1. Dataset:

Dataset_for_classification_cmc.csv, giải thích:

- Là dataset về con người. Các features:
  - age: tuổi
  - education: trình độ giáo dục
  - spouse_education: trình độ giáo dục của vợ/chồng
  - number_of_children: số lượng con
  - religion: tôn giáo
  - now_working: có đi làm hay không
  - spouse_occupation: nghề nghiệp của vợ/chồng
  - index_living_standard: chỉ số mức sống
  - media_exposure: chỉ số trải nghiệm mạng xã hội

- Cột cuối cùng **use_insurrance** là Label, gồm có 3 phân lớp cần phân loại:
  - not_use
  - use_long
  - use_short

- Các nhóm dùng 1 tập dataset này làm cho tất cả các giải thuật đã học về Classification
- Lưu ý:
  - Không phải data features nào cũng giúp cho Classification 😉 , do đó mỗi nhóm tùy ý chọn data features nào để dùng cho giải thuật
  - Nếu muốn vẽ plot phân lớp thì có thể chọn <= 3 features để plot trên không gian 3 chiều
  - Các giá trị của Data features:
    - Có một số giá trị bị thiếu trong một số features, thường được ghi bằng thông tin "unknown". Sinh viên có thể dùng các giá trị bị thiếu này như là 1 thông tin để dùng cho giải thuật Classification, hoặc có thể hoặc sử dụng các kỹ thuật xóa hoặc cắt bỏ tùy ý
    - Sinh viên xử lý theo cách nào thì ghi vào báo cáo
  - Các nhóm tùy ý chia tỷ lệ số lượng data points dành cho training và testing

Dataset_for_clustering_student.csv, giải thích:

- Là dataset dữ liệu sinh viên, gồm có:
  1. school - trường học của sinh viên
  2. sex - giới tính của sinh viên (nhị phân: 'F' - nữ hoặc 'M' - nam)
  3. age - tuổi của sinh viên (số: từ 15 đến 22)
  4. address - loại địa chỉ nhà của sinh viên (nhị phân: 'U' - đô thị hoặc 'R '- nông thôn)
  5. famsize - quy mô gia đình (nhị phân:' LE3 '- ít hơn hoặc bằng 3 người, hoặc' GT3 '- lớn hơn 3 người)

6. Pstatus - tình trạng chung sống của cha mẹ (nhị phân:' T '- sống chung hoặc' A '- không sống chung)
7. Medu – nền tảng giáo dục của mẹ (số: 0 - không, 1 - giáo dục tiểu học (lớp 4), 2 - lớp 5 đến lớp 9, 3 - Giáo dục trung học 3 , hoặc 4 - giáo dục đại học)
8. Fedu – nền tảng giáo dục của cha (số: 0 - không, 1 - giáo dục tiểu học (lớp 4), 2 - lớp 5 đến lớp 9, giáo dục trung học 3 hoặc giáo dục đại học 4)
9. Mjob - công việc của mẹ (danh nghĩa: 'giáo viên', 'liên quan đến chăm sóc sức khỏe,' dịch vụ dân sự '(ví dụ: hành chính hoặc cảnh sát),' at_home 'hoặc' khác ')
10. Fjob - công việc của cha (danh nghĩa:' giáo viên ',' liên quan đến chăm sóc sức khỏe ', dân sự 'dịch vụ' (ví dụ: hành chính hoặc cảnh sát), 'at_home' hoặc 'khác')
11. reason - lý do để chọn trường này (danh nghĩa: gần với 'nhà', trường 'danh tiếng', 'ưu tiên' hoặc 'khác')
12. guardian - người giám hộ của sinh viên (danh nghĩa: 'mẹ', 'cha' hoặc 'người khác')
13. traveltime - thời gian di chuyển từ nhà đến trường (số: 1 - <15 phút, 2 - 15 đến 30 phút, 3 - 30 phút đến 1 giờ hoặc 4 -> 1 giờ)
14. studytime - thời gian học hàng tuần (số: 1 - <2 giờ, 2 - 2 đến 5 giờ, 3 - 5 đến 10 giờ, hoặc 4 -> 10 giờ)
15. failures - số lần thi rớt trong các lớp trước đó (số: n nếu 1 <= n <3, other 4)
16. schoolup - hỗ trợ giáo dục bổ sung (nhị phân: có hoặc không)
17. famsup - hỗ trợ giáo dục gia đình (nhị phân: có hoặc không)
18. paid - lớp học có trả phíhay không (nhị phân: có hoặc không)
19. activities - hoạt động ngoại khóa (nhị phân: có hoặc không)
20. nursery – có học trường mẫu giáo không (nhị phân: có hoặc không)
21. higher - muốn học cao hơn (nhị phân: có hoặc không)
22. internet - Truy cập Internet tại nhà (nhị phân: có hoặc không)
23. romantic - có mối quan hệ bạn trai/bạn gái hay không (nhị phân: có hoặc không)
24. famrel - chất lượng mối quan hệ gia đình (số: từ 1 - rất tệ đến 5 - xuất sắc)
25. freetime - thời gian rảnh sau giờ học (số: từ 1 - rất thấp đến 5 - rất cao)
26. gout - đi chơi với bạn bè (số : từ 1 - rất thấp đến 5 - rất cao)
27. Dalc - mức tiêu thụ rượu trong ngày làm việc (số: từ 1 - rất thấp đến 5 - rất cao)
28. Walc - mức tiêu thụ rượu cuối tuần (số: từ 1 - rất thấp đến 5 - rất cao)
29. health - tình trạng sức khỏe hiện tại (số: từ 1 - rất tệ đến 5 - rất tốt)
30. absences - số lần nghỉ học (số: từ 0 đến 93)
31. G1 – điểm số của khi học môn 1
32. G2 – điểm số khi học môn 2
33. G3 – điểm số khi học môn 3

- Các nhóm dùng 1 tập dataset này làm cho tất cả các giải thuật đã học về Clustering để phân cụm sinh viên theo cách phù hợp
- Lưu ý:
    - Không phải data features nào cũng giúp cho Clustering 😉 , do đó mỗi nhóm tùy ý chọn data features nào để dùng cho giải thuật
    - Nếu muốn vẽ plot phân lớp thì có thể chọn <= 3 features để plot trên không gian 3 chiều
    - Giải thuật HC yêu cầu vẽ được cây Dendrogram

- Chia ra bao nhiêu phân cụm là tùy sinh viên lựa chọn
- Khi nộp report, yêu cầu các nhóm nộp lại dataset này (file CSV) kèm thêm 1 cột cuối cùng:
  - Đặt tên cột là Cluster
  - Với từng dòng thì ghi rõ dòng đó thuộc Cluster nào, ví dụ, 1, 2, 3, 4….

2. Phần Report giữa kì cần có:

2.1.Lời cam kết không đạo văn, không đạo code trong toàn bộ các phần của báo cáo (xem mục ***Note*** để hiểu rõ)

2.2.Trình bày phần tìm hiểu Scikit-learn theo giải thuật đã bốc thăm.

a. Nhóm nào present giữa kì thì trình bày trong Report giữa kì.

b. Nhóm nào present cuối kì thì trình bày trong Report cuối kì

2.3.Tất cả các nhóm làm bài tập cho tất cả 6 giải thuật giữa kì

a. Dataset: giảng viên cung cấp
b. Các nhóm làm bài tập gồm có:
   i. Code python tất cả 6 giải thuật
   ii. Trình bày cách làm
   iii. Đánh giá độ hiệu quả của mô hình học máy (Evaluate Model) trên dataset đó (Evaluation)
   iv. Đồ thị, plot, chart 2-3 chiều… nếu có
   v. Đối với các dataset dành cho bài toán Clustering thì phải chỉ ra 1 datapoint thuộc cluster nào (yêu cầu nộp lại file dataset theo dạng CSV của phần clustering với từng datapoint(datarow) đã được gán Cluster)

2.4.Điểm báo cáo:

a. Nhóm nào trình bày code tốt + rõ ràng, hiểu rõ về thư viện và nắm vững những thông số (parameters) của Scikit-learn dành cho các giải thuật, và tinh chỉnh các thông số (parameters) sao cho ra được Model càng hiệu quả (theo phần Evaluate Model) càng tốt thì càng được điểm cao
b. Điểm là điểm riêng của từng nhóm nhỏ
c. Khi chấm điểm có xét tới yếu tố số lượng thành viên nhóm, nghĩa là nhóm 3 người sẽ được/bị xem xét với yêu cầu cao hơn nhóm 2 người

2.5.Report: in 2 mặt và nộp cho giảng viên

a. Đồng thời phải nộp luôn bản mềm (dataset + source code)
b. Đối với các dataset dành cho bài toán Clustering thì phải nộp lại dataset để chỉ ra 1 datapoint thuộc cluster nào

2.6.Deadline nộp report môn ML lấy điểm giữa kì: 04/11/2019 (đến 12h đêm)

a. Nộp trễ:
    i. Từ 0h ngày 05/11/2019 bị tính là nộp trễ
    ii. Mỗi ngày nộp trễ trừ 1 điểm giữa kì cho toàn bộ nhóm
b. Report nộp bản giấy
c. Source code và dataset sau khi clustering gửi qua email (không nhận gửi qua facebook)
    i. Title emal bắt buộc: **MALE431085 – Nhóm [N] – Nộp bài giữa kì**
    ii. [N] là số thứ tự nhóm
    iii. Nhóm nào gửi không đúng title email như trên tự động bị xóa mail khỏi inbox và bị trừ điểm

Note:

Tất cả các nhóm:

Tìm hiểu thế nào là khái niệm "Đạo văn" và phải cam kết không đạo văn, không đạo code trong toàn bộ các phần của báo cáo

- Tổng kết phần tìm hiểu về "Đạo văn" (chỉ ghi tối đa: 1 trang A4)
- Lời cam kết không đạo văn và không đạo code lập trình trong toàn bộ các phần của báo cáo
- Link tham khảo:

- https://www.google.com.vn/search?q=th%E1%BA%BF+n%C3%A0o+l%C3%A0+%C4%91%E1%BA%A1o+v%C4%83n+trong+nghi%C3%AAn+c%E1%BB%A9u+khoa+h%E1%BB%8Dc&oq=the%E1%BA%BF+n%C3%A0o+l%C3%A0+%C4%91%E1%BA%A1o+v%C4%83n%2C&aqs=chrome.2.69i57j0l4.5834j0j7&sourceid=chrome&ie=UTF-8
- https://vnexpress.net/tin-tuc/giao-duc/sinh-vien-vo-tu-dao-van-3225930.html
- http://vietnamnet.vn/vn/giao-duc/khoa-hoc/sinh-vien-viet-nam-y-thuc-chong-dao-van-gan-nhu-bang-0-435116.html
- https://www.google.com.vn/search?q=sinh%20vi%C3%AAn%20v%C3%A0%20%C4%91%E1%BA%A1o%20v%C4%83n&oq=sinh%20vi%C3%AAn%20v%C3%A0%20%C4%91%E1%BA%A1o%20v%C4%83n&aqs=chrome..69i57j69i60.6564j0j7&sourceid=chrome&ie=UTF-8