TRƯỜNG ĐẠI HỌC VĂN LANG
Khoa Thương Mại

ĐỀ THI/ĐỀ BÀI, RUBRIC VÀ THANG ĐIỂM
THI KẾT THÚC HỌC PHẦN
Học kỳ 1, năm học 2024-2025

I. Thông tin chung

Tên học phần:	KHAI THÁC VÀ PHÂN TÍCH DỮ LIỆU (trong Marketing)		
Mã học phần:	71MISS40233	Số tín chỉ:	3
Mã nhóm lớp học phần:	71K28MARK; 71K29TMDT		
Hình thức thi: **Đồ án/Tiểu luận (Thuyết trình/Không thuyết trình)**		Thời gian làm bài: **14**	ngày
☐ GV giao đề bài trong thời gian giảng dạy lớp học phần	☒ TT. Khảo thí thiết lập và giao đề bài trên hệ thống thi CTE theo lịch thi Phòng Đào tạo công bố		
☐ Cá nhân	☒ Nhóm	Số SV/nhóm:	5-10
Quy cách đặt tên file	***Mã SV_Ho va ten SV_Tên học phần***		

Giảng viên nộp đề thi, đáp án bao gồm cả **Lần 1 và Lần 2 trước ngày 17/11/2024**.

1. Format đề thi

- Font: Times New Roman
- Size: 13
- Quy ước đặt tên file đề thi/đề bài:

+ Mã học phần_Tên học phần_Mã nhóm học phần_TIEUL_De 1

2. Giao nhận đề thi

Sau khi kiểm duyệt đề thi, đáp án/rubric. **Trưởng Khoa/Bộ môn** gửi đề thi, đáp án/rubric về Trung tâm Khảo thí qua email: khaothivanlang@gmail.com bao gồm file word và file pdf (***nén lại và đặt mật khẩu file nén***) và nhắn tin + họ tên người gửi qua số điện thoại **0918.01.03.09** (Phan Nhất Linh).

II. Các yêu cầu của đề thi nhằm đáp ứng CLO

(Phần này phải phối hợp với thông tin từ đề cương chi tiết của học phần)

Ký hiệu CLO	Nội dung CLO	Hình thức đánh giá	Trọng số CLO trong thành phần đánh giá (%)	Câu hỏi thi số	Điểm số tối đa	Lấy dữ liệu đo lường mức đạt PLO/PI
(1)	(2)	(3)	(4)	(5)	(6)	(7)
CLO1	Hiểu được quy trình và các phương pháp thường được sử dụng trong phân tích dữ liệu trong Marketing.	Báo cáo nhóm	20%	Task 1, Task 2, Task 3	Task 1: 35/100 Task 2: 35/100 Task 3: 30/100	M – PI 4.1 R – PI 5.1 R – PI 8.3 R – PI 9.2 R – PI 10.1
CLO2	Áp dụng các phương pháp và công cụ nghiên cứu định tính và định lượng để phân tích, tổng hợp, và đánh giá dữ liệu và thông tin về các hoạt động kinh doanh và marketing của doanh nghiệp.	Báo cáo nhóm	50%	Task 1, Task 2, Task 3	Task 1: 35/100 Task 2: 35/100 Task 3: 30/100	M – PI 4.1 R – PI 5.1 R – PI 8.3 R – PI 9.2 R – PI 10.1
CLO3	Áp dụng hiệu quả kỹ năng làm việc nhóm và kỹ năng làm việc độc lập để phát triển bản thân và thực hiện công việc hiệu quả.	Báo cáo nhóm	10%	Task 1, Task 2, Task 3	Task 1: 35/100 Task 2: 35/100 Task 3: 30/100	M – PI 4.1 R – PI 5.1 R – PI 8.3 R – PI 9.2 R – PI 10.1
CLO4	Hình thành ý thức học tập suốt đời để làm việc hiệu quả và phát triển con đường sự nghiệp.	Báo cáo nhóm	10%	Task 1, Task 2, Task 3	Task 1: 35/100 Task 2: 35/100 Task 3: 30/100	M – PI 4.1 R – PI 5.1 R – PI 8.3 R – PI 9.2 R – PI 10.1
CLO5	Thể hiện tinh thần trách nhiệm; khả năng chịu được áp lực trong công việc; trung thực và có đạo đức nghề nghiệp; có tính kỷ luật trong môi trường học tập và làm việc.	Báo cáo nhóm	10%	Task 1, Task 2, Task 3	Task 1: 35/100 Task 2: 35/100 Task 3: 30/100	M – PI 4.1 R – PI 5.1 R – PI 8.3 R – PI 9.2 R – PI 10.1

Chú thích các cột:

(1) Chỉ liệt kê các CLO được đánh giá bởi đề thi kết thúc học phần (tương ứng như đã mô tả trong đề cương chi tiết học phần). Lưu ý không đưa vào bảng này các CLO không dùng bài thi kết thúc học phần để đánh giá (có một số CLO được bố trí đánh giá bằng bài kiểm tra giữa kỳ, đánh giá qua dự án, đồ án trong quá trình học hay các hình thức đánh giá quá trình khác chứ không bố trí đánh giá bằng bài thi kết thúc học phần). Trường hợp một số CLO vừa được bố trí đánh giá quá trình hay giữa kỳ vừa được bố trí đánh giá kết thúc học phần thì vẫn đưa vào cột (1)

(2) Nêu nội dung của CLO tương ứng.

(3) Hình thức kiểm tra đánh giá có thể là: trắc nghiệm, tự luận, dự án, đồ án, vấn đáp, thực hành trên máy tính, thực hành phòng thí nghiệm, báo cáo, thuyết trình,..., phù hợp với nội dung của CLO và mô tả trong đề cương chi tiết học phần.

(4) Trọng số mức độ quan trọng của từng CLO trong đề thi kết thúc học phần do giảng viên ra đề thi quy định (mang tính tương đối) trên cơ sở mức độ quan trọng của từng CLO. Đây là cơ sở để phân phối tỷ lệ % số điểm tối đa cho các câu hỏi thi dùng để đánh giá các CLO tương ứng, bảo đảm CLO quan trọng hơn thì được đánh giá với điểm số tối đa lớn hơn. Cột (4) dùng để hỗ trợ cho cột (6).

(5) Liệt kê các câu hỏi thi số (câu hỏi số ... hoặc từ câu hỏi số... đến câu hỏi số...) dùng để kiểm tra người học đạt các CLO tương ứng.

(6) Ghi điểm số tối đa cho mỗi câu hỏi hoặc phần thi.

(7) Trong trường hợp đây là học phần cốt lõi - sử dụng kết quả đánh giá CLO của hàng tương ứng trong bảng để đo lường đánh giá mức độ người học đạt được PLO/PI - cần liệt kê ký hiệu PLO/PI có liên quan vào hàng tương ứng. Trong đề cương chi tiết học phần cũng cần mô tả rõ CLO tương ứng của học phần này sẽ được sử dụng làm dữ liệu để đo lường đánh giá các PLO/PI. Trường hợp học phần không có CLO nào phục vụ việc đo lường đánh giá mức đạt PLO/PI thì để trống cột này.

III. Nội dung đề bài

1. Đề bài

- **ĐỀ BÀI bao gồm 3 TASKS:**

  1. Hồi Quy (Regression);

  2. Phân Loại (Classification);

  3. Trực Quan Hóa (Visualization).

❖ TASK 1 (35 điểm): REGRESSION PROBLEM

Đề thi: Dự Báo Giá Nhà Sử Dụng Phương Pháp Hồi Quy

- **Mục Tiêu**

Sinh viên cần xây dựng một mô hình dự đoán giá nhà (SalePrice) sử dụng bộ dữ liệu được cung cấp. Nhiệm vụ của sinh viên:

1. Xử lý và tiền xử lý dữ liệu (Data Preprocessing).
2. Phân tích dữ liệu khám phá (EDA).
3. Tạo và Lựa chọn Biến Độc Lập (Feature Engineering and Feature Selection).
4. Xây dựng mô hình hồi quy (Modeling).
5. Đánh giá mô hình hồi quy (Evaluation).
6. Diễn giải và đưa ra các insight thực tiễn.

- **Hướng Dẫn**

1. Sử dụng bộ dữ liệu **train.csv** đã được cung cấp.

2. Sinh viên có thể sử dụng **bất kỳ công cụ nào** (Python, RapidMiner, Excel, R, SPSS...).

- **Giai Đoạn 1 (15 điểm): Tiền Xử Lý Dữ Liệu**

1. **Xác Định Giá Trị Khuyết Thiếu (Missing Data Detection)**:
   1.1. Xác định kiểu dữ liệu của từng biến (số liên tục, danh mục, thứ tự).
   1.2. Liệt kê tất cả các biến có giá trị khuyết thiếu.
2. **Xử Lý Giá Trị Khuyết Thiếu (Missing Data Imputation)**:
   2.1. Áp dụng phương pháp xử lý phù hợp:
      2.1.1. Biến số: Điền giá trị trung bình, trung vị, hoặc sử dụng phương pháp dự đoán (KNN, MICE, etc).
      2.1.2. Biến danh mục: Điền giá trị phổ biến nhất (mode) hoặc giá trị thay thế (None).
   2.2. Giải thích lý do lựa chọn phương pháp xử lý cho từng biến.
3. **Phát Hiện và Xử Lý Giá Trị Outliers**:
   3.1. Sử dụng các phương pháp như IQR hoặc z-score để phát hiện ngoại lai.
   3.2. Mô tả cách xử lý (loại bỏ, giữ lại hoặc thay thế) và lý do.
4. **Mã Hóa Dữ Liệu (Encoding)**:
   4.1. Chuyển đổi các biến Categorical (ví dụ: MSZoning) sang dạng số (mã hóa one-hot encoding hoặc thứ tự ordinal).
   4.2. Giải thích chiến lược mã hóa cho các biến thứ tự (ví dụ: OverallQual).
5. **Chuẩn Hóa Biến Số (Normalization/Standardization)**:
   5.1. Chuẩn hóa các biến số nếu mô hình yêu cầu.
   5.2. Ghi lại phương pháp sử dụng và lý do.

- **Giai Đoạn 2 (8 điểm): Phân Tích Dữ Liệu Khám Phá (EDA)**

1. **Thống Kê Mô Tả**:
   1.1. Cung cấp các thống kê mô tả (trung bình, trung vị, mode, độ lệch chuẩn) cho các biến số.
   1.2. Phân tích tần suất cho các biến danh mục.

2. **Phân Tích Đơn Biến**:

Vẽ biểu đồ phân phối cho các biến quan trọng (ví dụ: biểu đồ histogram cho biến phụ thuộc SalePrice).

3. **Phân Tích Song Biến (giữa các biến độc lập và biến phụ thuộc)**:

Phân tích mối quan hệ giữa SalePrice và các biến dự đoán:

3.1.Sử dụng biểu đồ scatterplot cho các biến số.

3.2.Sử dụng boxplot cho các biến danh mục.

4. **Phân Tích Tương Quan**:

4.1.Tính ma trận tương quan giữa các biến số (heatmap).

4.2.Xác định các biến có tương quan cao với SalePrice.

- **Giai Đoạn 3 (2 điểm): Tạo và Lựa chọn Biến Độc Lập**

1. **Tạo Biến Độc Lập Mới (nếu cần thiết)**:

1.1.Tạo Biến mới (ví dụ: Age = YrSold - YearBuilt).

2. **Lựa Chọn Biến Độc Lập**:

2.1.Xác định các biến dự đoán quan trọng nhất sử dụng dựa vào Phân tích tương quan với Biến Phụ Thuộc.

2.2.Loại bỏ các biến không liên quan hoặc dư thừa.

- **Giai Đoạn 4 (8 điểm): Xây Dựng Mô Hình Hồi Quy và Đánh giá mức độ hiệu quả**

1. Xây dựng một mô hình hồi quy tuyến tình trên training set.
2. Áp dụng mô hình vào testing set.

- **Giai Đoạn 5 (2 điểm): Đánh Giá và Diễn Giải Kết Quả**

1. Đánh giá độ chính xác của mô hình trên training set và testing set, sử dụng các chỉ số như RMSE, MAE, R^2.
2. Cung cấp các insight thực tiễn (ví dụ: "Tăng một đơn vị của GrLivArea làm tăng SalePrice thêm X đơn vị.").

❖ TASK 2 (35 điểm): CLASSIFICATION PROBLEM

Đề Thi: Phân Khúc Khách Hàng Sử Dụng Phương Pháp RFM và Phân Loại Không Giám Sát

- **Mục Tiêu**

Sinh viên cần:

1. Thực hiện toàn bộ quy trình phân tích dữ liệu trên bộ dữ liệu phân khúc khách hàng.

2. Áp dụng phương pháp RFM để phân loại khách hàng và diễn giải kết quả.
3. Thực hiện các phương pháp phân loại không giám sát (như k-means) và đánh giá kết quả phân nhóm bằng biểu đồ Elbow và Silhouette.

- **Hướng Dẫn**

1. Sử dụng bộ dữ liệu **RFM.csv** được cung cấp.
2. Sinh viên có thể sử dụng bất kỳ ngôn ngữ lập trình hoặc công cụ nào (Python, RapidMiner, Excel, SPSS, R...).

Yêu Cầu Chi Tiết

- **Giai Đoạn 1 (15 điểm): Tiền Xử Lý Dữ Liệu**

1. **Xử Lý Giá Trị Khuyết Thiếu (Missing Data Imputation)**:
   1.1. Xác định các cột chứa giá trị khuyết thiếu (nếu có), ví dụ như CustomerID, Quantity, UnitPrice.
   1.2. Áp dụng các phương pháp xử lý phù hợp:
      1.2.1. Điền giá trị trung bình hoặc trung vị cho các biến số.
      1.2.2. Điền giá trị phổ biến nhất (mode) hoặc giá trị thay thế (None) cho các biến danh mục.
   1.3. Giải thích lý do lựa chọn phương pháp xử lý.
2. **Phát Hiện và Xử Lý Giá Trị Ngoại Lai (Outliers)**:
   2.1. Sử dụng các kỹ thuật như IQR hoặc z-score để phát hiện các ngoại lai trong các cột như Quantity và UnitPrice.
   2.2. Quyết định giữ lại, loại bỏ hoặc thay thế giá trị ngoại lai, và giải thích lý do.
3. **Tạo Biến Recency, Frequency, Monetary**:
   3.1. **Recency (R):** Số ngày kể từ lần mua hàng gần nhất.
   3.2. **Frequency (F):** Tổng số giao dịch của mỗi khách hàng.
   3.3. **Monetary (M):** Tổng chi tiêu của mỗi khách hàng: Monetary = Quantity * UnitPrice.

- **Giai Đoạn 2 (10 điểm): Phân Loại Khách Hàng Bằng RFM**

1. **Tính Toán Điểm RFM**:
   1.1. Gán điểm từ 1 đến 4 cho từng biến Recency, Frequency, và Monetary dựa trên tứ phân vị (quartile).
   1.2. Tạo điểm RFM tổng hợp (ví dụ: R+F+M) để phân loại khách hàng.

2. **Phân Loại Khách Hàng**:

Phân chia khách hàng thành các nhóm như:

Champions: Khách hàng thường xuyên mua gần đây và chi tiêu cao.

Loyal Customers: Khách hàng trung thành, mua thường xuyên.

At-Risk: Khách hàng có nguy cơ rời bỏ.

…

- **Giai Đoạn 3 (10 điểm): Phân Loại Không Giám Sát**

1. **Áp Dụng Các Thuật Toán Phân Loại Không Giám Sát**:

Sử dụng ít nhất một thuật toán phân cụm (ví dụ: k-means, hierarchical clustering) để phân nhóm khách hàng dựa trên giá trị RFM.

2. **Xác Định Số Lượng Nhóm Tối Ưu**:

2.1 Biểu Đồ Elbow: Vẽ biểu đồ tổng bình phương khoảng cách trong nhóm (WCSS) theo số lượng nhóm.

2.2 Silhouette Score: Tính điểm Silhouette để đánh giá mức độ cô lập và liên kết của các nhóm.

❖ TASK 3 (30 điểm): DATA VISUALIZATION

- **Mục Tiêu**

Sinh viên phải sử dụng Tableau để khám phá và trực quan hóa bộ dữ liệu Walmart Retail. Mục tiêu là tạo ra nhiều loại biểu đồ nhất có thể để hiểu rõ dữ liệu và rút ra các insight thực tiễn.

**Tiêu chí tối thiểu mỗi nhóm sử dụng 10 biểu đồ để biểu đạt các insight ẩn sau tập dữ liệu:

a) Nhóm nào có thêm nhiều hình ảnh thì số điểm càng cao.
b) Nhóm nào có độ sáng tạo trong cách trực quan hóa dữ liệu (kết hợp đa thông tin, nhiều cột dữ liệu trong cùng một biệu đồ; hoặc cấu trúc biểu đồ có độ "lạ" và hiệu quả).

c) Những giải thích lập luận cho các biểu đồ có đi kèm trích dẫn nguồn, trình bày khoa học, độ sâu và nhạy bén của phần luận giải của từng biểu đồ.

1. **Giới thiệu về tập dữ liệu (4 điểm)**

1.1. Dataset này thể hiện dữ liệu của ngành công nghiệp nào?

1.2. Giới thiệu sơ lược về ngành công nghiệp đó, ở nơi mà tập dữ liệu này được tạo ra (1 trang).

1.3. Cấu trúc dataset này như thế nào? Ví dụ: liệt kê các Biến số (numeric: continuous, discrete), các Biến định danh (categorical: nominal, ordinal), các Biến dữ liệu tọa độ, etc.

1.4. Nêu ra các Biến chứa missing values? Chỉ rõ bao nhiều dòng, và bao nhiêu % số dòng thuộc Biến đó gặp tình trạng missing values?

1.5. Có xử lý missing values không? Nêu phương pháp Imputation (trám dữ liệu) cho từng Biến có chứa missing values.

2. **Các bước chuẩn bị (2 điểm)**

2.1. Bao nhiêu Biến được sử dụng trong bài phân tích? Liệt kê các Biến đó.

2.2. Nêu sơ lược các nội dung muốn truyền tải đến người đọc thông qua bài phân tích.

Ví dụ:

Bài phân tích nêu lên các nội dung chính bao gồm: (1) thống kê doanh thu tổng của chuỗi siêu thị theo (a) năm, (b) quý, (c) tháng; (2) doanh thu của từng category theo (a) năm, (b) quý, (c) tháng; (3) phân tích doanh thu/lợi nhuận của từng category hàng hóa, (4) lượng tiêu thụ từng loại hàng hóa theo khu vực, (n) etc.

2.3. Nêu ra các Biến tham gia trong từng mục (theo 2.2)

Ví dụ:

Visual chart 1 = cột A + cột B + cột C

Visual chart 2 = cột C + cột D + cột F

2.4. Nếu mục nào có tạo biến mới thì kê khai biến đó ra. Và nếu biến mới được tạo ra bởi hàm/syntax thì kê khai hàm/câu lệnh ra, hoặc nêu cách tạo biến mới.

Ví dụ:

Visual chart 1 có tạo ra biến mới, tên là Profit ratio (vốn không có trong dataset gốc). Vậy biến mới Profit ratio được tạo ra bằng cách:

Profit ratio = Profit (cột A trong dataset) / Revenue (Cột B trong dataset)

...

3. **Data visualization (20 điểm)**

3.1.Trình bày các Figures (hình ảnh) ứng với thứ tự đã kê khai trong 2.2

Ví dụ:

(Visual chart 1a) thống kê doanh thu tổng của chuỗi siêu thị theo năm

(Visual chart 1b) thống kê doanh thu tổng của chuỗi siêu thị theo quý

(Visual chart 1c) thống kê doanh thu tổng của chuỗi siêu thị theo tháng

3.2.Giải thích insight ứng với các Figures trong 3.1

4. **Kết luận và nhận định sau khi phân tích (4 điểm)**

4.1.Nhận định chung về tình hình kinh doanh đã quan sát được từ việc trực quan hóa tập dữ liệu

4.2.Đưa ra suggestions để cải thiện issues mà Walmart Retail gặp phải.

Ví dụ:

Theo tập dữ liệu về tình hình kinh doanh của siêu thị, nhóm chúng tôi phát hiện sự chậm trễ thường xảy ra trong việc giao những món hàng abcd, vậy cần cải thiện bằng cách xyz. Hoặc,

Vào tháng mưa, thì các mặt hàng nên được nhập chủ đạo bao gồm: a,b,c,d,...Vào tháng nắng thì các mặt hàng nên được nhập nhiều bao gồm; x,y,z,...

2. Hướng dẫn thể thức trình bày đề bài

- **HƯỚNG DẪN SINH VIÊN TRÌNH BÀY:**
  - ✓ **Trang bìa:** Tên trường, Tên môn học, Tiêu đề, Tên Giảng viên, Tên nhóm
  - ✓ **Trang 2:** Tên các thành viên + Đóng góp của từng thành viên trong bài report cuối kỳ (các thành viên ký tên xác nhận)
  - ✓ **Table of contents**

✓ **Nội dung bài báo cáo**

3. Thang điểm

Phần câu hỏi	Thang điểm	Ghi chú
Task 1	**35**	
Giai đoạn 1	15	
Giai đoạn 2	8	
Giai đoạn 3	2	
Giai đoạn 4	8	
Giai đoạn 5	2	
Task 2	**35**	
Giai đoạn 1	15	
Giai đoạn 2	10	
Giai đoạn 3	10	
Task 3	**10.0**	
Giai đoạn 1	4	
Giai đoạn 2	2	
Giai đoạn 3	20	
Giai đoạn 4	4	
Tổng	**100**	

******LƯU Ý******

a) Đây là báo cáo nhóm, điểm số sẽ được cho dựa trên độ hoàn thiện của bài báo cáo. Nhóm trưởng đánh mức độ hoàn thiện công việc của từng thành viên (theo thang 0-100%).

b) Trong trường hợp có mâu thuẫn đánh giá đối với thành viên X bất kỳ nào đó trong nhóm, các thành viên còn lại trong nhóm được phép voting. Nếu trên 50% số thành viên đồng thuận với đánh giá của nhóm trưởng cho thành viên X, thì sự đánh giá của nhóm trưởng dành cho thành viên X là không cần phải thay đổi.

c) Điểm số các cá nhân = điểm tổng chung của bài nhóm * % mức độ hoàn thiện công việc của từng thành viên

TP. Hồ Chí Minh, ngày 16 tháng 11 năm 2024

Người duyệt đề	**Giảng viên ra đề**
TS. Trần Nguyễn Hải Ngân	**TS. LƯƠNG THÁI HÀ**