**BÁO CÁO BÀI THỰC HÀNH TỔNG HỢP**

## **Mô tả về bộ dữ liệu**

1. Tên dữ liệu: German Credit data
2. Nguồn: Professor Dr. Hans Hofmann Institut f"ur Statistik und "Okonometrie Universit"at Hamburg FB Wirtschaft swissenschaften Von-Melle-Park 5 2000 Hamburg 13
3. Link: <https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)>
4. Số lượng mẫu: 1000
5. Đặc điểm của tập dữ liệu: đa biến
6. Đặc điểm thuộc tính: Categorical, Integer
7. Bài toán: phân loại
8. Lĩnh vực: tài chính
9. Giá trị bị mất: N/A
10. Ngày đóng góp: 17/11/1994
11. Số lượng dữ liệu đã được dùng: 712481
12. Tóm tắt: German credit là bộ dữ liệu về lịch sử tín dụng của 1000 tài khoản người dùng. Bài toán đặt ra là dựa vào lịch sử tín dụng này để dự đoán mức độ rủi ro tín dụng (khả năng khách hàng sẽ trả nợ hay không). Bộ dữ liệu bao gồm 20 thuộc tính (biến đầu vào), trong đó có 7 biến liên tục (numeric) và 13 biến phân loại (category).
13. Diễn giải các thuộc tính

*1. Trạng thái tài khoản hiện có.*

*2. Kỳ hạn theo tháng (Duration)*

*3. Lịch sử tín dụng*

*4. Mục đích*

*5. Dư nợ tín dụng (Credit amount)*

*6. Tài khoản tiết kiệm (Savings account)*

*7. Số năm làm việc liên tục gần nhất tới thời điểm hiện tại*

*8. Tỷ lệ trả góp (Installment rate) theo thu nhập khả dụng*

*9. Giới tính và trạng thái hôn nhân*

*10. Những người nợ khác*

*11. Thời gian cư trú tại nơi ở hiện tại*

*12. Bất động sản*

*13. Tuổi*

*14. Trả góp khác*

*15. Nhà ở*

*16. Số lượng khoản vay tại ngân hàng*

*17. Nghề nghiệp*

*18. Số người phụ thuộc*

*19. Điện thoại*

*20. Có phải là lao động nước ngoài*

## **Nêu ít nhất 3 bài toán phân loại (classification) và phân cụm (clustering) có thể thực hiện trên tập dữ liệu này**

Bài toán được thể hiện ở phần dưới

## **Khám phá, tiền xử lý dữ liệu, loại bỏ outliers (nếu có)**

1. Khám phá dữ liệu

|  | min | max | mean | std |
| --- | --- | --- | --- | --- |
| Creaditability | 0 | 1 | 0,7 | 0,458 |
| Account Balance | 1 | 4 | 2,577 | 1,258 |
| Duration of credit(month) | 4 | 72 | 20,903 | 12,059 |
| Payment status of previous credit | 0 | 4 | 2,545 | 1,083 |
| Purpose | 0 | 10 | 2,828 | 2,744 |
| Credit Amount | 250 | 18424 | 3271,248 | 2822,752 |
| value saving/stocks | 1 | 5 | 2,105 | 1,58 |
| Length of currrent employment | 1 | 5 | 3,384 | 1,208 |
| Instalment per cent | 1 | 4 | 2,973 | 1,119 |
| Sex&Maritial | 1 | 4 | 2,6682 | 0,708 |
| Guarantors | 1 | 3 | 1,145 | 0,478 |
| Duration in current address | 1 | 4 | 2,845 | 1,104 |
| Most valuable available asset | 1 | 4 | 2,358 | 1,05 |
| Age | 19 | 75 | 35,542 | 11,353 |
| Concurrent credits | 1 | 3 | 2,675 | 0,706 |
| Type of apartment | 1 | 3 | 1,928 | 0,53 |
| No of credits at this bank | 1 | 4 | 1,407 | 0,578 |
| Occupation | 1 | 4 | 2,904 | 0,654 |
| No of dependents | 1 | 2 | 1,155 | 0,362 |
| Foreign Worker | 1 | 2 | 1,037 | 0,189 |

## **Thực hiện phân lớp để đánh giá rủi ro tín dụng dùng mô hình cây quyết định**

Bài toán phân loại khả năng tín dụng với class phân lớp là Creditability

1. J48

| Thời gian build model | 0,15s | |
| --- | --- | --- |
| Tổng số bản ghi | 1000 | |
| Tổng số lớp | 340 | |
| Số mẫu phân lớp đúng | 221 | |
| Số mẫu phân lớp sai | 119 | |
| Tỷ lệ phân lớp đúng | 65% | |
| Tỷ lệ phân lớp sai | 35% | |
| Số lá | 83 | |
| Kích thước cây | 165 | |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | | |
| Ma trận nhầm lẫn | a b <-- classified as  175 48 | a = good  71 46 | b = bad | |
|
|
|
|
|
| 175 mẫu phân lớp đúng vào a | |
| 46 mẫu phân lớp đúng vào b | |
| 48 mẫu phân lớp sai vào a | |
| 71 mẫu phân lớp sai vào b | |

1. RandomTree

| Thời gian build model | 0,03s | |
| --- | --- | --- |
| Tổng số bản ghi | 1000 | |
| Tổng số lớp | 340 | |
| Số mẫu phân lớp đúng | 247 | |
| Số mẫu phân lớp sai | 93 | |
| Tỷ lệ phân lớp đúng | 72.6471 % | |
| Tỷ lệ phân lớp sai | 27.3529 % | |
| Kích thước cây | 579 | |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | | |
| Ma trận nhầm lẫn | a b <-- classified as  184 39 | a = good  54 63 | b = bad | |
|
|
|
|
|
| 184 mẫu phân lớp đúng vào a | |
| 63 mẫu phân lớp đúng vào b | |
| 39 mẫu phân lớp sai vào a | |
| 54 mẫu phân lớp sai vào b | |

1. RepTree

| Thời gian build model | 0,04s | |
| --- | --- | --- |
| Tổng số bản ghi | 1000 | |
| Tổng số lớp | 1000 | |
| Số mẫu phân lớp đúng | 748 | |
| Số mẫu phân lớp sai | 252 | |
| Tỷ lệ phân lớp đúng | 74.8 % | |
| Tỷ lệ phân lớp sai | 25.2 % | |
| Kích thước cây | 49 | |
| Cross-validation Fold = 10 | | |
| Ma trận nhầm lẫn | a b <-- classified as  611 89 | a = good  163 137 | b = bad | |
|
|
|
|
|
| 611 mẫu phân lớp đúng vào a | |
| 137 mẫu phân lớp đúng vào b | |
| 89 mẫu phân lớp sai vào a | |
| 163 mẫu phân lớp sai vào b | |

Kết luận: Sau khi thực hiện phân lớp trên một số cây với các tham số và tỷ lệ huấn luyện khác nhau, em đánh giá thuật toán **REPTREE** cho kết quả phân lớp chính xác nhiều nhất bằng phương pháp **Cross-validation** F = 10 nhưng dựa trên ma trận nhầm lẫn thì ta thấy phân lớp sai vào lớp b khá nhiều nên không thể dùng để đánh giá bài toán rủi ro tín dụng

| **Độ đo** | **Thuật toán** | | |
| --- | --- | --- | --- |
| **J48** | **RandomTree** | **RepTree** |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | Phân lớp đúng: 65% | 72.6471 % | 69.4118 % |
| Phân lớp sai: 35% | 27.3529 % | 30.5882 % |
| Cross-validation Fold = 10 | Phân lớp đúng: 73.9 % | 66.6 % | 74.8 % |
| Phân lớp sai: 26.1 % | 33.4 % | 25.2 % |

## **Thực hiện phân lớp để đánh giá rủi ro tín dụng sử dụng kỹ thuật Boosting**

Bài toán phân loại khả năng tín dụng với class phân lớp là Creditability

| Độ đo | | Thuật toán Adaboost M1 | | |
| --- | --- | --- | --- | --- |
| DecisionStump | J48 | RepTree |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | Số lần lặp | 100 | 300 | 100 |
| Weight | 0,01 | 2,74 | 0,9 |
| Tỷ lệ phân lớp đúng | 75.2941 % | 74.1176 % | 72.9412 % |
| Tỷ lệ phân lớp sai | 24.7059 % | 25.8824 % | 27.0588 % |
|  |  |  |  |
|  |  |  |  |

Kết luận: Khi sử dụng học tăng cường trên cây **DecisionStump** cho kết quả phân loại với độ chính tốt nhất

## **Thực hiện phân lớp để đánh giá rủi ro tín dụng dùng mô hình k-NN**

| Độ đo | | Thuật toán IBK | | |
| --- | --- | --- | --- | --- |
| K=3 | K=7 | K=17 |
| Cross-validation Fold = 10 | Ma trận nhầm lẫn | a b <-- classified as  594 106 | a = good  165 135 | b = bad | a b <-- classified as  620 80 | a = good  184 116 | b = bad | a b <-- classified as  644 56 | a = good  189 111 | b = bad |
|
|
| Tỷ lệ phân lớp đúng | 72.9 % | 73.6 % | 75.5 % |
| Tỷ lệ phân lớp sai | 27.1 % | 26.4 % | 24.5 % |
| Thời gian chạy | 0s | 0s | 0s |

Kết luận: Sau khi thử thuật toán **IBK** với các giá trị Fold và K em thấy rằng **Fold = 10 & K=17** cho kết quả phân loại chính xác nhất với class là Creditability

## **So sánh, phân tích các kết quả thu được từ 3 mô hình phân lớp nói trên**.

| Độ đo | | Thuật toán | | |
| --- | --- | --- | --- | --- |
| ReepTree | Adaboost M1(DT) | IBK(k=17) |
| Ma trận nhầm lẫn | | a b <-- classified as  611 89 | a = good  163 137 | b = bad | a b <-- classified as  594 106 | a = good  165 135 | b = bad | a b <-- classified as  644 56 | a = good  189 111 | b = bad |
|
|
| Tỷ lệ phân lớp đúng | | 74.8 % | 75.2941 % | 75.5 % |
| Tỷ lệ phân lớp sai | | 25.2 % | 24.7059 % | 24.5 % |
| Thời gian chạy | | 0,04s | 0s | 0s |

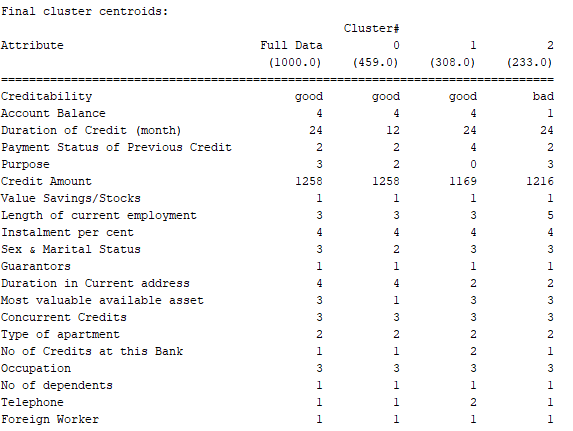
Nhận xét: các kết quả ở trên tỷ lệ phân lớp đúng cao nhưng tỷ lệ phân lớp sai của lớp b ở ma trận nhầm lẫn đều hơn 1 nửa nên ta không áp dụng được bài toán phân lớp ở trên để đánh giá rủi ro tín dụng

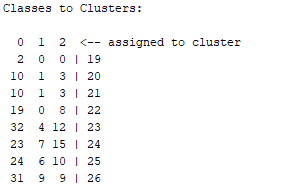
## **Thực hiện phân cụm để xác định các nhóm tuổi**

| Lược đồ phân cụm | SimpleKmeans |
| --- | --- |
| Số cụm | 3 |
| Tên dữ liệu | german credit |
| Số mẫu | 1000 |
| Số thuộc tính | 21 |
| Thuộc tính bỏ qua | Age (years) |
| Số lần lặp | 4 |
| Tổng số lỗi bình phương | 8162.0 |
|  |
| Số mẫu phân cụm không chính xác | 932 |
| 93,20% |
| Thời gian chạy | 0,04s |

| Phiên bản cụm | 0 459 ( 46%)  1 308 ( 31%)  2 233 ( 23%) | |
| --- | --- | --- |
|
|
| Cụm 0 với 459 đại diện, chiếm 46% tất cả các mẫu/ trường hợp được phân nhóm | | |
| Cụm 1 với 308 đại diện, chiếm 31% tất cả các mẫu/ trường hợp được phân nhóm | | |
| Cụm 2 với 233 đại diện, chiếm 23% tất cả các mẫu/ trường hợp được phân nhóm | | |

| Classes to Clusters | Cluster 0 <-- 23  Cluster 1 <-- 36  Cluster 2 <-- 24 | |
| --- | --- | --- |
|
|
| Đại diện cho lớp tuổi được chỉ định cho cụm 0 là 23 | | |
| Đại diện cho lớp tuổi được chỉ định cho cụm 1 là 36 | | |
| Đại diện cho lớp tuổi được chỉ định cho cụm 2 là 24 | | |





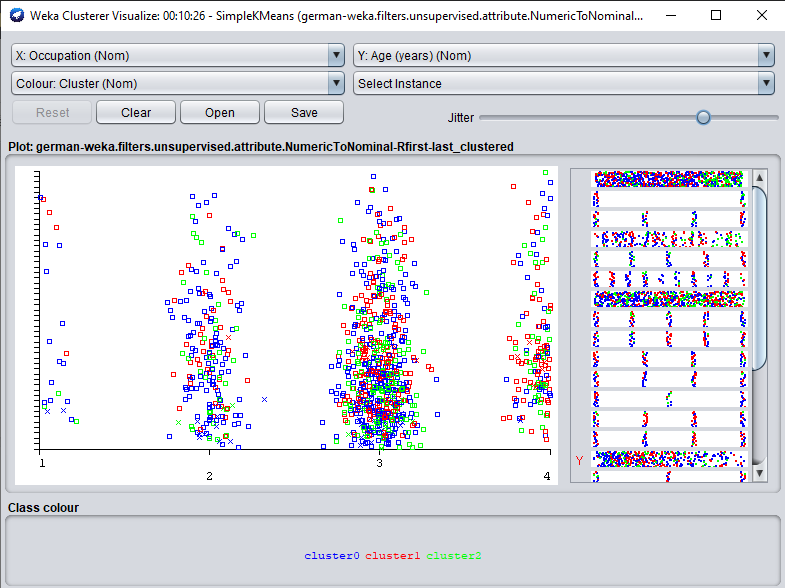
Nhận xét: Giả sử với kết quả như trên + kết quả cụm trung tâm em có nhận xét sau:

Ví dụ: Cụm 1 có 28 trường hợp: người này 19 tuổi, trạng thái tài khoản tốt, tài khoản tiết kiệm là 1169, nghề nghiệp skilled employee / official

Ví dụ: Cụm 0 có 161 trường hợp: người này 20 tuổi, giới tính và tình trạng hôn nhân là nam và độc thân, số năm làm việc liên tục đến thời điểm hiện tại: thất nghiệp

Ví dụ mô hình hoá nghề nghiệp với nhóm tuổi

Ta thấy % phân cụm không chính xác tới 93,2% nên ta không dùng để phân cụm các nhóm tuổi được

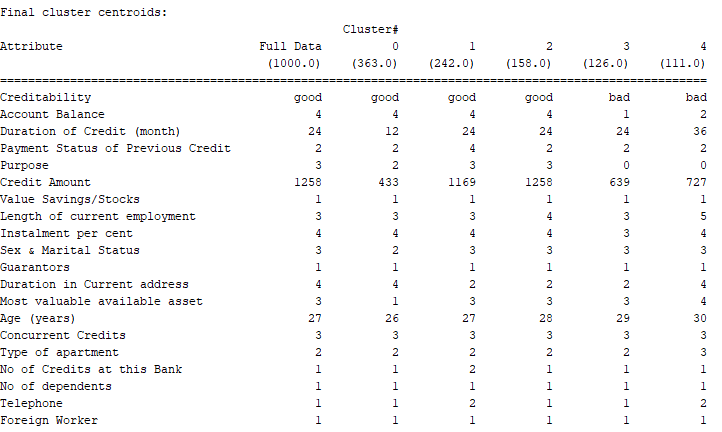


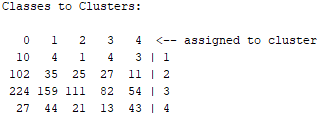
## **Thực hiện phân cụm để xác định các nhóm nghề nghiệp**

| Lược đồ phân cụm | SimpleKmeans |
| --- | --- |
| Số cụm | 5 |
| Tên dữ liệu | german credit |
| Số mẫu | 1000 |
| Số thuộc tính | 21 |
| Thuộc tính bỏ qua | Occupation |
| Số lần lặp | 5 |
| Tổng số lỗi bình phương | 8269.0 |
|  |
| Số mẫu phân cụm không chính xác | 629 |
| 69,20% |
| Thời gian chạy | 0,03 |

| Phiên bản cụm | 0 363 ( 36%)  1 242 ( 24%)  2 158 ( 16%)  3 126 ( 13%)  4 111 ( 11% | |
| --- | --- | --- |
|
|
|
|
| Cụm 0 với 363 đại diện, chiếm 36% tất cả các mẫu/ trường hợp được phân nhóm | | |
| Cụm 1 với 242 đại diện, chiếm 24% tất cả các mẫu/ trường hợp được phân nhóm | | |
| Cụm 2 với 158 đại diện, chiếm 16% tất cả các mẫu/ trường hợp được phân nhóm | | |
| Cụm 3 với 126 đại diện, chiếm 13% tất cả các mẫu/ trường hợp được phân nhóm | | |
| Cụm 4 với 111 đại diện, chiếm 11% tất cả các mẫu/ trường hợp được phân nhóm | | |

| Classes to Clusters | Cluster 0 <-- 2  Cluster 1 <-- 3  Cluster 2 <-- No class  Cluster 3 <-- 1  Cluster 4 <-- 4 | |
| --- | --- | --- |
|
|
|
|
| Đại diện cho lớp nghề nghiệp được chỉ định cho cụm 0 là 2(unskilled - resident) | | |
| Đại diện cho lớp nghề nghiệp được chỉ định cho cụm 1 là 3(skilled employee / official) | | |
| Đại diện cho lớp nghề nghiệp được chỉ định cho cụm 3 là 1(unemployed/ unskilled - non-resident) | | |
| Đại diện cho lớp nghề nghiệp được chỉ định cho cụm 4 là 4(management/ self-employed/) | | |

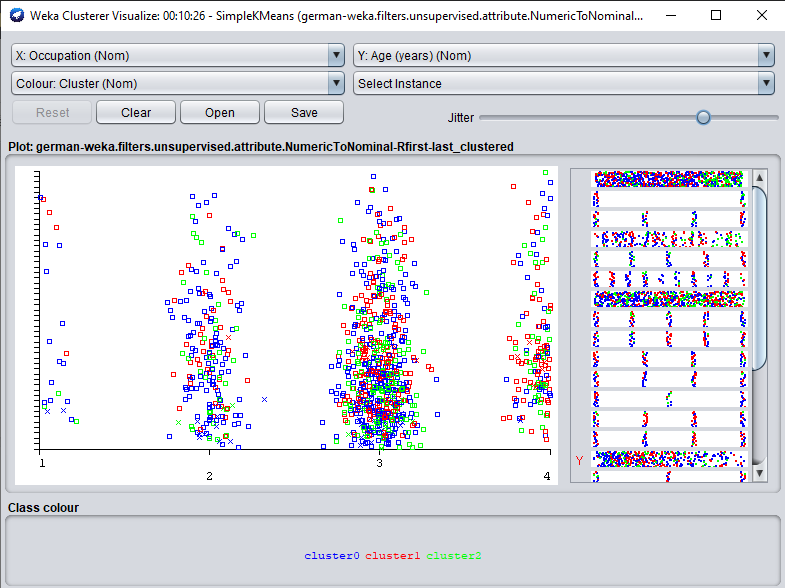




Nhận xét: Với kết quả như trên

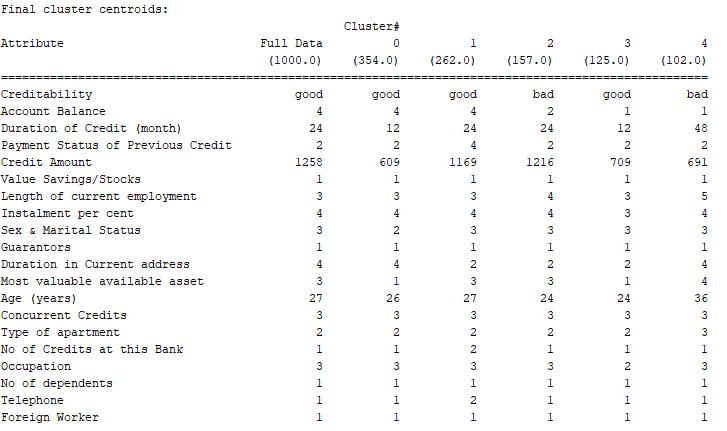
Ví dụ: Cụm 2 có 158 trường hợp, có 1 người thất nghiệp, trạng thái tài khoản tốt, là nam, độc thân, tuổi 28, …

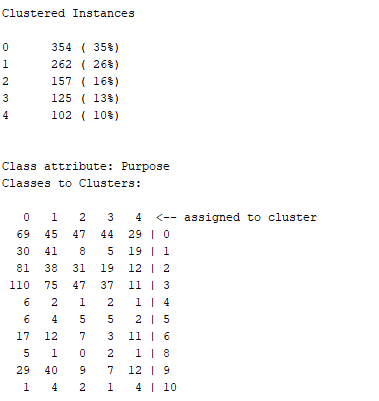
Ta thấy % phân cụm không chính xác khá cao 69,2% nên không giải quyết được bài toán xác định các nhóm nghề nghiệp

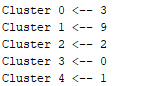


## **Thực hiện phân cụm để xác định các nhóm mục đích vay (purpose)**

| Lược đồ phân cụm | SimpleKmeans |
| --- | --- |
| Số cụm | 5 |
| Tên dữ liệu | german credit |
| Số mẫu | 1000 |
| Số thuộc tính | 21 |
| Thuộc tính bỏ qua | Purpose |
| Số lần lặp | 5 |
| Tổng số lỗi bình phương | 7938.0 |
|  |
| Số mẫu phân cụm không chính xác | 756 |
| 75,60% |
| Thời gian chạy | 0,02 |







Tương tự như các ví dụ trên ta thấy % phân cụm không chính xác sai rất nhiều 75,6% nên không dùng nó để xác định nhóm mục đích vay được