Bài thực hành số 6

Lưu ý: Chủ đề trong diễn đàn nhóm tạo ra dùng để nộp các bài tập, các bài báo cáo, các kết quả thảo luận,.. trong suốt học kỳ. Các nhóm có thể tham khảo nội dung của các nhóm khác nhưng tuyệt đối không sao chép cho các bài kết quả của nhóm mình (đề cao tính trung thực).

Trong các buổi làm việc: yêu cầu làm việc theo nhóm, các nhóm có thể trao đổi, hỗ trợ nhau khi cần thiết. Nhóm có thể mang máy tính cá nhân cùng làm việc.

Câu 1:

Tập tin Collected_Hr_data_performances.xls (trên LMS) chứa các thông tin đánh giá năng lực làm việc của khoảng 1200 nhân viên dựa trên các thông tin về gia đình, học vấn, kinh nghiệm,…Tách thành 2 tập tin: Train và Test để thực hiện các yêu cầu sau:

- Dùng thuật toán Cây Quyết định (J48) với tập dữ liệu Train để phát hiện các luật đánh giá năng lực nhân viên. Thực hiện nhiều lần với các cách huấn luyện, các tham số khác nhau,… để chọn ra một mô hình mà anh chị xem là tốt nhất.
- Quan sát các luật được tạo ra và cho vài nhận xét.
- Sử dụng tập tin Test để kiểm tra xếp loại các nhân viên có đúng/tốt hay không?
- Thực nghiệm lại các câu trên nhưng với thuật toán REP Tree (trong WEKA) – cho phép tỉa cây (giới hạn độ sâu Cây QĐ).

Câu 2:

- Sử dụng tập tin 1-TNTHPT-XepLoai.csv chứa kết quả thi THTHPT 2020 của TpHCM có điểm từng môn thi, điểm TB và kết quả xếp loại để
- Tách thành 2 tập tin Train và Test.
- Dùng thuật toán Cây QĐ để tìm ra mô hình tốt nhất. Xem và nhận xét bộ luật của cây QĐ.
- Thử nghiệm với tập tin Test để xem kết quả.
- Thực hiện tương tự với tập tin 2-TNTHPT-XepLoai.csv chứa kết quả thi THTHPT 2020 của TpHCM có khoảng điểm từng môn thi (đã rời rạc hóa) và kết quả xếp loại
- Cho nhận xét sau khi thực hiện xong câu 2.

Kết quả câu 1 và câu 2 sẽ nộp trong buổi thực hành.

Câu 3:

- Tập tin bank-data.csv trong LMS/Bộ dữ liệu thức hành chứa thông tin về các cá nhân vay tiền ngân hàng. Thông tin bao gồm: mã định danh, tuổi, giới tính, khu vực sinh sống, thu nhập (USD) 1 năm, tình trạng hôn nhân, số con, có xe hơi?, có mở tài khoản tiết kiệm?, có mở tài khoản thanh toán? Có nợ tiền mua nhà?, quyết định cho vay (Yes/No). Sử dụng dữ liệu trong tập tin này để thực hiện các yêu cầu sau:
- Tách thành 2 tập tin Train (80%) và Test (20%)
- Dùng tập tin Train để xây dựng mô hình phân lớp dựa vào thuật toán Cây QĐ cho phép đưa ra quyết định cho vay/không cho vay.
- Dùng tập tin Test để kiểm tra lại kết quả mô hình đã tìm ra.
- Sử dụng lại thuật toán Cây QĐ nhưng qui định chiều sâu của cây lần lượt là 4, 5 và 6. Với mỗi độ sâu: thực nghiệm lại với tập tin Test để thu thập kết quả.

Câu 4: Phần thu hoạch

Phân công thực hiện, thực hiện ở nhà, thời hạn 1 tuần và gửi vào chủ đề trong diễn đàn.

- Báo cáo các kết quả của câu 3 và cho nhận xét.
- Thực hiện câu 3 với chương trình Python (sữ dụng thư viện liên quan đến Machine Learning – sklearn) .

Lưu ý: Nhóm có thể dùng bất kỳ ứng dụng để thực hiện báo cáo (Power Point, Word, Clip/Video,…). Phần thu hoạch này cần nộp lên diễn đàn trước buổi học lý thuyết kế tiếp: GV có thể mời 1 nhóm lên trình bày.