5



**UỶ BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

—🙐🕮🙐—



**MÔN HỌC: PHÂN TÍCH DỮ LIỆU**

**BÀI TẬP NHÓM**

**ĐỀ TÀI: PHÂN TÍCH KHÁM PHÁ VỀ BỆNH ĐÁI THÁO ĐƯỜNG**

**THÀNH VIÊN**

LƯƠNG THANH TUẤN – 3122410447

**GIẢNG VIÊN HƯỚNG DẪN: ĐỖ NGỌC TÀI**

**TP. HỒ CHÍ MINH, 9/2025**

**MỤC LỤC**

[DANH MỤC HÌNH ẢNH 1](#_Toc209733609)

[DANH MỤC BẢNG BIỂU 2](#_Toc209733610)

[LỜI CẢM ƠN 3](#_Toc209733611)

[I. GIỚI THIỆU CHUNG 4](#_Toc209733612)

[1. Lý do chọn đề tài 4](#_Toc209733613)

[2. Mục tiêu xây dựng đề tài 5](#_Toc209733614)

[3. Bảng phân công công việc 5](#_Toc209733615)

[3.1. Bảng tiến độ làm việc 5](#_Toc209733616)

[3.2. Bảng phân công chi tiết 5](#_Toc209733617)

[II. PHÂN TÍCH DỮ LIỆU 6](#_Toc209733618)

[1. Lấy dữ liệu 6](#_Toc209733619)

[2. Khám phá dữ liệu 6](#_Toc209733620)

[3. Kiểm tra tính toàn vẹn của dữ liệu 7](#_Toc209733621)

[4. Các tính chất thống kê dữ liệu số 7](#_Toc209733622)

[4.1. Các cột chiếm tỷ lệ giá trị 0 cao 8](#_Toc209733623)

[5. Tần số xuất hiện trên dữ liệu phân lớp 10](#_Toc209733624)

[6. Mối liên hệ giữa các tính chất 10](#_Toc209733625)

[7. Hiển thị dữ liệu 11](#_Toc209733626)

[III. Các câu hỏi khám phá dữ liệu 12](#_Toc209733627)

[1. Các cột nào ảnh hưởng lớn đến tỷ lệ mắc bệnh? 12](#_Toc209733628)

[2. Glucose tăng thì các biến nào sẽ có xu hướng gia tăng, vì sao? 13](#_Toc209733629)

[3. So sánh phụ nữ mang bầu hay không mang bầu sẽ tỷ lệ mắc bệnh hơn? 14](#_Toc209733630)

[4. Độ tuổi nào thì sẽ có tỷ lệ mắc bệnh cao? 15](#_Toc209733631)

[TÀI LIỆU THAM KHẢO 16](#_Toc209733632)

# 

# DANH MỤC HÌNH ẢNH

[Hình 1. Thông tin dữ liệu 6](#_Toc209733445)

[Hình 2. Tính toàn vẹn của dữ liệu 7](#_Toc209733446)

[Hình 3. Tính chất thống kê dữ liệu số 7](#_Toc209733447)

[Hình 4. Tính chất thống kê dữ liệu số 8](#_Toc209733448)

[Hình 5. Xử lý cột chiếm tỷ lệ cao 8](#_Toc209733449)

[Hình 6. Xử lý cột chiếm tỷ lệ thấp 8](#_Toc209733450)

[Hình 7. Tính chất thống kê dữ liệu số sau khi xử lý 9](#_Toc209733451)

[Hình 8. Tỷ lệ sau khi xử lý 9](#_Toc209733452)

[Hình 9. Tần số xuất hiện 10](#_Toc209733453)

[Hình 10. Mối liên hệ giữa các biến 10](#_Toc209733454)

[Hình 11. Độ phân tán của các cột 11](#_Toc209733455)

[Hình 12. Phân phối trên Histogram 11](#_Toc209733456)

[Hình 13. Mức độ ảnh hưởng với Outcome 12](#_Toc209733457)

[Hình 14. Phân bố ảnh hưởng với Glucose 13](#_Toc209733458)

[Hình 15. So sánh tỷ lệ chưa/đã mang bầu 14](#_Toc209733459)

[Hình 16. So sánh tỷ lệ mắc bệnh theo nhóm tuổi 15](#_Toc209733460)

# DANH MỤC BẢNG BIỂU

[Bảng 1. Bảng tiến độ làm việc 3](#_Toc209733483)

[Bảng 2. Bảng phân công chi tiết 3](#_Toc209733484)

[Bảng 3. Bảng mô tả chi tiết dữ liệu 4](#_Toc209733485)

# 

# 

# LỜI CẢM ƠN

Nhóm em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đến thầy Đỗ Ngọc Tài, giảng viên khoa Công nghệ Thông tin – Trường Đại học Sài Gòn, vì đã tận tình giảng dạy và hướng dẫn nhóm em trong suốt quá trình học tập cũng như hoàn thành bài báo cáo cuối kỳ này. Những kiến thức chuyên môn cùng với những kinh nghiệm thực tế mà thầy đã truyền đạt cho chúng em trong suốt khóa học không chỉ là nền tảng vững chắc giúp nhóm em hoàn thành tốt bài báo cáo, mà còn là nguồn động lực để chúng em tiếp tục rèn luyện, phát triển trên con đường học tập và nghề nghiệp sau này.

Trong quá trình thực hiện bài báo cáo, thầy đã không chỉ giúp nhóm em định hướng đúng đắn, giải quyết những khó khăn vướng mắc, mà còn truyền đạt những kỹ năng quý báu trong cách tiếp cận vấn đề và phương pháp làm việc hiệu quả. Nhờ có sự chỉ bảo tận tình và sự quan tâm sát sao của thầy, nhóm em đã có cơ hội học hỏi thêm nhiều kiến thức mới, từ lý thuyết cho đến thực hành. Những gì nhóm em học được không chỉ nằm trong khuôn khổ của bài báo cáo này, mà còn là hành trang cần thiết để ứng dụng vào các dự án thực tế và trong suốt quá trình làm việc sau này.

Dù đã cố gắng hết sức, nhưng với sự hạn chế về kiến thức chuyên môn và kinh nghiệm thực tiễn, nhóm em khó tránh khỏi những thiếu sót trong bài báo cáo. Nhóm em xin kính mong thầy thông cảm và bỏ qua cho những sai sót đó. Nhóm em rất mong nhận được những góp ý, nhận xét từ thầy, không chỉ để hoàn thiện tốt hơn bài báo cáo này mà còn giúp nhóm em rút ra những bài học quý báu, bổ sung và củng cố thêm những kỹ năng, kiến thức cần thiết để chuẩn bị tốt hơn cho những dự án, bài tập lớn và đặc biệt là bài báo cáo tốt nghiệp sắp tới.

Một lần nữa, nhóm em xin gửi lời cảm ơn sâu sắc đến thầy Đỗ Ngọc Tài. Kính chúc thầy luôn dồi dào sức khỏe, hạnh phúc và thành công trong sự nghiệp giảng dạy, tiếp tục dẫn dắt nhiều thế hệ sinh viên trên con đường chinh phục tri thức và vươn tới những thành công mới. Chúng em luôn trân trọng và biết ơn sự tận tâm của thầy.

*Nhóm em trân thành cảm ơn thầy!*

# GIỚI THIỆU CHUNG

## Lý do chọn đề tài

Mặc dù đề tài này do giảng viên phân công, việc thực hiện phân tích khám phá dữ liệu (EDA) cho bộ dữ liệu Đái tháo đường của NIDDK – thu thập trên phụ nữ người Pima (Akimel O’odham) từ 21 tuổi trở lên tại Phoenix, Arizona (Hoa Kỳ) – vẫn có ý nghĩa học thuật và thực tiễn rõ rệt. Đái tháo đường là bệnh mạn tính phổ biến với gánh nặng biến chứng cao; việc khảo sát các biến lâm sàng nền tảng (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age) giúp nhận diện mẫu hình nguy cơ ban đầu và nâng cao hiểu biết về dữ liệu sức khỏe ở cấp độ mô tả.

Đây là bộ dữ liệu chuẩn mực, quy mô gọn, được cộng đồng dùng rộng rãi, thuận tiện cho tái lập thí nghiệm và so sánh, đồng thời phù hợp mục tiêu học phần là rèn luyện năng lực EDA: mô tả phân phối, trực quan hóa, phát hiện ngoại lệ, kiểm tra chất lượng dữ liệu, phân tích tương quan và diễn giải insight. Đặc biệt, sự hiện diện của các giá trị 0 bất thường ở một số biến đo sinh học tạo bối cảnh thực tế để người học đưa ra quyết định xử lý dữ liệu (coi như thiếu, thay thế bằng trung vị/giá trị hợp lý, hoặc áp dụng kỹ thuật bù khuyết thích hợp), từ đó rèn tư duy phản biện và hiểu tác động của tiền xử lý lên kết quả phân tích.

Bên cạnh giá trị học thuật, đề tài cũng nhấn mạnh ràng buộc phạm vi và đạo đức: mẫu chỉ gồm phụ nữ người Pima tại một khu vực cụ thể nên không suy rộng tùy tiện cho quần thể khác; các phân tích chỉ nhằm mục đích nghiên cứu–giáo dục, không thay thế chẩn đoán y khoa và cần tôn trọng bối cảnh cộng đồng bản địa. Hoàn thành đề tài kỳ vọng đem lại một quy trình EDA chặt chẽ, có thể tái sử dụng, bộ chỉ báo mô tả/biểu đồ minh họa cấu trúc–chất lượng dữ liệu, cùng khuyến nghị tiền xử lý sẵn sàng cho các bước tiếp theo, qua đó đáp ứng đúng trọng tâm kỹ năng của học phần dù không phải đề tài do người học tự lựa chọn.

## Mục tiêu xây dựng đề tài

* Mục tiêu chung
* Phân tích khám phá về bệnh đái tháo đường.
* Mục tiêu cụ thể
* Nhìn tổng quan về dữ liệu và ý nghĩa của các cột
* Đặt ra các câu hỏi để khai phá dữ liệu và trả lời dựa vào biểu đồ
* Xây dựng câu chuyện về dữ liệu

## Bảng phân công công việc

### Bảng tiến độ làm việc

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **MÃ SV** | **HỌ TÊN** | **Email** | **TỈ LỆ PHẦN TRĂM** |
| 1 | 3122410447 | Lương Thanh Tuấn | aquattda@gmail.com | 100% |

Bảng 1. Bảng tiến độ làm việc

### Bảng phân công chi tiết

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **HỌ TÊN** | **CÔNG VIỆC** | **TRẠNG THÁI** |
| 1 | Lương Thanh Tuấn | Xây dựng quy trình khai phá dữ liệu | Hoàn thành |
| Viết báo cáo | Hoàn thành |
| Tìm hiểu và đặt câu hỏi | Hoàn thành |
| Xác định vấn đề | Hoàn thành |

Bảng 2. Bảng phân công chi tiết

# PHÂN TÍCH DỮ LIỆU

## Lấy dữ liệu

* Bộ dữ liệu được thầy Đỗ Ngọc Tài cung cấp, bao gồm các paper về căn bệnh đái tháo đường.

## Khám phá dữ liệu

* **Thông tin dữ liệu:**

A screenshot of a computer

AI-generated content may be incorrect.

Hình 1. Thông tin dữ liệu

* Bộ dữ liệu bao gồm 768 dòng, 9 cột. Trong đó có cột BMI và DiabetesPedigreeFunction có kiểu dữ liệu float, còn lại là int. Tất cả các cột đều không Null.
* Dữ liệu có 8 tính chất để phân lớp: Pregnancies, Glucose(mg/dL), BloodPressure(mmHg), SkinThickness(mm), Insulin(U/mL), BMI(kg/m²), DiabetesPedigreeFunction và Age.
* Dữ liệu để phân lớp ở cột Outcome.
* **Mô tả chi tiết dữ liệu:**

|  |  |
| --- | --- |
| Tên cột | Ý nghĩa |
| Pregnancies | Số lần mang thai |
| Glucose | Đường huyết 2h (mg/dL) |
| BloodPressure | Huyết áp tâm trương (mmHg) |
| SkinThickness | Bề dày da tam đầu (mm) |
| Insulin | Insulin huyết tương 2h (U/mL) |
| BMI | Chỉ số khối cơ thể (kg/m²) |
| DiabetesPedigreeFunction | Chỉ số phả hệ (nguy cơ di truyền) |
| Age | Tuổi (năm) |
| Outcome | Phân loại mắc bệnh |

Bảng 3. Bảng mô tả chi tiết dữ liệu

## Kiểm tra tính toàn vẹn của dữ liệu

A screen shot of a computer program

AI-generated content may be incorrect.

Hình 2. Tính toàn vẹn của dữ liệu

* Dữ liệu không bị trùng và không có giá trị rỗng (NaN, Null).

## Các tính chất thống kê dữ liệu số

A screenshot of a computer screen

AI-generated content may be incorrect.

Hình 3. Tính chất thống kê dữ liệu số

* Các tính chất Glucose, BloodPressure, SkinThickness, Insulin, BMI xuất hiện giá trị 0 bất thường.

### 4.1. Các cột chiếm tỷ lệ giá trị 0 cao

A screen shot of a computer program

AI-generated content may be incorrect.

Hình 4. Tính chất thống kê dữ liệu số

- Xuất hiện nhiều values 0 có thể là "missing trá hình" hoặc missing data đối với SkinThickness và Insulin:

* Khó/không đo được ở lần khám (đo nếp gấp da cần kẹp caliper đúng vị trí; người đo chưa thành thạo; bệnh nhân không hợp tác…).
* Một số phòng xét nghiệm (ví dụ Insulin) không thực hiện quy trình này.

- Bản chất: các số 0 này là “giá trị thiếu được mã hoá thành 0” (disguised missing data) chứ không phải đo lường hợp lý về sinh học.

* Với các cột chiếm tỷ lệ cao sẽ thay bằng NaN:

A screen shot of a computer code

AI-generated content may be incorrect.

Hình 5. Xử lý cột chiếm tỷ lệ cao

* Với các cột chiếm tỷ lệ thấp sẽ thay bằng median:

A screen shot of a computer

AI-generated content may be incorrect.

Hình 6. Xử lý cột chiếm tỷ lệ thấp

Kết quả sau khi xử lý:

A screenshot of a black screen

AI-generated content may be incorrect.

Hình 7. Tính chất thống kê dữ liệu số sau khi xử lý

A black screen with white text

AI-generated content may be incorrect.

Hình 8. Tỷ lệ sau khi xử lý

* So với trước khi thay 0 thành NaN và median thì mean của các cột chiếm tỷ lệ 0 thay đổi. Nổi bật nhất là SkinThickness tăng từ ~20.5 → 29.15 và Insulin từ ~79.8 → 155.55. Điều này chứng tỏ giá trị 0 đã kéo lệch trung bình rất mạnh.
* Dữ liệu không còn phi lâm sàng nữa.

## Tần số xuất hiện trên dữ liệu phân lớp

A screen shot of a computer

AI-generated content may be incorrect.

Hình 9. Tần số xuất hiện

* Dữ liệu phục vụ bài toán phân loại nhị phân (mắc/không mắc đái tháo đường).
* Phân bố lớp mất cân bằng, trong đó nhóm không mắc chiếm tỷ trọng lớn khoảng gấp đôi nhóm mắc.

## Mối liên hệ giữa các tính chất

A screenshot of a graph

AI-generated content may be incorrect.

Hình 10. Mối liên hệ giữa các biến

* Glucose tương quan cao với Outcome. Cho thấy biến này có ảnh hưởng đối với căn bệnh.
* Ngoài ra còn Insulin và BMI cũng khá tương quan với Outcome.

## Hiển thị dữ liệu

A diagram of a graph

AI-generated content may be incorrect.

Hình 11. Độ phân tán của các cột

- Độ trải rộng giữa các tính chất khác nhau hoàn toàn.

- Độ lệch đa số sẽ phân phối lệch phải và theo hướng giá trị lớn.

- Phân bố giá trị của Glucose cân bằng.

A graph of blood glucose and insulin

AI-generated content may be incorrect.

Hình 12. Phân phối trên Histogram

- Hình dạng phân phối của Insulin có "đuôi" dài tới 800.

- Đối với Glucose, BMI, BloodPresure đơn đỉnh, có chú lệt phải nhẹ.

# Các câu hỏi khám phá dữ liệu

## Các cột nào ảnh hưởng lớn đến tỷ lệ mắc bệnh?

A chart with different colored bars

AI-generated content may be incorrect.

Hình 13. Mức độ ảnh hưởng với Outcome

- Glucose có mỗi tương quan cao nhất trong các features.

- Tiếp sau đó là BMI, Insulin, SkinThicknes và Age cũng tương quan tương đối

## Glucose tăng thì các biến nào sẽ có xu hướng gia tăng, vì sao?

A group of blue dots

AI-generated content may be incorrect.

Hình 14. Phân bố ảnh hưởng với Glucose

* Đa số các biến đều tương quan khá mạnh với Glucose. Điều này càng làm rõ tầm quan trọng của Glucose ảnh hưởng tương đối với các biến khác.
* Thể hiện rõ ràng cho tương quan mạnh đó là các biến Insulin, BMI và Age.

**Giải thích:**

- Insulin và Glucose tương quuan dương rõ rệt do được đo cùng thời điểm 2h sau nghiệm pháp dung nạp glucose.

- Với độ tuổi cao → dung nạp glucose kém, giảm nhạy insulin → Sau phương pháp dung nạp thì sẽ có xu hướng cao.

- BMI cao phản ánh mỡ trên cơ thể cao → dễ kháng insulin ⇒ nên Glucose ảnh hưởng theo do nghiệm pháp dung nạp glucose.

## So sánh phụ nữ mang bầu hay không mang bầu sẽ tỷ lệ mắc bệnh hơn?

A graph with blue squares

AI-generated content may be incorrect.

Hình 15. So sánh tỷ lệ chưa/đã mang bầu

* Tỷ lệ mắc bệnh giữa nhóm chưa từng mang thai và đã từng mang thai gần như tương đương, cho thấy không có khác biệt đáng kể ở mức mô tả. Kết quả này có thể bị ảnh hưởng bởi yếu tố nhiễu như tuổi và cách gộp nhóm.

## Độ tuổi nào thì sẽ có tỷ lệ mắc bệnh cao?

A graph with numbers and a bar

AI-generated content may be incorrect.

Hình 16. So sánh tỷ lệ mắc bệnh theo nhóm tuổi

* Dễ dàng thấy được các nhóm tuổi từ 30 trở đi tỷ lệ mắc bệnh cao.
* Trong khi đó thì với mức tuổi 25-29 cũng khá cao.

# TÀI LIỆU THAM KHẢO

[1] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2017.

[2] W. Shields, *SQL QuickStart Guide: The Simplified Beginner’s Guide to Managing, Analyzing, and Manipulating Data With SQL*, 1st ed. ClydeBank Media LLC, 2019.

[3] M. Shron, *Thinking with Data: How to Turn Information into Insights*, 1st ed. O'Reilly Media, 2014.

[4] C. N. Knaflic, *Storytelling with Data: A Data Visualization Guide for Business Professionals*, 1st ed. Wiley, 2015.

[5] A. Maheshwari, *Data Analytics Made Accessible*, 1st ed. DMM Direct, 2014.

[6] J. VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly Media, 2016.

[7] H. Chen, *Practical Data Science with Python*, Apress, 2019.

[8] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed., O'Reilly Media, 2019.

[9] N. C. Zakas, *Understanding ECMAScript 6: The Definitive Guide for JavaScript Developers*, No Starch Press, 2016.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.

[11] R. L. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. Namburu, *Data Mining for Scientific and Engineering Applications*, Springer, 2001.

[12] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.

[13] J. Heaton, *Deep Learning and Neural Networks: Python, Theano, and Keras*, Heaton Research, 2018.