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

ĐỀ CƯƠNG MÔN HỌC
DS105 – PHÂN TÍCH và TRỰC QUAN DỮ LIỆU

1. THÔNG TIN CHUNG (General information)

Tên môn học (tiếng Việt):	Phân tích và Trực quan Dữ liệu
Tên môn học (tiếng Anh):	Data Analysis and Visualization
Mã môn học:	DS105
Thuộc khối kiến thức:	Đại cương ☐; Cơ sở nhóm ngành ☐; Cơ sở ngành ☑; Chuyên ngành ☐; Tốt nghiệp ☐
Khoa, Bộ môn phụ trách:	Bộ môn Khoa học Dữ liệu
Giảng viên biên soạn:	ThS. Phạm Thế Sơn Email: sonpt@uit.edu.vn
Giảng viên tham gia giảng dạy:	
Số tín chỉ:	3
Lý thuyết:	2
Thực hành:	1
Tự học:	3
Môn học tiên quyết:	
Môn học trước:	Xác suất thống kê, Nhập môn lập trình Thu thập và tiền xử lý dữ liệu

2. MÔ TẢ MÔN HỌC (Course description)

Môn học tập trung vào các nội dung chính sau: (1) Giới thiệu các khái niệm, các quy trình, các bộ dữ liệu liên quan trong quá trình phân tích dữ liệu. (2) Nhập, xuất, sắp xếp, tiền xử lý bộ dữ liệu. (3) Phương pháp thăm dò dữ liệu. (4) Phát triển các mô hình phân tích dữ liệu, cách chọn mô hình phân tích dữ liệu sao cho thích hợp hiệu quả với nguồn dữ liệu, cung cấp các kiến thức nâng cao để người học có thể tự thiết kế phát triển các mô hình nghiên cứu trong phân tích dữ liệu. (5) Đánh giá mô hình phân tích dữ liệu. (6) Các kiến thức toán cơ bản thống kê trong phân tích dữ liệu. (7) Các công cụ và phương pháp trực quan hóa dữ liệu trong quá trình phân tích.

Trong môn học này, Python đóng vai trò chính hỗ trợ phân tích dữ liệu, chủ yếu tập trung vào các thư viện hỗ trợ sau: Pandas, NumPy, Scipy, MatlotLib, Seaborn, Scikit-learn, Statsmodels…

Ngoài ra, môn học trang bị thêm một số kỹ năng hướng dẫn đọc tài liệu thành thạo (đọc hiểu các project requirements document), kỹ năng tiến hành nghiên cứu, kỹ năng viết báo cáo phân tích dữ liệu, trình bày thuyết minh xây dựng đề tài môn học và làm việc nhóm, phối hợp với nhau để hoàn thành đề tài.

3. MỤC TIÊU MÔN HỌC (Course goals)

Sau khi hoàn thành môn học này, sinh viên có thể:

Bảng 1.

Ký hiệu	Mục tiêu môn học	Chuẩn đầu ra trong CTĐT
G1	Cung cấp kiến thức đóng vai trò cốt lõi quan trọng trong phân tích dữ liệu.	LO2
G2	Giúp sinh viên có khả năng chọn lựa, phát triển mô hình phân tích dữ liệu thích hợp với nguồn dữ liệu.	LO3
G3	Có khả năng đánh giá mô hình phân tích dữ liệu.	LO3, LO5,
G4	Định hướng vị trí việc làm cho sinh viên, hoàn thành đồ án của môn học qua các kỹ thuật lập trình trong phân tích dữ liệu.	LO5

CHUẨN ĐẦU RA MÔN HỌC (Course learning outcomes)

Bảng 2.

CĐRMH	Mô tả CĐRMH (Mục tiêu cụ thể)	Mức độ giảng dạy
G1.1 (LO2)	Phân biệt các khái niệm, các quy trình trong phân tích dữ liệu; mô tả, chọn lựa được các nguồn dữ liệu thích hợp.	I
G2.1 (LO3)	Áp dụng các thư viện có sẵn để phát triển mô hình phân tích dữ liệu.	IT
G3.1 (LO5)	Đánh giá các mô hình phân tích dữ liệu.	TU
G3.2 (LO3)	Áp dụng đúng các công cụ và phương pháp trực quan hóa dữ liệu trong quá trình phân tích.	ITU
G4.1 (LO5)	Đánh giá, phản biện các kết quả phân tích dữ liệu.	ITU

4. NỘI DUNG MÔN HỌC, KẾ HOẠCH GIẢNG DẠY (Course content, lesson plan)

a. Lý thuyết

Bảng 3.

Buổi học (4 tiết)	Nội dung	CĐR MH	Hoạt động dạy và học	Thành phần đánh giá
Buổi 1 (4 tiết)	**Chương 1**: Tổng quan về phân tích và trực quan dữ liệu - Phân tích dữ liệu? Tại sao phải phân tích dữ liệu. - Đặt vấn đề trong phân tích dữ liệu. - Phương pháp tìm nguồn dữ liệu, hiểu dữ liệu. - Nhập và xuất bộ dữ liệu từ nhiều nguồn lưu trữ. - Giới thiệu các thư viện (Pandas, NumPy, Scipy, MatlotLib, Seaborn, Scikit-learn, Statsmodels…) hỗ trợ lĩnh vực khoa học dữ liệu. - Giới thiệu Anaconda Navigator, Jupyter Notebook, Jupyter Lab, PyCharm.	G1.1 G2.1	Dạy: Thuyết giảng. Học ở lớp: Tiếp thu, thảo luận nhóm. Học ở nhà: Đọc thêm tài liệu, hoàn thành bài tập.	A1
	Chương 2: Sắp xếp dữ liệu - Tiền xử lý dữ liệu. - Xử lý các giá trị bị khuyết. - Định dạng dữ liệu. - Chuẩn hóa dữ liệu. - Binning. - Biến đổi các biến phân loại thành các biến định lượng.	G2.1	Dạy: Thuyết giảng. Học ở lớp: Tiếp thu, thảo luận nhóm. Học ở nhà: Đọc thêm tài liệu, hoàn thành bài tập.	A3
Buổi 2 (4 tiết)	**Chương 3**: Phân tích thăm dò dữ liệu - Giới thiệu phân tích thăm dò dữ liệu. - Thống kê mô tả. - Gom nhóm dữ liệu. - Độ tương quan. - Độ tương quan trong thống kê. - Phân tích phương sai.	G3.2	Dạy: Thuyết giảng. Học ở lớp: Tiếp thu, thảo luận nhóm. Học ở nhà: Đọc thêm tài liệu, hoàn thành bài tập.	A3
Buổi 3, 4 (8 tiết)	**Chương 4**: Phát triển mô hình - Giới thiệu phát triển mô hình. - Hồi qui tuyến tính đơn biến và đa biến. - Đánh giá mô hình dùng trực quan.	G2.1	Dạy: Thuyết giảng. Học ở lớp: Tiếp thu, thảo luận nhóm.	A3, A4

Buổi học (4 tiết)	Nội dung	CĐR MH	Hoạt động dạy và học	Thành phần đánh giá
	- Hồi quy đa thức. - Pipelines. - Thang đo (R-squared và MSE) dùng để đánh giá tập mẫu. - Dự đoán và Ra quyết định.		Học ở nhà: Đọc thêm tài liệu, hoàn thành bài tập.	
Buổi 5 (04 tiết)	**Chương 5**: Đánh giá mô hình - Đánh giá và sàng lọc mô hình. - Overfitting, underfitting và chọn lựa mô hình. - Ridge regression. - Grid search.	G3.1	Dạy: Thuyết giảng. Học ở lớp: Tiếp thu, thảo luận nhóm. Học ở nhà: Đọc thêm tài liệu, hoàn thành bài tập.	A3, A4
Buổi 6, 7 (08 tiết)	**Chương 6**: Trực quan hóa dữ liệu - Giới thiệu trực quan dữ liệu. - Giới thiệu Matplotlib, Seaborn. - Các loại biểu đồ thông dụng. - Các công cụ trực quan cơ bản và chuyên dụng. - Trực quan dữ liệu địa lý.	G3.2	Dạy: Thuyết giảng. Học ở lớp: Tiếp thu, thảo luận nhóm. Học ở nhà: Đọc thêm tài liệu, hoàn thành bài tập.	A1, A3
Buổi 8 (02 tiết)	**Chương 7**: Xây dựng các ứng dụng phân tích dữ liệu (Các chủ đề cơ bản và nâng cao trong các ứng dụng phân tích dữ liệu). - Phương pháp chọn domain dữ liệu. - Thiết kế quy trình phân tích dữ liệu. - Đánh giá kết quả phân tích dữ liệu.	G2.1 G3.1 G4.1	Dạy: Thuyết giảng. Học ở lớp: Tiếp thu, thảo luận nhóm. Học ở nhà: Đọc thêm tài liệu, hoàn thành bài tập.	A3, A4

b. Thực hành

Bảng 4.

Buổi học (5 tiết)	Nội dung	CĐR MH	Hoạt động dạy và học	Thành phần đánh giá
Buổi 1 (5 tiết)	Bài thực hành 1: Thực hành nhập và xuất bộ dữ liệu,	G2.1	Thực hành trên lớp theo	A3

Buổi học (5 tiết)	Nội dung	CĐR MH	Hoạt động dạy và học	Thành phần đánh giá
	cách sử dụng thư viện Pandas, NumPy, Statsmodels, Scikit-learn, MatlotLib, Seaborn …		hướng dẫn của GV.	
Buổi 2 (5 tiết)	Bài thực hành 2: Thực hành sắp xếp dữ liệu, và phân tích thăm dò dữ liệu.	G2.1	Thực hành trên lớp theo hướng dẫn của GV.	A3
Buổi 3 (5 tiết)	Bài thực hành 3: Thực hành phát triển mô hình phân tích dữ liệu.	G2.1	Thực hành trên lớp theo hướng dẫn của GV.	A3
Buổi 4 (5 tiết)	Bài thực hành 4: Thực hành đánh giá mô hình phân tích dữ liệu.	G3.1	Thực hành trên lớp theo hướng dẫn của GV.	A3
Buổi 5, 6 (10 tiết)	Bài thực hành 5: Tìm nguồn dữ liệu thích hợp cho đồ án. Sau đó, áp dụng phát triển mô hình phân tích dữ liệu, đánh giá mô hình phân tích.	G3.2 G4.1	Thực hành trên lớp theo hướng dẫn của GV.	A3

5. ĐÁNH GIÁ MÔN HỌC (Course assessment)

Bảng 5.

Thành phần đánh giá	CĐRMH	Tỷ lệ (%)
A1. Quá trình (Kiểm tra trên lớp, bài tập, đồ án, …)	G1.1, G2.1, G3.1	30%
A2. Giữa kỳ		
A3. Thực hành	G2.1, G2.2, G3.1, G4.1	20%
A4. Cuối kỳ (Đồ án)	G2.1, G2.2, G4.1	50%

Rubric của từng thành phần đánh giá trong Bảng 5

a. Rubric của thành phần đánh giá A1

Tiêu chí đánh giá	Dưới trung bình	Trung bình	Khá	Giỏi	Xuất sắc	Điểm
Quiz kiểm tra đầu buổi học	Làm 1 Quiz	Làm 2 Quiz	Làm 3 Quiz	Làm 4 Quiz	Làm 5 Quiz	5
Làm bài tập tuần	Làm 1 bài	Làm 2 bài	Làm 3 bài	Làm 4 bài	Làm 5 bài	5

b. Rubric của thành phần đánh giá A3

Tiêu chí đánh giá	Dưới trung bình	Trung bình	Khá	Giỏi	Xuất sắc	Điểm
Thực hiện bài phân tích dữ liệu	Không xác định được vấn đề cần phân tích	Xác định đúng vấn đề, Sắp xếp dữ liệu chưa hợp lý	Xác định đúng vấn đề, Sắp xếp dữ liệu hợp lý, và phát triển đúng mô hình	Đánh giá đúng mô hình phát triển để sàn lọc ra mô hình tốt	Mô hình phát triển có kết quả dự đoán cao, Tạo đúng các trực quan	10

c. Rubric của thành phần đánh giá A4

Tiêu chí đánh giá	Dưới trung bình	Trung bình	Khá	Giỏi	Xuất sắc	Điểm
Bố cục trình bày báo cáo	Sơ sài, không rõ ràng	Đầy đủ các phần trình bày theo yêu cầu	Bố cục trình bày rõ ràng	Bố cục trình bày khoa học	Bố cục trình bày khoa học và sáng tạo	1
File báo cáo	Sơ sài, nội dung không rõ ràng	Đầy đủ các nội dung theo yêu cầu	Có hình ảnh minh họa	Nội dung trình bày xúc tích	Nội dung trình bày xúc tích kết hợp với hình ảnh minh họa dễ hiểu	2
Slides trình bày	Sơ sài, nội dung không rõ ràng	Đầy đủ các nội dung theo yêu cầu	Có hình ảnh minh họa	Nội dung trình bày xúc tích kết hợp với hình ảnh minh họa dễ hiểu	Trình bày bám sát quy trình phân tích, có minh chứng, có so sánh kết quả	1
Kỹ năng trình bày	Nói nhỏ, không hiểu rõ nội dung trình bày	Nói nhỏ nhưng hiểu được nội dung trình bày	Trình bày rõ ràng và nắm được cấu trúc slides	Tự tin và trình bày tốt	Tự tin, trình bày cuốn hút, tương tác với người nghe và	1

					trả lời được các câu hỏi thắc mắc	
Sản phẩm	Sai hướng, không đúng với vấn đề đã xác định trong file báo cáo và slide	Code không rõ ràng, không thể hiện từng bước quy trình phân tích	Phát triển được mô hình khớp với vấn đề đã xác định	Xác định đúng thang đo để đánh giá mô hình	Chứng minh được mô hình có tỉ lệ cao, kèm các thang đo đánh giá.	5

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

- Dự lớp theo qui định chung của trường.
- Cách tính điểm A1: làm bài tập thực hành trên lớp theo nội dung giảng dạy, bài tập về nhà và trả lời các câu hỏi ngắn trên lớp, điểm bài tập sẽ được đánh giá tính chuyên cần của sinh viên. Điểm A1 chiếm trọng số 30% của điểm môn học. Sinh viên cần thực hiện đầy đủ yêu cầu về các loại bài tập dưới sự hướng dẫn của giảng viên.
- Điểm khuyến khích (A*): đối với các sinh viên giỏi có kỹ năng nghiên cứu khoa học, nếu trình bày đồ án môn học tốt thì được điểm khuyến khích.
- Nộp bài thu hoạch/đồ án môn học, bài tập trên lớp, bài tập về nhà đúng thời gian quy định.
- Điểm môn học = 30%Điểm A1 + 20%Điểm A3 + 50%Điểm A4 + Điểm A*. Quy về thang điểm đối đa 10 điểm.
- Sinh viên không nộp đồ án môn học cuối kỳ và báo cáo đúng hạn sẽ không được tính điểm cuối kỳ; vắng thực hành 2 buổi sẽ không được tính điểm thực hành.

7. TÀI LIỆU HỌC TẬP, THAM KHẢO

Giáo trình

1. Wes McKinney (2017). *Python for Data Analysis, 2nd Edition*. O'Reilly Media, Inc.
2. David Paper (2020). *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python*. Apress, Berkeley, CA
3. Trent Hauck, Julian Avila (2017). *scikit-learn Cookbook, 2nd Edition*. Packt Publishing
4. Samuel Burns (2019). *Python Data Visualization: An Easy Introduction to Data Visualization in Python with Matplotlip, Pandas, and Seaborn*. Independently Published

Tài liệu tham khảo

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). *Scikit-learn: Machine learning in Python*. Journal of machine learning research, 12(Oct), 2825-2830.
2. Wes McKinney (2015). *pandas: a Python data analysis library*. Đường dẫn: http://pandas.pydata.org
3. John Hunter, Michael Droettboom. *Matplotlib*. Đường dẫn: https://www.aosabook.org/en/matplotlib.html

8. PHẦN MỀM HAY CÔNG CỤ HỖ TRỢ THỰC HÀNH

1. Anaconda. Đường dẫn: https://www.anaconda.com/open-source
2. PyCharm Community Edition. Đường dẫn: https://www.jetbrains.com/pycharm/

Tp.HCM, ngày 26 tháng 08 năm 2020

Trưởng khoa/bộ môn	**Giảng viên biên soạn**
(Ký và ghi rõ họ tên)	(Ký và ghi rõ họ tên)
Nguyễn Văn Kiệt	**Phạm Thế Sơn**