

# Phân tích hồi quy bộ dữ liệu Diabetes \*

Thái Minh Triết<sup>1</sup>, Chu Hà Thảo Ngân<sup>2</sup>, Võ Tuấn Anh<sup>3</sup>, and Đỗ Trọng Hợp<sup>4</sup>

<sup>1</sup> Trường Đại học Công nghệ thông tin, Đại học quốc gia Thành phố Hồ Chí Minh

<sup>2</sup> Khoa Khoa học và Kỹ thuật Thông tin  
{<sup>1</sup>19522397, <sup>2</sup>19521882, <sup>3</sup>19521226}@gm.uit.edu.vn  
<sup>4</sup>hopdt@uit.edu.vn

Tóm tắt nội dung ...

Keywords: ...

## 1 Giới thiệu chung

## 2 Tổng quan bộ dữ liệu

### 2.1 Giới thiệu chung về bộ dữ liệu

Bộ dữ liệu được xây dựng bao gồm thuộc tính cơ bản của 422 bệnh nhân bệnh tiểu đường và kết quả là một giá trị đáng giá quá trình tiến triển của bệnh nhân sau 1 năm ghi nhận. Dưới đây là ví dụ các điểm dữ liệu được trích ra từ bộ dữ liệu và codebook của bộ dữ liệu.

**Bảng 1.** Ví dụ về các điểm dữ liệu

AGE	SEX	BMI	BF	S1	S2	S3	S4	S5	S6	Y
59	2	32.1	101	157	93.2	38	4	4.8598	87	151
48	1	21.6	87	183	103.2	70	3	3.8918	69	75
72	2	30.5	93	156	93.6	41	4	4.6728	85	141
24	1	25.3	84	198	131.4	40	5	4.8903	89	206
50	1	23	101	192	125.4	52	4	4.2905	80	135

---

\* Hướng dẫn bởi TS. Đỗ Trọng Hợp

**Bảng 2.** Codebook của bộ dữ liệu

Thông tin	Nội dung
Tên bộ dữ liệu	Diabetes dataset
Nguồn thu nhập	<a href="https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset">https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset</a>
Số điểm dữ liệu	422
Số thuộc tính	11
Thông tin thuộc tính	<p><b>AGE:</b> độ tuổi của bệnh nhân</p> <p><b>SEX:</b> giới tính của bệnh nhân</p> <ul style="list-style-type: none"> <li>- Giá trị 1 tương ứng với giới tính là</li> <li>- Giá trị 2 tương ứng với giới tính là</li> </ul> <p><b>BMI:</b> Body Mass Index (chỉ số khối cơ thể)</p> <p><b>BF:</b> average blood pressure (giá trị huyết áp trung bình)</p> <p><b>S1:</b> tổng lượng cholesterol trong huyết thanh (tc - total serum cholesterol)</p> <p><b>S2:</b> giá trị lipoprotein tỷ trọng thấp (ldl - low-density lipoproteins)</p> <p><b>S3:</b> giá trị lipoprotein tỷ trọng cao (hdl - high-density lipoproteins)</p> <p><b>S4:</b> tỉ lệ giữa cholesterol toàn phần so với lượng HDL (tch - total cholesterol / HDL)</p> <p><b>S5:</b> mức triglycerides có trong huyết thanh có thể ghi nhận (ltg - possibly log of serum triglycerides level)</p> <p><b>S6:</b> chỉ số mức đường huyết (glu - blood sugar level)</p> <p><b>Y:</b> giá trị định lượng về tiến triển của bệnh nhân sau 1 năm kể từ thời điểm ghi nhận</p>

## 2.2 Phân tích và trực quan bộ dữ liệu

## 3 Tiền xử lý dữ liệu

### 3.1 Xử lý ngoại lệ

### 3.2 Chuyển đổi dữ liệu

### 3.3 Phân chia tập train và test

## 4 Phân tích hồi quy

### 4.1 Tổng quan về Phân tích Hồi quy (Regression Analysis)

### 4.2 Phân tích ảnh hưởng của yếu tố

### 4.3 Tương tác giữa các yếu tố

### 4.4 Các độ đo đánh giá mô hình hồi quy

### 4.5 Xây dựng mô hình hồi quy

#### Simple Linear Regression

#### Multiple Linear Regression

#### Polynomial Regression

#### Elastic Net Regression

#### Ridge Regression

#### Lasso Regression

## 5 Đánh giá mô hình hồi quy

### 5.1 Kết quả đánh giá mô hình

### 5.2 Mô hình tốt nhất

## 6 Kết luận

## Tài liệu

## 7 First Section

### 7.1 A Subsection Sample

Please note that the first paragraph of a section or subsection is not indented. The first paragraph that follows a table, figure, equation etc. does not need an indent, either.

Subsequent paragraphs, however, are indented.

**Sample Heading (Third Level)** Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

*Sample Heading (Fourth Level)* The contribution should contain no more than four levels of headings. Table 3 gives a summary of all heading levels.

**Bảng 3.** Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	<b>Lecture Notes</b>	14 point, bold
1st-level heading	<b>1 Introduction</b>	12 point, bold
2nd-level heading	<b>2.1 Printing Area</b>	10 point, bold
3rd-level heading	<b>Run-in Heading in Bold.</b> Text follows	10 point, bold
4th-level heading	<i>Lowest Level Heading.</i> Text follows	10 point, italic

Displayed equations are centered and set on a separate line.

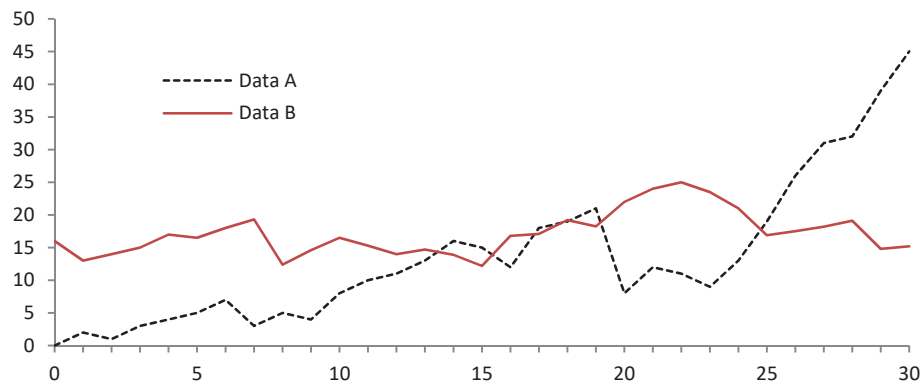
$$x + y = z \tag{1}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

**Theorem 1.** *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

*Chứng minh.* Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], and a homepage [5]. Multiple citations are grouped [1–3], [1, 3–5].



**Hình 1.** A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

## Tài liệu

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). <https://doi.org/10.1007/1234567890>
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, <http://www.springer.com/lncs>. Last accessed 4 Oct 2017