Machine Learning

Bộ môn Khoa học dữ liệu
Khoa Công nghệ thông tin
Trường Đại học Công nghiệp thành phố Hồ Chí Minh-IUH

Bài 1:
Một giám đốc nhân sự mong muốn so sánh tính hiệu quả của hai phương pháp huấn luyện các nhân viên công nghiệp nhằm thực hiện một hoạt động lắp ráp nào đó. Một số lượng nhân viên được chia thành hai nhóm bằng nhau, n hóm thứ nhất nhận được phương pháp huấn luyện 1 và nhóm thứ hai được huấn luyện bằng phương pháp 2. Mỗi nhóm sẽ thực hiện hoạt động lắp ráp này, và độ dài của thời gian lắp ráp sẽ được ghi nhận. Người ta kỳ vọng rằng các đại lượng cho cả hai nhóm sẽ có một khoảng xấp xỉ 8 phút. Để cho ước lượng về sự khác biệt về thời gian trung bình để lắp ráp chính xác trong giới hạn 1 phút với xác suất bằng với 0.95, thì cần phải đưa bao nhiêu công nhân vào mỗi nhóm huấn luyện?

Bài 2:
Một công ty kiểm toán mong muốn ước lượng sai số trung bình mỗi tài khoản trong các khoản phải thu cho một công ty cung cấp hệ thống ống nước chính xác trong giới hạn $20 với xác suất bằng 0.99. Một mẫu nhỏ trước đó gợi ý rằng sai số mỗi tài khoản sở hữu một độ lệch chuẩn xấp xỉ bằng với $58. Nếu công ty kiểm toán đó mong muốn ước lượng sai số trung bình mỗi tài khoản chính xác trong giới hạn $20, thì có bao nhiêu tài khoản ắt sẽ phải được chọn mẫu? Mẫu này phải sở hữu (các) thuộc tính nào?

Bài 3:

Cho dataset điểm PISA năm 2015 của học sinh Việt Nam tham gia kỳ thi đánh giá năng lực Toán học, Đọc hiểu và Khoa học. Bạn hãy thực hiện các yêu cầu sau:

a. Đọc file dữ liệu PISA_VN_2015.csv và thực hiện:

- Cho biết bảng dữ liệu có bao nhiêu dòng và bao nhiêu cột?
- Cho biết thông tin về các cột dữ liệu và kiểu của từng cột trong bảng
- Cho biết các dữ liệu thiếu
- Trực quan hóa dữ liệu bằng ít nhất 3 loại biểu đồ và nhận xét về các biểu đồ đó

b. Tự viết các hàm tính Trung binh (mean), Độ lệch chuẩn (standard deviation) và khoảng dữ liệu (range) và áp dụng các hàm này với cột dữ liệu pv1MATH.

c. Dựa vào dữ liệu cột Gender, bạn hãy thực hiện đếm số lượng học sinh Nam (Boys) và học sinh Nữ (Girls) và vẽ đồ thị Pie để so sánh mối tương quan giữa Boy và Girl.

d. Bạn hãy liệt kê các học sinh có điểm Toán (pv1MATH) cao nhất và các học sinh có điểm Khoa học (pv1SCIE) thấp nhất, lưu kết quả này vào 1 file csv.

e. Phân tích mỗi tương quan giữa các đặc trưng và xác định các đặc trưng quan trọng trong bộ dữ liệu này, trực quan bằng các hình vẽ.

f. Thống kê để hoàn thành bảng số liệu dưới đây, từ đó trực quan hóa dữ liệu Bảng 1 bằng đồ thị Bar.

Bảng 1

Vùng (REGION)	Số học sinh
SOUTH	
CENTRAL	
NORTH	

<u>Bài 4</u>:

Một công ty muốn dự đoán số phút gọi vào tổng đài dịch vụ của khách dựa vào số linh kiện cần sửa chữa (Units), hãy tính β_o và β_1 và xây dựng mô hình hồi quy theo mô hình sau:

Số phút = $\beta_o + \beta_1$Units

Dữ liệu được cho như bảng dưới đây.

STT	Số phút	Units	STT	Số phút	Units
1	49	3	8	145	9
2	64	4	9	154	10
3	23	1	10	166	10
4	29	2	11	97	6
5	87	5	12	109	7
6	96	6	13	111	8
7	74	4	14	149	9

Viết chương trình xây dựng một ứng dụng với chức năng nhập vào một dữ liệu và đưa ra dự đoán số phút gọi vào dịch vụ tổng đài.