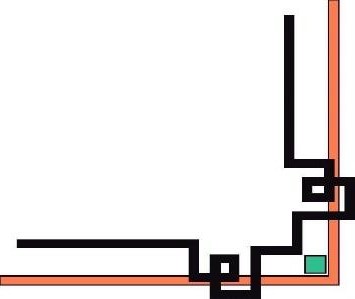
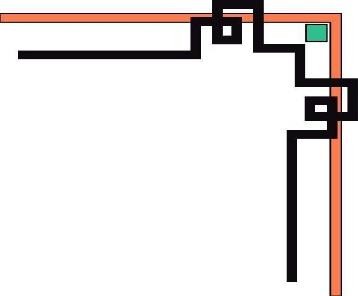
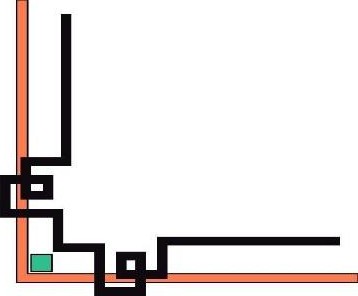
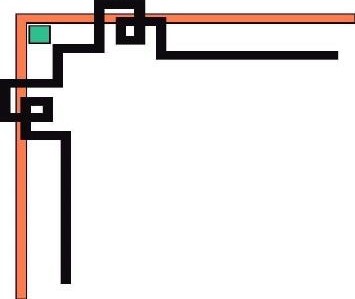
1



### TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH KHOA HỆ THỐNG THÔNG TIN

**BÁO CÁO ĐỒ ÁN MÔN HỌC DỮ LIỆU LỚN**

**ĐỀ TÀI**

**ỨNG DỤNG KỸ THUẬT MÁY HỌC XÂY DỰNG HỆ THỐNG DỰ ĐOÁN BỆNH ĐỘT QUỴ**

**GIẢNG VIÊN HƯỚNG DẪN**

#### ThS. Nguyễn Hồ Duy Tri

|  |  |  |
| --- | --- | --- |
| **Nhóm sinh viên thực hiện:** | | |
| STT | Họ tên | MSSV |
| 1 | Đặng Văn Hiệp | 16520373 |
| 2 | Nguyễn Trần Hà Đức | 18520615 |
| 3 | Hoàng Tuấn Anh | 18520447 |
| 4 | Nguyễn Hữu Nhật Tân | 18520354 |

**Thành phố Hồ Chí Minh, tháng 12 năm 2021**

## DANH MỤC HÌNH ẢNH

[Hình 3: Code import các thư viện cần thiết. 14](#_bookmark13)

[Hình 4: Code đọc dữ liệu đầu vào và cấu trúc lại dữ liệu. 14](#_bookmark15)

[Hình 5: Kết quả cấu trúc dữ liệu. 15](#_bookmark16)

[Hình 6: Kết quả dataframe của dữ liệu. 15](#_bookmark17)

[Hình 7: Kết quả phân phối biến Outcome. 15](#_bookmark18)

[Hình 8: Code tính tỉ lệ dữ liệu train và các giá trị thống kê 15](#_bookmark19)

[Hình 9: Kết quả tính tỉ lệ dữ liệu train và các giá trị thống kê. 16](#_bookmark20)

[Hình 10: Code trực quan hóa dữ liệu. 17](#_bookmark21)

[Hình 11: Kết quả trực quan hóa dữ liệu. 17](#_bookmark22)

[Hình 12: Code lọc dữ liệu null và dữ liệu có giá trị N/A. 18](#_bookmark24)

[Hình 13: Kết quả lọc dữ liệu null và dữ liệu có giá trị N/A. 18](#_bookmark25)

[Hình 14: Code chia dữ liệu thành dữ liệu train và test. 18](#_bookmark27)

[Hình 15: Code đánh giá thuật toán 22](#_bookmark42)

[Hình 16: Hàm sigmoid. 23](#_bookmark44)

[Hình 17: Hàm compute\_cost. 24](#_bookmark45)

[Hình 18: Hàm gradient\_descent. 24](#_bookmark46)

[Hình 19: Hàm predict. 24](#_bookmark47)

[Hình 20: Hàm score. 24](#_bookmark48)

[Hình 21: Tìm tham số cho phương trình. 25](#_bookmark49)

[Hình 22: Code đánh giá thuật toán 25](#_bookmark50)

#### **MỤC LỤC**

[**CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 8**](#_bookmark0)

* 1. [**Lý do chọn đề tài 8**](#_bookmark1)
  2. [**Mô tả dữ liệu 8**](#_bookmark2)
     1. [**Giới thiệu về nguồn dữ liệu 8**](#_bookmark3)
     2. [**Các thuộc tính của bộ dữ liệu 8**](#_bookmark4)
  3. [**Mô tả bài toán 9**](#_bookmark5)
  4. [**Kỹ thuật tiền xử lý dữ liệu được lựa chọn 9**](#_bookmark6)
  5. [**Thuật toán khai thác dữ liệu được lựa chọn 10**](#_bookmark7)
     1. [**Hồi quy logistic (Logistic Regression) 11**](#_bookmark9)

[**1.5 Phương pháp phân chia dữ liệu 12**](#_bookmark10)

[**CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU 14**](#_bookmark11)

* 1. [**Import các thư viện cần thiết 14**](#_bookmark12)
  2. [**Đọc dữ liệu đầu vào thực hiện Exploratory Data Analysis (EDA) 14**](#_bookmark14)
  3. [**Lọc những dòng dữ liệu chứa rỗng và xóa dòng trùng 18**](#_bookmark23)

[**2.3 Chia tập dữ liệu 18**](#_bookmark26)

[**CHƯƠNG 3: THUẬT TOÁN KHAI THÁC DỮ LIỆU 19**](#_bookmark28)

* 1. [**Thuật toán Logistic Regression 19**](#_bookmark29)

[**CHƯƠNG 4: KẾT QUẢ ĐẠT ĐƯỢC 27**](#_bookmark51)

* 1. [**Kết quả đạt được 27**](#_bookmark52)
  2. [**So sánh, đánh giá 27**](#_bookmark53)

[**CHƯƠNG 5: KẾT LUẬN 28**](#_bookmark54)

* 1. [**Ưu điểm 28**](#_bookmark55)
  2. [**Hạn chế 28**](#_bookmark56)
  3. [**Hướng phát triển 28**](#_bookmark57)

[**CHƯƠNG 6: TÀI LIỆU THAM KHẢO 29**](#_bookmark58)

## LỜI CẢM ƠN

Đầu tiên, em xin gởi lời cảm ơn chân thành đến tập thể quý Thầy Cô Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM và quý Thầy Cô khoa Hệ thống thông tin đã giúp cho em có những kiến thức cơ bản làm nền tảng để thực hiện đề tài này. Đặc biệt, em xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới Thầy Nguyễn Hồ Duy Tri (Giảng viên môn Dữ liệu lớn). Các Thầy Cô đã trực tiếp hướng dẫn tận tình, sửa chữa và đóng góp nhiều ý kiến quý báu giúp em hoàn thành tốt báo cáo môn học của mình. Trong thời gian một học kỳ thực hiện đề tài, em đã vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới. Từ đó, em vận dụng tối đa những gì đã thu thập được để hoàn thành một báo cáo đồ án tốt nhất. Tuy nhiên, trong quá trình thực hiện, em không tránh khỏi những thiếu sót. Chính vì vậy, em rất mong nhận được những sự góp ý từ phía các Thầy Cô nhằm hoàn thiện những kiến thức mà em đã học tập và là hành trang để em thực hiện tiếp các đề tài khác trong tương lai. Xin chân thành cảm ơn các quý Thầy Cô!

Thành phố Hồ Chí Minh, tháng 1 năm 2021

Nhóm sinh viên thực hiện

## NHẬN XÉT CỦA GIẢNG VIÊN

##### …………………………………………………………………………………………

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**…………………………………………………………………………………………**

**………………………………………………………………………………………**

**CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI**

* 1. **Lý do chọn đề tài**

Bệnh đột quỵ hay còn gọi là tai biến mạch máu não. Là căn bệnh Đột quỵ xảy ra khi mạch máu nuôi não bị tắc nghẽn hoặc vỡ làm cho não không nhận đủ oxy. Khi đó, một phần não bắt đầu chết đi và gây tổn thương não. Đột quỵ là một trong những nguyên nhân gây tử vong cao và để lại một số di chứng vĩnh viễn cho bệnh nhân.

Theo thống kê, có khoảng 50% bệnh nhân đột quỵ tử vong. Đây là nguyên nhân tử vong hàng thứ 3 trên thé giới, chỉ sau đột quỵ và tim mạch. Thứ hạng này có sự thay đổi hàng năm nhưng đột quỵ chưa bao giờ “rơi” khỏi top 3. Người đột quỵ có nguy cơ tử vong cao, nhất là khi không được cấp cứu kịp thời.

Mỗi năm, ở Việt Nam có hơn 200.000 người bị đột quỵ, hơn 50% trong số đó tử vong và chỉ có 10% những người sống sót là bình phục hoàn toàn (không có di chứng và không phải phụ thuộc vào người khác). Tỷ lệ bệnh nhân đột quỵ ở Việt Nam chia theo giới tính thì nam nhiều gấp 4 lần so với nữ giới.

Giống như nhiều nước khác, số ca đột quỵ tại Việt Nam tăng lên khoảng 2% trong vòng 3 năm và ngày càng có xu hướng trẻ hóa. Bệnh nhân trong độ tuổi 40-45 chiếm 1/3 tổng số ca đột quỵ. Đáng chú ý, mỗi năm có tới khoảng 83.000 người ở độ tuổi 20 hoặc trẻ hơn cùng mắc bệnh nà. Thông tin từ Trung tâm Đột quỵ não - Bệnh viện Trung ương Quân đội 108 cho biết, năm 202- đã ghi nhận nhiều trường hợp thanh thiếu niên bị đột quỵ, trong đs nhỏ nhất là bệnh nhân 12 tuổi. Hay mới đây, đầu năm 2021, trường hợp một bé trai 3 tuổi ở tỉnh Vĩnh Long bị đột quỵ cũng đã được ghi nhận.

**1.2 Mô tả dữ liệu**

**1.2.1 Giới thiệu về nguồn dữ liệu**

**Tên nguồn dữ liệu:** Dữ liệu chẩn đoán về bệnh đột quỵ

**Nguồn dữ liệu:** <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>

**Mô tả dữ liệu:** Dữ liệu nhóm sử dụng cho việc đánh giá xem có thể dự doán khi nào phát sinh và xảy ra của căn bệnh đột quỵ. Dữ liệu bao gồm 5110 dòng và 12 cột

### 1.2.2 Các thuộc tính của bộ dữ liệu

Bảng mô tả dữ liệu:

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa** |
| **1** | Id | Mã ID bệnh nhân |
| **2** | Gender | Giới tính |
| **3** | Age | Độ tuổi |
| **4** | Hypertension | Bệnh nhân có bị tăng huyết áp không  1: có  0: không |
| **5** | Heart\_disease | Bệnh nhân có bị đau tim không  1: có  0: không |
| **6** | Ever\_married | Bệnh nhân đã từng kết hôn chưa |
| **7** | Work\_type | Loại hình công việc của bệnh nhân |
| **8** | Residence\_type | Khu vực sinh sống của bệnh nhân |
| **9** | Avg\_glucose\_level | Chỉ số glucose trung bình của bệnh nh |
| **10** | Bmi | Chỉ số BMI của bệnh nhân |
| **11** | Smoking\_status | Tình trạng bênh nhân có hút thuốc không |
| **12** | Stroke | Kết luận bệnh nhân có đột quỵ hay không |

### Mô tả bài toán

Mục tiêu của project là xây dựng mô hình học máy để dự đoán xem một bệnh nhân có mắc bệnh đột quỵ hay không. Đây là project phân loại nhị phân (2 lớp) với học có giám sát.

Biểu đồ một số thuộc tính:

Chart, bar chart

Description automatically generated

***Biểu đồ Histogram của thuộc tính Age***

***Chart, bar chart, histogram

Description automatically generated***

***Biểu đồ Histogram của thuộc tính Hypetension***

***Chart, bar chart, histogram

Description automatically generated***

***Biểu đồ Histogram của thuộc tính Heart\_disease***

***Chart, bar chart

Description automatically generated***

***Biểu đồ Histogram của thuộc tính Ever\_married***

***Chart, bar chart

Description automatically generated***

***Biểu đồ Histogram của thuộc tính Work\_type***

***Chart, bar chart

Description automatically generated***

***Biểu đồ Histogram của thuộc tính Residence\_type***

***Chart, bar chart

Description automatically generated***

***Biểu đồ Histogram của thuộc tính Avg\_glucose\_level***

***Chart, bar chart

Description automatically generated***

***Biểu đồ Histogram của thuộc tính bmi***

***Chart, bar chart

Description automatically generated***

***Biểu đồ Histogram của smoking\_status***

***Chart, bar chart

Description automatically generated***

***Biểu đồ Histogram của gender***

### Kỹ thuật tiền xử lý dữ liệu được lựa chọn

Kỹ thuật xử lý dữ liệu đơn giản gồm các bước:

* Đọc dữ liệu.
* Gán kiểu dữ liệu thích hợp cho các thuộc tính.
* Chuẩn hóa dữ liệu.
* Loại bỏ những dòng dữ liệu chứa dữ liệu rỗng.
* Loại bỏ dòng dữ liệu bị trùng.

### Thuật toán khai thác dữ liệu được lựa chọn

### Hồi quy logistic (Logistic Regression)

Hồi quy Logistic là một phương pháp thống kê để dự đoán các lớp nhị phân và thường được sử dụng nhất để phân loại hai lớp. Ví dụ như thường được chuẩn đoán bệnh đột quỵ trong y khoa,… Nó tính toán xác suất xảy ra bằng hàm sigmoid và là trường hợp đặc biệt của hồi quy tuyến tính. Hồi quy Logistic mô tả và ước tính mối quan hệ giữ một biến nhị phân phụ thuộc vào một hoặc nhiều biến độc lập.

##### Công thức thuật toán:

##### Hàm Logistic được biểu diễn dưới dạng:

##### Diagram Description automatically generated with low confidence

##### Chart, line chart Description automatically generated

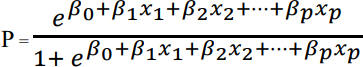
##### Giá trị của hàm g(z) logistic luôn nằm trong khoảng từ [0,1]

* + - * Phương trình tổng quát của hồi quy Logistic đơn biến:



***Hình 2: Hồi quy Logistic đơn biến.***

* + - * Phương trình tổng quát của hồi quy Logistic đa biến:



***Hình 3: Hồi quy Logistic đa biến.***

* + - * Phương trình Cost Function cho Logistic Regression:

Text, letter

Description automatically generated

* Để tối ưu hóa mô hình dự đoán ta cần hàm Cost Function đạt giá trị cực tiểu dựa theo giá trị của đối số Theta “θ”.
* Sử dụng thuật toán Gradient Descent để tìm các giá trị Theta hay hàm Cost Function đạt cực tiểu:
* Dạng tổng quát của Gradient Descent :

A picture containing calendar

Description automatically generated

* Lấy đạo hàm của Cost Function ta được:

Text

Description automatically generated

##### Ưu điểm:

* + - * Bởi vì tính hiệu quả và đơn giản của nó, không đòi hỏi sức mạnh tính toán cao.
      * Dễ thực hiện, dễ hiểu nên được sử dụng rộng rãi bởi các nhà phân tích dữ liệu và nhà khoa học.

##### Nhược điểm:

* + - * Không thể xử lý với số lượng các biến phân loại.
      * Không thể giải quyết vấn đề phi tuyến tính bằng hồi quy logistic.
      * Không hoạt động tốt với các biến độc lập không tương quan với biến mục tiêu.

### Phương pháp đánh giá mô hình dự đoán

### Accuracy: Chỉ đơn giản đánh giá mô hình dự báo thường xuyên dự đoán đúng đến mức nào. Accuracy là tỉ lệ giữa số điểm dữ liệu dự đoán đúng và tổng số điểm dữ liệu. Tuy nhiên giá trị accuracy có giá trị cao chưa hẳn đã tốt bởi accuracy sẽ bị hạn chế trên các tập dữ liệu không cân bằng (imbalanced dataset) nghĩa là có sự chênh lệnh lớn giữa số lượng phần tử trong mỗi lớp.(Số bệnh nhân không bị đột quỵ nhiều hơn hẳn số không bị đột quỵ).

### Text Description automatically generated with medium confidence

### Confusion Matrix: Là một kỹ thuật đánh giá hiệu năng của mô hình dự đoán cho các bài toán phân lớp thể hiện số lượng điểm dữ liệu thuộc vào 1 lớp và được dự đoán thuộc về 1 lớp. Confusion Matrix cung cấp thêm thông tin về tỉ lệ phân lớp đúng giữa các lớp, hay giúp phát hiện các lớp có tỉ lệ phân lớp nhầm cao nhờ vào các khái niệm True (False), Positive (Negative).

### Confusion Matrix for Your Multi-Class Machine Learning Model | by Joydwip Mohajon | Towards Data Science

### True Positive (TP): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Positive (dự đoán đúng).

### True Negative (TN): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Negative (dự đoán đúng).

### False Positive (FP): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Positive (dự đoán sai) – Type I Error.

### False Negative (FN): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Negative (dự đoán sai) – Type II Error.

### Recall: Cho biết có bao nhiêu điểm dữ liệu thực sự ở lớp Positive được mô hình phân lớp đúng trong mọi điểm dữ liệu thực sự ở lớp Positive.

### Text Description automatically generated

### Precision: Cho biết trong số các điểm dữ liệu được mô hình phân loại hay dự đoán vào lớp Positive có bao nhiêu điểm thực sự thuộc về lớp Positive.

### Text Description automatically generated with medium confidence

### Precision và Recall có giá trị trong khoảng [0,1], hai giá trị này càng gần với 1 thì mô hình càng chính xác.

### Precision càng cao đồng nghĩa với các điểm được phân loại càng chính xác.

### Recall càng cao cho thể hiện cho việc ít bỏ sót các điểm dữ liệu đúng.

### F-Score: Một mô hình tốt khi cả Precision và Recall đều cao, thể hiện cho việc mô hình ít phân loại nhầm giữa các lớp cũng như tỉ lệ bỏ sót các đối tượng thuộc lớp cần quan tâm là thấp. Tuy nhiên, hai giá trị Precision và Recall thường không cân bằng với nhau (giá trị này tăng thì giá trị kia thường có xu hướng giảm). Để đánh giá cùng lúc cả Precision và Recall, ta sử dụng độ đo F-Score.

### Text Description automatically generated with low confidence

# CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU

### Đọc dữ liệu bằng Spark

**Bước 1:** Sử dụng lệnh spark.read.format("csv").option("header", "true").load() để đọc tập dữ liệu “healthcare-dataset-stroke-data” có 5110 dòng với 11 thuộc tính dự báo và 1 thuộc tính dùng để phân lớp (1 bị đột quỵ và 0 không bị đột quỵ)



### Đọc dữ liệu đầu vào thực hiện Exploratory Data Analysis (EDA)

**Bước 2**: Hiển thị dữ liệu với lệnh show() và xem cấu trúc dữ liệu của từng thuộc tính bằng lệnh printSchema()

Graphical user interface

Description automatically generated

##### Kết quả:

##### Text Description automatically generated

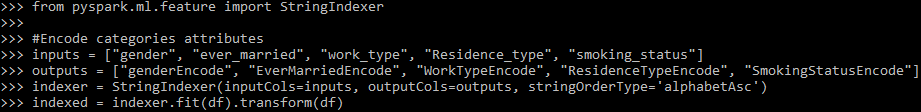
**Bước 3:** Loại bỏ thuộc tính id không mang ý nghĩa cho việc dự báo và phân tích tập dữ liệu



**Bước 4:** Bởi vì việc sử dụng thuật toán Logistic Regression để phân tích tập dữ liệu yêu cầu các biến đầu vào (dự báo) phải mang kiểu số (numeric). Do đó ta cần mã hóa giá trị của các thuộc tính categories sang dạng số.

Graphical user interface, application, table, Excel

Description automatically generated



**Bước 5:** Xóa bỏ các thuộc tính category

Text

Description automatically generated

**Bước 6:** Thuộc tính BMI có nhiều dòng dữ liệu “N/A” sẽ được thay thể bằng giá trị meanText

Description automatically generated

Text

Description automatically generated

**Bước 7:** Chuyển đổi kiểu dữ liệu của các thuộc tính sang Double

Text

Description automatically generated

**Bước 8:** Kiểm tra phân phối của biến phân lớp

Text

Description automatically generated

Chart, bar chart

Description automatically generated

**Bước 9:** Mean, Count, Min, Max, Độ Lệnh Chuẩn của các thuộc tính numeric

A computer screen capture

Description automatically generated with medium confidence

**Bước 10:** Thêm cột X0 và sắp xếp lại thứ tự thuộc tính

Text

Description automatically generated

**Bước 11:** Chia tập dữ liệu 80% để train và 20% để test

Text

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

# CHƯƠNG 3: THUẬT TOÁN KHAI THÁC DỮ LIỆU

### Thuật toán Logistic Regression

**Bước 1:** Viết hàm **sigmoid**. Hàm nhận đầu vào là một giá trị và trả về giá trị nằm trong khoảng từ [0,1].

Text

Description automatically generated

***Hình 29: Hàm sigmoid.***

**Bước 2:** Viết hàm **classification**. Hàm nhận đầu vào là kết quả trả về của hàm sigmoid để phân lớp thuộc về mắc bệnh đột quỵ (1) hay không mắc bệnh đột quỵ (0).

Text

Description automatically generated

***Hình 30: Hàm Classification***

**Bước 3:** Viết hàm **gradient\_descent**. Hàm nhận đầu vào là tập dữ liệu, các giá trị theta khởi tạo ban đầu, giá trị step để tiến đến điểm cực tiểu của Cost Function và số lần lặp tối đa để tìm ra các giá trị theta được trả về.

Text

Description automatically generated

*Hình 31: Hàm gradient\_descent.*

**Bước 4:** Viết hàm **cal\_confusion\_matrix** để so sánh giá trị dự đoán với giá trị thực tế nhằm đánh giá hiệu năng của mô hình dự báo.

Text

Description automatically generated

***Hình 32: Hàm cal\_confusion\_matrix.***

**Bước 5:** Viết hàm **getValueByKey**. Hàm để lấy giá trị của Key trong một mảng các cặp key-value.

Text

Description automatically generated

***Hình 33: Hàm getValueByKey****.*

**Bước 6:** Viết hàm **score**. Hàm nhận đầu vào là tập dữ liệu và các tham số theta để đánh giá độ chính xác của thuật toán và confusion matrix.

Text

Description automatically generated

***Hình 34: Hàm score.***

**Bước 6:** Tiến hành tính các giá trị thete của mô hình dự đoán.

Text

Description automatically generated

Text

Description automatically generated

***Hình 35: Tính các giá trị theta***

**Bước 7:** Đánh giá mô hình dự đoán.

Text

Description automatically generated

**Confusion Matrix**

Chart

Description automatically generated

# CHƯƠNG 4: KẾT QUẢ ĐẠT ĐƯỢC

### Kết quả đạt được

Kết quả mà nhóm đạt được sau khi hoàn thành đồ án:

* + - Hiểu được vấn đề và cách giải quyết bài toán phân loại.
    - Áp dụng được các phương pháp để tiền xử xử lý dữ liệu.
    - So sánh và đánh giá được kết quả các mô hình.
    - Tối ưu hóa được các thuật toán sử dụng
    - Đưa ví dụ thực tế vào trong bài toán.

### So sánh, đánh giá

**Bảng so sánh, đánh giá:**

|  |
| --- |
| **Thuật toán hồi quy Logistic** |
| Giá trị Precision = 0.5 và Recall = 0.017 của mô hình còn thấp do đó độ chính xác của mô hình dự đoán hồi quy Logistic chưa cao. |
| Có 968 bệnh nhân được chuẩn đoán không mắc bệnh đột quỵ đúng với thực tế. |
| Chỉ có 1 bệnh nhân bị chuẩn đoán mắc bệnh đột quỵ đúng so với thực tế. |
| Có 56 bệnh nhân thực tế mắc bệnh đột quỵ nhưng mô hình dự báo Logistic đưa ra không mắc bệnh. (FN) (Type II Error).  Có 1 bệnh nhân thực tế không mắc bệnh đột quỵ nhưng mô hình dự báo kết luận bị đột quỵ. (FP) (Type I Error). |

**Nhận xét:**

Thuật toán hồi quy Logistic cho kết quả dự đoán chưa tối ưu.

### Ưu điểm

**CHƯƠNG 5: KẾT LUẬN**

Những ưu điểm của đồ án:

* + - Áp dụng được nhiều phương pháp so sánh và đánh giá mô hình.
    - Ứng dụng được ví dụ thực tế vào trong đồ án là dự đoán bệnh đột quỵ.

### Hạn chế

Vì thời gian nghiên cứu và hiện thực đề tài còn giới hạn nên đề tài còn có một số nhược điểm, như sau:

* + - Độ chính xác thuật toán không cao.
    - Thời gian thực thi thuận toán chậm.
    - Chưa nghiên cứu tìm ra các thuật toán tối ưu hơn.

### Hướng phát triển

Nhóm sẽ tiếp tục phát triển đồ án để:

* Đối với thuật toán hồi quy Logistic cần sử dụng các thuật toán như: “Conjugate Gradient”, “BFGS”, và “L-BFGS” nhằm tối ưu các tham số theta giúp rút ngắn thời gian tính toán và không cần phải điều chỉnh step hay rate learning để tìm ra theta.
  + - Ứng dụng hệ thống vào các phần mềm chuẩn đoán bệnh đột quỵ.
    - Sử dụng nhiều tập dataset để đánh giá hiệu năng của mô hình.
    - Tiếp tục tối ưu các thuật toán sử dụng để đưa ra kết quả chính xác và thời gian nhanh hơn.
    - Tiếp tục nguyên cứu tìm ra các thuật toán mới tối ưu tốt hơn.
    - Kết hợp các thuật toán lại thành một mô hình hoàn chỉnh.

# CHƯƠNG 6: TÀI LIỆU THAM KHẢO

Link Dataset: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

1. https://www.baeldung.com/cs/gradient-descent-logistic-regression
2. <http://tutorials.aiclub.cs.uit.edu.vn/index.php/2021/05/18/evaluation/>
3. https://www.coursera.org/learn/machine-learning