Machine Learning

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

Cho bộ dữ liệu ở đường link

Heart Disease - UCI Machine Learning Repository

hãy thực hiện các yêu cầu sau:

1. Sử dụng thống kê mô tả, mô tả về bộ dữ liệu trên với min, max, std, avg,…
2. Trực quan hóa dữ liệu với các biểu đồ grid line, box, histogram, và scatter matrix
3. Cho biết những dữ liệu còn thiếu và đề xuất cách xử lý dữ liệu thiếu đó một cách tự động, lưu bộ dữ liệu đã xử lý dữ liệu thiếu sử dụng cho các ý tiếp theo
4. Chia bộ dữ liệu thành 2 phần Train/Test với các tỉ lệ 8:2 và 7:3
5. Đánh giá bộ dữ liệu bằng phương pháp 5 fold và 10-Fold (k-fold) với giải thuật Bayes, nhận xét về các kết quả đạt được
6. Huấn luyện dữ liệu cho bài toán phân lớp sử dụng với giải thuật Linear Regression với bộ dữ liệu với 2 tỉ lệ ở trên.
7. Huấn luyện dữ liệu cho bài toán phân lớp sử dụng với giải thuật Bayes với bộ dữ liệu với 2 tỉ lệ ở trên.
8. Huấn luyện dữ liệu cho bài toán phân lớp sử dụng với giải thuật SVM với bộ dữ liệu với 2 tỉ lệ ở trên.
9. Thay đổi các tham số của mô hình SVM và Bayes và nhận xét về các kết quả đạt được trên bộ dữ liệu với 2 tỉ lệ ở trên.
10. Tính độ đo F1 score cho các mô hình Linear Regression, SVM, Bayes với bộ dữ liệu với 2 tỉ lệ ở trên, trực quan hóa kết quả đạt được.
11. Tính độ đo Accuracy cho các mô hình Linear Regression, SVM, Bayes với bộ dữ liệu với 2 tỉ lệ ở trên, trực quan hóa kết quả đạt được.
12. Tính độ đo Confusion Matrix cho các mô hình Linear Regression, SVM, Bayes với bộ dữ liệu với 2 tỉ lệ ở trên, trực quan hóa kết quả đạt được.
13. So sánh các kết quả ở câu 10, 11, 12 bằng các biểu đồ, nhận xét về các biểu đồ này
14. Lưu model với giải thuật đạt kết quả tốt nhất ở mỗi tỉ lệ dữ liệu.
15. Xây dựng ứng dụng với đầu vào là 1 dữ liệu hay từ 1 file, in kết quả ra màn hình.