INTRODUCTION OF DATA SCIENCE

Giảng viên: TS. Bùi Thanh Hùng
Bộ môn Khoa học dữ liệu, Khoa Công nghệ thông tin
Đại học Công nghiệp thành phố Hồ Chí Minh
Email: buithanhhung@iuh.edu.vn
Website: https://sites.google.com/site/hungthanhbui1980/

Bài 1:

Cho bộ dữ liệu về bệnh tiểu đường ở file đính kèm. Bộ dữ liệu gồm 768 dòng và 9 cột với các cột được mô tả chi tiết như sau:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral
glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Hãy viết code để thực hiện các yêu cầu sau:

1.1 Đọc dữ liệu

1.2 Trực quan hóa bằng biểu đồ histogram, plotbox

1.3 Dùng lệnh để mô tả thông tin chi tiết về dữ liệu với các thống kê: min, max, std, 25%, 50%, 75%

1.4 Cho biết kích thước dữ liệu, có dữ liệu nào trống không, và mô tả loại dữ liệu của từng cột

1.5 Cho biết trong dữ liệu có bao nhiêu lượt mang thai và và bệnh tiểu đưởng theo gợi ý như bảng sau:

Số lượt	Có bệnh	Không có	Tổng
0	4	5	9
1	3	7	10
2	6	5	11

1.6 Lưu kết quả ở câu 1.5 vào 1 file csv đồng thời trực quan hóa kết quả ở câu 1.5 qua các hình, lưu các hình vào các file khác nhau.

1.7 Phân tích chi tiết độ phân bố của bệnh tiểu đường/không có bệnh tiểu đường theo độ tuổi?

1.8 Trực quan hóa kết quả ở câu 7?

1.9 Tính mối tương quan giữa số lần mang thai và bệnh tiểu đường theo độ đo Pearson

1.10 Tính mỗi liên quan giữa các đặc trưng khác với bệnh tiểu đường, cho biết đâu là đặc trưng có ảnh hưởng nhất tới bệnh tiểu đường?