|  |
| --- |
| **TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**  **TRUNG TÂM ĐÀO TẠO SAU ĐẠI HỌC**  **---------------------------------------**  HỌC PHẦN: PHÂN TÍCH DỮ LIỆU  **PHÂN TÍCH DỰ ĐOÁN BỆNH TIỂU ĐƯỜNG**  **SỬ DỤNG MÔ HÌNH HỒI QUY LOGISTIC**  **CBHD : TS. Nguyễn Mạnh Cường**  **Học viên : Nguyễn Tiến Duy**  **Mã học viên : 2020601607**  Hà Nội – Năm 2024 |

LỜI CẢM ƠN

Trước tiên, tôi xin gửi lời cảm ơn chân thành đến TS. Nguyễn Mạnh Cường vì sự hướng dẫn tận tâm và những ý kiến quý báu trong suốt quá trình thực hiện tiểu luận này. Sự giúp đỡ của giảng viên không chỉ cung cấp cho tôi kiến thức sâu rộng mà còn khuyến khích tôi suy nghĩ một cách phân tích và sáng tạo hơn.

Tôi cũng xin cảm ơn các bạn đồng học đã chia sẻ những ý kiến và góp ý hữu ích, giúp tôi hoàn thiện và nâng cao chất lượng của bài viết. Sự hỗ trợ và sự trao đổi ý kiến từ các bạn đã đóng góp không nhỏ vào sự thành công của tiểu luận này.

Sản phẩm là kết quả tốt nhất trong khả năng của quá trình nghiên cứu. Rất mong nhận được sự đóng góp của thầy để sản phẩm có thể hoàn thiện hơn.

Tôi xin chân thành cảm ơn.

Học viên

Nguyễn Tiến Duy

MỤC LỤC

[DANH MỤC HÌNH ẢNH 1](#_Toc175011794)

[DANH MỤC BẢNG BIỂU 2](#_Toc175011795)

[MỞ ĐẦU 3](#_Toc175011796)

[CHƯƠNG 1. TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU 5](#_Toc175011797)

[1.1. Khái niệm phân tích dữ liệu 5](#_Toc175011798)

[1.2. Quy trình phân tích dữ liệu 5](#_Toc175011799)

[1.3. Một số kỹ thuật phân tích dữ liệu 6](#_Toc175011800)

[1.3.1. Phân tích mô tả 6](#_Toc175011801)

[1.3.2. Phân tích hồi quy 7](#_Toc175011802)

[1.3.3. Phân tích sự khác biệt 8](#_Toc175011803)

[1.3.4. Phân tích dữ liệu time series 8](#_Toc175011804)

[1.4. Một số công cụ phân tích dữ liệu 9](#_Toc175011805)

[1.4.1. Python 9](#_Toc175011806)

[1.4.2. Microsoft Excel 9](#_Toc175011807)

[1.4.3. OpenRefine 9](#_Toc175011808)

[1.4.4. Tableau 10](#_Toc175011809)

[1.4.5. WEKA 10](#_Toc175011810)

[CHƯƠNG 2. MÔ HÌNH HỒI QUY LOGISTIC 12](#_Toc175011811)

[2.1. Tổng quan về mô hình hồi quy logistic 12](#_Toc175011812)

[2.2. Ưu, nhược điểm của mô hình hồi quy logistic 14](#_Toc175011813)

[CHƯƠNG 3. THỰC NGHIỆM, ĐÁNH GIÁ VÀ XÂY DỰNG ỨNG DỤNG THỬ NGHIỆM 16](#_Toc175011814)

[3.1. Mô tả dữ liệu 16](#_Toc175011815)

[3.2. Xử lý giá trị khuyết 26](#_Toc175011816)

[3.3. Huấn luyện và đánh giá mô hình hồi quy Logistic 27](#_Toc175011817)

[3.4. Xây dựng ứng dụng dự đoán bệnh tiểu đường 29](#_Toc175011818)

[3.4.1. Công cụ Flutter 29](#_Toc175011819)

[3.4.2. Giao diện ứng dụng 30](#_Toc175011820)

[KẾT LUẬN 31](#_Toc175011821)

[TÀI LIỆU THAM KHẢO 33](#_Toc175011822)

DANH MỤC HÌNH ẢNH

[Hình 1.1 Biểu đồ phân tích mô tả 6](#_Toc175010616)

[Hình 2.1 Các giả định cơ bản để thực hiện hồi quy logistic 12](#_Toc175010617)

[Hình 3.1 Một phần của bộ dữ liệu 17](#_Toc175010618)

[Hình 3.2 Tổng quan bộ dữ liệu 17](#_Toc175010619)

[Hình 3.3 Thống kê về kiểu dữ liệu của các thuộc tính 18](#_Toc175010620)

[Hình 3.4 Tóm tắt bộ dữ liệu 18](#_Toc175010621)

[Hình 3.5 Dữ liệu trùng lặp trong bộ dữ liệu 21](#_Toc175010622)

[Hình 3.6 Mức độ tương quan giữa các thuộc tính 21](#_Toc175010623)

[Hình 3.7 Biểu đồ nhiệt thể hiện sự tương quan 22](#_Toc175010624)

[Hình 3.8 Ma trận biểu đồ phân tán của bộ dữ liệu 23](#_Toc175010625)

[Hình 3.9 Mối quan hệ giữa Glucose và bệnh tiểu đường 24](#_Toc175010626)

[Hình 3.10 Mối quan hệ giữa tuổi và bệnh tiểu đường 24](#_Toc175010627)

[Hình 3.11 Mối quan hệ giữa chỉ số BMI và bệnh tiểu đường 24](#_Toc175010628)

[Hình 3.12 Mối quan hệ giữa Insulin và bệnh tiểu đường 24](#_Toc175010629)

[Hình 3.13 Mối quan hệ giữa huyết áp tâm trương và bệnh tiểu đường 25](#_Toc175010630)

[Hình 3.14 Xử lý dữ liệu khuyết của bộ dữ liệu 26](#_Toc175010631)

[Hình 3.15 Tổng quan bộ dữ liệu sau khi xử lý 26](#_Toc175010632)

[Hình 3.16 Biểu đồ nhiệt của bộ dữ liệu sau khi xử lý giá trị khuyết 27](#_Toc175010633)

[Hình 3.17 Huấn luyện mô hình với bộ dữ liệu gốc 28](#_Toc175010634)

[Hình 3.18 Huấn luyện mô hình với bộ dữ liệu đã xử lý giá trị khuyết 28](#_Toc175010635)

[Hình 3.19 Giao diện ứng dụng 30](#_Toc175010636)

DANH MỤC BẢNG BIỂU

[Bảng 3.1 So sánh kết quả huấn luyện với hai bộ dữ liệu 28](#_Toc175010637)

MỞ ĐẦU

1. **Tên đề tài**

Phân tích dự đoán bệnh tiểu đường sử dụng mô hình hồi quy logistic.

1. **Lý do chọn đề tài**

Bệnh tiểu đường là một trong những bệnh lý phổ biến và có ảnh hưởng nghiêm trọng đến sức khỏe cộng đồng toàn cầu. Việc phát hiện sớm và dự đoán nguy cơ mắc bệnh tiểu đường đóng vai trò quan trọng trong việc giảm thiểu tác động của bệnh, cải thiện chất lượng cuộc sống và giảm chi phí y tế. Một trong những công cụ mạnh mẽ trong việc dự đoán bệnh tiểu đường là mô hình hồi quy logistic.

Mô hình hồi quy logistic là một phương pháp thống kê được sử dụng để phân tích mối quan hệ giữa biến phụ thuộc nhị phân và một hoặc nhiều biến độc lập. Trong trường hợp này, biến phụ thuộc nhị phân là sự hiện diện hay không của bệnh tiểu đường, trong khi các biến độc lập có thể bao gồm các yếu tố như tuổi tác, giới tính, chỉ số khối cơ thể (BMI), mức độ glucose trong máu và các chỉ số sinh học khác.

Với những lý do trên, việc lựa chọn mô hình hồi quy logistic để phân tích và dự đoán nguy cơ bệnh tiểu đường không chỉ mang lại lợi ích trong việc cải thiện khả năng dự đoán mà còn góp phần vào việc phát triển các phương pháp phòng ngừa hiệu quả.

1. **Mục tiêu của đề tài**

* Xây dựng mô hình dự đoán bệnh tiểu đường dựa trên bộ dữ liệu thu thập được.
* Xây dựng ứng dụng thử nghiệm đánh giá khả năng mắc bệnh tiểu đường thực tế

1. **Phương pháp nghiên cứu**

* Phương pháp nghiên cứu lý thuyết.
* Phương pháp thu thập số liệu.
* Phương pháp toán học.
* Phương pháp thực nghiệm.

1. **Đối tượng nghiên cứu**

* Ngôn ngữ lập trình python
* Mô hình hồi quy Logistic
* Bộ dữ liệu bao gồm 768 mẫu thu thập trên các bệnh nhân là phụ nữ từ 21 tuổi trở lên là người gốc da đỏ Pima.
* Framework Flutter

# TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU

## Khái niệm phân tích dữ liệu

Phân tích dữ liệu là quá trình kiểm tra, làm sạch, chuyển đổi và mô hình hóa dữ liệu với mục tiêu khám phá thông tin hữu ích, đưa ra kết luận và hỗ trợ việc ra quyết định.

Phân tích dữ liệu có nhiều khía cạnh và cách tiếp cận, bao gồm các kỹ thuật đa dạng dưới nhiều tên khác nhau và được sử dụng trong các lĩnh vực kinh doanh, khoa học và khoa học xã hội khác nhau. Trong thế giới kinh doanh ngày nay, phân tích dữ liệu đóng vai trò giúp đưa ra quyết định khoa học hơn và giúp doanh nghiệp hoạt động hiệu quả hơn.[1]

## Quy trình phân tích dữ liệu

A yellow and white sign with black text

Description automatically generated

Hình 1.1 Quy trình phân tích dữ liệu

Quy trình phân tích dữ liệu thường bao gồm các bước chính sau:

1. **Xác định mục tiêu**: Trước khi bắt tay vào phân tích, cần phải rõ ràng về mục tiêu và vấn đề cần giải quyết. Điều này giúp xác định các dữ liệu cần thu thập và phương pháp phân tích phù hợp.
2. **Thu thập dữ liệu**: Dữ liệu có thể được thu thập từ nhiều nguồn khác nhau, như cơ sở dữ liệu nội bộ, khảo sát, hoặc dữ liệu công khai trên Internet. Việc thu thập dữ liệu cần được thực hiện một cách chính xác để đảm bảo chất lượng và tính đại diện của dữ liệu.
3. **Tiền xử lý dữ liệu**: Đây là bước quan trọng để làm sạch và chuẩn bị dữ liệu cho phân tích. Tiền xử lý bao gồm các hoạt động như loại bỏ dữ liệu bị thiếu, xử lý các giá trị bất thường, và chuẩn hóa dữ liệu.
4. **Phân tích dữ liệu**: Sau khi dữ liệu đã được chuẩn bị, bước tiếp theo là phân tích để tìm ra các mẫu, mối liên hệ, và thông tin hữu ích. Phân tích có thể bao gồm việc sử dụng các kỹ thuật thống kê, mô hình hóa dữ liệu, hoặc máy học để đưa ra các kết luận.
5. **Kết luận và dự đoán**: Dựa trên kết quả phân tích, đưa ra các kết luận và dự đoán về vấn đề đã đặt ra. Kết luận cần phải được trình bày rõ ràng và có thể kèm theo các khuyến nghị hoặc hành động cần thực hiện dựa trên kết quả phân tích.

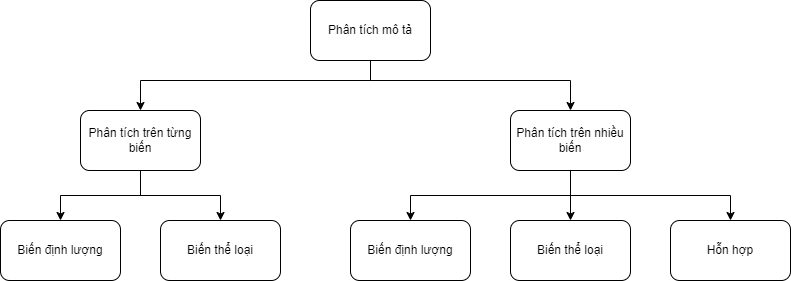
Các bước này có thể lặp lại nhiều lần trong quá trình phân tích để điều chỉnh và cải thiện kết quả.

## Một số kỹ thuật phân tích dữ liệu

### Phân tích mô tả

Phân tích mô tả là kỹ thuật giúp chuyển đổi dữ liệu để dễ hiểu, dễ diễn giải, sắp đặt, sắp xếp lại và thao tác trên đó để tạo ra thông tin mô tả.

Tùy theo dữ liệu, công việc mà ta lựa chọn công việc phù hợp cho phân tích mô tả.



Hình 1.2 Biểu đồ phân tích mô tả

* Phân tích trên 1 biến (1 thuộc tính): giúp xác định các giá trị ngoại lai hoặc các giá trị bất thường để có thể kiểm tra lỗi nhập dữ liệu.
* **Dữ liệu số:** Vẽ biểu đồ Histogram, Box & Whisker, Tính các đại lượng thống kê (mean, stdev, median, quartile...)
* **Dữ liệu phi số:** tạo bảng tần suất (frequency table), vẽ biểu đồ thanh, biểu đồ hình tròn hoặc biểu đồ doughnut
* Phân tích trên nhiều biến: tìm ra thông tin sơ bộ về mối liên hệ giữa các biến
* **Dữ liệu số:** tạo biểu đồ scatter
* **Dữ liệu phi số:** tạo các bảng dữ liệu thống kê
* **Dữ liệu hỗn hợp:** tạo bảng thống kê tổng hợp, vẽ đồ thị box-and-whisker của dữ liệu định lượng trên từng thể loại.

### Phân tích hồi quy

Phân tích hồi quy là một phương pháp thống kê được sử dụng để xác định mối quan hệ giữa một biến phụ thuộc (biến đáp ứng) và một hoặc nhiều biến độc lập (biến dự đoán). Mục tiêu chính của phân tích hồi quy là xây dựng một mô hình toán học có thể dự đoán giá trị của biến phụ thuộc dựa trên các biến độc lập.

Ưu điểm:

* Cung cấp cái nhìn sâu sắc về mối quan hệ giữa các biến.
* Dễ dàng dự đoán và đưa ra các quyết định dựa trên mô hình hồi quy.
* Giúp nhận diện và loại bỏ những yếu tố không quan trọng hoặc gây nhiễu.

Nhược điểm:

* Giả định rằng mối quan hệ giữa các biến là tuyến tính có thể không đúng trong mọi trường hợp.
* Mô hình hồi quy có thể bị ảnh hưởng bởi các điểm ngoại lai và đa cộng tuyến giữa các biến độc lập.
* Yêu cầu kích thước mẫu đủ lớn để có thể ước lượng chính xác các tham số.

### Phân tích sự khác biệt

Phân tích sự khác biệt là một phương pháp thống kê được sử dụng để phân loại các đối tượng hoặc quan sát vào các nhóm khác nhau dựa trên các đặc điểm hoặc biến số dự đoán của chúng. Mục tiêu chính của phân tích sự khác biệt là tìm ra những hàm phân biệt có khả năng tối đa hóa sự khác biệt giữa các nhóm, giúp phân loại chính xác các đối tượng vào đúng nhóm.

Ưu điểm:

* Hiệu quả trong phân loại và dự đoán khi các giả định được thỏa mãn.
* Giúp hiểu rõ hơn về các yếu tố nào ảnh hưởng mạnh nhất đến sự khác biệt giữa các nhóm.
* Có thể xử lý nhiều biến số và nhóm phân loại cùng lúc.

Nhược điểm:

* Hiệu quả của mô hình phụ thuộc vào việc các giả định (như phân phối chuẩn và đồng phương sai) có được thỏa mãn hay không.
* Nhạy cảm với dữ liệu ngoại lai, điều này có thể ảnh hưởng đến kết quả phân loại.

### Phân tích dữ liệu time series

Phân tích dữ liệu chuỗi thời gian là một phương pháp thống kê tập trung vào việc phân tích dữ liệu được thu thập và sắp xếp theo trình tự thời gian. Mục tiêu của phân tích chuỗi thời gian là hiểu và mô hình hóa các cấu trúc, xu hướng, và các mẫu hình trong dữ liệu để dự báo tương lai hoặc nhận diện các đặc điểm nổi bật trong quá khứ.

## Một số công cụ phân tích dữ liệu

### Python

Python là một ngôn ngữ lập trình mạnh mẽ và linh hoạt, được sử dụng rộng rãi trong cộng đồng khoa học dữ liệu và phân tích dữ liệu. Python nổi bật nhờ cú pháp dễ đọc, thư viện phong phú, và khả năng tích hợp với nhiều công cụ khác nhau, khiến nó trở thành công cụ lý tưởng cho nhiều nhiệm vụ phân tích dữ liệu, từ xử lý dữ liệu cơ bản đến học máy và trí tuệ nhân tạo.

Các thư viện chính:

* Pandas: Xử lý và phân tích dữ liệu dạng bảng.
* NumPy: Tính toán khoa học và xử lý mảng.
* Matplotlib/Seaborn: Vẽ đồ thị và biểu đồ.
* Scikit-learn: Học máy và mô hình hóa dữ liệu.
* Statsmodels: Phân tích thống kê và mô hình hóa dữ liệu.

### Microsoft Excel

Microsoft Excel là một trong những công cụ phổ biến nhất để phân tích dữ liệu, đặc biệt là trong các môi trường kinh doanh. Excel nổi bật nhờ giao diện người dùng thân thiện và khả năng xử lý dữ liệu bảng tính trực quan, khiến nó trở thành lựa chọn ưu tiên cho những ai mới bắt đầu với phân tích dữ liệu hoặc cần thực hiện các phân tích đơn giản và nhanh chóng.

Chức năng chính:

* Bảng tính và xử lý dữ liệu cơ bản.
* Công cụ phân tích dữ liệu (Data Analysis Toolpak).
* Pivot Table: Tạo báo cáo tóm tắt và phân tích đa chiều.
* Vẽ biểu đồ và đồ thị.

### OpenRefine

OpenRefine (trước đây gọi là Google Refine) là một công cụ mã nguồn mở dùng để làm sạch và biến đổi dữ liệu. Nó cho phép người dùng dễ dàng khám phá, chuẩn hóa, và chuẩn bị dữ liệu trước khi đưa vào phân tích sâu hơn.

Chức năng chính:

* Làm sạch dữ liệu: OpenRefine có khả năng xử lý và làm sạch các tập dữ liệu lớn, bao gồm việc phát hiện và sửa các lỗi như chính tả, định dạng không nhất quán, và dữ liệu trùng lặp.
* Biến đổi dữ liệu: Người dùng có thể dễ dàng chuyển đổi dữ liệu từ các định dạng khác nhau, gộp các trường dữ liệu, và thực hiện các phép biến đổi phức tạp bằng các biểu thức mạnh mẽ.
* Khám phá dữ liệu: OpenRefine cung cấp các công cụ trực quan giúp người dùng khám phá các mẫu hình, tần suất, và phân phối trong dữ liệu.
* Tích hợp và liên kết dữ liệu: Nó hỗ trợ liên kết và so sánh dữ liệu với các nguồn dữ liệu khác như các cơ sở dữ liệu mở hoặc các API trực tuyến.

### Tableau

Tableau là một công cụ trực quan hóa dữ liệu mạnh mẽ, được thiết kế để giúp người dùng dễ dàng hiểu và khám phá dữ liệu thông qua các biểu đồ và bảng điều khiển tương tác. Tableau không yêu cầu người dùng có kỹ năng lập trình, và nhờ giao diện kéo-thả thân thiện, nó cho phép người dùng nhanh chóng tạo ra các hình ảnh trực quan hóa phức tạp từ dữ liệu.

### WEKA

WEKA (Waikato Environment for Knowledge Analysis) là một bộ công cụ mã nguồn mở dành cho học máy và khai phá dữ liệu, phát triển bởi Đại học Waikato, New Zealand. WEKA cung cấp một bộ sưu tập các thuật toán học máy để khai thác dữ liệu, cũng như các công cụ để tiền xử lý dữ liệu, đánh giá mô hình, và trực quan hóa kết quả.

Chức năng chính:

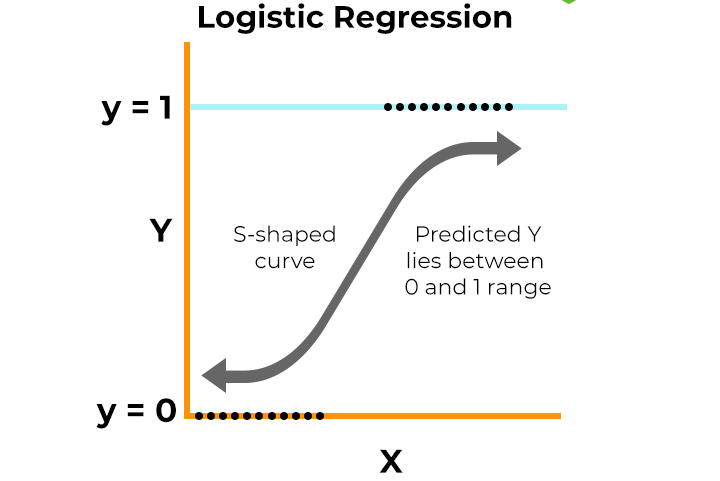
* Thuật toán học máy: WEKA bao gồm một loạt các thuật toán học máy phổ biến như cây quyết định, hồi quy, phân cụm, và mạng nơ-ron.
* Tiền xử lý dữ liệu: WEKA cung cấp các công cụ để chuẩn bị dữ liệu, bao gồm lọc dữ liệu, chuẩn hóa, chọn đặc trưng, và giảm nhiễu.
* Đánh giá mô hình: Công cụ này cho phép đánh giá và kiểm định các mô hình học máy với các phương pháp như k-fold cross-validation.
* Trực quan hóa: WEKA có các công cụ trực quan hóa dữ liệu và mô hình, giúp người dùng hiểu rõ hơn về cấu trúc và hiệu suất của mô hình.

# MÔ HÌNH HỒI QUY LOGISTIC

## Tổng quan về mô hình hồi quy logistic

Mô hình hồi quy logistic là một trong những công cụ quan trọng trong thống kê và machine learning, đặc biệt được sử dụng rộng rãi trong các vấn đề phân loại và dự đoán. Mô hình này là một phần của họ các mô hình hồi quy và được áp dụng chủ yếu cho các vấn đề phân loại, trong đó chúng ta cố gắng dự đoán xác suất của một sự kiện xảy ra hoặc không xảy ra dựa trên các biến đầu vào.

Logistic Regression là một thuật toán học có giám sát (supervised learning), nó được sử dụng cho các bài toán phân lớp. Đây trong những phương pháp quan trọng trong machine learning và thống kê, được sử dụng để dự đoán kết quả phân loại dựa trên biến độc lập. Mô hình này thường được sử dụng khi biến phụ thuộc là biến phân loại, tức nhị phân hay đa lớp. Mục tiêu của mô hình là nhằm xác định xác suất xảy ra của một sự kiện, từ đó xác định sự kiện là có xảy ra hay là không.



Hình 2.1 Các giả định cơ bản để thực hiện hồi quy logistic

Hồi quy Logistic là một mô hình thống kê được sử dụng để phân loại nhị phân, tức dự đoán một đối tượng thuộc vào một trong hai nhãn. Mô hình hồi quy Logistic hoạt động dựa trên nguyên lý cơ bản sử dụng hàm sigmoid, hay còn gọi là hàm logistic chuẩn để ước lượng xác suất của một sự kiện xảy ra hoặc không xảy ra. Xét bài toán phân lớp nhị phân. Cho trước một tập dữ liệu với n mẫu:

(2.1.a)

Trong đó, là một vector trong không gian và là tập nhãn lớp. Đầu ra của mô hình có dạng với là hàm logistic.

Hàm logistic có dạng:

(2.1.b)

Hay:

(2.1.c)

* là xác xuất của một sự kiện xảy ra khi biết giá trị của biến độc lập .
* là hệ số của mô hình.
* là các biến độc lập.

Hàm sigmoid nhận đầu vào là một giá trị bất kì, và đầu ra trả về là một giá trị xác suất trong khoảng . Chẳng hạn, khi đầu vào của mô hình là ma trận dữ liệu và trọng số , ta có đầu vào dạng .

Mục đích của việc huấn luyện mô hình là tìm ra một bộ trọng số w sao chonkết quả trả về của hàm Sigmoid là gần với kết quả thực tế nhất. Trong các bài toán Machine Learning, điều này được thực hiện bằng cách tối ưu hóa hàm mất mát, hay còn gọi là Loss Function. Hàm mất mát được sử dụng để đo lường mức độ lỗi của mô hình khi dự đoán các kết quả đầu vào. Tức, hàm mất mát có giá trị càng nhỏ thì mức độ lỗi càng thấp, chứng tỏ việc huấn luyện mô hình càng tốt. Trong bài toán hồi quy Logistic, hàm mất mát được sử dụng để đánh giá hiệu năng của mô hình, được xây dựng với công thức:

(2.1.d)

Trong đó, là số lượng mẫu dữ liệu của tập huấn luyện, là nhãn thực của đầu ra thứ , là xác suất dự đoán thuộc lớp của mô hình với đầu vào thứ . là để biểu diễn logarit tự nhiên.

Hàm mất mát trên còn được gọi là Cross-Entropy, đo lường khoảng cách giữa 2 phân phối xác suất và . và cùng đồng thời tiến về 0 hoặc 1. Hàm mất mát sẽ tiến dần về 0.

Mặc dù sử dụng hàm Sigmoid là hàm liên tục để biểu diễn xác suất, nhưng trong thực tế, nó được áp dụng cho các bài toán dự đoán rời rạc, sử dụng nhiều trong các bài toán phân lớp (Classification). Sau khi tìm được mô hình, việc xác định nhãn lớp (class) y cho một điểm dữ liệu x được xác định thông qua việc so sánh hai biểu thức xác suất

(2.1.e)

Nếu giá trị của biểu thức thứ nhất lớn hơn biểu thức thứ hai, ta nói điểm dữ liệu thuộc vào class , và ngược lại, điểm dữ liệu thuộc vào class . Tuy nhiên, giả sử hai giá trị xác suất đó bằng nhau, ta có thể xác định điểm dữ liệu đó thuộc một trong hai nhãn lớp.

## Ưu, nhược điểm của mô hình hồi quy logistic

Mô hình Logistic là một công cụ mạnh mẽ được sử dụng rộng rãi trong lĩnh vực dự đoán và phân loại dựa trên dữ liệu đầu vào chính vì những ưu điểm của nó như:

* Phù hợp với dữ liệu rời rạc: Mô hình hồi quy logistic hoạt động tốt với dữ liệu rời rạc, tức là các biến đầu vào có thể là dạng nhóm hoặc các giá trị định danh.
* Xác định được xác suất: Mô hình này có khả năng dự đoán xác suất xảy ra của sự kiện quan trọng. Điều này rất hữu ích trong việc đưa ra quyết định dựa trên mức độ rủi ro.
* Khả năng diễn giải tham số: Các tham số của mô hình hồi quy logistic có thể diễn giải được, giúp hiểu rõ hơn về mối quan hệ giữa biến đầu vào và xác suất dự đoán.
* Ít yêu cầu về giả định: Mô hình này ít yêu cầu các giả định về phân phối của dữ liệu so với một số phương pháp khác như mô hình hồi quy tuyến tính.

Tuy nhiên, mô hình này cũng có những nhược điểm phải kể đến:

* Dễ bị ảnh hưởng bởi nhiễu và quan sát outlier: Mô hình hồi quy logistic có thể nhạy cảm với dữ liệu nhiễu hoặc quan sát outlier, có thể làm giảm độ chính xác của mô hình.
* Khả năng không hiệu quả khi có tương quan cao giữa biến đầu vào: Nếu có tương quan cao giữa các biến đầu vào, mô hình có thể gặp vấn đề về hiệu suất và sự chính xác của dự đoán.
* Khó khăn trong việc xử lý các dữ liệu không cân bằng: Khi dữ liệu không cân bằng, tức là tỷ lệ các lớp trong dữ liệu không đồng đều, mô hình có thể dự đoán chủ quan và chịu ảnh hưởng của việc mất cân bằng này.

# THỰC NGHIỆM, ĐÁNH GIÁ VÀ XÂY DỰNG ỨNG DỤNG THỬ NGHIỆM

Toàn bộ các thao tác trên dữ liệu được thực hiện bằng ngôn ngữ lập trình Python và các thư viện có liên quan.

## Mô tả dữ liệu

Bộ dữ liệu được sử dụng cho đề tài được công bố bởi Viện Quốc gia về Bệnh tiểu đường, Tiêu hóa và Bệnh thận thuộc về Viện Y tế Quốc gia Hoa Kỳ. Bộ dữ liệu bao gồm 768 mẫu thu thập trên các bệnh nhân là phụ nữ từ 21 tuổi trở lên là người gốc da đỏ Pima.

Mỗi mẫu bao gồm 9 thuộc tính bao gồm:

* Pregnancies: Số lần mang thai
* Glucose: Nồng độ glucose trong máu được đo 2 giờ sau khi tiêu thụ dung dịch glucose
* BloodPressure: Huyết áp tâm trương (mm Hg)
* SkinThickness: Độ dày da phía sau bắp tay (mm)
* Insulin: Insulin huyết thanh 2 giờ (mu U/ml)
* BMI: Chỉ số BMI (Được tính bằng )
* DiabetesPedigreeFunction: Chỉ số DPF (chỉ số đánh giá nguy cơ mắc bệnh tiểu đường dựa trên yếu tố di truyền)
* Age: Số tuổi
* Outcome: Biến lớp (0 hoặc 1, tương ứng với âm tính hoặc dương tính với bệnh tiểu đường)

A screenshot of a graph

Description automatically generated

Hình 3.1 Một phần của bộ dữ liệu

A computer code with text

Description automatically generated

Hình 3.2 Tổng quan bộ dữ liệu

A screenshot of a computer

Description automatically generated

Hình 3.3 Thống kê về kiểu dữ liệu của các thuộc tính

Từ *Hình 3.3* ta thấy dữ liệu đều là kiểu số và không có dữ liệu bị khuyết.

A screenshot of a computer

Description automatically generated

Hình 3.4 Tóm tắt bộ dữ liệu

Từ *Hình 3.4*, ta có thống kê mô tả như sau:

* **Pregnancies**:
* Trung bình: 3.85
* Giá trị lớn nhất: 17
* Giá trị nhỏ nhất: 0
* **Glucose**:
* Trung bình: 120.89
* Giá trị lớn nhất: 199
* Giá trị nhỏ nhất: 0
* **BloodPressure**:
* Trung bình: 69.11
* Giá trị lớn nhất: 122
* Giá trị nhỏ nhất: 0
* **SkinThickness**:
* Trung bình: 20.54
* Giá trị lớn nhất: 99
* Giá trị nhỏ nhất: 0
* **Insulin**:
* Trung bình: 79.80
* Giá trị lớn nhất: 846
* Giá trị nhỏ nhất: 0
* **BMI**:
* Trung bình: 31.99
* Giá trị lớn nhất: 67.10
* Giá trị nhỏ nhất: 0
* **DiabetesPedigreeFunction**:
* Trung bình: 0.47
* Giá trị lớn nhất: 2.42
* Giá trị nhỏ nhất: 0.078
* **Age**:
* Trung bình: 33.24
* Giá trị lớn nhất: 81
* Giá trị nhỏ nhất: 21
* **Outcome**:
* Trung bình: 0.35
* Giá trị lớn nhất: 1
* Giá trị nhỏ nhất: 0
* Cho thấy tỷ lệ mắc bệnh tiểu đường trong bộ dữ liệu khoảng 35%.

Nhận xét:

* Độ phân bố dữ liệu:
  + Một số biến như Glucose, Insulin, và BMI có giá trị cực trị cao, điều này là tương đối phù hợp với bộ dữ liệu được đo trên nhiều bệnh nhân với các tình trạng bệnh lý phức tạp.
  + Nhiều biến như Glucose, BloodPressure, SkinThickness, Insulin, và BMI có giá trị 0, điều này là không hợp lý đối với phép đo thông thường, điều này chỉ ra dữ liệu thiếu hoặc sai sót trong việc ghi chép dữ liệu. Các giá trị này có thể coi là giá trị khuyết và cần được xử lý.
* Phân phối:
  + Pregnancies và Age có phân phối rộng với các giá trị lớn nhất và nhỏ nhất cho thấy độ phân tán cao.
  + Outcome cho thấy tỷ lệ mắc bệnh tiểu đường là 35%, cho thấy có sự không cân bằng giữa các lớp (0 và 1).

A screenshot of a computer code

Description automatically generated

Hình 3.5 Dữ liệu trùng lặp trong bộ dữ liệu

Từ thống kê *Hình 3.5* ta thấy dữ liệu không có sự trùng lặp và lớp đầu ra (Outcome) chỉ bao gồm 0 hoặc 1.

A screenshot of a computer

Description automatically generated

Hình 3.6 Mức độ tương quan giữa các thuộc tính

A screenshot of a graph

Description automatically generated

Hình 3.7 Biểu đồ nhiệt thể hiện sự tương quan

Từ *Hình 3.6 và Hình 3.7*, ta có nhận xét về mức độ tương quan giữa các thuộc tính như sau:

* **Pregnancies và Age**: Có một mối quan hệ tích cực mạnh (0.54). Số lần mang thai có xu hướng tăng khi tuổi tác cao hơn, có thể phản ánh rằng những người lớn tuổi hơn thường có nhiều lần mang thai hơn trong mẫu này.
* **Glucose và Outcome**: Tương quan khá mạnh (0.47), cho thấy rằng nồng độ glucose cao có liên quan đến nguy cơ mắc bệnh tiểu đường cao hơn.
* **SkinThickness và Insulin**: Tương quan khá mạnh (0.44), cho thấy mối liên hệ tích cực mạnh giữa độ dày lớp da và nồng độ insulin, có thể liên quan đến việc tăng lượng mỡ cơ thể.
* **Insulin và Outcome**: Tương quan yếu (0.13), cho thấy mối liên hệ nhẹ giữa nồng độ insulin và kết quả tiểu đường.
* **BMI và Outcome**: Tương quan vừa (0.29), chỉ ra rằng chỉ số khối cơ thể có ảnh hưởng đáng kể đến nguy cơ mắc bệnh tiểu đường.
* **Age và Outcome**: Tương quan yếu (0.24), có mối liên hệ nhẹ giữa tuổi tác và nguy cơ mắc bệnh tiểu đường.

A grid of blue dots

Description automatically generated

Hình 3.8 Ma trận biểu đồ phân tán của bộ dữ liệu

Từ *Hình 3.8*, ta thấy một số cặp thuộc tính có mức độ tương quan tuyến tính dương như Glucose và Isnsulin, BMI và SkinThickness. Một số mối quan hệ sáng có sự phân tán cao, tương quan không rõ ràng.

Một số biểu đồ thể hiện mối quan hệ giữa các thuộc tính với kết quả xét nghiệm bệnh tiểu đường.

A graph with red lines and numbers

Description automatically generated

Hình 3.9 Mối quan hệ giữa Glucose và bệnh tiểu đường

A graph with red lines and numbers

Description automatically generated

Hình 3.10 Mối quan hệ giữa tuổi và bệnh tiểu đường

A graph with red lines

Description automatically generated

Hình 3.11 Mối quan hệ giữa chỉ số BMI và bệnh tiểu đường

A graph with red lines and numbers

Description automatically generated

Hình 3.12 Mối quan hệ giữa Insulin và bệnh tiểu đường

A graph with red lines and numbers

Description automatically generated

Hình 3.13 Mối quan hệ giữa huyết áp tâm trương và bệnh tiểu đường

*Hình 3.9* cho thấy, khi lượng đường huyết tăng, xác suất mắc bệnh tiểu đường cũng tăng theo. Dựa trên biểu đồ, có thể xác định được một ngưỡng đường huyết nhất định, vượt qua ngưỡng này, xác suất mắc bệnh tiểu đường tăng đáng kể.

Ở *Hình 3.10*, ta thấy giữa tuổi và khả năng mắc tiểu đường không rõ ràng, các dữ liệu phân bố ngẫu nhiên, thể hiện tuổi và khả năng mắc bệnh tiểu đường ít có mối quan hệ trực tiếp.

*Hình 3.11* cũng cho thấy khi chỉ số BMI tăng, tỉ lệ mắc bệnh tiểu đường tăng đáng kể, *Hình 3.12* cũng có kết quả tương tự khi mức insulin tăng, khả năng mắc bệnh tiểu đường cũng tăng theo, điều này phù hợp với kiến thức y khoa về vai trò của insulin trong việc điều hòa đường huyết.

## Xử lý giá trị khuyết

A screenshot of a computer

Description automatically generated

Hình 3.14 Xử lý dữ liệu khuyết của bộ dữ liệu

Nhóm tiến hành xóa bỏ các dữ liệu có giá trị khuyết và lưu ra một biến khác. Sau khi xử lý bộ dữ liệu còn 392 dòng.

A screenshot of a computer

Description automatically generated

Hình 3.15 Tổng quan bộ dữ liệu sau khi xử lý

A screenshot of a graph

Description automatically generated

Hình 3.16 Biểu đồ nhiệt của bộ dữ liệu sau khi xử lý giá trị khuyết

Nhận xét: Việc xóa bỏ các giá trị khuyết không làm thay đổi đáng kể tính chất của bộ dữ liệu.

Dữ liệu hiện tại đã sẵn sàng để thực hiện hồi quy.

## Huấn luyện và đánh giá mô hình hồi quy Logistic

Để đánh giá được thêm nhiều thông tin, ta tiến hành thực hiện huấn luyện mô hình hồi quy với cả hai bộ dữ liệu là bộ dữ liệu gốc và bộ dữ liệu sau khi xóa bỏ các giá trị khuyết.

Đối với việc thực hiện với hai bộ dữ liệu, các tham số được thiết lập cho hai mô hình là như nhau, bao gồm: Chia dữ liệu thành tập train và test với tỉ lệ 75/25, solver = "liblinear", penalty = 12.

A screenshot of a computer program

Description automatically generated

Hình 3.17 Huấn luyện mô hình với bộ dữ liệu gốc

A screenshot of a computer code

Description automatically generated

Hình 3.18 Huấn luyện mô hình với bộ dữ liệu đã xử lý giá trị khuyết

Sau khi tiến hành huấn luyện, ta thu được kết quả.

Bảng 3.1 So sánh kết quả huấn luyện với hai bộ dữ liệu

|  |  |  |
| --- | --- | --- |
| **Bộ dữ liệu** | **Test accuracy** | **F1 Score** |
| Bộ dữ liệu gốc | 0.8073 | 0.7695 |
| Bộ dữ liệu được xử lý giá trị khuyết | 0.8061 | 0.7883 |

Mô hình sử dụng bộ dữ liệu gốc có độ chính xác cao hơn một chút (0.8073 so với 0.8061) nhưng sự chênh lệch rất nhỏ và không đáng kể.

Với Score cao hơn (0.7883 so với 0.7685), mô hình sử dụng bộ dữ liệu được xử lý giá trị khuyết hoạt động tốt hơn với bộ dữ liệu của mình.

Mô hình xử dụng bộ dữ liệu đã được xử lý giá trị khuyết cho ra kết quả có thể coi là tốt hơn nhưng khá ít. Trong nội dung nghiên cứu, thử nghiệm với bộ dữ liệu đầy đủ thông tin, tôi quyết định chọn mô hình này để xây dựng ứng dụng thử nghiệm.

## Xây dựng ứng dụng dự đoán bệnh tiểu đường

### Công cụ Flutter

Flutter là một framework phát triển ứng dụng mã nguồn mở do Google phát triển. Nó cho phép xây dựng ứng dụng di động, web và desktop từ một cơ sở mã nguồn duy nhất. Dưới đây là một số điểm nổi bật về Flutter:

* **Hiệu suất cao**: Flutter sử dụng ngôn ngữ lập trình Dart và cung cấp các công cụ để biên dịch mã nguồn trực tiếp thành mã máy, giúp ứng dụng chạy mượt mà và nhanh chóng.
* **Giao diện người dùng tùy chỉnh**: Flutter cung cấp một bộ widget phong phú và có thể tùy chỉnh cao, cho phép các nhà phát triển dễ dàng tạo ra giao diện người dùng đẹp và độc đáo.
* **Hot Reload**: Tính năng "Hot Reload" cho phép các nhà phát triển xem ngay lập tức các thay đổi trong mã nguồn mà không cần khởi động lại ứng dụng, giúp tăng tốc quá trình phát triển.
* **Hỗ trợ đa nền tảng**: Flutter có khả năng xây dựng ứng dụng cho nhiều nền tảng khác nhau như Android, iOS, web và desktop từ cùng một cơ sở mã nguồn.
* **Dart Language**: Flutter sử dụng ngôn ngữ lập trình Dart, được thiết kế để dễ học và sử dụng. Dart hỗ trợ các tính năng như lập trình đồng thời và lập trình hướng đối tượng.
* **Cộng đồng và hỗ trợ**: Flutter có một cộng đồng phát triển lớn và ngày càng mở rộng, cùng với tài liệu phong phú và các công cụ hỗ trợ từ Google và các bên thứ ba.

Nhờ những tính năng này, Flutter đã trở thành một lựa chọn phổ biến cho các nhà phát triển muốn tạo ra ứng dụng đa nền tảng với hiệu suất cao và giao diện người dùng đẹp mắt.

### Giao diện ứng dụng

A screenshot of a computer

Description automatically generated

Hình 3.19 Giao diện ứng dụng

KẾT LUẬN

Nghiên cứu về việc phân tích dự đoán bệnh tiểu đường sử dụng mô hình hồi quy logistic đã chỉ ra rằng phương pháp này là một công cụ mạnh mẽ trong việc xử lý và phân tích dữ liệu y tế. Trong chương 1, chúng tôi đã trình bày tổng quan về phân tích dữ liệu, bao gồm khái niệm, quy trình và các kỹ thuật phân tích dữ liệu cơ bản như phân tích mô tả, hồi quy, sự khác biệt, và dữ liệu thời gian. Các công cụ phân tích dữ liệu như Python, Microsoft Excel, OpenRefine, Tableau và WEKA đã được giới thiệu, nhấn mạnh tầm quan trọng của việc lựa chọn công cụ phù hợp cho từng loại phân tích.

Chương 2 đã cung cấp cái nhìn sâu rộng về mô hình hồi quy logistic, một phương pháp phân tích thống kê rất hiệu quả trong việc dự đoán các sự kiện nhị phân, chẳng hạn như nguy cơ mắc bệnh tiểu đường. Chúng tôi đã phân tích các ưu điểm và nhược điểm của mô hình hồi quy logistic, từ đó hiểu rõ hơn về khả năng và giới hạn của phương pháp này trong việc xử lý dữ liệu y tế.

Chương 3 tập trung vào thực nghiệm, đánh giá và xây dựng ứng dụng thử nghiệm dự đoán bệnh tiểu đường. Việc mô tả dữ liệu và xử lý giá trị khuyết đã được thực hiện để đảm bảo chất lượng dữ liệu đầu vào cho mô hình. Sau khi huấn luyện và đánh giá mô hình hồi quy logistic, kết quả cho thấy mô hình này có khả năng dự đoán chính xác nguy cơ mắc bệnh tiểu đường, cung cấp một công cụ hữu ích trong việc hỗ trợ các quyết định y tế. Cuối cùng, việc xây dựng ứng dụng dự đoán bệnh tiểu đường bằng công cụ Flutter đã cho phép chúng tôi phát triển một giải pháp thực tiễn và dễ sử dụng cho người dùng.

Tóm lại, mô hình hồi quy logistic đã chứng tỏ sự hiệu quả của nó trong việc phân tích và dự đoán bệnh tiểu đường. Những kết quả thu được từ nghiên cứu này không chỉ cung cấp cái nhìn sâu sắc về các yếu tố nguy cơ mà còn mở ra cơ hội cho việc phát triển các công cụ hỗ trợ quyết định y tế dựa trên dữ liệu, góp phần vào việc cải thiện sức khỏe cộng đồng.

Trong tương lai, việc phát triển và cải tiến các phương pháp dự đoán bệnh tiểu đường có thể tập trung vào một số hướng quan trọng để nâng cao độ chính xác và hiệu quả của các công cụ phân tích. Đầu tiên, việc tích hợp các mô hình học máy tiên tiến như mạng nơ-ron sâu và học máy tăng cường có thể giúp cải thiện khả năng dự đoán bằng cách khai thác những mẫu dữ liệu phức tạp hơn mà các mô hình hồi quy logistic truyền thống có thể chưa phát hiện ra. Thứ hai, mở rộng dữ liệu đầu vào bằng cách kết hợp dữ liệu gen, lối sống và môi trường có thể cung cấp cái nhìn toàn diện hơn về nguy cơ mắc bệnh tiểu đường. Hơn nữa, việc ứng dụng công nghệ dữ liệu lớn (big data) và phân tích dữ liệu thời gian thực có thể tạo ra những giải pháp dự đoán chính xác và cá nhân hóa hơn, cho phép theo dõi và can thiệp sớm. Cuối cùng, phát triển các ứng dụng di động và nền tảng dựa trên điện toán đám mây có thể giúp người dùng dễ dàng tiếp cận các công cụ dự đoán và theo dõi sức khỏe của họ bất cứ lúc nào và ở bất kỳ đâu. Những tiến bộ này không chỉ góp phần nâng cao chất lượng dự đoán mà còn hỗ trợ quản lý sức khỏe toàn diện và hiệu quả hơn trong cộng đồng.

TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | Xia, B. S., & Gong, P. (2015). Review of business intelligence through data analysis. *Benchmarking*, *21*(2), 300-311. doi:10.1108/BIJ-08-2012-0050 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |