



TRƯỜNG ĐẠI HỌC VĂN HIẾN CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT
NAM
KHOA CÔNG NGHỆ THÔNG TIN Độc lập – Tự do – Hạnh phúc

ĐỀ CƯƠNG CHI TIẾT HỌC PHẦN

1. Thông tin về học phần

- Tên học phần: **KHAI PHÁ DỮ LIỆU – DATA MINING**
- Mã học phần: INT418
- Số tín chỉ: 3 (2/0/4)
- Cấp đào tạo: Đại học
- Loại học phần: Bắt buộc
- Học phần tiên quyết/ Học phần trước: Toán rời rạc, Cấu trúc dữ liệu và giải thuật, Lập trình C++
- Đơn vị phụ trách: Khoa Công Nghệ Thông Tin
- Số giờ tín chỉ: 60 tiết , trong đó:
- Lý thuyết: 30 tiết (1 tín chỉ LT = 15 tiết)
- Thực hành: 30 tiết (1 tín chỉ TH = 30 tiết)
- Thực tập: 0 (1 tín chỉ TT = 60 giờ TT tại cơ sở)
- Đồ án/ Khóa luận: 0 (1 tín chỉ ĐA/KL = 45 giờ làm ĐA/KL)

2. Thông tin về giảng viên

GIẢNG VIÊN 1:

- Họ và tên: Nguyễn Hữu Hương Xuân
- Chức danh, học vị: Thạc sĩ
- Thời gian làm việc: Giờ hành chính (8:00 -16:00)
- Địa điểm làm việc: Khoa CNTT - Trường Đại học Văn Hiến
- Điện thoại: 0907981812
- Email: xuannhh@vhu.edu.vn

GIẢNG VIÊN 2:

- Họ và tên: Nguyễn Thị Diệu Anh

- Chức danh, học vị: Thạc sĩ
- Thời gian làm việc: Giờ hành chính (8:00 -16:00)
- Địa điểm làm việc: Khoa CNTT - Trường Đại học Văn Hiến
- Điện thoại: 0949.844.355
- Email: anhntd@vhu.edu.vn

3. Tóm tắt nội dung học phần

Học phần này giới thiệu về quy trình và các kỹ thuật khai phá dữ liệu, nhằm mục đích phân tích và khám phá tri thức từ các dữ liệu lớn. Các phương pháp khai phá dữ liệu: Bao gồm các kỹ thuật như phân tích phân cụm, phân tích chuỗi thời gian, học máy, kỹ thuật xử lý ngôn ngữ tự nhiên và kỹ thuật mạng lưới neuron. Sinh viên hiểu rõ quy trình từ việc thu thập dữ liệu, tiền xử lý, lựa chọn phương pháp, đến việc đánh giá và tinh chỉnh mô hình. Cuối cùng, sinh viên sẽ được thảo luận về các thách thức đặt ra bởi dữ liệu lớn và các cơ hội mà khai phá dữ liệu mang lại cho doanh nghiệp và cộng đồng..

4. Mục tiêu của học phần

Học phần có những mục tiêu:

- Hiểu biết về Dữ liệu và Tri thức: Sinh viên sẽ hiểu được sự khác biệt giữa dữ liệu, thông tin và tri thức, cũng như vai trò của chúng trong tổ chức và nền kinh tế.
- Nắm vững Các Bước Cơ Bản trong Quá trình Phát hiện Tri thức: Sinh viên sẽ nắm bắt được các bước cơ bản trong quá trình phát hiện tri thức từ dữ liệu, từ thu thập dữ liệu đến áp dụng các phương pháp khai phá dữ liệu để rút ra tri thức.
- Vận dụng Các Phương pháp và Thuật toán Khai phá Dữ liệu: Sinh viên sẽ vận dụng được các phương pháp và thuật toán khai phá dữ liệu vào các bài toán cụ thể, bao gồm phân loại, phân cụm, khám phá luật kết hợp và các phương pháp tiền xử lý dữ liệu.

Về kỹ năng:

- Phát biểu được một bài toán KPDL thực tiễn và thiết kế được mô hình giải quyết bài toán cụ thể
- Khai thác được tài nguyên KPDL để giải quyết một bài toán KPDL cụ thể có quy mô nhỏ.

5. Chuẩn đầu ra của học phần:

TT	Mã CDR	Nội dung chuẩn đầu ra
4.1	Về kiến thức	
4.1.1	CLO1	Hiểu biết về dữ liệu, thông tin và trích xuất thông tin trong khai phá dữ liệu, cần tập trung vào: Hiểu về dữ liệu là gì và các loại dữ liệu khác nhau, nhận biết sự khác biệt giữa dữ liệu và thông tin, quá trình trích xuất thông tin từ dữ liệu, tìm hiểu về các vấn đề và thách thức phổ biến trong khai phá dữ liệu, cùng với các phương pháp giải quyết chúng
4.1.2	CLO2	Thao tác thành thạo các ngôn ngữ lập trình phổ biến, kỹ thuật lập trình, phương pháp và công nghệ lập trình, cũng như có thể áp dụng chúng để giải quyết các vấn đề cơ bản trong KPD.
4.1.3	CLO3	Đánh giá và lựa chọn các kỹ thuật thích hợp với dữ liệu mẫu
4.1.4	CLO4	Thử nghiệm các thuật toán để trích xuất thông tin cơ bản của môn KPD.
4.2	Về kỹ năng	
4.2.1	CLO5	Hiểu biết các bước trong quy trình KPD và vận dụng phù hợp với tình hình thực tế của yêu cầu KPD
4.2.2	CLO6	Thành thạo thao tác công cụ như Weka, Python, Excel...để triển khai một số lớp thuật toán KPD cơ bản
4.2.3	CLO7	Trình bày trước đám đông sử dụng phương tiện trình chiếu
4.2.4	CLO8	Nâng cao kỹ năng phân công công việc và làm việc theo nhóm
4.3	Mức tự chủ và trách nhiệm	
4.3.1	CLO9	Tuân thủ bảo vệ tính riêng tư trong KPD

4.3. 2	CL10	Nhận thức những kỹ năng mềm, những đam mê trong quá trình thực hiện các bài tập KPDL
4.3. 3	CL11	Trang bị tinh thần học tập suốt đời

Ma trận liên kết giữa Chuẩn đầu ra chương trình đào tạo và Chuẩn đầu ra học phần

Chuẩn đầu ra	PL O1	PL O2	PL O3	PL O4	PL O5	PL O6	PL O7	PL O8	PL O9	PL O10
CLO1	X									
CLO2	X	X								
CLO3	X	X	X	X						
CLO4			X							
CLO5		X	X							
CLO6		X	X				X			
CLO7		X		X	X	X				
CLO8					X	X				
CLO9				X					X	
CL10										X
CL11							X			

Ghi chú: *PLOs (Programme Learning Outcomes): Chuẩn đầu ra cấp CTĐT*

CLOs (Course Learning Outcomes): Chuẩn đầu ra học phần

6. Nội dung chi tiết của học phần

6.1. Lý thuyết

Chương	Nội dung	Đáp ứng CLOs
Chương 1.	Giới thiệu về khai phá dữ liệu	CLO1, CLO9, CLO10, CL11
1.1.	Giới thiệu môn học	
1.2.	Nhu cầu phát hiện tri thức từ dữ liệu	
1.3.	Quá trình phát hiện tri thức trong Cơ sở dữ liệu (KDD)	
1.4.	Khai phá dữ liệu và xử lý C DL truyền thống	
1.5.	Kiểu dữ liệu trong KPDL	
1.6.	Một số lĩnh vực ứng dụng KPDL điển hình	
1.7.	Các bài toán KPDL điển hình	
Chương 2.	Xử lý dữ liệu	CLO1, CLO2, CLO9
2.1.	Giới thiệu	
2.2.	Hiểu về phân bố dữ liệu	
2.3.	Tiền xử lý dữ liệu	
2.4.	Làm sạch dữ liệu	
2.5.	Tích hợp dữ liệu	
2.6.	Chuyển đổi dữ liệu	
2.7.	Rút trích dữ liệu	
2.8.	Phương pháp phân chia dữ liệu	
Chương 3.	Phương pháp khai phá tập mục phổ biến	CLO1, CLO3, CLO5, CLO6, CLO7, CLO8
3.1.	Các khái niệm cơ bản 3.1.1. Tập phổ biến và các đơn vị đo cơ bản 3.1.2. Các tính chất của tập phổ biến 3.1.3. Tập phổ biến cực đại 3.1.4. Tập phổ biến đóng	
3.2	Thuật toán Apriori	

Chương	Nội dung	Đáp ứng CLOs
	3.2.1. Phương pháp xây dựng ứng viên 3.2.2. Phương pháp loại bỏ các ứng viên 3.2.3. Thuật toán 3.2.4. Phân tích ưu và nhược điểm	
3.3.	Thuật toán FP-Growth 3.3.1. Phương pháp tổ chức và lưu trữ của cây FP 3.3.2. Phương pháp duyệt cây FP 3.3.3. Thuật toán 3.3.4. Nhận xét và đánh giá	
3.4.	Sử dụng công cụ Weka để phát hiện luật kết hợp 3.4.1. Chuẩn bị dữ liệu 3.4.2. Khai phá luật kết hợp bằng thuật toán Apriori 3.4.3. Khai phá luật kết hợp bằng thuật toán FPGrowth	
3.4.		
Chương 4.	Phân lớp	CLO1, CLO2, CLO3, CLO5, CLO6, CLO7, CLO8
4.1.	Tổng quan 4.1.1. Các tình huống phân lớp dữ liệu 4.1.2. Các khái niệm cơ bản 4.1.3. Các giải thuật phân lớp	
4.2.	Giải thuật phân lớp bằng cây quyết định 4.3.1. Các khái niệm cơ bản 4.3.2. Cây ID3	
4.3.	Giải thuật phân lớp bằng mạng Bayesian 4.4.1. Các khái niệm cơ bản 4.4.2. Giải thuật phân lớp Bayesian	

Chương	Nội dung	Đáp ứng CLOs
4.4.	Các vấn đề về overfitting và underfitting	
4.5.	Các vấn đề về độ chính xác của mô hình	
Chương 5.	Phân cụm	CLO3, CLO4, CLO5, CLO6, CLO7, CLO8
5.1.	Giới thiệu về Phân cụm: 5.1.1. Định nghĩa về phân cụm trong lĩnh vực khai phá dữ liệu. 5.1.2. Mục tiêu và ý nghĩa của việc phân cụm dữ liệu. 5.1.3. Ứng dụng của phân cụm trong thế giới thực.	
5.2.	Phương pháp K-means 5.1.1. Cơ sở lý thuyết 5.1.2. Giải thuật phân cụm K-mean	
5.3.	Các vấn đề và thách thức: 5.3.1. Vấn đề về lựa chọn số lượng cụm: Thảo luận về cách chọn số lượng cụm phù hợp cho dữ liệu. 5.3.2. Vấn đề về độ đo khoảng cách: Xem xét các độ đo phổ biến như Euclidean, Manhattan, và Cosine similarity. 5.3.3. Xử lý dữ liệu thiếu và nhiễu: Cách xử lý dữ liệu bị thiếu hoặc nhiễu để đảm bảo tính chính xác của quá trình phân cụm.	
6.1.	Xu hướng phát triển KPDL trong những năm gần đây	
6.2.	Phát triển và hoàn thiện luật	
6.3.	Khai phá dữ liệu tác động đến xã hội	
6.4.	Xu hướng phát triển	

6.2. Thực hành

Modul	Nội dung	Đáp ứng CLOs
6.2.1	Bài tập thực hành	
1	Giới thiệu về khai phá dữ liệu – Python	CLO5, CLO6, CLO7, CLO8, CLO9, CLO10, CLO11, CLO12
2	Xử lý dữ liệu - Phân loại bằng Công cụ ước tính scikit-learn	
3	Dự đoán người chiến thắng trong các môn thể thao bằng cây quyết định	
4	Dự đoán rừng ngẫu nhiên	
5	Trích xuất văn bản từ các trang web tùy ý	
6.2.2	Bài tập nhóm	
6	Sinh viên chia nhóm và chuẩn bị các bài tập sau: Các nhóm báo cáo kết quả làm việc trước lớp. Bài tập: Dự đoán doanh số bán hàng với mô hình kết hợp Mục tiêu: Xây dựng một mô hình dự đoán doanh số bán hàng dựa trên dữ liệu lịch sử bán hàng và các yếu tố khác như quảng cáo, mùa vụ, và giá cả. Dữ liệu: Sử dụng bộ dữ liệu lịch sử bán hàng từ một cửa hàng hoặc doanh nghiệp, bao gồm thông tin về doanh số bán hàng hàng ngày hoặc hàng tuần, số tiền quảng cáo đã chi tiêu, mùa vụ (nếu có), và giá cả. Bài tập: Dự đoán khả năng nghỉ việc của nhân viên Mục tiêu: Xây dựng một mô hình cây quyết định để dự đoán khả năng nghỉ việc của nhân viên dựa trên các yếu tố như độ tuổi, mức lương, trình độ học vấn, và thời gian làm việc. Dữ liệu: Sử dụng bộ dữ liệu về thông tin nhân viên, bao gồm thông tin về độ tuổi, mức lương, trình độ học vấn, số năm kinh nghiệm, và kết quả (có nghỉ việc hay không).	CLO5, CLO6, CLO7, CLO8, CLO9, CLO10

	Bài tập: Phân loại khách hàng thành các nhóm dựa trên hành vi mua sắm	
--	--	--

7. Phân bổ thời gian theo tiết và điều kiện thực hiện:

Chương	Tên chương	Số tiết tín chỉ					Ghi chú
		Lý thuyết	Bài tập	Thực hành	Tự học	Tổng g	
1	Giới thiệu về khai phá dữ liệu	2	0	0	4	6	
2	Xử lý dữ liệu	3	0	6	6	15	
3	Phương pháp khai phá tập mục phổ biến	7	0	6	14	27	
4	Phân cụm	7	0	6	14	27	
5	Phân lớp	7	0	6	14	27	
6	Ứng dụng và xu hướng phát triển	4	0	6	8	18	
Tổng		30	0	0	60	90	

8. Phương pháp giảng dạy:

Giảng viên giảng dạy với sự kết hợp của một số phương pháp sau:

Gợi ý:

- Thuyết trình
- Phát vấn
- Giải quyết vấn đề

Ma trận liên kết giữa Chuẩn đầu ra với phương pháp giảng dạy

Phương pháp giảng dạy	CL O1	CL O2	CL O3	CL O4	CL O5	CL O6	CL O7	CL O8	CL O9	CL O10
Thuyết trình	X	X			X	X	X	X	X	X
Phát vấn	X		X	X						

Phương pháp giảng dạy	CL O1	CL O2	CL O3	CL O4	CL O5	CL O6	CL O7	CL O8	CL O9	CL O10
Giải quyết vấn đề	X				X		X	X		X

9. Phương pháp học tập

Sinh viên học tập với sự kết hợp của một số phương pháp sau:

Gợi ý:

- Làm việc nhóm
- Tự học, tự nghiên cứu

Ma trận liên kết giữa Chuẩn đầu ra với phương pháp học tập

Phương pháp học tập	CL O1	CL O2	CL O3	CL O4	CL O5	CL O6	CL O7	CL O8	CL O9	CL O10
Làm việc nhóm	X		X	X				X	X	X
Tự học, tự nghiên cứu	X	X	X	X			X	X	X	X

10. Nhiệm vụ của sinh viên

Gợi ý:

- Chủ động tổ chức thực hiện giờ tự học.
- Tham gia đầy đủ các giờ lên lớp và giờ thuyết trình (sinh viên chỉ được vắng mặt tối đa 20% thời gian lên lớp của học phần).
- Đọc tài liệu tham khảo bắt buộc và bổ trợ do giảng viên giới thiệu.
- Hoàn thành đầy đủ các bài tập cá nhân, bài tập nhóm.
- Tham gia kỳ thi kết thúc học phần.

11. Thang điểm đánh giá: Điểm đánh giá quá trình và điểm thi kết thúc học phần được chấm theo thang điểm 10 (từ 0 đến 10), làm tròn đến 1 chữ số thập phân.

12. Phương pháp kiểm tra, đánh giá kết quả học tập

Sinh viên được đánh giá kết quả học tập trên cơ sở hai điểm thành phần như sau:

1. Sinh viên được đánh giá kết quả học tập trên cơ sở hai điểm thành phần như sau:

Điểm đánh giá quá trình: trọng số 40% bao gồm:

a. Điểm chuyên cần: trọng số 10%

b. Điểm kiểm tra thực hành: trọng số 30%

2. Điểm thi kết thúc học phần: trọng số 60%

Hình thức thi: Tự luận 60 phút theo ngân hàng đề thi

Ma trận quan hệ giữa Chuẩn đầu ra và phương pháp kiểm tra, đánh giá

Hình thức đánh giá	CL O1	CL O2	CL O3	CL O4	CL O5	CL O6	CL O7	CL O8	CL O9	CL O10
Chuyên cần								X	X	X
Giữa kỳ	X				X	X	X	X		X
Tự luận	X	X	X	X	X	X	X	X		X

13. Tài liệu phục vụ cho học phần (các tài liệu xuất bản trong 05 năm trở lại đây và cung cấp được cho Trung tâm Học liệu nơi đặt tài liệu)

13.1. Tài liệu chính

[1]. Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann.

[2]. Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2019). *Data mining for business analytics: concepts, techniques and applications in Python*. John Wiley & Sons.

13.2. Tài liệu tham khảo

[1]. Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods (Business in the Digital Economy) (July 1, 2014) by S. Finlay (Author)

[2]. Rosa, G. J. (2010). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* by HASTIE, T., TIBSHIRANI, R., and FRIEDMAN, J

13.3. Phần mềm yêu cầu

[1]. Weka

[2]. Python V3.0

- [3]. Jupyter Notebook
- [4]. Visual Code Studio

TP.Hồ Chí Minh, ngày 12 tháng 04 năm 2024

Hiệu trưởng
Duyệt

Phó Trưởng Khoa
(Ký và ghi rõ họ tên)



NGUYỄN THỊ DIỆU ANH

Giảng viên biên soạn
(Ký và ghi rõ họ tên)



**NGUYỄN HỮU HƯƠNG
XUÂN**