Dataset_for_classification.csv, giải thích:

- Là dataset của ngành Tiếp thị ngân hàng, cho biết dữ liệu khách hàng của ngân hàng. Mục tiêu của chiến dịch tiếp thị là thuyết phục khách hàng gửi tiền tiết kiệm có kì hạn

- Số lượng data point (số hàng): trên 41K

- Số lượng data features: 20, gồm các cột từ A→ T, giải thích ý nghĩa như sau (lưu ý có 1 số chữ được dịch ra tiếng Việt cho dễ hiểu nhưng sinh viên không cần chuyển qua tiếng Việt trong bài làm)
    1. tuoi: tuổi
    2. nghe_nghiep: loại công việc (gồm có các loại nghề nghiệp trong tiếng Anh)
    3. hon_nhan: tình trạng hôn nhân (phân loại: "đã ly dị", "đã kết hôn", "độc thân", "không xác định")
    4. hoc_van: tình trạng học vấn
    5. co_the_tin_dung: có, hoặc không, hoặc không rõ (unknown)
    6. co_nha_o: có nhà ở không? Gồm có: có, không, không rõ
    7. vay_ca_nhan: các khoản vay cá nhân, gồm có: có, không, không rõ
    8. kenh_lien_lac: liên lạc với khách hàng qua điện thoại di động hoặc dt bàn
    9. thang_lien_lac: tháng mà nhân viên ngân hàng đã liên hệ với khách hàng
    10. ngay_lien_lac: ngày mà nhân viên ngân hàng đã liên hệ với khách hàng
    11. thoi_luong_lien_lac: thời lượng của lần liên lạc cuối cùng, tính bằng giây
    12. so_luong_lien_lac: số lượng lần liên lạc được thực hiện trong chiến dịch này và cho khách hàng này
    13. ngay: số ngày trôi qua sau khi khách hàng được liên hệ lần cuối từ một chiến dịch tiếp thị trước đó (số; 999 có nghĩa là khách hàng không được liên hệ trước đó)
    14. so_luong_lien_lac_truoc_day: số lượng liên hệ được thực hiện cho các chiến lịch tiếp thị lần trước và cho khách hàng này
    15. ket_qua_lan_truoc: kết quả của chiến dịch tiếp thị trước đó (gồm có: "thất bại", "không tồn tại", "thành công")
    16. ti_le_thay_doi_viec_lam: tỷ lệ thay đổi việc làm theo quý trong năm
    17. CPI: chỉ số giá tiêu dùng tính theo hàng tháng
    18. CCI: chỉ số niềm tin của người tiêu dùng theo hàng tháng
    19. lai_suat_3thang: lãi suất chuẩn liên ngân hàng tại Khu vực sử dụng đồng euro
    20. so_luong_nhan_vien: số lượng nhân viên làm việc cho ngân hàng

- Cột cuối cùng là Label, gồm có 2 phân lớp cần phân loại:
    - 0: “thất bại”: khách hàng không chịu gửi tiết kiệm có kì hạn
    - 1: “thành công” : khách hàng đã gửi tiết kiệm có kì hạn

- 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.csv, giải thích:

- Là dataset miêu tả các thông số kỹ thuật của xe hơi, gồm có:
  1. ten_xe
  2. luong_hao_xang
  3. so_luong_xi_lanh
  4. the_tich_dong_co
  5. ma_luc
  6. ty_le_truc_sau
  7. khoi_luong_xe
  8. gia_toc_xe
  9. loai_xy_lanh_dong_co
  10. loai_truyen_dong
  11. so_luong_banh_rang
  12. so_luong_bo_che_hoa_khi

- Các nhóm dùng 1 tập dataset này làm cho tất cả các giải thuật đã học về Clustering
- 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….