Trường ĐH CNTP TP.HCM Khoa: CNTT Bộ môn: CNPM Phát Triển Phần Mềm Và Ứng Dụng Thông Minh	**BÀI 11** PHÂN TÍCH ỨNG DỤNG THUẬT TOÁN **ỨNG DỤNG THUẬT TOÁN CÂY QUYẾT ĐỊNH VÀO TƯ VẤN CHỌN CHUYÊN NGÀNH CỦA SINH VIÊN**	

A. MỤC TIÊU:

- Phân tích và ứng dụng thuật toán thông minh vào sản phẩm phần mềm

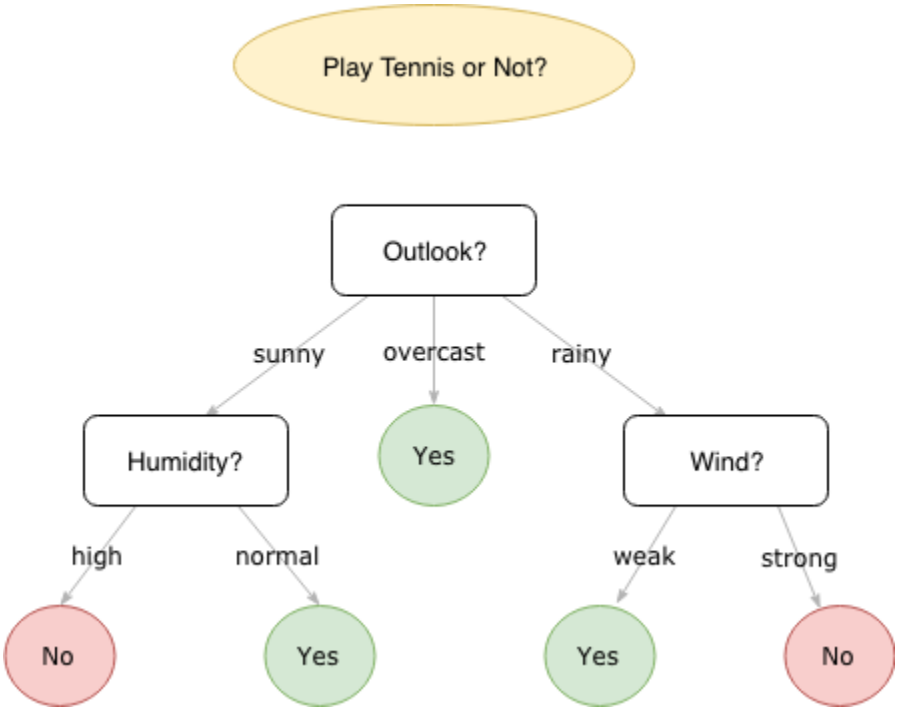
B. DỤNG CỤ - THIẾT BỊ THỰC HÀNH CHO MỘT SV:

STT	Chủng loại – Quy cách vật tư	Số lượng	Đơn vị	Ghi chú
1	Computer	1	1	

C. NỘI DUNG THỰC HÀNH

1. Cây quyết định là gì?

Cây quyết định (gọi tắt là *DT*) là mô hình đưa ra quyết định dựa trên các câu hỏi.
Dưới đây là mô hình DT về một ví dụ kinh điển.
Câu hỏi có chơi tennis hay không? Quyết định đưa ra dựa trên các yếu tố về thời tiết: outlook, humidity, wind.



DT được áp dụng vào cả 2 bài toán: Phân loại (*Classification*) và Hồi quy (*Regression*). Tuy nhiên bài toán phân loại được sử dụng nhiều hơn.

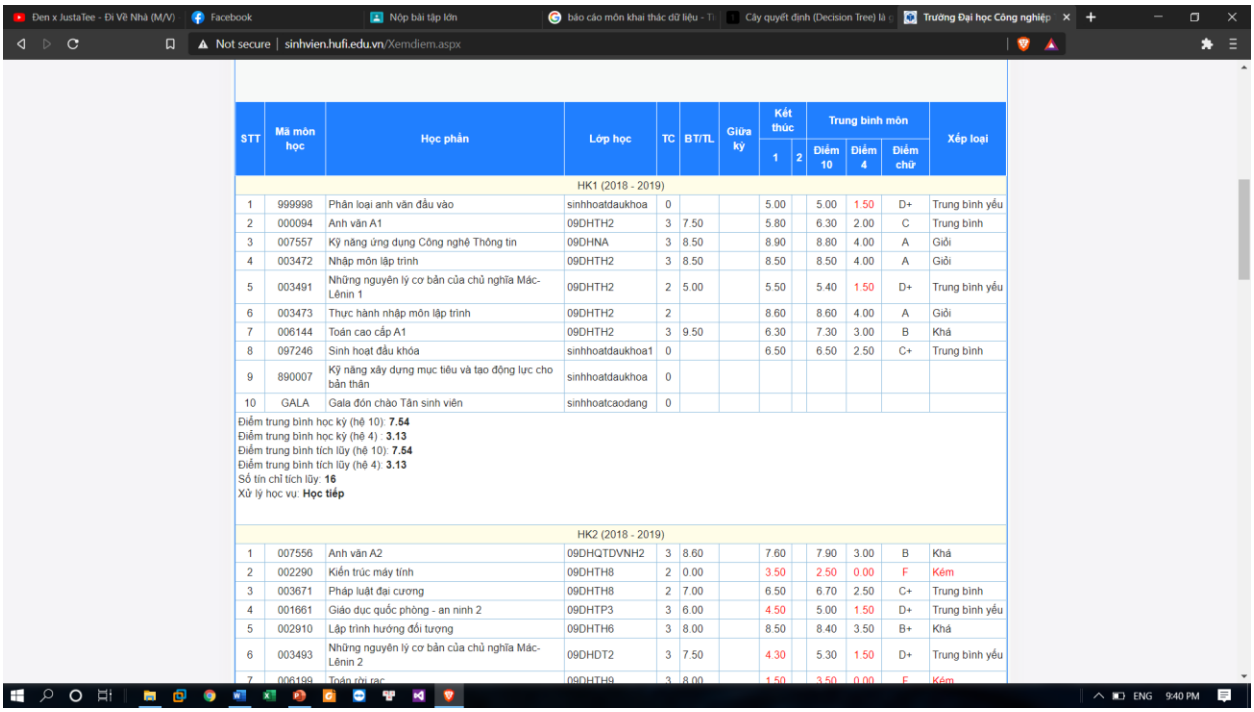
Có nhiều thuật toán để xây dựng DT, trong bài này chúng ta tìm hiểu một thuật toán nổi tiếng và cơ bản nhất của DT là thuật toán ID3.

Thuật toán ID3

Iterative Dichotomiser 3 (ID3) C45 là thuật toán nổi tiếng để xây dựng Decision Tree, áp dụng cho bài toán Phân loại (*Classification*) mà tất các các thuộc tính để ở dạng category.

Ứng dụng gợi ý chuyên ngành cho sinh viên:

Bước 1: Collect Data



STT	Mã môn học	Học phần	Lớp học	TC	BT/TL	Giữa kỳ	Kết thúc 1	Kết thúc 2	Trung bình môn Điểm 10	Trung bình môn Điểm 4	Trung bình môn Điểm chữ	Xếp loại
HK1 (2018 - 2019)												
1	999998	Phân loại anh văn đầu vào	sinhhoatdaukhoa	0			5.00		5.00	1.50	D+	Trung bình yếu
2	000094	Anh văn A1	09DHTH2	3	7.50		5.80		6.30	2.00	C	Trung bình
3	007557	Kỹ năng ứng dụng Công nghệ Thông tin	09DHNA	3	8.50		8.90		8.80	4.00	A	Giỏi
4	003472	Nhập môn lập trình	09DHTH2	3	8.50		8.50		8.50	4.00	A	Giỏi
5	003491	Những nguyên lý cơ bản của chủ nghĩa Mác-Lênin 1	09DHTH2	2	5.00		5.50		5.40	1.50	D+	Trung bình yếu
6	003473	Thực hành nhập môn lập trình	09DHTH2	2			8.60		8.60	4.00	A	Giỏi
7	006144	Toán cao cấp A1	09DHTH2	3	9.50		6.30		7.30	3.00	B	Khá
8	097246	Sinh hoạt đầu khóa	sinhhoatdaukhoa1	0			6.50		6.50	2.50	C+	Trung bình
9	890007	Kỹ năng xây dựng mục tiêu và tạo động lực cho bản thân	sinhhoatdaukhoa	0								
10	GALA	Gala đón chào Tân sinh viên	sinhhoatcaodang	0								

Điểm trung bình học kỳ (hệ 10): **7.54**
Điểm trung bình học kỳ (hệ 4) : **3.13**
Điểm trung bình tích lũy (hệ 10): **7.54**
Điểm trung bình tích lũy (hệ 4): **3.13**
Số tín chỉ tích lũy: **16**
Xử lý học vụ: **Học tiếp**

HK2 (2018 - 2019)												
1	007556	Anh văn A2	09DHQTDVNH2	3	8.60		7.60		7.90	3.00	B	Khá
2	002290	Kiến trúc máy tính	09DHTH8	2	0.00		3.50		2.50	0.00	F	Kém
3	003671	Pháp luật đại cương	09DHTH8	2	7.00		6.50		6.70	2.50	C+	Trung bình
4	001661	Giáo dục quốc phòng - an ninh 2	09DHTP3	3	6.00		4.50		5.00	1.50	D+	Trung bình yếu
5	002910	Lập trình hướng đối tượng	09DHTH6	3	8.00		8.50		8.40	3.50	B+	Khá
6	003493	Những nguyên lý cơ bản của chủ nghĩa Mác-Lênin 2	09DHDT2	3	7.50		4.30		5.30	1.50	D+	Trung bình yếu
7	006199	Toán rời rạc	09DHTH9	3	8.00		1.50		3.50	0.00	F	Kém

Data sẽ được collect từ trang sinh viên sinhvien.hufi.edu.vn của khóa 09. Gồm 9 môn(input)

- Nhập môn lập trình
- Thực hành nhập môn lập trình
- Lập trình hướng đối tượng
- Thực hành lập trình hướng đối tượng

- Cấu trúc dữ liệu và giải thuật
- Thực hành cấu trúc dữ liệu và giải thuật
- Cơ dỡ dữ liệu
- Thực hành cơ sở dữ liệu
- Toán cao cấp

Chuyên ngành(ouput)

- Công nghệ phầm mềm
- Hệ thống thông tin
- Mạng máy tính
- Khoa học phân tích dữ liệu

Bước 2: Làm sạch dữ liệu

Dự định ban đầu collect 100 record

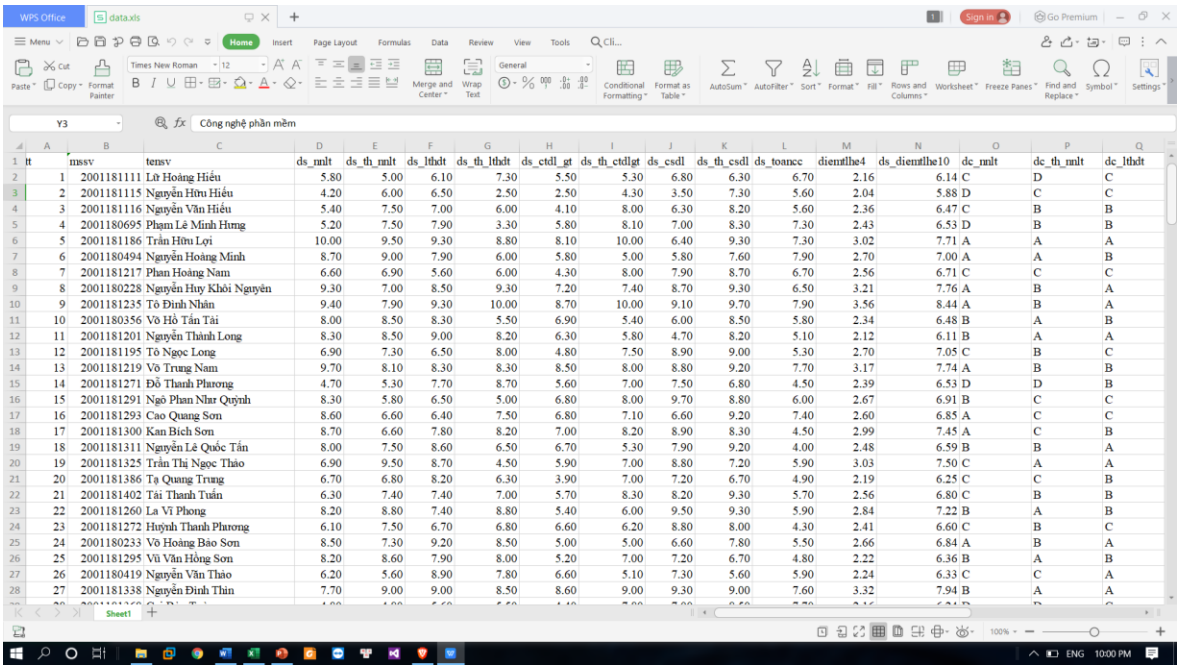
Những sinh viên có quá nhiều điểm F sẽ loại bỏ

Sau khi chạy thử -> kết quả còn sai lệch quá nhiều

Nhóm quyết định collect thêm data, collect luôn cả những sinh viên dù nhiều điểm F nhưng trạng thái là đang học và đã được phân chuyên nghành. Chỉ loại bỏ những sinh viên có trạng thái thôi học và được kết quả là data có 187 records training 27 records test

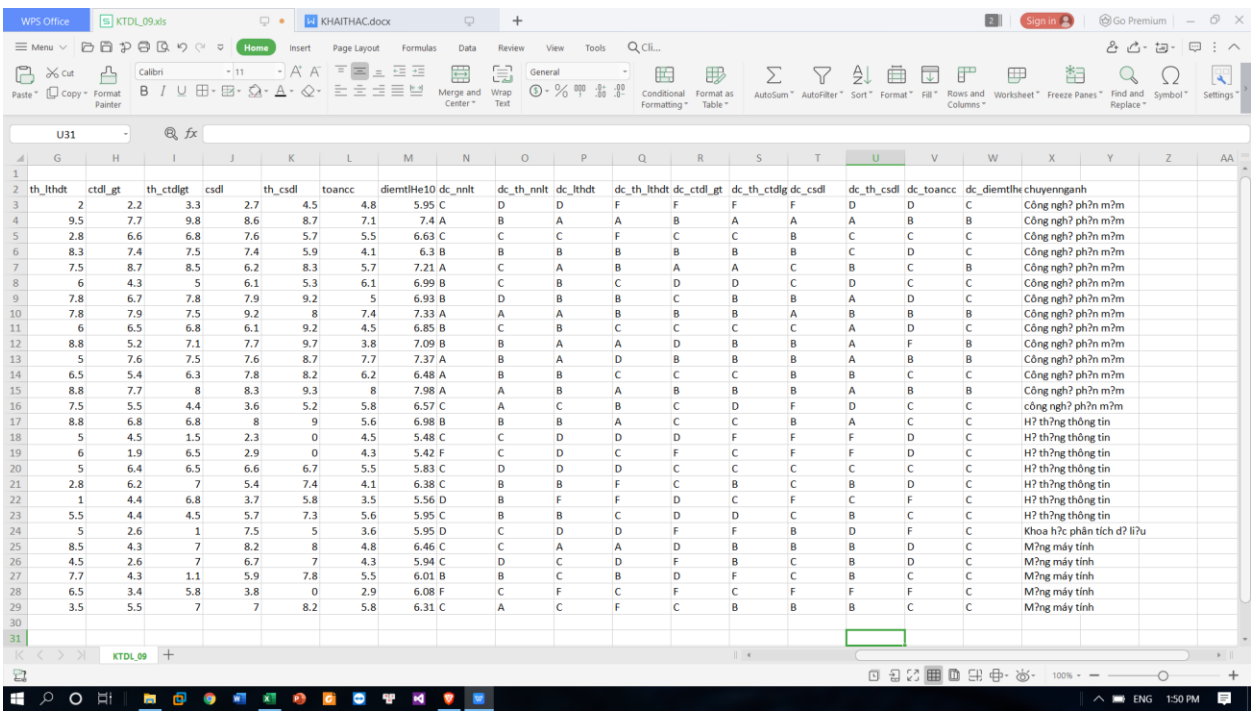
Đối với những sinh viên còn thiếu 1-2 môn , sẽ được vào điểm của các môn còn lại tự động điền vào.

Bước 3: Trainning



tt	mssv	tensv	ds_nnlt	ds_th_nnlt	ds_lthdt	ds_th_lthdt	ds_ctdl_gt	ds_th_ctdlgt	ds_csdl	ds_th_csdl	ds_toancc	diemtlhe4	ds_diemtlhe10	dc_nnlt	dc_th_nnlt	dc_lthdt
1	2001181111	Lữ Hoàng Hiếu	5.80	5.00	6.10	7.30	5.50	5.30	6.80	6.30	6.70	2.16	6.14	C	D	C
2	2001181115	Nguyễn Hữu Hiếu	4.20	6.00	6.50	2.50	2.50	4.30	3.50	7.30	5.60	2.04	5.88	D	C	C
3	2001181116	Nguyễn Văn Hiếu	5.40	7.50	7.00	6.00	4.10	8.00	6.30	8.20	5.60	2.36	6.47	C	B	B
4	2001180695	Phạm Lê Minh Hưng	5.20	7.50	7.90	3.30	5.80	8.10	7.00	8.30	7.30	2.43	6.53	D	B	B
5	2001181186	Trần Hữu Lợi	10.00	9.50	9.30	8.80	8.10	10.00	6.40	9.30	7.30	3.02	7.71	A	A	A
6	2001180494	Nguyễn Hoàng Minh	8.70	9.00	7.90	6.00	5.80	5.00	5.80	7.60	7.90	2.70	7.00	A	A	B
7	2001181217	Phan Hoàng Nam	6.60	6.90	5.60	6.00	4.30	8.00	7.90	8.70	6.70	2.56	6.71	C	C	C
8	2001180228	Nguyễn Huy Khôi Nguyên	9.30	7.00	8.50	9.30	7.20	7.40	8.70	9.30	6.50	3.21	7.76	A	B	A
9	2001181235	Tô Đình Nhân	9.40	7.90	9.30	10.00	8.70	10.00	9.10	9.70	7.90	3.56	8.44	A	B	A
10	2001180356	Võ Hồ Tấn Tài	8.00	8.50	8.30	5.50	6.90	5.40	6.00	8.50	5.80	2.34	6.48	B	A	B
11	2001181201	Nguyễn Thành Long	8.30	8.50	9.00	8.20	6.30	5.80	4.70	8.20	5.10	2.12	6.11	B	A	A
12	2001181195	Tô Ngọc Long	6.90	7.30	6.50	8.00	4.80	7.50	8.90	9.00	5.30	2.70	7.05	C	B	C
13	2001181219	Võ Trung Nam	9.70	8.10	8.30	8.30	8.50	8.00	8.80	9.20	7.70	3.17	7.74	A	B	B
14	2001181271	Đỗ Thanh Phương	4.70	5.30	7.70	8.70	5.60	7.00	7.50	6.80	4.50	2.39	6.53	D	D	B
15	2001181291	Ngô Phan Như Quỳnh	8.30	5.80	6.50	5.00	6.80	8.00	9.70	8.80	6.00	2.67	6.91	B	C	C
16	2001181293	Cao Quang Sơn	8.60	6.60	6.40	7.50	6.80	7.10	6.60	9.20	7.40	2.60	6.85	A	C	C
17	2001181300	Kan Bích Sơn	8.70	6.60	7.80	8.20	7.00	8.20	8.90	8.30	4.50	2.99	7.45	A	C	B
18	2001181311	Nguyễn Lê Quốc Tấn	8.00	7.50	8.60	6.50	6.70	5.30	7.90	9.20	4.00	2.48	6.59	B	B	A
19	2001181325	Trần Thị Ngọc Thảo	6.90	9.50	8.70	4.50	5.90	7.00	8.80	7.20	5.90	3.03	7.50	C	A	A
20	2001181386	Tạ Quang Trung	6.70	6.80	8.20	6.30	3.90	7.00	7.20	6.70	4.90	2.19	6.25	C	C	B
21	2001181402	Tái Thanh Tuấn	6.30	7.40	7.40	7.00	5.70	8.30	8.20	9.30	5.70	2.56	6.80	C	B	B
22	2001181260	La Vĩ Phong	8.20	8.80	7.40	8.80	5.40	6.00	9.50	9.30	5.90	2.84	7.22	B	A	B
23	2001181272	Huỳnh Thanh Phương	6.10	7.50	6.70	6.80	6.60	6.20	8.80	8.00	4.30	2.41	6.60	C	B	C
24	2001180233	Võ Hoàng Bảo Sơn	8.50	7.30	9.20	8.50	5.00	5.00	6.60	7.80	5.50	2.66	6.84	A	B	A
25	2001181295	Vũ Văn Hồng Sơn	8.20	8.60	7.90	8.00	5.20	7.00	7.20	6.70	4.80	2.22	6.36	B	A	B
26	2001180419	Nguyễn Văn Thảo	6.20	5.60	8.90	7.80	6.60	5.10	7.30	5.60	5.90	2.24	6.33	C	C	A
27	2001181338	Nguyễn Đình Thìn	7.70	9.00	9.00	8.50	8.60	9.00	9.30	9.00	7.60	3.32	7.94	B	A	A

Bước 4: Test

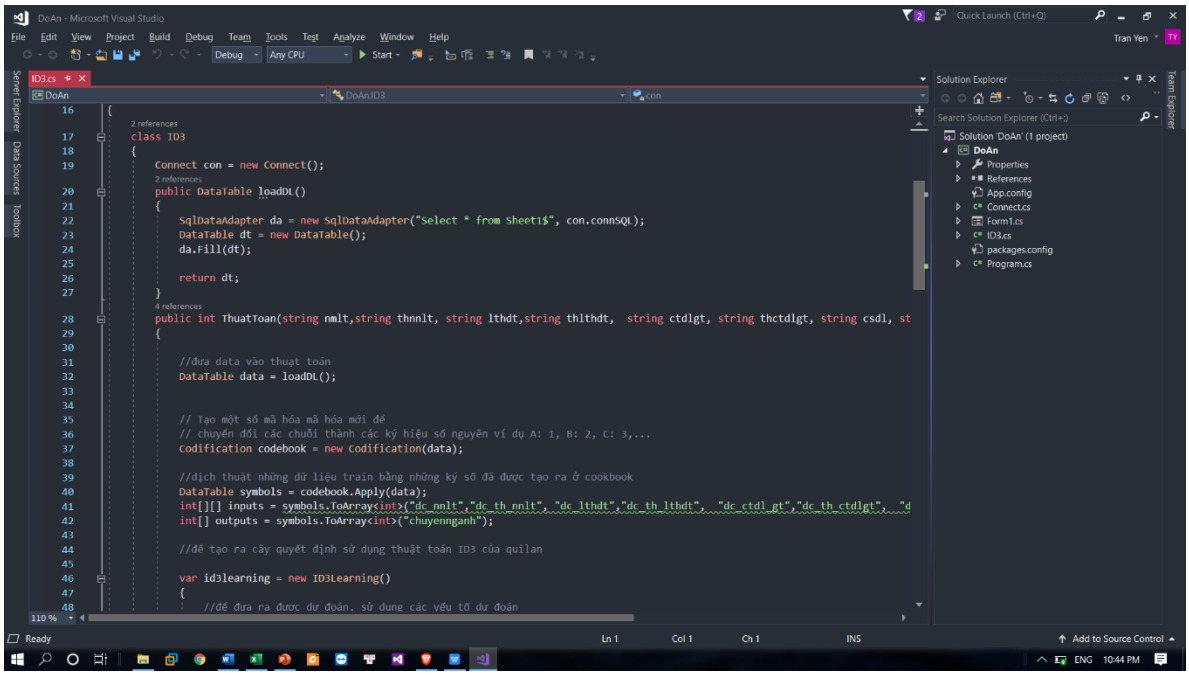


th_lthdt	ctdl_gt	th_ctdlgt	csdl	th_csdl	toancc	diemtlHe10	dc_nnlt	dc_th_nnlt	dc_lthdt	dc_th_lthdt	dc_ctdl_gt	dc_th_ctdlg	dc_csdl	dc_th_csdl	dc_toancc	dc_diemtlhe	chuyennganh
2	2.2	3.3	2.7	4.5	4.8	5.95	C	D	D	F	F	F	F	D	D	C	Công ngh? ph?n m?m
9.5	7.7	9.8	8.6	8.7	7.1	7.4	A	B	A	A	B	A	A	A	B	B	Công ngh? ph?n m?m
2.8	6.6	6.8	7.6	5.7	5.5	6.63	C	C	C	F	C	C	B	C	C	C	Công ngh? ph?n m?m
8.3	7.4	7.5	7.4	5.9	4.1	6.3	B	B	B	B	B	B	B	C	D	C	Công ngh? ph?n m?m
7.5	8.7	8.5	6.2	8.3	5.7	7.21	A	C	A	B	A	A	C	B	C	B	Công ngh? ph?n m?m
6	4.3	5	6.1	5.3	6.1	6.99	B	C	B	C	D	D	C	D	C	C	Công ngh? ph?n m?m
7.8	6.7	7.8	7.9	9.2	5	6.93	B	D	B	B	C	B	B	A	D	C	Công ngh? ph?n m?m
7.8	7.9	7.5	9.2	8	7.4	7.33	A	A	A	B	B	B	A	B	B	B	Công ngh? ph?n m?m
6	6.5	6.8	6.1	9.2	4.5	6.85	B	C	B	C	C	C	C	A	D	C	Công ngh? ph?n m?m
8.8	5.2	7.1	7.7	9.7	3.8	7.09	B	B	A	A	D	B	B	A	F	B	Công ngh? ph?n m?m
5	7.6	7.5	7.6	8.7	7.7	7.37	A	B	A	D	B	B	B	A	B	B	Công ngh? ph?n m?m
6.5	5.4	6.3	7.8	8.2	6.2	6.48	A	B	B	C	C	C	B	B	C	C	Công ngh? ph?n m?m
8.8	7.7	8	8.3	9.3	8	7.98	A	A	B	A	B	B	B	A	B	B	Công ngh? ph?n m?m
7.5	5.5	4.4	3.6	5.2	5.8	6.57	C	A	C	B	C	D	F	D	C	C	công ngh? ph?n m?m
8.8	6.8	6.8	8	9	5.6	6.98	B	B	B	A	C	C	B	A	C	C	H? th?ng thông tin
5	4.5	1.5	2.3	0	4.5	5.48	C	C	D	D	D	F	F	F	D	C	H? th?ng thông tin
6	1.9	6.5	2.9	0	4.3	5.42	F	C	D	C	F	C	F	F	D	C	H? th?ng thông tin
5	6.4	6.5	6.6	6.7	5.5	5.83	C	D	D	D	C	C	C	C	C	C	H? th?ng thông tin
2.8	6.2	7	5.4	7.4	4.1	6.38	C	B	B	F	C	B	C	B	D	C	H? th?ng thông tin
1	4.4	6.8	3.7	5.8	3.5	5.56	D	B	F	F	D	C	F	C	F	C	H? th?ng thông tin
5.5	4.4	4.5	5.7	7.3	5.6	5.95	C	B	B	C	D	D	C	B	C	C	H? th?ng thông tin
5	2.6	1	7.5	5	3.6	5.95	D	C	D	D	F	F	B	D	F	C	Khoa h?c phân tích d? li?u
8.5	4.3	7	8.2	8	4.8	6.46	C	C	A	A	D	B	B	B	D	C	M?ng máy tính
4.5	2.6	7	6.7	7	4.3	5.94	C	D	C	D	F	B	C	B	D	C	M?ng máy tính
7.7	4.3	1.1	5.9	7.8	5.5	6.01	B	B	C	B	D	F	C	B	C	C	M?ng máy tính
6.5	3.4	5.8	3.8	0	2.9	6.08	F	C	F	C	F	C	F	F	F	C	M?ng máy tính
3.5	5.5	7	7	8.2	5.8	6.31	C	A	C	F	C	B	B	B	C	C	M?ng máy tính

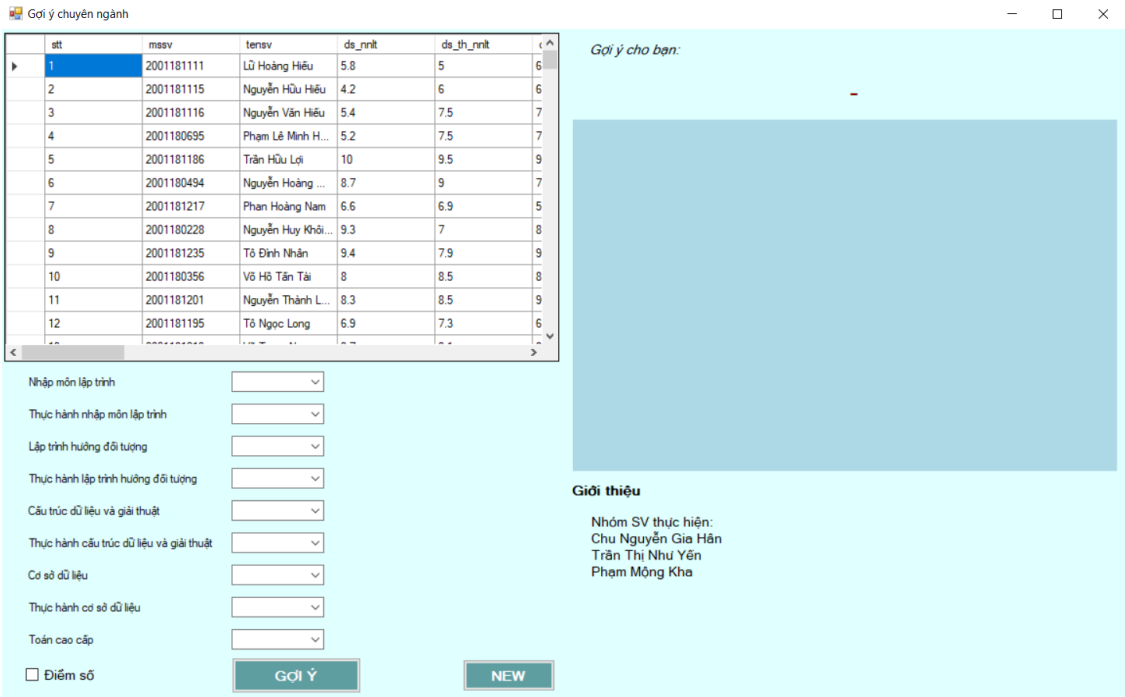
Input thuật toán

Sử dụng framework của sẵn từ Machine Learning

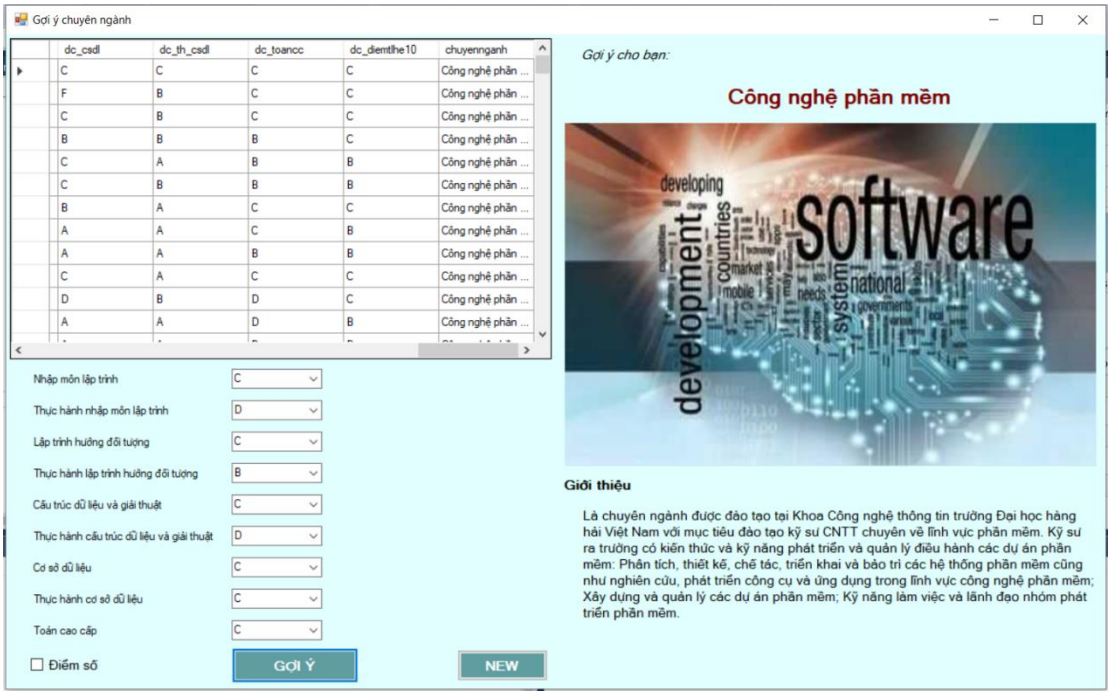
Đầu tiên, chuyển file excel thành sql và load data vào thuật toán để train



Giao diện

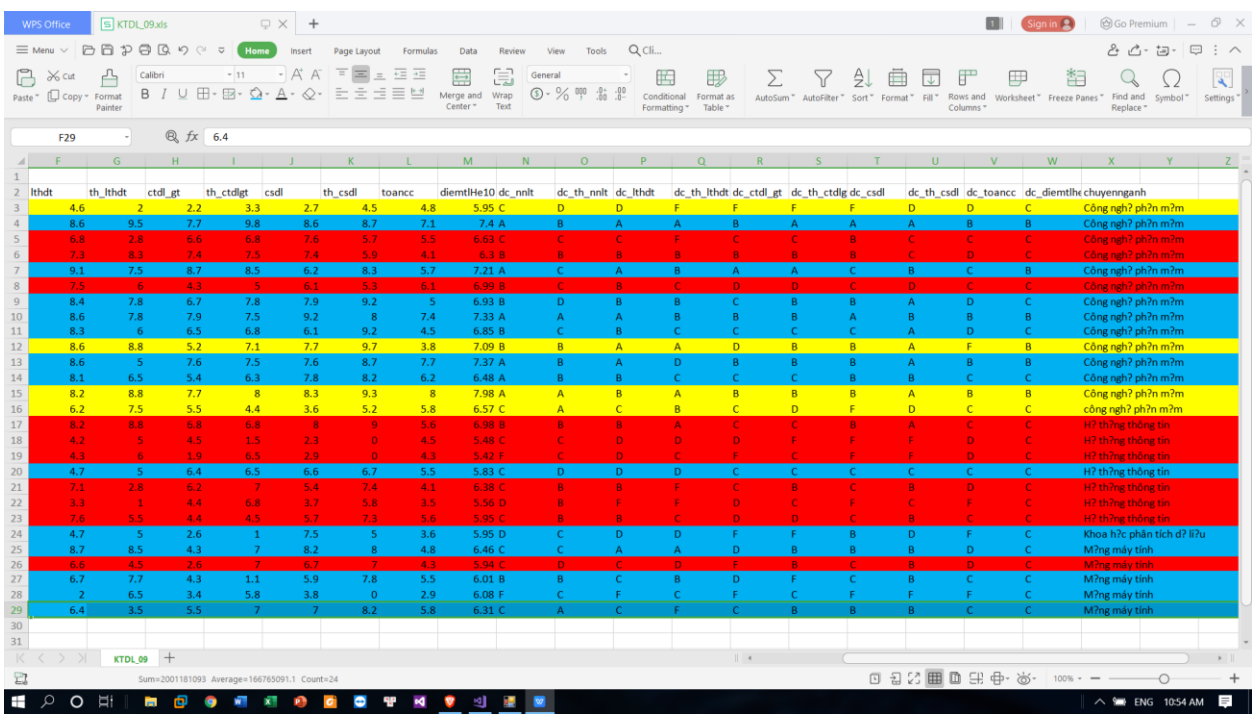


Khi sinh viên nhập điểm các môn vào theo 2 loại điểm: điểm chữ hoặc điểm số -> nhấn vào GỢI Ý -> sẽ thấy kết quả gợi ý ở bên phải



TEST: 27 records

Sau khi nhập dữ liệu vào test có được kết quả:



lthdt	th_lthdt	ctdl_gt	th_ctdlgt	csdl	th_csdl	toancc	diemtlHe10	dc_nnlt	dc_th_nnlt	dc_lthdt	dc_th_lthdt	dc_ctdl_gt	dc_th_ctdlg	dc_csdl	dc_th_csdl	dc_toancc	dc_diemtlhe	chuyennganh
4.6	2	2.2	3.3	2.7	4.5	4.8	5.95	C	D	D	F	F	F	F	D	D	C	Công ngh? ph?n m?m
8.6	9.5	7.7	9.8	8.6	8.7	7.1	7.4	A	B	A	A	B	A	A	A	B	B	Công ngh? ph?n m?m
6.8	2.8	6.6	6.8	7.6	5.7	5.5	6.63	C	C	C	F	C	C	B	C	C	C	Công ngh? ph?n m?m
7.3	8.3	7.4	7.5	7.4	5.9	4.1	6.3	B	B	B	B	B	B	B	C	D	C	Công ngh? ph?n m?m
9.1	7.5	8.7	8.5	6.2	8.3	5.7	7.21	A	C	A	B	A	A	C	B	C	B	Công ngh? ph?n m?m
7.5	6	4.3	5	6.1	5.3	6.1	6.99	B	C	B	C	D	D	C	D	C	C	Công ngh? ph?n m?m
8.4	7.8	6.7	7.8	7.9	9.2	5	6.93	B	D	B	B	C	B	B	A	D	C	Công ngh? ph?n m?m
8.6	7.8	7.9	7.5	9.2	8	7.4	7.33	A	A	A	B	B	B	A	B	B	B	Công ngh? ph?n m?m
8.3	6	6.5	6.8	6.1	9.2	4.5	6.85	B	C	B	C	C	C	C	A	D	C	Công ngh? ph?n m?m
8.6	8.8	5.2	7.1	7.7	9.7	3.8	7.09	B	B	A	A	D	B	B	A	F	B	Công ngh? ph?n m?m
8.6	5	7.6	7.5	7.6	8.7	7.7	7.37	A	B	A	D	B	B	B	A	B	B	Công ngh? ph?n m?m
8.1	6.5	5.4	6.3	7.8	8.2	6.2	6.48	A	B	B	C	C	C	B	B	C	C	Công ngh? ph?n m?m
8.2	8.8	7.7	8	8.3	9.3	8	7.98	A	A	B	A	B	B	B	A	B	B	Công ngh? ph?n m?m
6.2	7.5	5.5	4.4	3.6	5.2	5.8	6.57	C	A	C	B	C	D	F	D	C	C	công ngh? ph?n m?m
8.2	8.8	6.8	6.8	8	9	5.6	6.98	B	B	B	A	C	C	B	A	C	C	H? th?ng thông tin
4.2	5	4.5	1.5	2.3	0	4.5	5.48	C	C	D	D	D	F	F	F	D	C	H? th?ng thông tin
4.3	6	1.9	6.5	2.9	0	4.3	5.42	F	C	D	C	F	C	F	F	D	C	H? th?ng thông tin
4.7	5	6.4	6.5	6.6	6.7	5.5	5.83	C	D	D	D	C	C	C	C	C	C	H? th?ng thông tin
7.1	2.8	6.2	7	5.4	7.4	4.1	6.38	C	B	B	F	C	B	C	B	D	C	H? th?ng thông tin
3.3	1	4.4	6.8	3.7	5.8	3.5	5.56	D	B	F	F	D	C	F	C	F	C	H? th?ng thông tin
7.6	5.5	4.4	4.5	5.7	7.3	5.6	5.95	C	B	B	C	D	D	C	B	C	C	H? th?ng thông tin
4.7	5	2.6	1	7.5	5	3.6	5.95	D	C	D	D	F	F	B	D	F	C	Khoa h?c phân tích d? li?u
8.7	8.5	4.3	7	8.2	8	4.8	6.46	C	C	A	A	D	B	B	B	D	C	M?ng máy tính
6.6	4.5	2.6	7	6.7	7	4.3	5.94	C	D	C	D	F	B	C	B	D	C	M?ng máy tính
6.7	7.7	4.3	1.1	5.9	7.8	5.5	6.01	B	B	C	B	D	F	C	B	C	C	M?ng máy tính
2	6.5	3.4	5.8	3.8	0	2.9	6.08	F	C	F	C	F	C	F	F	F	C	M?ng máy tính
6.4	3.5	5.5	7	7	8.2	5.8	6.31	C	A	C	F	C	B	B	B	C	C	M?ng máy tính

Đúng (xanh dương): 13 records

- Sai (đỏ): 10 records

- Không đủ dữ liệu train (vàng): 4 records

Kết luận:

Data còn ít nên độ sai lệch còn cao -> Data càng lớn sẽ cải thiện kết quả gợi ý chính xác hơn.

D. BÀI TẬP TỰ LÀM

1. Thực hiện phân tích cơ sở dữ liệu bài toán và ứng dụng cây quyết định vào tư vấn chọn chuyên ngành.