ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh Phúc

ĐỀ CƯƠNG MÔN HỌC
IS252 – KHAI THÁC DỮ LIỆU

1. THÔNG TIN CHUNG

Tên môn học (tiếng Việt):	**Khai thác dữ liệu**
Tên môn học (tiếng Anh):	Data Mining
Mã môn học:	IS252
Thuộc khối kiến thức:	Chuyên ngành Hệ thống Thông tin (HTTT)
Khoa/Bộ môn phụ trách:	Hệ thống Thông tin
Website môn học	
Giảng viên phụ trách:	PGS. TS. Nguyễn Đình Thuân
	Email: thuannd@uit.edu.vn
Giảng viên tham gia giảng dạy:	PGS. TS. Nguyễn Đình Thuân, PGS. TS. Đỗ Phúc, TS. Cao Thị Nhạn, ThS. Mai Xuân Hùng, ThS. Trịnh Minh Tuấn.
	Email: phucd@uit.edu.vn, nhanct@uit.edu.vn, hungmx@uit.edu.vn, tuantm@uit.edu.vn
Số tín chỉ:	4
	TC lý thuyết: 3 TC thực hành: 1
Lý thuyết: (tiết)	45
Thực hành: (tiết)	30
Tự học: (tiết)	
Tính chất của môn	Bắt buộc đối với sinh viên ngành HTTT
Điều kiện đăng ký: *(môn học trước)*	Xác suất thống kê, Cơ sở dữ liệu.

2. MÔ TẢ MÔN HỌC (Course description)

Môn học nhằm cung cấp các kiến thức cơ bản về khai thác dữ liệu, quá trình khám phá tri thức, các giai đoạn chính của quá trình khai thác dữ liệu, một số kỹ thuật khai thác dữ liệu đã và đang được sử dụng rộng rãi hiện nay. Ngoài ra môn học còn giới thiệu đến sinh viên các xu hướng nghiên cứu cũng như những thách thức trong lĩnh vực khai thác dữ liệu hiện nay.

Sinh viên được trang bị và thực hành để hiểu rõ các kỹ thuật chính trong khai thác dữ liệu như: tiền xử lý dữ liệu, tập phổ biến và luật kết hợp, phân lớp, gom cụm, các phần tử đặc biệt. Thông qua các bài thực hành, sinh viên được rèn luyện để hiểu rõ những nội dung lý thuyết và biết cách sử dụng những công cụ khai thác dữ liệu. Bên cạnh đó, sinh viên thực hiện một đồ án nhóm để giải quyết bài toán khai thác dữ liệu thực tế.

3. MỤC TIÊU MÔN HỌC (Course Goals)

Mục tiêu	Mô tả [1]	Mục tiêu (Theo CĐR cấp 2) [2]
G1	Kết hợp làm việc cá nhân và nhóm để thảo luận và thực hiện đề tài theo nội dung môn học.	7.1, 7.2
G2	Hiểu được khái niệm cơ bản của Khai thác dữ liệu, sự cần thiết của việc khai thác dữ liệu. Biết ứng dụng của việc khai thác dữ liệu trong các lĩnh vực của đời sống.	2.7
G3	Hiểu được các bước trong quy trình khai thác dữ liệu. Hiểu và áp dụng được kỹ thuật tìm tập phổ biến và luật kết hợp, khai thác dữ liệu dãy phổ biến, tập thô và ứng dụng để rút gọn chiều dữ liệu, phân lớp dữ liệu, gom cụm dữ liệu.	3.1, 3.2, 3.3, 9.2
G4	Phân tích, đánh giá kết quả thuật toán	3.3, 3.4, 9.2
G5	Biết vận dụng kiến thức, kỹ năng để giải quyết bài toán khai thác dữ liệu thực tế.	10.2

4. CHUẨN ĐẦU RA MÔN HỌC (Course learning outcomes)

CĐRMH [1]	CĐR cấp 3 của CTĐT [3]	Mô tả CĐRMH (mục tiêu cụ thể) [2]	Mức độ giảng dạy [4]
G1.1	7.1.1, 7.1.2	Hình thành và điều hành nhóm	U
G1.2	7.2.1, 7.2.3, 7.2.4, 7.2.5, 7.2.6	Tham gia thảo luận và làm việc theo từng nhóm trên từng chủ đề của môn học.	U
G2.1	2.7	Hiểu được khái niệm khai thác dữ liệu là gì, sự cần thiết của việc khai thác dữ liệu trong đời sống thực tế. Hiểu khả năng hỗ trợ khai thác dữ liệu của công nghệ cơ sở dữ liệu. Hiểu xu hướng nghiên cứu và những thách thức trong lĩnh vực khai thác dữ liệu hiện nay.	T
G3.1	3.1.1, 3.2.1, 3.1.3	Hiểu được các bước trong quy trình khai thác dữ liệu. Xác định và phát biểu bài toán khai phá dữ liệu: - Phân tích sơ bộ các dữ kiện - Lựa chọn bài toán giải quyết dựa trên phân tích, đánh giá tổng thể các dữ kiện. - Đề xuất giải pháp tiền xử lý dữ liệu	T, U

G3.2	3.2.1, 3.2.2, 3.2.3, 3.3.1, 3.3.2, 3.4.1 9.2.1, 9.2.2	Hiểu các khái niệm và áp dụng được các kỹ thuật: - Tìm tập phổ biến, tìm tập phổ biến tối đại, tìm luật kết hợp dựa trên tập phổ biến tối đại, cách tính độ tin cậy - Tập thô, quan hệ bất khả phân biệt, xấp xỉ trên, xấp xỉ dưới, quan hệ bất khả phân biệt, cách rút gọn hàm phân biệt - Phân lớp dữ liệu, các kỹ thuật phân lớp dữ liệu bằng mạng Bayes, cây quyết định, Support Vector Machine, K-nearest neighbors, mạng Neural. - Gom nhóm dữ liệu, các kỹ thuật gom nhóm dữ liệu bằng mạng K-Means, K-Medoids, Kohonen, DBSCAN.	T,U
G4.1	3.3.3, 3.4.2. 3.4.3	Có khả năng đề xuất giải pháp phù hợp cho một bài toán cụ thể. Áp dụng thành thạo kỹ thuật phân tích, đánh giá kết quả khi dùng công cụ	T,U
G4.2	3.3.3, 3.4.1, 3.4.2. 3.4.3, 9.2.1, 9.2.2	Đối với các nhóm thuật toán phân nhóm dữ liệu, gom cụm dữ liệu, có khả năng: - Phân tích ưu, nhược điểm của các thuật toán. - So sánh, đánh giá kết quả qua các thực nghiệm - Tổng hợp giải pháp và khuyến nghị	T,U
G5.1	10.2.4, 10.2.5, 10.2.6	Biết vận dụng kiến thức, kỹ năng để giải quyết bài toán khai thác dữ liệu thực tế: từ xác định bài toán, tiền xử lý dữ liệu, chọn lựa thuật toán khai thác dữ liệu phù hợp, đánh giá kết quả, đề xuất cải tiến (nếu có), và xây dựng ứng dụng dựa trên tri thức khai phá được.	T,U

5. NỘI DUNG MÔN HỌC, KẾ HOẠCH GIẢNG DẠY (Course content, Lesson plan)

a. Lý thuyết:

Buổi (4 tiết) [1]	**Nội dung**	**CĐRMH [3]**	**Hoạt động dạy và học [4]**	**Thành phần đánh giá**
1	**Chương 1: Tổng quan về khai thác dữ liệu** 1.1. Khai thác dữ liệu 1.2. Quá trình khai thác dữ liệu 1.3. Các loại dữ liệu được dùng để khai thác dữ liệu	G1.1, G1.2, G2.1	- *Giảng viên*: Thuyết giảng, nêu câu hỏi thảo luận, góp ý về các nội dung thảo luận nhóm - *Sinh viên:* nghe giảng, thảo luận	- Kết quả thảo luận nhóm.

	1.4. Ứng dụng của khai thác dữ liệu 1.5. Các kỹ thuật khai thác dữ liệu		nhóm, trình bày kết quả nhóm trước lớp.	
2	**Chương 2: Tiền xử lý dữ liệu** 2.1. Giới thiệu về tiền xử lý dữ liệu 2.2. Đối tượng dữ liệu và các loại thuộc tính 2.3. Làm sạch dữ liệu 2.4. Tích hợp dữ liệu 2.5. Thu giảm dữ liệu 2.6. Biến đổi dữ liệu 2.7. Rời rạc hóa dữ liệu	G3.1	- *Giảng viên*: Thuyết giảng, nêu câu hỏi thảo luận, góp ý về các nội dung thảo luận nhóm - Sinh *viên:* nghe giảng, thảo luận nhóm, trình bày kết quả nhóm trước lớp	- Thảo luận nhóm và trình bày kết quả thảo luận nhóm tại lớp.
3	**Chương 3: Tập phổ biến và Luật kết hợp** 3.1. Khái niệm cơ bản - Tập phổ biến - Tập phổ biến tối đại - Luật kết hợp 3.2. Khám phá các kết hợp với giải thuật Apriori và các biến thể của giải thuật Apriori 3.3. Khám phá các kết hợp dựa trên tập phổ biến tối đại. 3.4. Cách tính độ tin cậy của luật	G3.2, G4.1	- *Giảng viên*: Thuyết giảng, nêu câu hỏi thảo luận, góp ý về các nội dung thảo luận nhóm - Sinh *viên:* nghe giảng, thảo luận nhóm, trình bày kết quả nhóm trước lớp; làm bài tập về nhà 1.	- Thảo luận nhóm và trình bày kết quả thảo luận nhóm tại lớ p. - Bài tập về nhà 1.
4	**Chương 4: Dãy Phổ biến** 4.1. Tổng quan về dãy phổ biến 4.2. Cách tìm dãy phổ biến song song 4.3. Cách tìm dãy phổ biến tuần tự 4.4. Khám phá luật kết hợp từ dãy phổ biến, cách tính độ tin cậy của luật	G3.2, G4.1	- *Giảng viên*: Thuyết giảng, nêu câu hỏi thảo luận, góp ý về các nội dung thảo luận nhóm - Sinh *viên:* nghe giảng, thảo luận nhóm, trình bày kết quả nhóm trước lớp; làm bài tập về nhà 2.	- Thảo luận nhóm và trình bày kết quả thảo luận nhóm tại lớ p. - Bài tập về nhà 2
4,5	**Chương 4: Tập thô** 4.1. Khái niệm cơ bản - Tập thô - Quan hệ bất khả phân biệt - Xấp xỉ trên - Xấp xỉ dưới - Sự phụ thuộc thuộc tính 4.2. Rút gọn thuộc tính - Thành lập ma trận phân biệt	G3.2, G4.1	- *Giảng viên*: Thuyết giảng, nêu câu hỏi thảo luận, góp ý về các nội dung thảo luận nhóm - Sinh *viên:* nghe giảng, thảo luận nhóm, trình bày kết quả nhóm trước lớp; làm bài tập về nhà 3.	- Thảo luận nhóm và trình bày kết quả thảo luận nhóm tại lớ p. - Bài tập về nhà 3.

	- Tìm các rút gọn 4.3. Tìm luật			
6, 7	**Chương 5: Phân lớp dữ liệu** 5.1. Tổng quan về phân lớp dữ liệu 5.2. Một số cách tiếp cận cơ bản - Cây quyết định - Naïve Bayesian - Mạng Neural	G3.2, G4.1, G4.2	- *Giảng viên*: Thuyết giảng, nêu câu hỏi thảo luận, góp ý về các nội dung thảo luận nhóm - Sinh *viên:* nghe giảng, thảo luận nhóm, trình bày kết quả nhóm trước lớp; làm bài tập về nhà 4.	- Thảo luận nhóm và trình bày kết quả thảo luận nhóm tại lớ p. - Bài tập về nhà 4.
8, 9	**Chương 6: Gom cụm dữ liệu** 6.1. Tổng quan về gom cụm dữ liệu 6.2 Một số cách tiếp cận cơ bản - K – Means - K – Medoids - Kohonen (SOM – Self-Organizing Map) - DBSCAN	G3.2, G4.1, G4.2	- *Giảng viên*: Thuyết giảng, nêu câu hỏi thảo luận, góp ý về các nội dung thảo luận nhóm - Sinh *viên:* nghe giảng, thảo luận nhóm, trình bày kết quả nhóm trước lớp; làm bài tập về nhà 5.	- Thảo luận nhóm và trình bày kết quả thảo luận nhóm tại lớ p. - Bài tập về nhà 5.
10	**Chương 7: Khai thác văn bản** 7.1 Mở đầu 7.2 Kiến trúc của khai thác văn bản 7.3 Ứng dụng tập phổ biến và luật kết hợp vào bài toán phân loại văn bản 7.4 Gom cụm bằng mạng Kohonen hỗ trợ truy vấn tương tự trong khối ngữ liệu bài báo khoa học	G3.2, G4.1, G4.2	- *Giảng viên*: Thuyết giảng, nêu câu hỏi thảo luận, góp ý về các nội dung thảo luận nhóm - Sinh *viên:* nghe giảng, thảo luận nhóm, trình bày kết quả nhóm trước lớp; làm bài tập về nhà 6.	- Thảo luận nhóm và trình bày kết quả thảo luận nhóm tại lớ p. - Bài tập về nhà 5.
11 (5 tiết)	Ôn tập		- *Giảng viên*: Ôn tập, nhắc nhở những lỗi sai thường gặp. - *Sinh viên:* đặt câu hỏi và tham gia trao đổi thảo luận.	

b. Thực hành

Buổi học (5 tiết) [1]	Nội dung [2]	CĐRMH [3]	Hoạt động dạy và học [4]	Hoạt động đánh giá
1	Các công cụ hỗ trợ bài toán khai phá dữ liệu Hiểu các kiểu dữ liệu	G2.1 G3.1	❖ *Giảng viên:* – Hướng dẫn sử dụng các công cụ hỗ trợ bài toán khai phá dữ liệu: Weka, R, Python. ❖ *Sinh viên học ở lớp:* – Sử dụng các chức năng của phần các phần mềm hỗ trợ. Qua đó nắm được định dạng dữ liệu đầu vào trước khi chạy một thuật toán cụ thể, các kiểu dữ liệu, cách thức chạy một thuật toán, đọc và hiểu kết quả. ❖ *Sinh viên học ở nhà*: – Cài đặt và tìm hiểu các công cụ: Weka, R, Python.	Đánh giá dựa trên việc sử dụng các chức năng của phần mềm hỗ trợ
	Tiền xử lý dữ liệu	G3.1	❖ *Giảng viên:* – Cách tìm kiếm dữ liệu cho bài toán Khai thác dữ liệu. – Hướng dẫn cách tiền xử lý dữ liệu trên dữ liệu. – Xác định các loại dữ liệu khác nhau và chọn lựa cách tiền xử lý dữ liệu phù hợp. ❖ *Sinh viên học ở lớp:* – Làm theo hướng dẫn của giảng viên – Hoàn thành phần ***Thực hành*** trong bài Lab 1 (Tiền xử lý dữ liệu) ❖ *Sinh viên học ở nhà*: – Tìm hiểu trước các data set thường dùng trong khai thác dữ liệu. – Đọc trước phần ***Hướng dẫn chung*** trong bài Lab 1 – Hoàn thành phần ***Bài tập làm thêm*** trong bài Lab 1	Thực hành tại lớp và hoàn thành bài tập làm thêm ở nhà
2	- Bài toán tập phổ biến và luật kết hợp - Đọc tài liệu cho bài thực hành phân lớp	G3.1, G3.2, G4.1	❖ *Giảng viên:* – Hướng dẫn tạo dữ liệu (tiền xử lý dữ liệu, chọn đặc trưng phù hợp để áp dụng thuật toán tìm luật kết hợp và ngôn ngữ	Thực hành tại lớp và hoàn thành bài tập làm thêm ở nhà

	dữ liệu		áp dụng. ❖ *Sinh viên học ở lớp*: – Làm theo hướng dẫn của giảng viên – Hoàn thành phần ***Thực hành*** trong bài Lab 2 (Tập phổ biến và luật kết hợp) – Hoàn thành một phần phần ***Thực hành*** trong bài lab 3 (Phân lớp dữ liệu) ❖ *Sinh viên học ở nhà*: – Đọc trước phần ***Hướng dẫn chung*** trong bài Lab 2 – Hoàn thành phần ***Bài tập làm thêm*** trong bài Lab 2 – Đọc trước phần ***Hướng dẫn chung trong bài lab 3***	
3	Phân lớp dữ liệu	G3.1, G3.2, G4.1, G4.2	❖ *Giảng viên:* – Hướng dẫn tạo dữ liệu (tiền xử lý dữ liệu, chọn đặc trưng phù hợp để áp dụng thuật toán phân lớp dữ liệu và ngôn ngữ áp dụng. ❖ *Sinh viên học ở lớp*: – Làm theo hướng dẫn của giảng viên – Hoàn thành phần ***Thực hành*** trong bài Lab 3 (Phân lớp dữ liệu) – Đánh giá kết quả thực hiện của mỗi thuật toán. ❖ *Sinh viên học ở nhà*: – Đọc trước phần ***Hướng dẫn chung*** trong bài Lab 3 – Hoàn thành phần ***Bài tập làm thêm*** trong bài Lab 3	Thực hành tại lớp và hoàn thành bài tập làm thêm ở nhà
	Hướng dẫn chung Đồ án môn học	G1.1, G1.2 G5.1	❖ *Giảng viên:*: – Hướng dẫn sinh viên làm Đồ án nhóm: góp ý Bài toán, data set, tiền xử lý dữ liệu. ❖ *Sinh viên học ở lớp*: – Đặt câu hỏi liên quan đến đồ án và nhận góp ý từ giảng viên. ❖ *Sinh viên học ở nhà*: – Làm đồ án nhóm và chuẩn bị các câu hỏi.	Kết quả sơ khởi về bài toán, data set, tiền xử lý dữ liệu

4	Gom cụm dữ liệu	G3.1, G3.2, G4.1, G4.2	❖ *Giảng viên:* – Hướng dẫn tạo dữ liệu (tiền xử lý dữ liệu, chọn đặc trưng phù hợp để áp dụng thuật toán gom cụm dữ liệu và ngôn ngữ áp dụng. ❖ *Sinh viên học ở lớp*: – Làm theo hướng dẫn của giảng viên – Hoàn thành phần ***Thực hành*** trong bài Lab 4 (Gom cụm dữ liệu) – Đánh giá kết quả thực hiện của mỗi thuật toán. ❖ *Sinh viên học ở nhà*: – Đọc trước phần ***Hướng dẫn chung*** trong bài Lab 4 – Hoàn thành phần ***Bài tập làm thêm*** trong bài Lab 4	Thực hành tại lớp và hoàn thành bài tập làm thêm ở nhà
	Đồ án môn học	G1.2 G5.1	❖ *Giảng viên:*: – Hướng dẫn sinh viên làm Đồ án nhóm: + Rà soát lại Bài toán, data set, tiền xử lý dữ liệu; + Góp ý lựa chọn thuật toán, ngôn ngữ sử dụng. ❖ *Sinh viên học ở lớp*: – Đặt câu hỏi liên quan đến đồ án và nhận góp ý từ giảng viên. ❖ *Sinh viên học ở nhà*: – Hoàn thiện đề tài về phát biểu bài toán, tiền xử lý dữ liệu, chạy thuật toán, đánh giá kết quả và đề xuất cải tiến (nếu có)	Hoàn thiện góp ý buổi trước và chuẩn bị tiếp theo cho đồ án
5	Rà soát, hoàn tất các nội dung còn lại của Phân lớp dữ liệu, Gom cụm dữ liệu và Đồ án môn học	G3.1, G3.2, G4.1, G4.2 G1.2, G5.1	– Rà soát, hoàn thiện các nội dung thực hành và đồ án môn học	Thực hành tại lớp và trao đổi để hoàn thiện bài làm

6. ĐÁNH GIÁ MÔN HỌC (Course assessment)

Thành phần đánh giá [1]	CĐRMH (Gx) [2]	Tỷ lệ (%) [3]
Thực hành	G1.1, G1.2, G3.1, G3.2, G4.1, G4.2, G5.1	**50%** (20% bài thực

		hành + 30% đồ án)
Thi lý thuyết cuối kỳ	G2.1, G3.1, G3.2, G4.1, G4.2	**50%**

7. QUY ĐỊNH CỦA MÔN HỌC (Course requirements and expectations)

- *Đăng ký và làm Đồ án theo nhóm từ 2 đến 3 sinh viên.*
- *Sinh viên lắng nghe khi giảng viên giảng và tham gia trao đổi, thảo luận, báo cáo trong quá trình làm việc nhóm trên lớp. Sinh viên phải đọc slide bài giảng của buổi học trước khi đến lớp, thực hiện nghiêm túc Đồ án đã đăng ký cùng các bạn trong nhóm.*
- *Sinh viên phải hoàn thiện các bài thực hành ở phần thực hành.*
- *Sinh viên phải tham gia từ 80% các buổi học trên lớp, phải tham gia báo cáo Đồ án của nhóm.*

8. TÀI LIỆU HỌC TẬP, THAM KHẢO

[1]. Đỗ Phúc, *Giáo trình Khai thác dữ liệu*, NXB. Đại học Quốc Gia Tp. Hồ Chí Minh, 2020.
[2]. Vũ Hữu Tiệp, *Machine Learning cơ bản*, NXB Khoa học và Kỹ thuật, 2019.
[3]. Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining Concepts and Techniques, 3rd edition, Morgan Kaufmann Publishers, Elsevier, 2012.*

9. PHẦN MỀM HAY CÔNG CỤ HỖ TRỢ THỰC HÀNH

1. R Programming Language
2. Python Programming Language
3. Phần mềm WEKA
4. SQL Server

Tp. Hồ Chí Minh, ngày 14 tháng 02 năm 2021

Trưởng khoa/ bộ môn **Giảng viên**