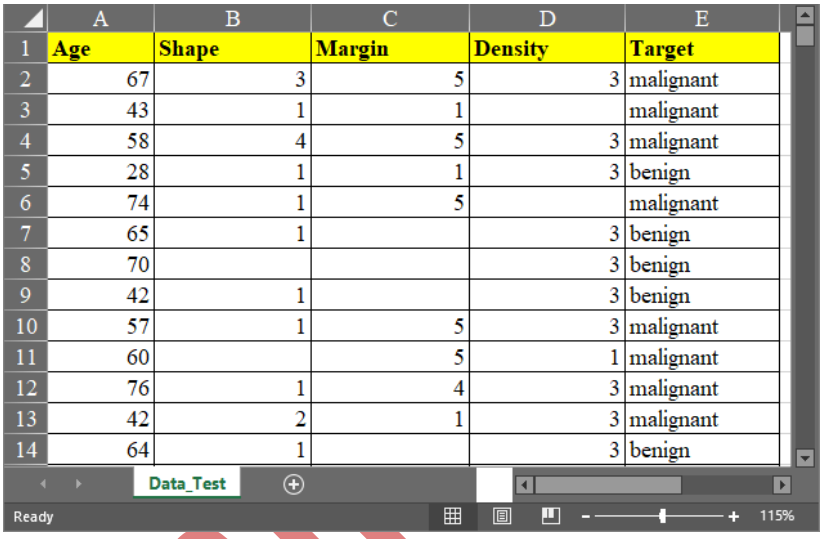
NỘI DUNG CHUẨN BỊ KẾT THÚC HỌC PHẦN
MÔN: "KỸ NGHỆ TRI THỨC VÀ HỌC MÁY"

MỤC TIÊU:

Xây dựng một mô hình học máy dự đoán một khối u là lành tính hay ác tính dựa vào các thông tin liên quan tới bệnh nhân.

PHẦN 1: CHUẨN BỊ DỮ LIỆU

a) Mô tả tập dữ liệu:



	A	B	C	D	E
1	**Age**	**Shape**	**Margin**	**Density**	**Target**
2	67	3	5	3	malignant
3	43	1	1		malignant
4	58	4	5	3	malignant
5	28	1	1	3	benign
6	74	1	5		malignant
7	65	1		3	benign
8	70			3	benign
9	42	1		3	benign
10	57	1	5	3	malignant
11	60		5	1	malignant
12	76	1	4	3	malignant
13	42	2	1	3	malignant
14	64	1		3	benign

Tập dữ liệu **Data_test.csv** chứa thông tin của 961 bệnh nhân có khối u với 4 thuộc tính độc lập (Age, Shape, Margin, Density) và 1 thuộc tính phụ thuộc (Target):

- **Age:** Thuộc tính cho biết tuổi của bệnh nhân [18-96 tuổi] (Dữ liệu Interger)
- **Shape:** Thuộc tính cho biết hình dạng khối u, bao gồm các giá trị:
  - 1 – round
  - 2 – oval
  - 3 – lobular
  - 4 – irregular
- **Margin:** Thuộc tính cho biết kích thước, đường biên của khối u, bao gồm các giá trị:
  - 1 – Circumscribed
  - 2 – Microlobulated
  - 3 – Obscured
  - 4 – ill-defined
  - 5 – Spiculated
- **Density:** Thuộc tính cho biết mật độ của khối u, bao gồm các giá trị:
  - 1 – High

- 2 – Iso
- 3 – Low
- 4 – Fat-containing

➢ **Target:** Thuộc tính phụ thuộc (label) cho biết khối u là lành tính – hay ác tính, trong đó:

- benign: Khối u lành tính
- malignant: Khối u ác tính

b) Chuẩn bị dữ liệu

Đây là tập dữ liệu thô, cần phải được chuẩn hóa và xử lý trước khi sử dụng. Sinh viên nghiên cứu tập dữ liệu, sử dụng các kỹ thuật tiền xử lý dữ liệu để chuẩn hóa tập dữ liệu này. Trong đó lưu ý một số vấn đề:

- Hiểu tập dữ liệu và các thuộc tính
- Kiểm tra và xử lý giá trị missing trong tập dữ liệu
- Chuyển đổi dữ liệu Categorical về dạng số
- Phân tách dữ liệu thành các thuộc tính độc lập và thuộc tính phụ thuộc
- Chia tập dữ liệu gốc thành Tập Train (80%) – Test (20%)

PHẦN 2: XÂY DỰNG MODEL

Sinh viên xác định bài toán học máy và thuật toán phù hợp để giải quyết bài toán: Sinh viên lựa chọn thuật toán KNN, hoặc Decision Tree, hoặc một thuật toán khác… phù hợp để xây dựng model giải quyết bài toán.

- **Bước 1:** Thực hiện huấn luyện model với tập dữ liệu Train
- **Bước 2:** Chạy kiểm thử model với tập dữ liệu Test
- **Bước 3:** Tùy chỉnh các tham số của model và lặp lại bước 1, 2 để thu được một model có độ chính xác cao nhất. Lưu model lại để sư dụng cho bài kiểm tra trên lớp.

NOTE:

- **SINH VIÊN THAM KHẢO PHẦN VÍ DỤ MẪU CÁC BƯỚC XÂY DỰNG MÔ HÌNH HỌC MÁY (phần 01 – Example – ML : Chương 3)**
- **HOÀN THÀNH TỪ BƯỚC 1 ĐẾN BƯỚC 6 CỦA QUY TRÌNH; SINH VIÊN TRÌNH BÀY CÁC BƯỚC THỰC HIỆN TRONG FILE CODE TƯƠNG NHƯ TRONG VÍ DỤ MẪU.**
- **ĐẶT TÊN FILE CODE BÀI LÀM TRÊN JUPYTER NOTEBOOK: MÃ SINH VIÊN _ TÊN SINH VIÊN_4080540_A**
- **SINH VIÊN NGHIÊN CỨU TỰ CHUẨN, KHÔNG SAO CHÉP CỦA NHAU (CÁC BÀI GIỐNG NHAU = 0 ĐIỂM)**