PHÂN TÍCH KHÁM PHÁ VỀ BỆNH ĐÁI THÁO ĐƯỜNG

Sử dụng Database Pima Indians Diabetes UCI

MỤC LỤC

- 3.2.4 SkinThickness (Độ dày da)
- 3.2.5 Insulin (Nồng độ insulin)
- 3.2.6 BMI (Chỉ số khối cơ thể)
- 3.2.7 DiabetesPedigreeFunction (Yếu tố di truyền)
- 3.2.8 Age (Tuổi)
- 3.2.9 Outcome (Kết quả chẩn đoán)

- 3.3 Thống kê mô tả cơ bản
- 3.4 Kiểm tra chất lượng dữ liệu

4. PHÂN TÍCH KHÁM PHÁ DỮ LIỆU (EDA)

4.1 PHÂN TÍCH ĐƠN BIẾN

- 4.1.1 Phân bố của từng biến số
- 4.1.2 Phát hiện các giá trị ngoại lệ và bất thường
- 4.1.3 Xử lý giá trị thiếu và giá trị zero

4.2 PHÂN TÍCH BIẾN MỤC TIÊU

- 4.2.1 Tỷ lệ mắc bệnh đái tháo đường trong dataset
- 4.2.2 Đặc điểm nhóm bệnh nhân và nhóm khỏe mạnh

4.3 PHÂN TÍCH THEO NHÓM TUỔI

- 4.3.1 Phân bố bệnh theo độ tuổi
- 4.3.2 Các chỉ số sinh lý thay đổi theo tuổi
- 4.3.3 Yếu tố nguy cơ ở từng nhóm tuổi

4.4 PHÂN TÍCH THEO CHỈ SỐ BMI

- 4.4.1 Phân loại BMI và tỷ lệ mắc bệnh
- 4.4.2 Mối quan hệ giữa béo phì và đái tháo đường
- 4.4.3 BMI kết hợp với các yếu tố khác

4.5 PHÂN TÍCH GLUCOSE VÀ INSULIN

- 4.5.1 Phân bố nồng độ glucose
- 4.5.2 Mối quan hệ glucose-insulin
- 4.5.3 Ngưỡng glucose trong chẩn đoán (so với tài liệu nghiên cứu)

4.6 PHÂN TÍCH YẾU TỐ DI TRUYỀN

- 4.6.1 Hàm phả hệ đái tháo đường và ý nghĩa
- 4.6.2 Tác động của yếu tố di truyền
- 4.6.3 Kết hợp yếu tố di truyền với các yếu tố khác

4.7 PHÂN TÍCH THEO SỐ LẦN CÓ THAI

- 4.7.1 Mối quan hệ giữa thai kỳ và đái tháo đường
- 4.7.2 Đái tháo đường thai kỳ
- 4.7.3 Tiến triển nguy cơ theo số lần có thai

5. PHÂN TÍCH TƯƠNG QUAN VÀ MỐI QUAN HỆ

- 5.1 Ma trận tương quan
- 5.2 Phân tích các cặp biến quan trọng
- 5.3 Phân tích tầm quan trọng của đặc trưng
- 5.4 Phát hiện đa cộng tuyến

6. PHÂN TÍCH NÂNG CAO

6.1 PHÂN TÍCH PHÂN NHÓM (PHÂN CỤM)

- 6.1.1 Xác định các nhóm bệnh nhân tiềm ẩn
- 6.1.2 Đặc điểm của từng cụm
- 6.1.3 Giải thích lâm sàng

6.2 PHÂN TÍCH YẾU TỐ NGUY CƠ

- 6.2.1 Xếp hạng các yếu tố nguy cơ
- 6.2.2 Tạo hệ thống chấm điểm nguy cơ
- 6.2.3 Phân tích ngưỡng cho sàng lọc

6.3 SO SÁNH VỚI TIÊU CHUẨN CHẨN ĐOÁN

- 6.3.1 Tiêu chí chẩn đoán của WHO/ADA
- 6.3.2 Độ chính xác của các giá trị cắt
- 6.3.3 Cân bằng giữa độ nhạy và độ đặc hiệu

7. MÔ HÌNH DỰ ĐOÁN SƠ BỘ

- 7.1 Các mô hình cơ sở
- 7.2 Lựa chọn và thiết kế đặc trưng
- 7.3 Đánh giá hiệu suất mô hình
- 7.4 So sánh với thuật toán ADAP (từ Paper 2)

8. THẢO LUẬN VÀ HIỂU BIẾT SÂU SẮC

8.1 CÁC PHÁT HIỆN CHÍNH

- 8.1.1 Yếu tố nguy cơ quan trọng nhất
- 8.1.2 Các mẫu và xu hướng đáng chú ý
- 8.1.3 Những phát hiện bất ngờ

8.2 SO SÁNH VỚI TÀI LIỆU NGHIÊN CỨU

- 8.2.1 Tính nhất quán với các nghiên cứu trước
- 8.2.2 Những phát hiện mới từ phân tích khám phá dữ liệu
- 8.2.3 Ý nghĩa lâm sàng

8.3 HẠN CHẾ VÀ XU HƯỚNG

- 8.3.1 Hạn chế của bộ dữ liệu (đặc thù của người Pima Indians)
- 8.3.2 Xu hướng chọn mẫu và khả năng tổng quát hóa
- 8.3.3 Tác động của dữ liệu thiếu

9. KẾT LUẬN VÀ KHUYẾN NGHỊ

9.1 TÓM TẮT CÁC PHÁT HIỆN

- 9.1.1 Các yếu tố nguy cơ chính được xác định
- 9.1.2 Ngưỡng lâm sàng được đề xuất
- 9.1.3 Những hiểu biết đặc thù về dân số nghiên cứu

9.2 ỨNG DỤNG THỰC TIỄN

- 9.2.1 Quy trình sàng lọc
- 9.2.2 Chiến lược phòng ngừa
- 9.2.3 Hỗ trợ quyết định lâm sàng

9.3 HƯỚNG NGHIÊN CỨU TIẾP THEO

- 9.3.1 Nhu cầu mở rộng bộ dữ liệu
- 9.3.2 Các phương pháp mô hình hóa nâng cao
- 9.3.3 Nghiên cứu theo chiều dọc

10. PHỤ LỤC

- 10.1 Statistical tests results
- 10.2 Additional visualizations
- 10.3 Code snippets (nếu cần)

FRAMEWORK PHÂN TÍCH CHI TIẾT

Biến đầu vào (Input Variables):

1. **Pregnancies** - Số lần có thai (rời rạc)
2. **Glucose** - Nồng độ glucose huyết tương (mg/dL)
3. **BloodPressure** - Huyết áp tâm trương (mmHg)
4. **SkinThickness** - Độ dày nếp gấp da triceps (mm)
5. **Insulin** - Insulin huyết thanh 2 giờ (mu U/ml)
6. **BMI** - Chỉ số khối cơ thể (kg/m^2)
7. **DiabetesPedigreeFunction** - Hàm phả hệ đái tháo đường
8. **Age** - Tuổi (năm)

Biến đầu ra (Output Variable):

- **Outcome** - Kết quả chẩn đoán (0: Không mắc, 1: Mắc đái tháo đường)

Mục tiêu phân tích:

1. **Mô tả**: Hiểu đặc điểm dân số nghiên cứu
2. **Chẩn đoán**: Xác định các mẫu và yếu tố nguy cơ
3. **Dự đoán**: Đánh giá khả năng dự đoán của các biến
4. **Lâm sàng**: Đưa ra những hiểu biết có ý nghĩa y tế

Câu hỏi nghiên cứu chính:

1. Yếu tố nào có tác động mạnh nhất đến nguy cơ mắc đái tháo đường?
2. Có tồn tại các giá trị ngưỡng để sàng lọc không?

3. Các yếu tố có tương tác với nhau như thế nào?
4. Bộ dữ liệu có đại diện cho dân số chung không?
5. Làm thế nào để tối ưu hóa các chiến lược phát hiện sớm?

Phương pháp phân tích:

- **Thống kê mô tả**: Trung bình, trung vị, độ lệch chuẩn, phân vị
- **Trực quan hóa**: Biểu đồ tần suất, biểu đồ hộp, biểu đồ tán xạ, bản đồ nhiệt
- **Kiểm định thống kê**: Kiểm định t, chi-bình phương, phương sai một chiều
- **Phân tích tương quan**: Tương quan Pearson, Spearman
- **Nâng cao**: Phân tích thành phần chính, phân cụm, tầm quan trọng của đặc trưng