[2.4. Áp dụng thuật toán vào bài toán. 10](#_Toc10851)

[2.5. Thực nghiệm với thư viện. 11](#_Toc24483)

[CHƯƠNG 3 XÂY DỰNG ỨNG DỤNG NGÔN NGỮ PYTHON 14](#_Toc23031)

[3.1 Xây dựng ứng dụng và giải thích 14](#_Toc11763)

[3.1.1. Logistic Regression 14](#_Toc30087)

[3.1.2. K-Nearest Neighbors 16](#_Toc28389)

[3.1.3. Random Forest 17](#_Toc618)

[3.1.4. Support Vector Classifier 18](#_Toc14641)

[3.1.5. Decision Tree Classifier 19](#_Toc12365)

[3.1.6. So sanh các phương pháp dự đoán 20](#_Toc27223)

[3.2. Một số ảnh trực phân hóa được phân tích trong dự án 21](#_Toc18702)

[3.2.1. Trực quan theo độ tuổi 21](#_Toc1430)

[3.2.2. Trực quan theo khả năng mắc bệnh tiểu đường ở các độ tuổi 21](#_Toc2148)

[3.2.3. Trực quan hóa về Mức đường huyết 22](#_Toc20343)

[3.2.4. Trực quan hóa về Mức đường huyết ảnh hưởng tới bệnh tiểu đường 22](#_Toc25016)

[3.2.5. Trực quan hóa về Huyết áp 23](#_Toc12985)

[3.2.6. Trực quan hóa về Huyết áp ảnh hưởng tới bệnh tiểu đường 23](#_Toc11741)

[3.2.7. Trực quan hóa về Nồng độ Insulin 24](#_Toc8615)

[3.2.8. Trực quan hóa về Nồng độ insulin ảnh hưởng tới bệnh tiểu đường 24](#_Toc18323)

[3.2.9. Trực quan hóa về Chỉ số khối cơ thể(BMI) 25](#_Toc32546)

[3.2.10. Trực quan hóa về Chỉ số khối cơ thể(BMI) ảnh hưởng tới bệnh tiểu đường 25](#_Toc5355)

[KẾT LUẬN 27](#_Toc13534)

[Kết quả đạt được. 27](#_Toc11531)

[Hạn chế và hướng phát triển . 28](#_Toc2212)

[TÀI LIỆU THAM KHẢO 29](#_Toc1256)

**DANH MỤC HÌNH**

[Hình 1 : Bảng dữ liệu 9](#_Toc6417)

[Hình 2 : Bảng thông tin dữ liệu 10](#_Toc1827)

[Hình 3 : Các thư viện đã được sử dụng 12](#_Toc9679)

[Hình 4 : Dữ liệu chuẩn bị cho phần học máy 14](#_Toc25749)

[Hình 5 : Khởi tạo mô hình Logistic Regression 15](#_Toc22687)

[Hình 6 : Kết quả mô hình Logistic Regreession 15](#_Toc22314)

[Hình 7 : Khởi tạo K-Nearest Neighbors 16](#_Toc19515)

[Hình 8 : Kết quả mô hình K-Nearest Neighbors 16](#_Toc30735)

[Hình 9 : Khởi tạo Random Forest 17](#_Toc20702)

[Hình 10 : Kết quả mô hình Random Forest 17](#_Toc29676)

[Hình 11 : Khởi tạo Support Vector Classifier 18](#_Toc10243)

[Hình 12 : Kết quả mô hình Support Vector Classifier 18](#_Toc17603)

[Hình 13 : Khởi tạo mô hình Decision Tree Classifier 19](#_Toc11631)

[Hình 14 : Kết quả mô hình Decision Tree Classifier 20](#_Toc17180)

[Hình 15 : Xuất ra phần trăm sai lệch của các dự báo 20](#_Toc4465)

[Hình 16 : Trực quan hóa của phần trăm sai lệch của các dự báo 21](#_Toc21087)

[Hình 17 : Trực quan hóa về tuổi 21](#_Toc2153)

[Hình 18 : Trực quan hóa về tuổi ảnh hưởng tới bệnh tiểu đường 22](#_Toc11393)

[Hình 19 : Trực quan hóa về nồng độ enzyme trong máu 22](#_Toc4999)

[Hình 20 : Trực quan hóa về Mức đường huyết ảnh hưởng tới bệnh tiểu đường 23](#_Toc4322)

[Hình 21 : Trực quan hóa về Huyết áp 23](#_Toc18133)

[Hình 22 : Trực quan hóa về Huyết áp ảnh hưởng tới bệnh tiểu đường 24](#_Toc23158)

[Hình 23 : Trực quan hóa về Nồng độ insulin huyết thanh 24](#_Toc12408)

[Hình 24 : Trực quan hóa về Nồng độ insulin ảnh hưởng tới bệnh tiểu đường 25](#_Toc16676)

[Hình 25 : Trực quan hóa về Chỉ số khối cơ thể(BMI) 25](#_Toc17715)

[Hình 26 :Trực quan hóa về Chỉ số khối cơ thể(BMI) ảnh hưởng tới bệnh tiểu đường 26](#_Toc19051)

**KÍ HIỆU CÁC CỤM TỪ VIẾT TẮT**

|  |  |
| --- | --- |
| Chữ viết tắt | Ý nghĩa |
| KNN | K-nearest neighbors |
| SVM | Support Vector Machines |

# CHƯƠNG 1: GIỚI THIỆU

* 1. **Giới thiệu đề tài.**

Trong kỷ nguyên hiện đại, bệnh tiểu đường, đặc biệt là tiểu đường type 2, đã không chỉ dừng lại ở một căn bệnh mạn tính thông thường mà đã trở thành một mối đe dọa nghiêm trọng đối với sức khỏe cộng đồng trên toàn cầu. Được mệnh danh là "kẻ giết người thầm lặng," bệnh tiểu đường không chỉ tác động tiêu cực đến cuộc sống của hàng triệu người mà còn gây ra những gánh nặng tài chính khổng lồ cho các hệ thống y tế trên toàn thế giới. Các biến chứng nguy hiểm liên quan đến tiểu đường như suy thận, bệnh tim mạch, và các tổn thương thần kinh không chỉ làm suy giảm chất lượng cuộc sống mà còn có thể dẫn đến tử vong nếu không được quản lý và điều trị kịp thời.

Trong bối cảnh đó, việc phát hiện sớm và dự đoán chính xác nguy cơ mắc bệnh tiểu đường trở thành một trong những thách thức lớn nhất của ngành y tế hiện đại. Tuy nhiên, các phương pháp truyền thống thường dựa vào dữ liệu lịch sử và các yếu tố nguy cơ đã biết, đôi khi không đủ nhanh chóng hoặc không đủ chính xác để đối phó với sự phát triển phức tạp của căn bệnh này.

Sự xuất hiện của công nghệ phân tích dữ liệu lớn (Big Data) đã mang đến những bước tiến vượt bậc trong y học. Với khả năng xử lý lượng dữ liệu khổng lồ từ nhiều nguồn khác nhau như hồ sơ bệnh án điện tử, dữ liệu di truyền, dữ liệu sinh trắc học, và các thông tin từ thiết bị theo dõi sức khỏe, Big Data không chỉ giúp chúng ta hiểu rõ hơn về các yếu tố nguy cơ tiểu đường mà còn giúp phát hiện ra những mô hình tiềm ẩn chưa từng được khám phá trước đây. Đặc biệt, khi kết hợp với Machine Learning – một nhánh của trí tuệ nhân tạo (AI), chúng ta có thể xây dựng những mô hình dự đoán tiên tiến, có khả năng học hỏi từ dữ liệu và liên tục cải thiện độ chính xác của mình.

## Đề tài "Nhận diện bệnh tiểu đường bằng phân tích dữ liệu lớn" không chỉ dừng lại ở việc sử dụng các thuật toán Machine Learning để dự đoán bệnh tiểu đường, mà còn tập trung vào việc tối ưu hóa các mô hình này để chúng có thể hoạt động hiệu quả trong các môi trường thực tế, nơi mà các yếu tố biến động và đa dạng của dữ liệu có thể gây ra nhiều thách thức. Với tiềm năng mở rộng và ứng dụng trong thực tiễn, đề tài này không chỉ mang lại giá trị nghiên cứu học thuật mà còn hứa hẹn những đóng góp quan trọng vào công cuộc cải thiện hệ thống chăm sóc sức khỏe cộng đồng, mở ra những triển vọng mới trong việc quản lý và phòng ngừa bệnh tiểu đường.

## 1.2.Lý do chọn đề tài

Nếu phải nói lý do nào để khiến bọn em chọn đề tài "Dự báo về tình trạng những người mắc bệnh tiểu đường" thì cự thể cũng không có nhiều lý do nhưng có thể kể đến những lý do như sau đó là:

* **Tính cấp thiết của vấn đề y tế toàn cầu:**

Bệnh tiểu đường đang ngày càng trở nên phổ biến, đặc biệt là tiểu đường type 2, với tỷ lệ mắc bệnh không ngừng gia tăng trên toàn thế giới. Theo các báo cáo của Tổ chức Y tế Thế giới (WHO), số người mắc bệnh tiểu đường dự kiến sẽ tiếp tục tăng cao trong các thập kỷ tới, dẫn đến những gánh nặng lớn đối với hệ thống y tế và nền kinh tế. Việc phát hiện sớm và quản lý bệnh tiểu đường không chỉ giúp giảm thiểu nguy cơ biến chứng mà còn cải thiện chất lượng cuộc sống của bệnh nhân. Điều này tạo ra nhu cầu cấp thiết về các giải pháp tiên tiến trong dự đoán và phòng ngừa bệnh tiểu đường.

* **Tiềm năng của Machine Learning trong dự đoán bệnh lý:**

Machine Learning, một nhánh quan trọng của trí tuệ nhân tạo (AI), đã chứng tỏ hiệu quả vượt trội trong việc phân tích và dự đoán từ các tập dữ liệu lớn, phức tạp. Các thuật toán học máy có khả năng tự học từ dữ liệu, tìm ra các mô hình tiềm ẩn và đưa ra các dự đoán chính xác mà các phương pháp truyền thống khó có thể đạt được. Việc áp dụng Machine Learning trong dự đoán bệnh tiểu đường giúp chúng ta không chỉ dự báo nguy cơ mắc bệnh mà còn cá nhân hóa các biện pháp can thiệp cho từng bệnh nhân cụ thể, nâng cao hiệu quả điều trị.

* **Tính thực tiễn và khả năng ứng dụng cao:**

Đề tài không chỉ dừng lại ở mức độ nghiên cứu lý thuyết mà còn có tiềm năng ứng dụng cao trong thực tế. Kết quả của nghiên cứu có thể được áp dụng trực tiếp trong việc cải tiến hệ thống chăm sóc y tế, giúp các bác sĩ và chuyên gia y tế đưa ra quyết định chính xác hơn trong việc dự đoán và quản lý bệnh tiểu đường. Ngoài ra, mô hình dự đoán cũng có thể được tích hợp vào các hệ thống quản lý y tế hiện đại, góp phần giảm thiểu tỷ lệ mắc bệnh và nâng cao chất lượng chăm sóc sức khỏe cho cộng đồng.

* **Đóng góp vào cộng đồng khoa học và xã hội:**

Bên cạnh những giá trị thực tiễn, đề tài này còn đóng góp vào lĩnh vực nghiên cứu khoa học trong y học và công nghệ. Việc phát triển các mô hình Machine Learning từ Big Data không chỉ giúp mở rộng hiểu biết về bệnh tiểu đường mà còn mang lại những phương pháp mới để đối phó với các bệnh lý khác. Những nghiên cứu về dự đoán bệnh lý dựa trên dữ liệu lớn không chỉ có ích cho cộng đồng y tế mà còn góp phần cải thiện chất lượng cuộc sống và sức khỏe của người dân nói chung.

* **Định hướng phát triển bền vững trong y học:**

Việc dự đoán và phòng ngừa bệnh tiểu đường sớm không chỉ giúp giảm gánh nặng cho các hệ thống y tế mà còn đóng góp vào phát triển bền vững trong chăm sóc sức khỏe. Khi các biện pháp dự đoán chính xác được triển khai, chi phí điều trị và quản lý bệnh tiểu đường sẽ được giảm thiểu đáng kể, tạo ra những lợi ích lâu dài cho cá nhân và xã hội. Chọn đề tài này là hướng tới mục tiêu phát triển một hệ thống y tế hiệu quả và bền vững hơn trong tương lai.

## 1.3.Mục tiêu đề tài.

Mục tiêu mà em đề ra cho đề tài "Dự đoán tình trạng bệnh tiểu đường" bao gôm các mục tiêu sau:

* **Xây dựng mô hình dự báo chính xác:** Phát triển các mô hình dựa trên dữ liệu y tế và công nghệ dữ liệu để định lượng và xác định nguy cơ mắc bệnh tiểu đường. Mục tiêu là tạo ra các công cụ có khả năng dự đoán chính xác nguy cơ mắc bệnh dựa trên các yếu tố nguy cơ đã được nhận diện.
* **Đánh giá các yếu tố nguy cơ chính:** Xác định và phân tích tác động của các yếu tố nguy cơ khác nhau đối với bệnh tim mạch. Việc loại bỏ hoặc cải thiện các yếu tố nguy cơ này có thể góp phần làm giảm nguy cơ mắc bệnh.
* **Hỗ trợ ra quyết định lâm sàng:** Cung cấp thông tin quan trọng để hỗ trợ bác sĩ và bệnh nhân trong việc đưa ra những quyết định lâm sàng hợp lý. Dự báo nguy cơ mắc bệnh tiểu đường giúp nâng cao nhận thức và khuyến khích quản lý sức khỏe chủ động.
* **Phòng ngừa và can thiệp sớm:** Sử dụng dữ liệu từ mô hình dự báo để đưa ra các biện pháp phòng ngừa bệnh tiểu đường, chẳng hạn như khuyến nghị thay đổi lối sống hoặc can thiệp y tế sớm để giảm thiểu nguy cơ mắc bệnh.
* **Nghiên cứu trên nhiều đối tượng:** Tiến hành nghiên cứu để đánh giá tác động của các yếu tố nguy cơ trên các nhóm đối tượng khác nhau, từ thanh thiếu niên đến người cao tuổi, từ các nhóm dân tộc đến các vùng địa lý khác nhau.

Tóm lại, mục tiêu của nghiên cứu này là phát triển các công cụ dự báo hiệu quả, cung cấp thông tin hữu ích và hỗ trợ trong việc ra quyết định cho cả bác sĩ và bệnh nhân, từ đó giảm nguy cơ mắc bệnh tim mạch và cải thiện sức khỏe tim mạch của cộng đồng.

## 1.4.Phương pháp đề tài.

Có một số phương pháp tiếp cận đề tài "Dự đoán nguy cơ mắc bệnh tiểu đường" mà em sử dụng để đạt được mục tiêu của nó

Thu thập và xử lý dữ liệu: Tiến hành thu thập dữ liệu từ nhiều nguồn khác nhau như bệnh viện, các nghiên cứu y tế, hoặc thiết bị đeo thông minh. Dữ liệu bao gồm các thông tin như tuổi tác, giới tính, các chỉ số y khoa (huyết áp, cholesterol, đường huyết, v.v.), lịch sử y tế cá nhân, và các yếu tố nguy cơ khác. Quá trình xử lý dữ liệu bao gồm các bước như làm sạch dữ liệu, xử lý giá trị thiếu, và chuẩn hóa dữ liệu để đảm bảo dữ liệu sẵn sàng cho các bước phân tích tiếp theo

Phân tích dữ liệu và xây dựng mô hình: Sử dụng các phương pháp như học máy, học sâu, hoặc các mô hình thống kê để phân tích dữ liệu và xây dựng mô hình dự đoán. Những mô hình phổ biến bao gồm hồi quy logistic, máy vector hỗ trợ, cây quyết định, rừng ngẫu nhiên, mạng nơ-ron, và nhiều mô hình dự đoán khác.

Đánh giá và cải thiện mô hình: Áp dụng các phương pháp như cross-validation, confusion matrix, đường cong ROC để đánh giá hiệu suất của mô hình. Quá trình tinh chỉnh mô hình có thể bao gồm điều chỉnh các siêu tham số và sử dụng các kỹ thuật chọn lọc đặc trưng (feature selection) để cải thiện độ chính xác của mô hình.

Xác định các yếu tố nguy cơ chính: Sử dụng các kỹ thuật như feature importance, SHAP values, hoặc permutation importance để nhận diện những yếu tố nguy cơ có tác động lớn nhất đến khả năng dự đoán nguy cơ mắc bệnh tim mạch.

Ứng dụng và triển khai: Sau khi hoàn thành mô hình dự đoán hiệu quả, ứng dụng mô hình này để dự báo nguy cơ mắc bệnh cho các bệnh nhân mới. Thông tin từ mô hình có thể hỗ trợ trong việc ra quyết định lâm sàng và đề xuất các biện pháp can thiệp sớm.

Bằng cách kết hợp hợp lý các phương pháp này, một quy trình phân tích toàn diện và tin cậy có thể được tạo ra để dự đoán nguy cơ mắc bệnh tiểu đường.

## 1.5.Đối tượng và phạm vi nghiên cứu.

Đối tượng và phạm vi nghiên cứu trong dự án về dự đoán nguy cơ mắc bệnh tiểu đường có thể được xác định như sau:

### **1.5.1. Đối Tượng Nghiên Cứu**

Đối tượng nghiên cứu của chúng tôi là nhóm bệnh nhân có nguy cơ cao phát triển bệnh tiểu đường, bao gồm những người có tiền sử gia đình mắc bệnh, những người thừa cân, ít hoạt động thể chất, hoặc những người có mức đường huyết cao nhưng chưa được chẩn đoán chính thức. Việc tập trung nghiên cứu nhóm đối tượng này nhằm giúp mô hình dự đoán nhận diện sớm và chính xác các trường hợp có nguy cơ cao.

### **1.5.2. Phạm Vi Nghiên Cứu**

**Thu thập và sử dụng dữ liệu**: Dữ liệu được thu thập từ nhiều nguồn khác nhau, bao gồm cơ sở y tế, bệnh viện, các nghiên cứu trước đó, hoặc các kho dữ liệu y tế khác. Dữ liệu này sẽ được sử dụng để huấn luyện và đánh giá các mô hình dự đoán.

**Xây dựng mô hình dự đoán**: Nghiên cứu sẽ tập trung vào việc phát triển mô hình dự đoán nguy cơ mắc bệnh tiểu đường dựa trên dữ liệu thu thập được. Các mô hình có thể áp dụng các thuật toán như hồi quy logistic, rừng ngẫu nhiên, KNN, mạng nơ-ron, và các thuật toán khác.

**Đánh giá và tinh chỉnh mô hình**: Hiệu suất của mô hình sẽ được đánh giá và các siêu tham số sẽ được tinh chỉnh để tối ưu hóa khả năng dự đoán.

**Ứng dụng trong y học thực tế**: Khám phá tiềm năng áp dụng của mô hình dự đoán trong thực tiễn y tế, bao gồm hỗ trợ ra quyết định lâm sàng và cung cấp thông tin hữu ích cho quá trình chẩn đoán và can thiệp y tế.

Nghiên cứu tập trung vào các cá nhân có nguy cơ mắc bệnh tiểu đường và phạm vi nghiên cứu bao gồm xử lý và phân tích dữ liệu y tế, xây dựng mô hình dự đoán nhằm cải thiện khả năng dự báo và ứng dụng thực tế trong y tế và chăm sóc sức khỏe.

# CHƯƠNG 2 ỨNG DỤNG VÀ THUẬT TOÁN

## 2.1 Giới thiệu bài toán.

Bệnh tiểu đường, đặc biệt là tiểu đường type 2, ngày càng trở nên phổ biến và nghiêm trọng trên toàn cầu. Với sự gia tăng nhanh chóng của số lượng bệnh nhân cùng với những biến chứng nguy hiểm như bệnh tim mạch, suy thận và tổn thương thần kinh, tiểu đường không chỉ ảnh hưởng đến sức khỏe cá nhân mà còn tạo ra gánh nặng lớn cho hệ thống y tế. Mặc dù đã có các phương pháp điều trị và quản lý bệnh, việc phát hiện sớm vẫn đóng vai trò thiết yếu trong việc kiểm soát và giảm thiểu biến chứng.

Trong bối cảnh này, việc dự đoán nguy cơ mắc bệnh tiểu đường trở nên cấp bách. Mục tiêu là xây dựng mô hình dự đoán có khả năng nhận diện sớm các trường hợp có nguy cơ cao, dựa trên dữ liệu y tế phong phú. Bài toán đòi hỏi không chỉ xem xét các chỉ số thông thường như đường huyết hay chỉ số khối cơ thể (BMI), mà còn khai thác dữ liệu phức tạp như thông tin di truyền và thói quen sinh hoạt.

Machine Learning nổi bật lên như một công cụ mạnh mẽ để giải quyết bài toán này. Với khả năng xử lý và phân tích dữ liệu lớn, chúng ta có thể phát hiện ra những mối liên hệ tiềm ẩn mà các phương pháp truyền thống khó nhận diện được. Việc áp dụng các thuật toán tiên tiến sẽ giúp xây dựng mô hình dự đoán chính xác, phát hiện sớm bệnh tiểu đường và đề xuất các biện pháp can thiệp kịp thời, không chỉ cải thiện chất lượng cuộc sống cho người bệnh mà còn giảm áp lực lên hệ thống y tế.

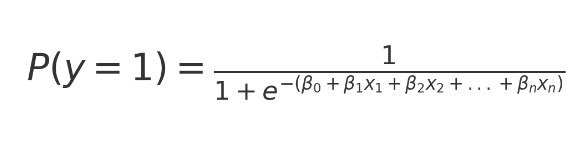
Bài toán dự đoán bệnh tiểu đường không chỉ mang giá trị học thuật mà còn có tiềm năng ứng dụng lớn trong thực tiễn. Việc phát triển các công cụ hỗ trợ bác sĩ ra quyết định và cung cấp các giải pháp chăm sóc sức khỏe cá nhân hóa có thể góp phần quan trọng trong việc nâng cao hiệu quả điều trị và quản lý bệnh tiểu đường trong tương lai.

## 2.2 Mô tả thuật toán.

### **2.2.1. Thuật toán Logistic Regression**

Logistic Regression được sử dụng để ước tính xác suất một sự kiện xảy ra, chẳng hạn như nguy cơ mắc bệnh tiểu đường. Dựa trên các yếu tố như tuổi tác, huyết áp, cholesterol, và tiền sử bệnh, thuật toán này giúp dự đoán khả năng một người mắc bệnh. Logistic Regression cũng cung cấp thông tin về mức độ ảnh hưởng của từng yếu tố nguy cơ đến kết quả dự đoán. Ngoài ra, thuật toán này có thể điều chỉnh ngưỡng quyết định để phù hợp với yêu cầu cụ thể của bài toán, tối ưu hóa độ chính xác, độ nhạy, hoặc độ đặc hiệu trong dự đoán.

Cụ thể, công thức của Logistic Regression có dạng như sau:



Trong đó:

* P(y=1) là xác suất một người mắc bệnh tiểu đường.
* β0,β1,...,βn​ là các tham số của mô hình mà Logistic Regression học được từ dữ liệu huấn luyện.
* x1,x2,...,xn​ là các biến số đầu vào đại diện cho các yếu tố nguy cơ như chỉ số đường huyết, BMI, v.v

### **2.2.2. Thuật toán Random Forest**

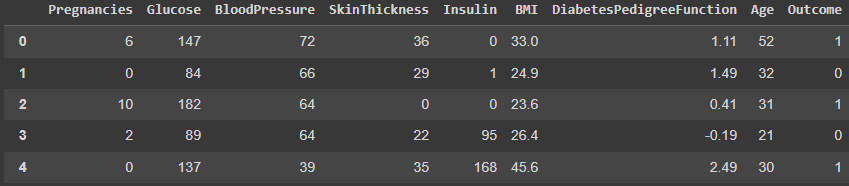
Random Forest là một phương pháp học máy kết hợp nhiều cây quyết định để cải thiện độ chính xác trong dự đoán. Thuật toán này hoạt động hiệu quả với dữ liệu lớn và có khả năng giảm thiểu hiện tượng quá khớp. Random Forest cũng cho phép đánh giá tầm quan trọng của các yếu tố đầu vào, giúp xác định yếu tố nguy cơ nào ảnh hưởng lớn nhất đến dự đoán bệnh tiểu đường. Bên cạnh đó, các kỹ thuật như đánh giá ngoài túi (Out-of-Bag Evaluation) và điều chỉnh siêu tham số giúp tối ưu hóa hiệu suất mô hình.

### **2.2.3. Thuật toán K-Nearest Neighbors**

KNN phân loại các cá nhân dựa trên các điểm dữ liệu gần nhất trong không gian đặc trưng. Trong dự đoán nguy cơ mắc bệnh tiểu đường, KNN xác định các láng giềng gần nhất và dựa vào đó để quyết định kết quả dự đoán. Tuy không trực tiếp cung cấp xác suất, nhưng có thể ước tính xác suất dựa trên tỷ lệ các láng giềng thuộc mỗi lớp. Việc chọn số láng giềng (K) là yếu tố quan trọng trong hiệu suất của mô hình, cần cân nhắc để tránh hiện tượng quá khớp hoặc chưa đủ khớp.

## 2.3. Xây dựng bộ dữ liệu.

### **2.3.1. Bộ dữ liệu gồm các cột**

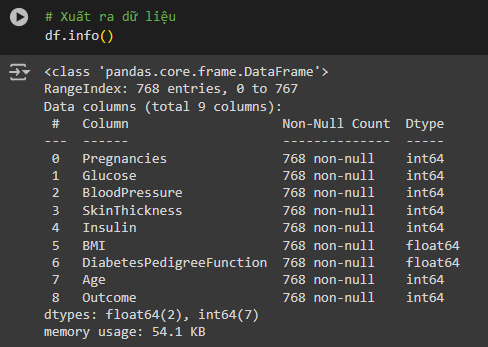


Hình 1: Bảng dữ liệu

1. Pregnancies: Số lần mang thai
2. Glucose: Nồng độ glucose trong máu
3. BloodPressure: Chỉ số huyết áp
4. SkinThickness: Độ dày lớp da
5. Insulin: Nồng độ insulin trong máu
6. BMI (Body Mass Index): Chỉ số khối cơ thể
7. DiabetesPedigreeFunction: Chỉ số nguy cơ tiểu đường theo di truyền
8. Age: Tuổi tác
9. Outcome: Kết quả bệnh tiểu đường

* 0: Không mắc bệnh tiểu đường
* 1: Đã được chẩn đoán mắc bệnh tiểu đường.

### **2.3.2. Tiền xử lý dữ liệu**



Hình 2: Bảng thông tin dữ liệu

## 2.4. Áp dụng thuật toán vào bài toán.

Trong đề tài về “dự báo khả năng mắc bệnh tiểu đường,” nhiều thuật toán machine learning và deep learning có thể được sử dụng để phát triển mô hình dự đoán hiệu quả. Dưới đây là một số thuật toán mà tôi đã áp dụng trong dự án này:

**Logistic Regression:** Đây là một thuật toán phân loại phổ biến, cho phép dự đoán các kết quả nhị phân dựa trên các biến đầu vào. Với đề tài này, hồi quy logistic có thể giúp ước tính xác suất phát sinh bệnh tiểu đường dựa trên các yếu tố nguy cơ cụ thể.

**Decision Trees và Random Forest:** Các thuật toán này nổi bật với khả năng phân loại dữ liệu linh hoạt, đặc biệt hiệu quả khi xử lý dữ liệu phi tuyến tính và nhiều chiều. Random Forest cũng giúp giảm thiểu tình trạng overfitting nhờ vào việc kết hợp nhiều cây quyết định khác nhau.

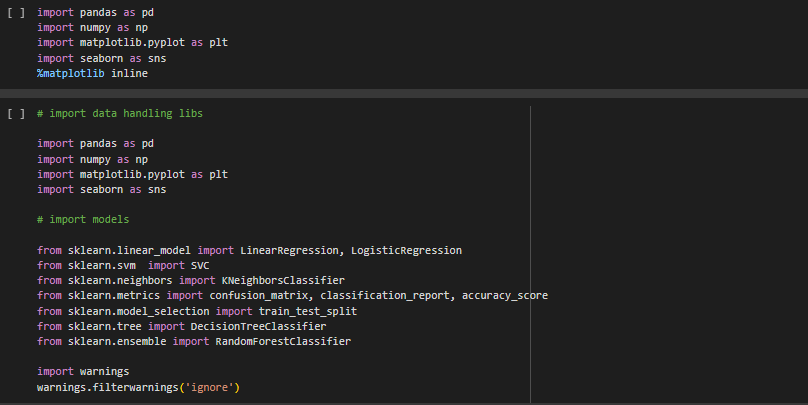
**Support Vector Machines (SVM):** SVM là một phương pháp mạnh mẽ để phân loại dữ liệu, tạo ra các đường ranh giới phức tạp hơn, giúp phân biệt rõ ràng giữa các nhóm nguy cơ mắc bệnh.

1. **nearest neighbors (KNN):** Dựa trên khoảng cách giữa các điểm dữ liệu trong không gian đặc trưng, KNN có thể xác định những trường hợp có nguy cơ mắc bệnh cao thông qua việc so sánh với các dữ liệu tương tự.

**Mạng nơ-ron (Neural Networks) và Học sâu (Deep Learning):** Đây là những phương pháp rất mạnh mẽ trong việc học từ các bộ dữ liệu phức tạp, đặc biệt là khi dữ liệu có nhiều yếu tố phi tuyến tính. Mạng nơ-ron sâu có khả năng tự động trích xuất các đặc trưng ẩn từ dữ liệu và tối ưu hóa quá trình dự báo. Với khả năng học hỏi từ dữ liệu lớn và đa chiều, học sâu tỏ ra rất phù hợp trong việc dự đoán nguy cơ mắc bệnh tiểu đường.

## 2.5. Thực nghiệm với thư viện.

Chắc hẳn bất kỳ ai học lập trình đều quen thuộc với thuật ngữ "Modules" – đề cập đến các thư viện lập trình chứa những đoạn mã thường xuyên được sử dụng, được các lập trình viên phát triển. Để chuẩn bị cho đề tài "Dự đoán nguy cơ mắc bệnh tiểu đường," việc sử dụng các thư viện lập trình là điều không thể thiếu. Sau đây, em xin trình bày một số thư viện đã được sử dụng trong quá trình thực hiện dự án này. Những thư viện này không chỉ hỗ trợ tối ưu hóa thời gian phát triển mà còn cung cấp các công cụ hữu ích cho việc xử lý dữ liệu và xây dựng mô hình dự đoán.



Hình 3: Các thư viện đã được sử dụng

**1.pandas (pd):**

Chúng ta bắt đầu với một thư viện rất quen thuộc với những người mới học Python. Đúng rồi, thư viện mà mình muốn nhắc đến chính là Pandas. Đây là một công cụ mạnh mẽ để xử lý và phân tích dữ liệu, cung cấp các cấu trúc dữ liệu như DataFrame và Series, giúp việc thao tác dữ liệu và trực quan hóa trở nên dễ dàng hơn.

**2.matplotlib.pyplot (plt):**

Thư viện tiếp theo mà em muốn giới thiệu là matplotlib, một công cụ hữu ích trong việc vẽ đồ thị trên Python.

Pyplot là một module của matplotlib, cung cấp các hàm tạo và hiển thị biểu đồ với các giao diện như Tkinter và wxPython.

**3.seaborn (sns):**

Seaborn là một thư viện trực quan hóa dữ liệu được xây dựng dựa trên matplotlib. Thư viện này cung cấp giao diện thân thiện, giúp dễ dàng tạo ra các biểu đồ thống kê với tính trực quan cao.

**4.numpy (np):**

Numpy là một thư viện quan trọng trong Python, chuyên dùng cho các phép tính số học. Nó cung cấp các đối tượng mảng và nhiều hàm toán học mạnh mẽ, giúp quá trình xử lý và thao tác dữ liệu trở nên nhanh chóng và hiệu quả hơn.

**5.scikit-learn**

Scikit-learn, hay còn gọi tắt là sklearn, là một trong những thư viện phổ biến nhất trong lĩnh vực machine learning và thống kê. Nó cung cấp các công cụ mạnh mẽ để xây dựng, đánh giá mô hình học máy, và được sử dụng rộng rãi trên các hệ điều hành như Linux nhờ vào giấy phép bản quyền FreeBSD.

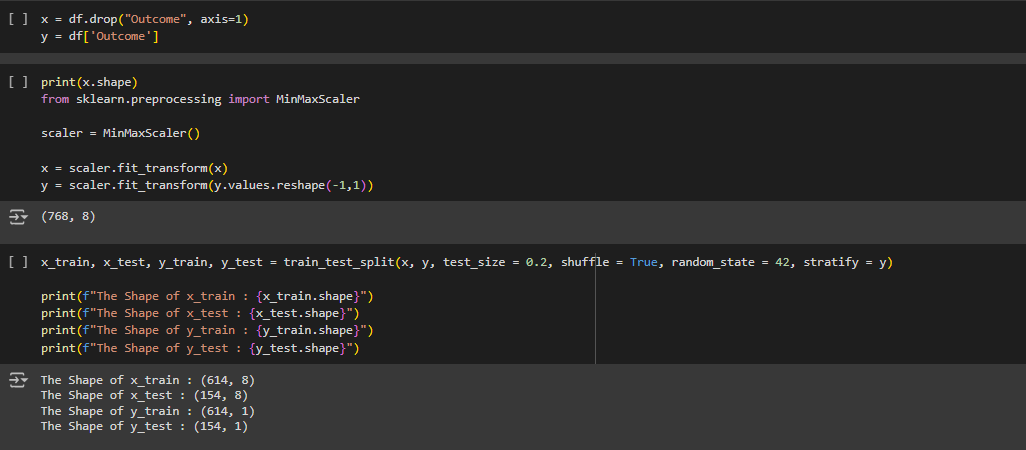
**6. plotly**

Plotly là một thư viện đồ họa tương tác, mã nguồn mở, giúp tạo ra các biểu đồ tương tác trên nền tảng trình duyệt. Thư viện này rất phù hợp cho việc trực quan hóa dữ liệu với khả năng tương tác cao.

# CHƯƠNG 3 XÂY DỰNG ỨNG DỤNG NGÔN NGỮ PYTHON

## 3.1 Xây dựng ứng dụng và giải thích

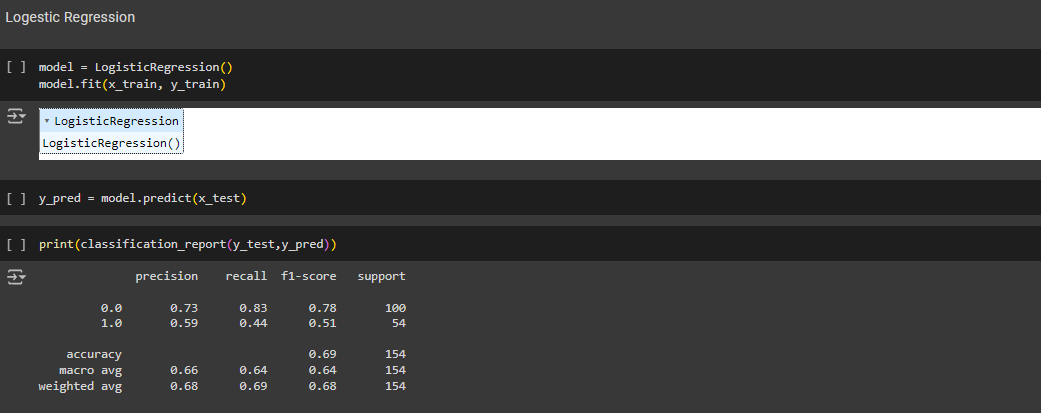
Trước khi đi vào các thuật toán dự đoán thì em xin phép giải thích một ít về phần dữ liệu mà em sẽ dự đoán như sau:



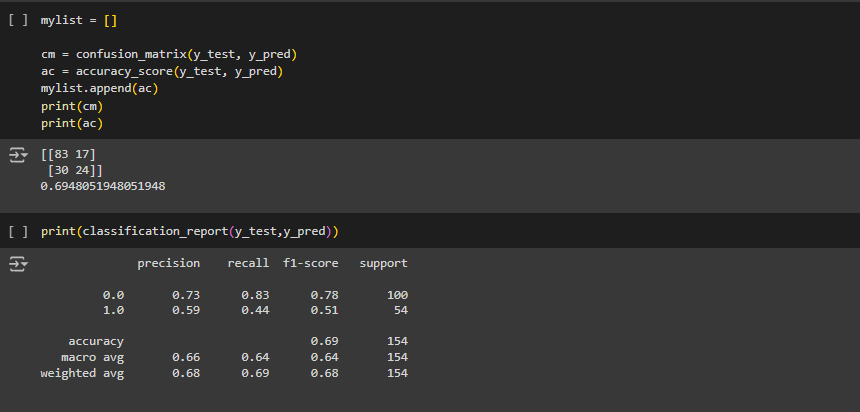
Hình 4: Dữ liệu chuẩn bị cho phần học máy

### **3.1.1. Logistic Regression**

Thuật toán Logistic Regression là một công cụ mạnh mẽ trong cả thống kê và học máy, đặc biệt hữu ích cho các bài toán phân loại. Khi cần dự đoán lớp (class) của một đối tượng dựa trên các đặc điểm (features), Logistic Regression sẽ giúp chúng ta tìm ra mối quan hệ giữa các yếu tố đầu vào và kết quả dự đoán. Trong đề tài "Dự đoán nguy cơ mắc bệnh tiểu đường," thuật toán này có thể được sử dụng để xác định những cá nhân có khả năng phát triển bệnh tiểu đường, dựa trên các yếu tố nguy cơ như tiền sử gia đình, lối sống, và các chỉ số y tế khác.



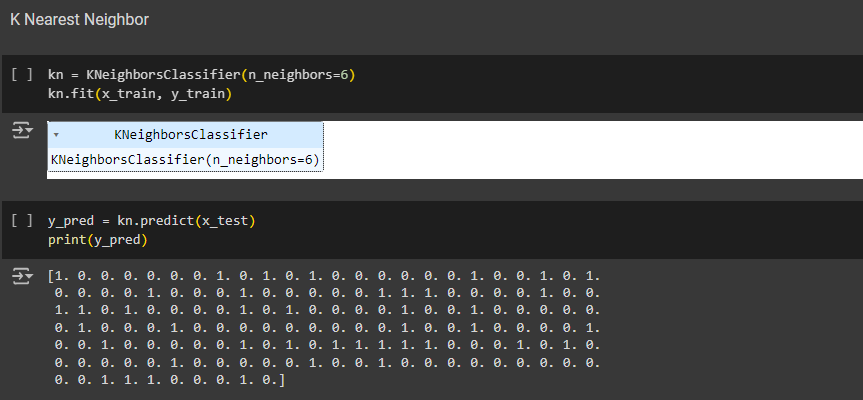
Hình 5: Khởi tạo mô hình Logistic Regression



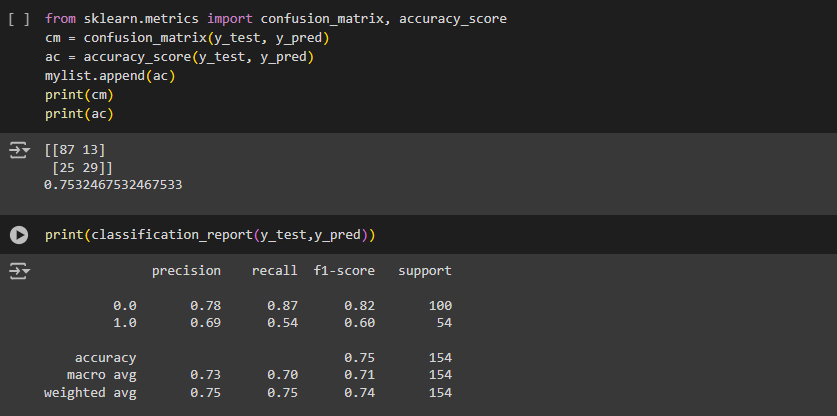
Hình 6: Kết quả mô hình Logistic Regreession

Như mọi người có thể thấy được ở phần kết quả của bài dự đoán này thì dự đoán của mô hình Logistic Regreesion chiếm 0.69

### **3.1.2. K-Nearest Neighbors**



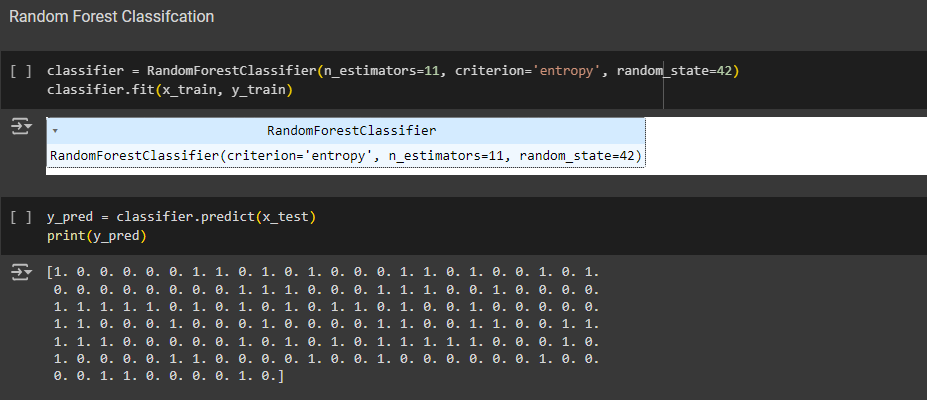
Hình 7: Khởi tạo K-Nearest Neighbors



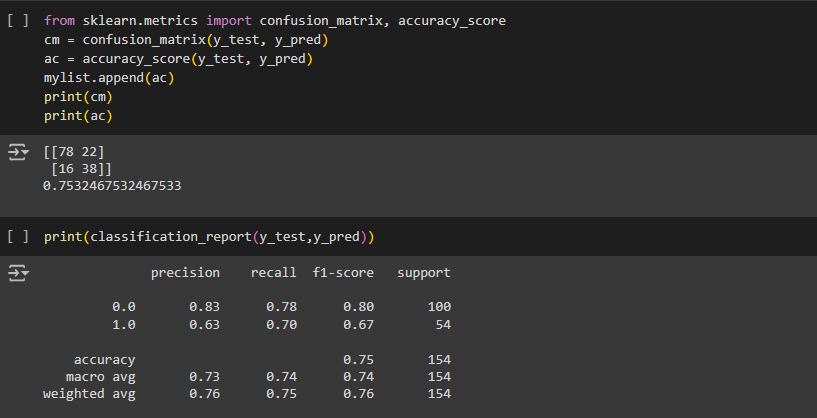
Hình 8: Kết quả mô hình K-Nearest Neighbors

Như mọi người có thể thấy thì theo dự đoán K-Nearest Neighbors chiếm 0.75

### **3.1.3. Random Forest**



Hình 9: Khởi tạo Random Forest

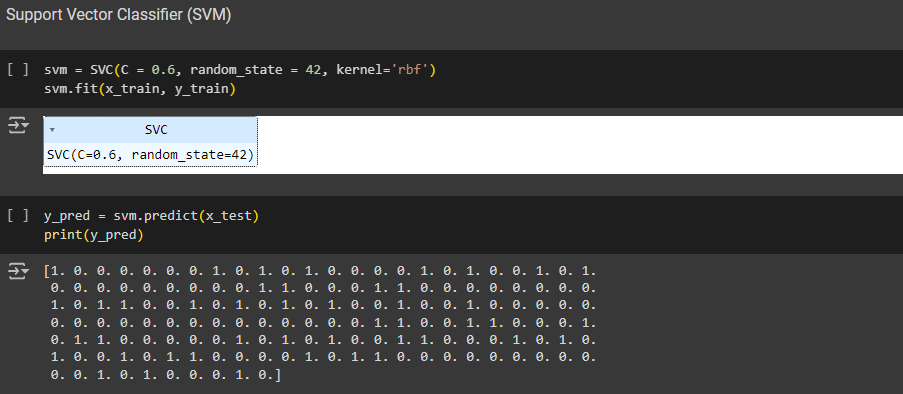


Hình 10: Kết quả mô hình Random Forest

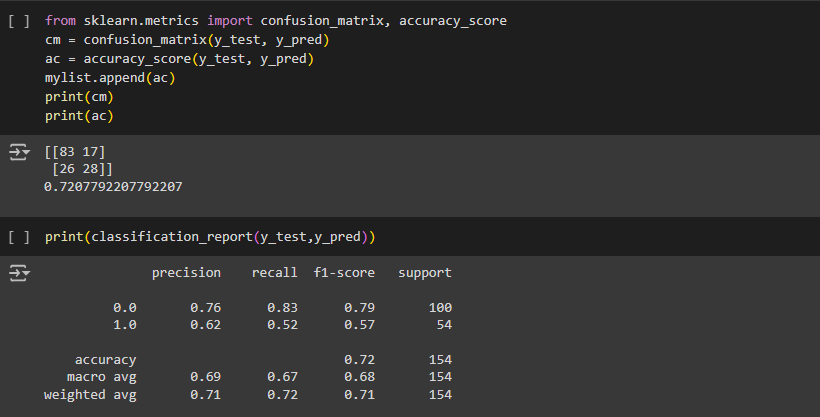
Kết quả của dự đoán này cũng không có sự trên lệch quá nhiều so với dự đoán Random Forest là mấy cũng chiếm 0.75

### **3.1.4. Support Vector Classifier**

Support Vector Classifier (SVC) tìm kiếm mặt phẳng phân cách tối ưu giữa các nhóm dữ liệu bằng cách tối đa hóa khoảng cách (margin) giữa mặt phân cách và các điểm dữ liệu gần nhất, giúp phân loại an toàn và chính xác hơn.



Hình 11: Khởi tạo Support Vector Classifier

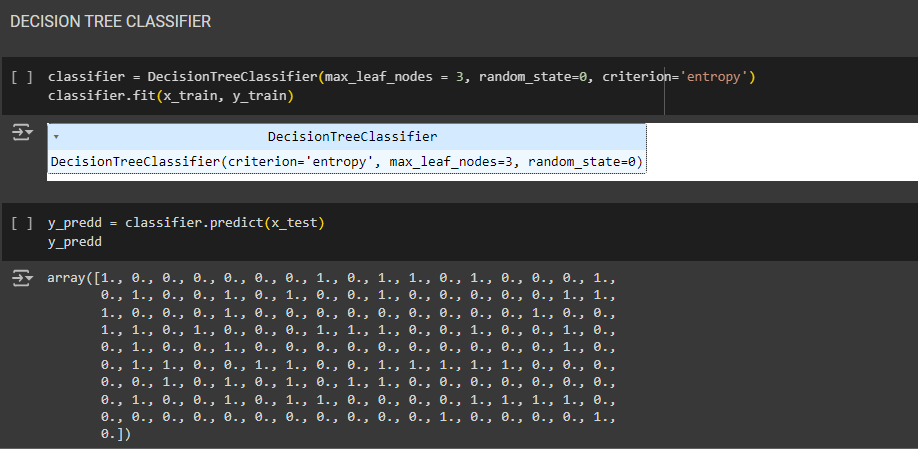


Hình 12: Kết quả mô hình Support Vector Classifier

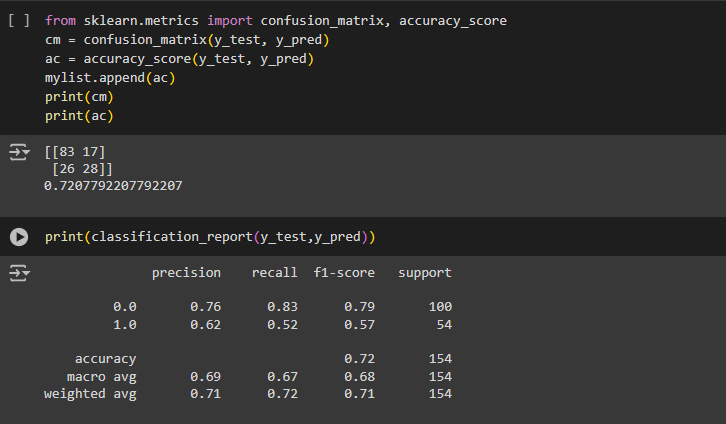
Kết quả của dự đoán này cũng không có sự trên lệch quá nhiều so với dự đoán Support Vector Classifier là mấy cũng chiếm 0.72

### **3.1.5. Decision Tree Classifier**

Với việc sử dụng thuật toán Decision Tree Classifier, em có thể phân tích các đặc trưng này để dự đoán xem một người có mắc bệnh tiểu đường hay không. Thuật toán này tỏ ra đặc biệt hữu ích trong việc phát hiện ra các mẫu và mối quan hệ phức tạp giữa các yếu tố, từ đó đưa ra các dự đoán chính xác về khả năng mắc bệnh của các bệnh nhân dựa trên các yếu tố như đã liệt kê.



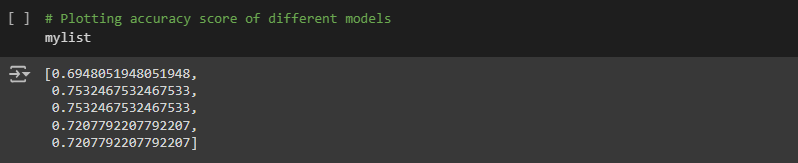
Hình 13: Khởi tạo mô hình Decision Tree Classifier



Hình 14: Kết quả mô hình Decision Tree Classifier

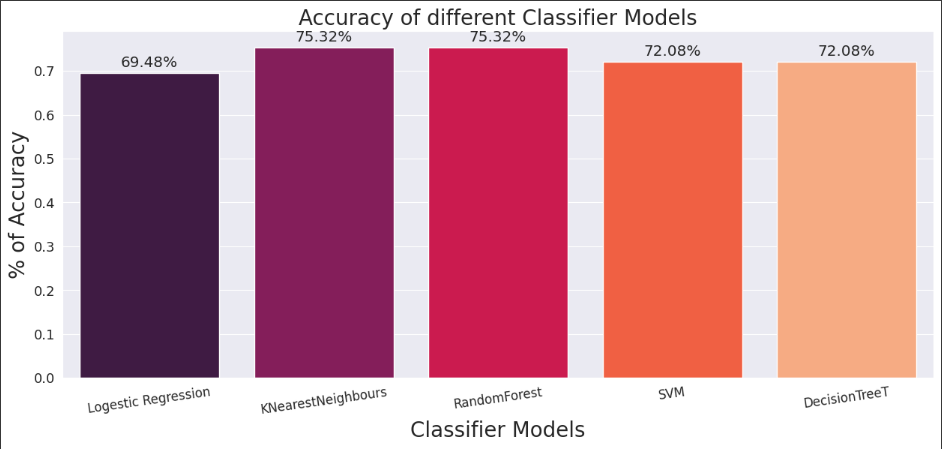
Kết quả của dự đoán này cũng không có sự trên lệch quá nhiều so với dự đoán Decision Tree Classifier là mấy cũng chiếm 0.72

**3.1.6. So sanh các phương pháp dự đoán**



Hình 15: Xuất ra phần trăm sai lệch của các dự báo

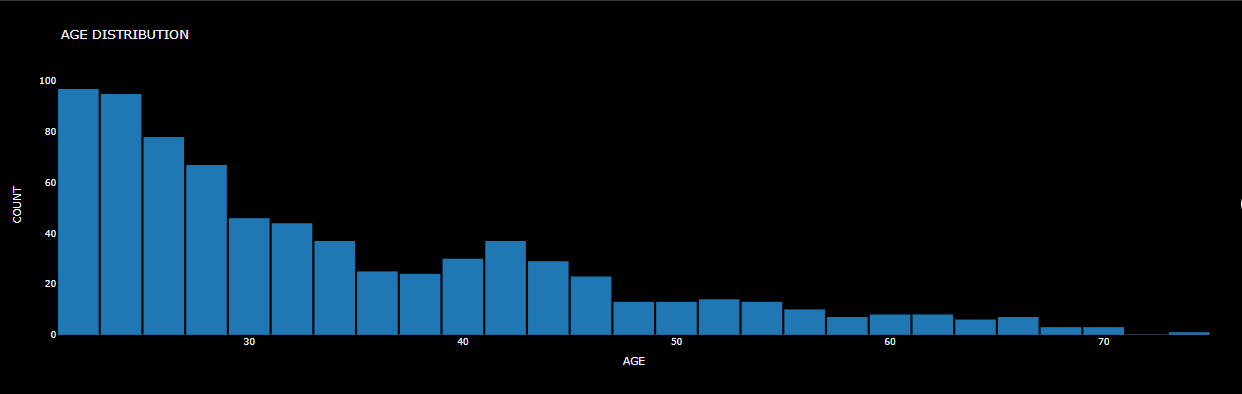
Từ bảy phương pháp dự đoán bao gồm: Logistic Regression, K-Nearest Neighbors, Random Forest, Support Vector Classifier, Decision Tree Classifier. Ta có thể thấy được rằng mô hình Logistic Regression là mô hình có độ dự báo sai lệch thấp nhất



Hình 16: Trực quan hóa của phần trăm sai lệch của các dự báo

**3.2. Một số ảnh trực phân hóa được phân tích trong dự án**

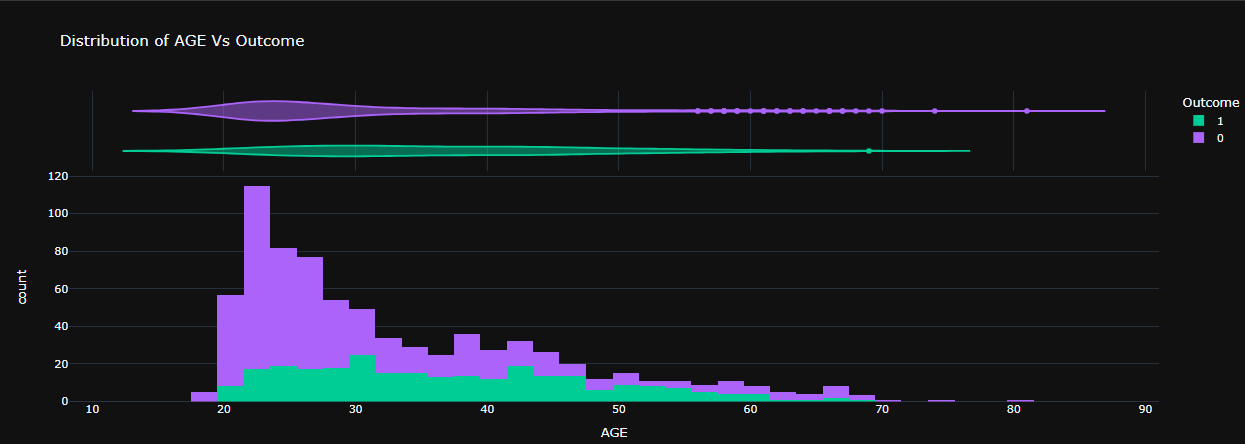
**3.2.1. Trực quan theo độ tuổi**



Hình 17: Trực quan hóa về tuổi

Như bạn có thể thấy, phần lớn các bệnh nhân nằm trong độ tuổi từ 20 đến 30, với số lượng giảm dần theo độ tuổi. Đặc biệt, sau tuổi 50, số lượng bệnh nhân giảm đáng kể, và số lượng rất ít ở độ tuổi từ 70 trở lên

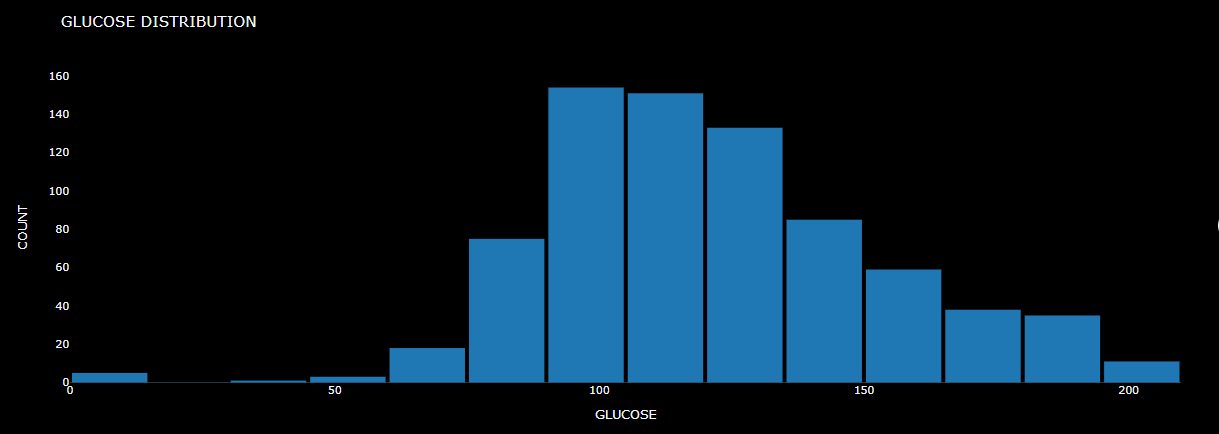
**3.2.2. Trực quan theo khả năng mắc bệnh tiểu đường ở các độ tuổi**



Hình 18: Trực quan hóa về tuổi ảnh hưởng tới bệnh tiểu đường

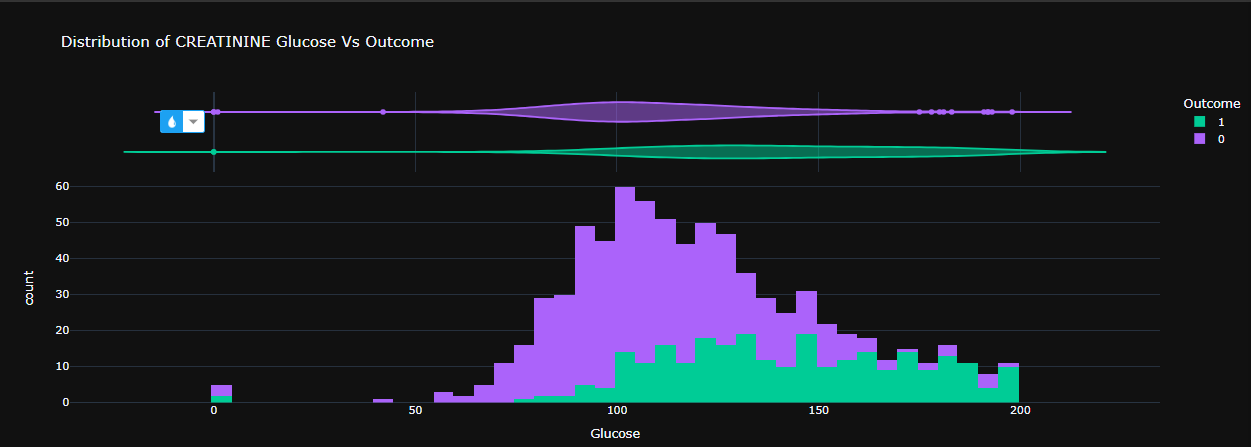
Nhìn vào dữ liệu đã được trực quan hóa, câu nói của người xưa "Qua tứ tuần đối với người trẻ thì đã già, qua ngũ tuần thì đối với người già vẫn còn trẻ, qua lục tuần là bước vào giai đoạn khó khăn nhất của một đời người" vẫn có sự liên hệ nhất định. Điều này thể hiện rõ trong phân bố độ tuổi của bệnh nhân mắc bệnh tiểu đường, khi số người mắc bệnh ở độ tuổi trung niên (khoảng 20-40 tuổi) là phổ biến nhất, nhưng tỷ lệ mắc bệnh dần giảm khi tuổi tác tăng cao.

**3.2.3. Trực quan hóa về Mức đường huyết**



Hình 19: Trực quan hóa về nồng độ enzyme trong máu

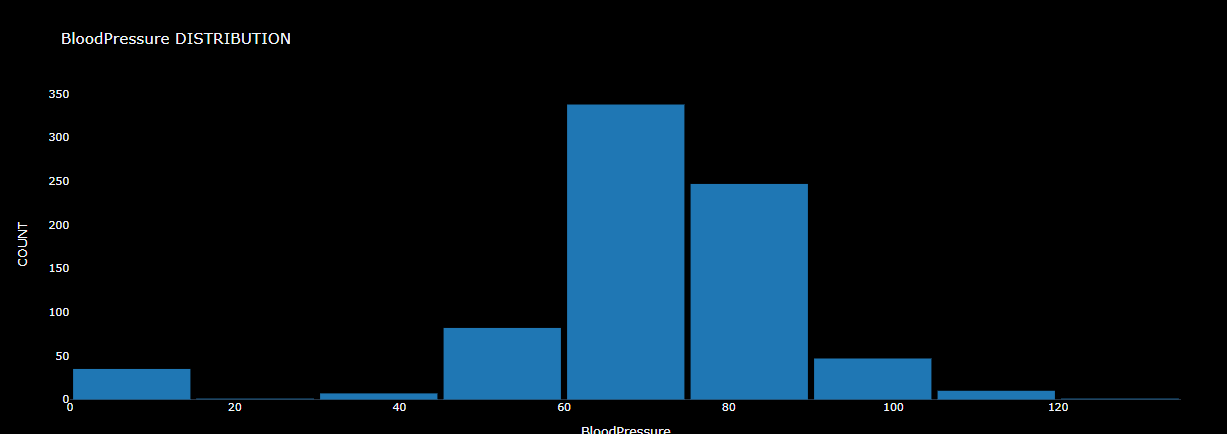
**3.2.4. Trực quan hóa về Mức đường huyết ảnh hưởng tới bệnh tiểu đường**



Hình 20: Trực quan hóa về Mức đường huyết ảnh hưởng tới bệnh tiểu đường

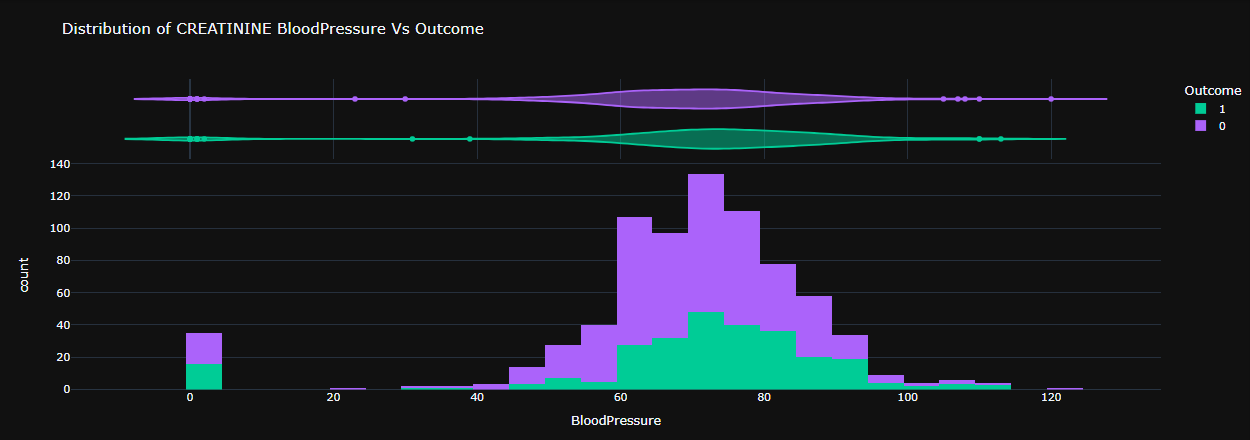
Dựa vào biểu đồ trực quan hóa, ta có thể nhận thấy mối liên hệ rõ ràng giữa nồng độ glucose trong máu và khả năng mắc bệnh tiểu đường. Những người có mức glucose thấp thường ít mắc bệnh, trong khi tỷ lệ mắc bệnh tăng dần ở những người có mức glucose cao. Điều này chứng tỏ rằng lượng glucose trong máu đóng vai trò quan trọng trong việc xác định nguy cơ tiểu đường.

**3.2.5. Trực quan hóa về Huyết áp**



Hình 21: Trực quan hóa về Huyết áp

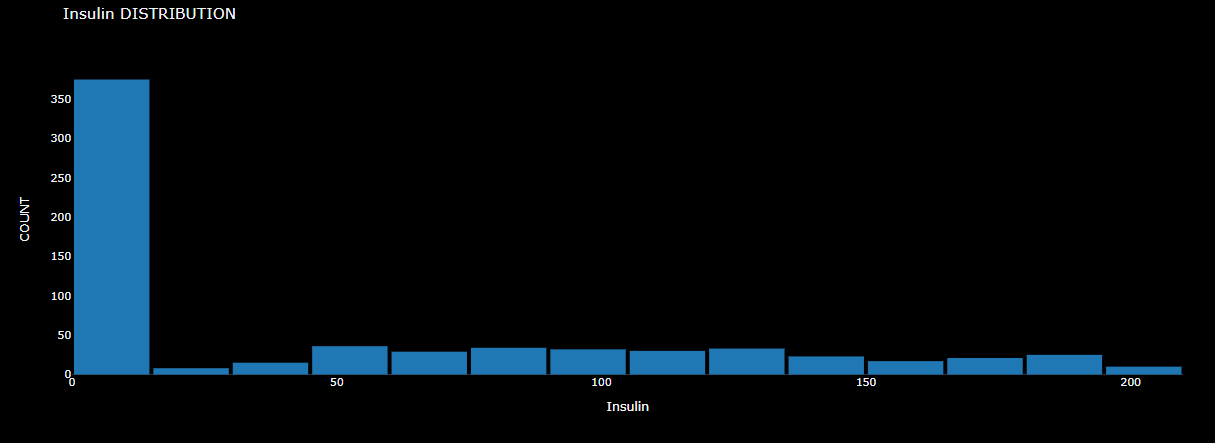
**3.2.6. Trực quan hóa về Huyết áp ảnh hưởng tới bệnh tiểu đường**



Hình 22: Trực quan hóa về Huyết áp ảnh hưởng tới bệnh tiểu đường

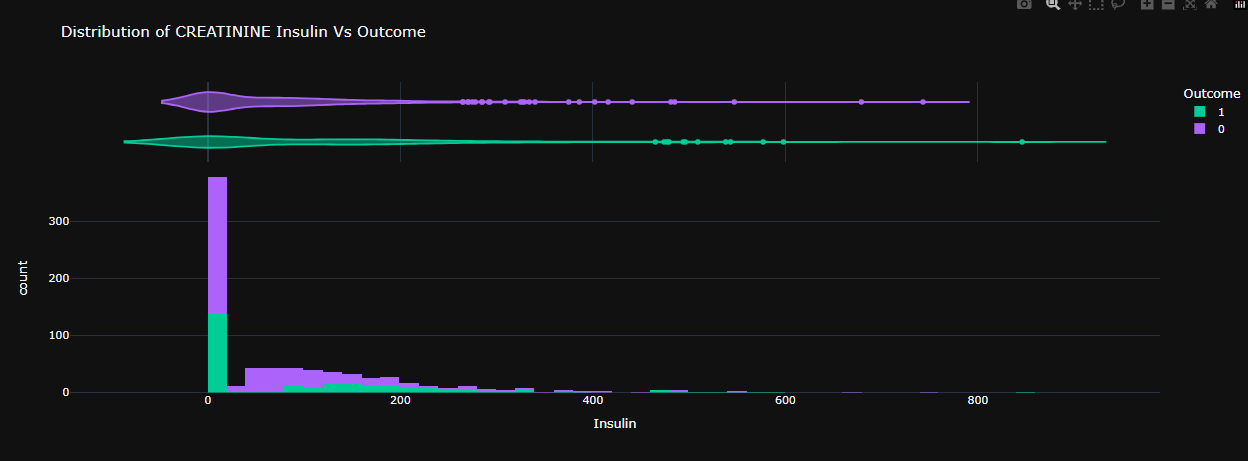
Tương tự, biểu đồ về huyết áp cũng cho thấy sự liên quan với nguy cơ mắc bệnh. Nhóm người có huyết áp nằm trong khoảng 60 đến 80 có tỷ lệ mắc bệnh tiểu đường cao hơn hẳn. Ngược lại, những người có huyết áp rất thấp hoặc rất cao lại có xu hướng ít mắc bệnh hơn. Điều này gợi ý rằng huyết áp cũng là một yếu tố đáng quan tâm trong việc đánh giá rủi ro tiểu đường.

**3.2.7. Trực quan hóa về Nồng độ Insulin**



Hình 23: Trực quan hóa về Nồng độ insulin huyết thanh

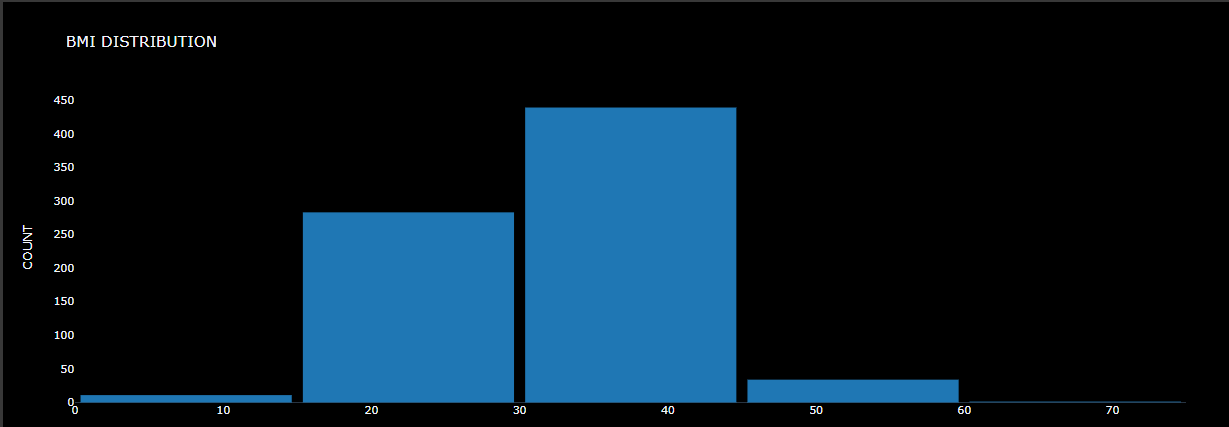
**3.2.8. Trực quan hóa về Nồng độ insulin ảnh hưởng tới bệnh tiểu đường**



Hình 24: Trực quan hóa về Nồng độ insulin ảnh hưởng tới bệnh tiểu đường

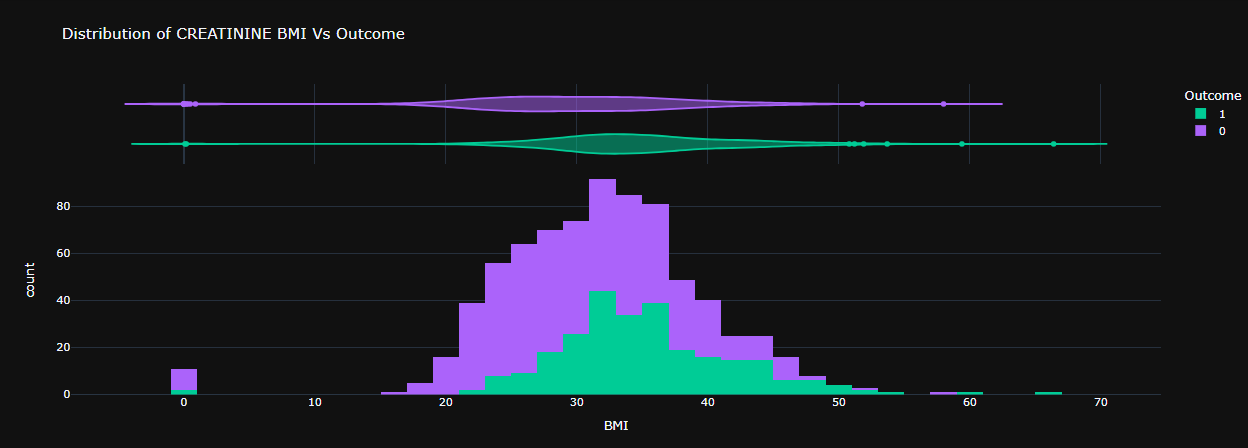
Biểu đồ về insulin cho thấy sự phân bố không đều giữa các bệnh nhân. Đáng chú ý là những bệnh nhân có mức insulin rất thấp thường nằm trong nhóm có nguy cơ mắc bệnh cao hơn. Điều này ám chỉ rằng nồng độ insulin có thể là một yếu tố then chốt để xác định khả năng mắc bệnh tiểu đường.

**3.2.9. Trực quan hóa về Chỉ số khối cơ thể(BMI)**



Hình 25: Trực quan hóa về Chỉ số khối cơ thể(BMI)

**3.2.10. Trực quan hóa về Chỉ số khối cơ thể(BMI) ảnh hưởng tới bệnh tiểu đường**



Hình 26:Trực quan hóa về Chỉ số khối cơ thể(BMI) ảnh hưởng tới bệnh tiểu đường

Cuối cùng, khi xem xét chỉ số BMI, biểu đồ cho thấy chỉ số này có ảnh hưởng lớn đến khả năng mắc bệnh. Đặc biệt, những người có BMI từ 30 đến 40 chiếm tỷ lệ cao trong nhóm mắc bệnh tiểu đường, nhấn mạnh mối quan hệ giữa cân nặng và nguy cơ phát triển bệnh tiểu đường.

# KẾT LUẬN

## Kết quả đạt được.

Dự án của chúng tôi đã đạt được những kết quả đáng kể trong việc sử dụng mô hình dự đoán để hỗ trợ các cơ sở y tế tăng độ chính xác và khả năng tiên đoán nguy cơ mắc bệnh tiểu đường. Nhờ phát hiện bệnh sớm, bệnh nhân có thể tiếp cận các phương pháp điều trị một cách nhanh chóng, từ đó giảm thiểu nguy cơ biến chứng về sau và nâng cao khả năng hồi phục.

Mô hình này không chỉ đóng góp vào lĩnh vực y tế mà còn thể hiện sự kết hợp hiệu quả giữa khoa học dữ liệu và công nghệ thông tin trong y học. Thông qua việc sử dụng dữ liệu từ bệnh nhân, chúng ta có thể phát triển các công cụ hỗ trợ y tế hiện đại, cung cấp thông tin giúp bệnh nhân nhận thức rõ hơn về tình trạng sức khỏe của mình. Dù các con số có thể khô khan, việc trình bày chúng một cách trực quan sẽ giúp bệnh nhân hiểu rõ hơn về tình hình sức khỏe hiện tại.

Bên cạnh đó, mô hình này còn có khả năng dự đoán xác suất mắc bệnh tiểu đường dựa trên các yếu tố cá nhân như tuổi tác, giới tính và các yếu tố khác. Điều này giúp bệnh nhân đưa ra các biện pháp phòng ngừa và điều trị kịp thời, đồng thời nâng cao nhận thức về mức độ nguy hiểm của bệnh tiểu đường trong cộng đồng.

Một điểm mạnh của mô hình là khả năng đánh giá tác động của các yếu tố nguy cơ đối với bệnh tiểu đường. Mô hình giúp xác định những yếu tố quan trọng nhất, cung cấp thông tin hữu ích cho việc phòng ngừa và can thiệp. Việc áp dụng mô hình vào thực tế có thể giúp nâng cao chất lượng chăm sóc sức khỏe cộng đồng thông qua việc phát hiện sớm các trường hợp có nguy cơ cao và từ đó giảm tỷ lệ mắc bệnh tiểu đường.

Kết quả từ dự án này hứa hẹn mang lại lợi ích lớn trong việc cải thiện dự đoán và quản lý bệnh tiểu đường, đồng thời góp phần nâng cao chất lượng cuộc sống cho bệnh nhân và cộng đồng.

## Hạn chế và hướng phát triển .

**Hạn chế**

Mặc dù dự án đã đạt được một số kết quả tích cực, nhưng cũng tồn tại một số hạn chế đáng kể. Do thời gian thực hiện dự án khá ngắn, chúng em không có đủ điều kiện để tự tổng hợp hoặc thu thập một bộ dữ liệu đủ lớn và chất lượng. Bộ dữ liệu sử dụng trong dự án còn thiếu sự đa dạng và không đủ phong phú, dẫn đến mô hình không thể tổng quát hóa tốt trên các trường hợp mới. Điều này làm giảm độ chính xác của dự đoán và khiến kết quả chưa đạt được sự hoàn thiện mong muốn.

**Hướng phát triển**

Trong tương lai, nếu có thêm thời gian và nguồn lực, chúng em dự định mở rộng và cải thiện chất lượng dữ liệu. Điều này có thể thực hiện bằng cách nghiên cứu và tích hợp thêm dữ liệu từ nhiều nguồn khác nhau, đảm bảo sự phong phú và đa dạng hơn, từ đó cải thiện khả năng dự đoán của mô hình.

Ngoài ra, chúng em cũng hướng đến việc tinh chỉnh mô hình dự đoán. Bằng cách điều chỉnh các tham số và thử nghiệm nhiều cấu hình khác nhau, chúng em sẽ tìm ra phiên bản mô hình tối ưu nhất, đồng thời tránh tình trạng overfitting hoặc underfitting. Những cải tiến này sẽ giúp nâng cao hiệu quả và tính chính xác của mô hình trong việc dự đoán bệnh tiểu đường.

Đó là những hạn chế và hướng phát triển mà nhóm đã suy nghĩ. Rất mong nhận được ý kiến đóng góp từ thầy cô để hoàn thiện dự án hơn nữa.

Chúng em xin chân thành cảm ơn!

# TÀI LIỆU THAM KHẢO

[1] DATA MINING: CONCEPTS AND TECHNIQUES 3RD EDITION

[2] Nguyễn Thanh Tuấn(2009), Deeplearning cơ bản The Legrand Orange Book Template by Mathias Legrand is used : <https://nttuan8.com/sach-deep-learning-co-ban/>

[3] Lecturer in deep learning at the Faculty of Information Technology