Phân tích hồi quy bộ dữ liệu Diabetes *

Thái Minh Triết[1], Chu Hà Thảo Ngân[2], Võ Tuấn Anh[3], and Đỗ Trọng Hợp[4]

[1] Trường Đại học Công nghệ thông tin, Đại học quốc gia Thành phố Hồ Chí Minh
[2] Khoa Khoa học và Kỹ thuật Thông tin
{[1]19522397,[2]19521882,[3]19521226}@gm.uit.edu.vn
[4]hopdt@uit.edu.vn

Tóm tắt nội dung Bệnh đái tháo đường theo giới y khoa là một trong những mối đe doạ sức khoẻ toàn cầu đối với cộng đồng có tỉ lệ người mắc ngày càng tăng đều theo từng năm với nhiều nguyên nhân ảnh hưởng, trong đó nguyên nhân chính là rối loạn chức năng chuyển hóa insulin. Nhằm tìm hiểu được các yếu tố bên trong cơ thể có tác động như thế nào đến sự rối loạn này, nhóm chúng tôi đã tìm hiểu các chỉ số y tế mang tính thống kê có liên quan bằng bộ dữ liệu có sẵn và tiến hành phân tích, xử lí và đánh giá để có được những kết quả phù hợp cho mục đích của bài nghiên cứu. Trong quá trình phân tích có sử dụng phương pháp thống kê phân tích hồi quy để tìm ra sự ảnh hưởng và tương tác lẫn nhau của các yếu tố này. Sau khi thực hiện xây dựng các mô hình máy học phù hợp và đánh giá chúng, kết quả thu được là <KẾT QUẢ ĐÁNH GIÁ> cho mô hình phân tích tốt nhất <MÔ HÌNH TỐT NHẤT>. <Lược ý kết luận>

Keywords: Đái tháo đường · Bệnh tiểu đường · Phân tích hồi quy

1 Giới thiệu chung

Đái tháo đường được xem là một căn bệnh phổ biến trong cộng đồng, đặc biệt là Mỹ với xu hướng gia tăng ngày càng cao với con số ước tính lên đến 79 triệu người trưởng thành mắc bệnh và trong đó 50% người mắc đái tháo đường không biết mình có bệnh vì chưa có triệu chứng rõ ràng.

Đái tháo đường là một căn bệnh liên quan đến sự rối loạn chuyển hóa trong hormone tuyến tuỵ dẫn đến thiếu hụt lượng insuline trong máu. Các tiêu chuẩn để xác định, chuẩn đoán bệnh đái tháo đường hiện nay được Hiệp hội Đái tháo đường Mỹ công bố cần dựa vào nhiều chỉ số và quá trình phức tạp. Ví dụ khi kiểm tra chỉ số HbA1c của bệnh nhân cần được thực hiện ở phòng thí nghiệm được chuẩn hóa cao. Tuy nhiên, cũng nhờ đó mà việc chuẩn đoán đái tháo đường hay tiền đái tháo đường sẽ phần lớn phụ thuộc vào các chỉ số này từ cơ thể. Chính vì đó, ta cần phải có được một bộ dữ liệu với các chỉ số y tế liên quan để có thể đánh giá phần nào tình trạng hiện tại của bệnh, giúp tăng hiệu quả cũng như giảm đi chi phí cho quá trình chuẩn đoán.

* Hướng dẫn bởi TS. Đỗ Trọng Hợp

Trong bài nghiên cứu này, bộ dữ liệu thu thập được từ các bệnh nhân đái tháo đường với các chỉ số y tế liên quan sẽ được đánh giá bằng các mô hình phân tích hồi quy. Từ đó không chỉ hỗ trợ tích cực cho việc chuẩn đoán mà còn có thể phân tích được sự ảnh hưởng và tương tác giữa các chỉ số trên với nhau.

Cấu trúc bài báo cáo được trình bày như sau: Ở mục 2, chúng tôi trình bày thông tin chi tiết của bộ dữ liệu được sử dụng trong bài nghiên cứu. Ở mục 3, chúng tôi tiếp cận bộ dữ liệu và thực hiện tiền xử lý dữ liệu. Sau khi bộ dữ liệu đã được xử lí, chúng sẽ được đưa vào phân tích bằng các mô hình hồi quy khác nhau, quá trình này sẽ được trình bày ở mục 4. Mục 5 sẽ trình bày kết quả đánh giá các mô hình, từ đó có được mô hình tốt nhất trên bộ dữ liệu. Và cuối cùng ở mục 6 là kết luận và hướng phát triển.

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

Bảng 2. Codebook của bộ dữ liệu

Thông tin	**Nội dung**
Tên bộ dữ liệu	Diabetes dataset
Nguồn thu nhập	`https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset`
Số điểm dữ liệu	442
Số thuộc tính	11
Thông tin thuộc tính	**AGE:** độ tuổi của bệnh nhân **SEX:** giới tính của bệnh nhân - Giá trị 1 tương ứng với giới tính là Nam - Gái trị 2 tương ứng với giới tính là Nữ **BMI:** Body Mass Index (chỉ số khối cơ thể) **BP:** average blood pressure (giá trị huyết áp trung bình) **S1:** tổng lượng cholesterol trong huyết thanh (tc - total serum cholesterol) **S2:** giá trị lipoprotein tỷ trọng thấp (ldl - low-density lipoproteins) **S3:** giá trị lipoprotein tỷ trọng cao (hdl - high-density lipoproteins) **S4:** tỉ lệ giữa cholesterol toàn phần so với lượng HDL (tch - total cholesterol / HDL) **S5:** mức triglyceries có trong huyết thanh có thể ghi nhận (ltg - possibly log of serum triglycerides level) **S6:** chỉ số mức đường huyết (glu - blood sugar level) **Y:** giá trị định lượng về tiến triển của bệnh nhân sau 1 năm kể từ thời điểm ghi nhận

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)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

Displayed equations are centered and set on a separate line.

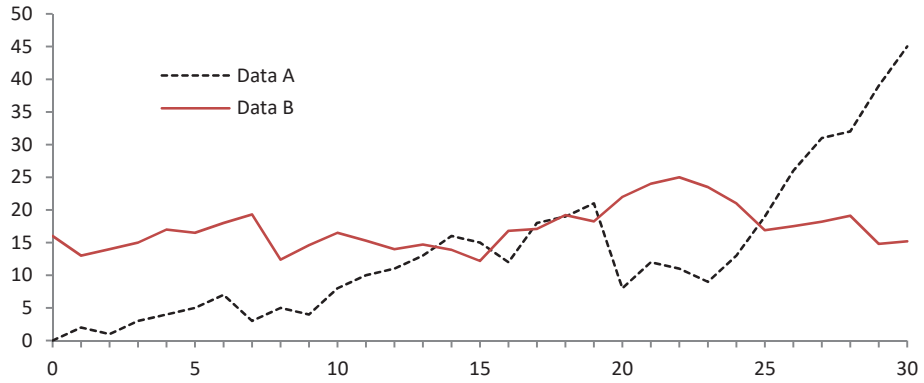
$$x + y = z \tag{1}$$

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/10.10007/1234567890
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