TRƯỜNG ĐẠI HỌC Y TẾ CÔNG CỘNG
CHUYÊN NGÀNH: KHOA HỌC DỮ LIỆU

TIỂU LUẬN KẾT THÚC HỌC PHẦN: HỆ KHUYẾN NGHỊ

Họ và tên	**:**	**Đinh Lê Quỳnh Phương**
MSSV	**:**	**2211090031**
Lớp	**:**	**CNCQ KHDL1-1A**
Giảng viên	**:**	**Trần Minh Quân**
Năm học	**:**	**2024-2025**

Năm 2025

MỤC LỤC

LỜI MỞ ĐẦU

Trong bối cảnh dữ liệu và công nghệ phát triển mạnh mẽ hiện nay, hệ thống khuyến nghị ngày càng đóng vai trò quan trọng trong việc hỗ trợ người dùng đưa ra quyết định—từ lựa chọn sản phẩm, nội dung giải trí cho đến học tập. Thông qua học phần *Hệ thống khuyến nghị*, em đã được trang bị kiến thức nền tảng cùng các kỹ năng thực hành cần thiết để xây dựng những hệ thống thông minh dựa trên học máy và khai phá dữ liệu.

Trong quá trình học, em đã tìm hiểu và triển khai nhiều phương pháp khuyến nghị như lọc cộng tác theo người dùng, theo sản phẩm, các mô hình học máy như cây quyết định, rừng ngẫu nhiên, cũng như kỹ thuật khai phá luật kết hợp để đưa ra các gợi ý có ý nghĩa. Bên cạnh đó, em cũng rèn luyện được kỹ năng tiền xử lý dữ liệu—một bước quan trọng trong mọi hệ thống trí tuệ nhân tạo.

Bài tiểu luận này được thực hiện nhằm tổng hợp, phân tích và đánh giá những nội dung em đã học và thực hành trong suốt học phần, đồng thời làm rõ sự tiến bộ của bản thân thông qua các minh chứng cụ thể từ notebook và các mô hình đã triển khai.

MỤC TIÊU VÀ PHƯƠNG PHÁP

A. Mục tiêu

Bài tiểu luận được thực hiện với mục tiêu:

- Đánh giá quá trình học tập và thực hành của bản thân trong suốt học phần *Hệ thống khuyến nghị*.
- Làm rõ các kỹ thuật và mô hình đã được triển khai trong từng giai đoạn học tập.
- Sử dụng bài tập và dự án cuối kỳ làm minh chứng cụ thể cho sự phát triển kỹ năng và tư duy thực hành.

B. Phương pháp

Bài tiểu luận sử dụng phương pháp tổng hợp và phân tích nội dung từ notebook thực hành, cụ thể gồm các khía cạnh:

- **Tiền xử lý dữ liệu:** Làm sạch và chuẩn hóa dữ liệu đầu vào để phục vụ cho các mô hình khuyến nghị.
- **Xây dựng hệ thống lọc cộng tác:** Bao gồm lọc cộng tác dựa trên người dùng và sản phẩm.
- **Ứng dụng mô hình học máy:** Sử dụng các mô hình như cây quyết định (Decision Tree) và rừng ngẫu nhiên (Random Forest) để tạo và đánh giá hệ thống khuyến nghị.
- **Khai phá luật kết hợp:** Triển khai thuật toán Apriori để phát hiện các mối quan hệ giữa các mục dữ liệu và đưa ra gợi ý khuyến nghị.
- **Công cụ và ngôn ngữ:** Các kỹ thuật trên được thực hiện bằng ngôn ngữ lập trình Python, sử dụng các thư viện như pandas, scikit-surprise, sklearn, và mlxtend nhằm đảm bảo tính hiệu quả và khả năng minh họa tiến bộ cá nhân một cách rõ ràng.

I. KỸ NĂNG TIỀN XỬ LÍ DỮ LIỆU

Trong quá trình tham gia học phần *Hệ thống khuyến nghị*, em nhận ra rằng tiền xử lý dữ liệu không chỉ là bước đầu tiên mà còn là yếu tố then chốt quyết định chất lượng của toàn bộ mô hình. Đây cũng là kỹ năng mà em cảm nhận rõ sự tiến bộ nhờ vào việc được trực tiếp thao tác với các bộ dữ liệu thực tế.

Khi thực hành trong notebook, em thực hiện nhiều bước như đọc dữ liệu từ tệp CSV, kiểm tra dữ liệu trống hoặc sai định dạng, mã hóa các biến phân loại và chuẩn hóa các biến dạng số. Thời gian đầu, em khá lúng túng trong việc lựa chọn công cụ xử lý phù hợp, nhưng dần dần em đã làm quen và sử dụng hiệu quả các thư viện như pandas để xử lý dữ liệu dạng bảng, hay sklearn.preprocessing để chuẩn hóa dữ liệu thông qua OneHotEncoder và StandardScaler.

Minh chứng rõ nhất nằm ở phần *CHẶNG 1: CHUẨN BỊ DỮ LIỆU* trong notebook, trong đó em đã:

- Lựa chọn và mã hóa hợp lý các biến phân loại để mô hình có thể tiếp nhận đầu vào đúng định dạng.
- Xử lý các giá trị bị thiếu thay vì bỏ qua toàn bộ dòng dữ liệu, giúp giữ lại thông tin quan trọng mà vẫn đảm bảo tính chính xác.
- Thực hiện chia dữ liệu thành tập huấn luyện và kiểm thử nhằm đánh giá mô hình một cách có hệ thống.

Việc được làm việc với nhiều tập dữ liệu có cấu trúc và nội dung khác nhau đã giúp em cải thiện rõ rệt khả năng xử lý và làm sạch dữ liệu. Quan trọng hơn, em dần biết cách kiểm soát quy trình này một cách chủ động, thay vì chỉ làm theo hướng dẫn từng bước như lúc đầu. Đây là một trong những kỹ năng mà em cảm thấy mình tiến bộ rõ nhất trong suốt học phần.

II. KỸ NĂNG ÁP DỤNG PHƯƠNG PHÁP LỌC CỘNG TÁC DỰA TRÊN NGƯỜI DÙNG

Lọc cộng tác dựa trên người dùng (User-based Collaborative Filtering – UBCF) là một trong những phương pháp khuyến nghị phổ biến, dựa trên giả định rằng người dùng có hành vi tương đồng sẽ có xu hướng yêu thích các sản phẩm giống nhau. Trước khi tham gia học phần, em chưa từng làm việc với kỹ thuật này nên ban đầu còn khá bỡ ngỡ trong việc hiểu cơ chế tính toán độ tương đồng.

Trong quá trình thực hành, em đã triển khai UBCF bằng thư viện scikit-surprise, sử dụng lớp KNNBasic với độ đo Pearson để xác định độ tương đồng giữa các người dùng. Sau khi xây dựng mô hình, em tiến hành dự đoán điểm đánh giá và quan sát được rằng hệ thống có khả năng đề xuất những bộ phim phù hợp dựa trên hành vi của những người dùng có sở thích tương tự.

Minh chứng cụ thể nằm trong phần thực hành về *lọc cộng tác người dùng*, trong đó em:

- Lựa chọn và xử lý bộ dữ liệu đánh giá phim.
- Áp dụng thuật toán KNN với tham số tối ưu để xây dựng mô hình đề xuất.
- Đánh giá khả năng khuyến nghị bằng cách so sánh dự đoán với dữ liệu thực.

Thông qua quá trình này, em dần hiểu rõ hơn cách mô hình UBCF hoạt động, từ giai đoạn xử lý dữ liệu đến việc tính toán và đưa ra gợi ý. Kỹ năng này đã được củng cố đáng kể nhờ việc trực tiếp triển khai và điều chỉnh mô hình theo từng bước cụ thể.

III. KỸ NĂNG ÁP DỤNG PHƯƠNG PHÁP LỌC CỘNG TÁC DỰA TRÊN SẢN PHẨM

Lọc cộng tác dựa trên sản phẩm (Item-based Collaborative Filtering – IBCF) là kỹ thuật dựa vào sự tương đồng giữa các sản phẩm để đưa ra gợi ý cho người dùng. Thay vì tìm kiếm những người dùng có hành vi tương tự, phương pháp này tập trung vào việc phát

hiện những sản phẩm có xu hướng được đánh giá giống nhau bởi cùng một nhóm người dùng.

Trong quá trình thực hành, em đã triển khai IBCF bằng cách sử dụng lớp KNNBasic trong thư viện scikit-surprise, với tham số user_based=False nhằm tính toán độ tương đồng giữa các sản phẩm. Việc trực tiếp phân tích ma trận tương đồng giúp em hình dung rõ hơn về cách các sản phẩm có thể được nhóm lại với nhau dựa trên hành vi đánh giá.

Minh chứng cụ thể nằm trong phần thực hành khuyến nghị theo sản phẩm, với các bước:

- Chuẩn bị và xử lý dữ liệu đầu vào theo định dạng tương thích với mô hình.
- Áp dụng thuật toán KNN và phân tích các chỉ số tương đồng giữa sản phẩm.
- Đánh giá khả năng khuyến nghị bằng cách thử nghiệm đầu vào và so sánh kết quả đầu ra.

Kỹ năng này được cải thiện rõ rệt nhờ việc em có cơ hội so sánh trực tiếp hiệu quả giữa lọc theo người dùng và sản phẩm, từ đó hiểu sâu hơn về ưu – nhược điểm của từng phương pháp và khi nào nên lựa chọn phù hợp.

IV. KỸ NĂNG LỰA CHỌN VÀ ÁP DỤNG CÁC MÔ HÌNH HỌC MÁY NHƯ CÂY QUYẾT ĐỊNH, RỪNG NGẪU NHIÊN

Trong quá trình học phần *Hệ thống khuyến nghị*, em được làm quen với việc sử dụng các mô hình học máy nhằm mục đích phân loại và dự đoán hành vi người dùng dựa trên dữ liệu thu thập được. Đây là một phần kiến thức khá mới với em, đặc biệt là khi phải trực tiếp xử lý dữ liệu thực tế để triển khai và đánh giá mô hình.

Em đã áp dụng hai mô hình phổ biến là **Cây quyết định (Decision Tree)** và **Rừng ngẫu nhiên (Random Forest)**. Với Cây quyết định, em sử dụng DecisionTreeClassifier từ thư viện sklearn.tree, còn mô hình Rừng ngẫu nhiên được triển khai bằng RandomForestClassifier từ sklearn.ensemble. Cả hai mô hình đều được huấn luyện trên

tập dữ liệu đã qua tiền xử lý, sau đó đánh giá hiệu quả trên tập kiểm thử. Các chỉ số như **độ chính xác**, **F1-score**, **ma trận nhầm lẫn**, và báo cáo phân loại (classification report) được sử dụng để phân tích kết quả đầu ra.

Trong giai đoạn đầu, em gặp khó khăn khi lựa chọn tham số phù hợp và phát hiện mô hình bị **overfitting**, đặc biệt là với Random Forest. Sau khi tìm hiểu thêm tài liệu và thử nghiệm nhiều cách, em đã áp dụng train_test_split để chia dữ liệu hợp lý, đồng thời sử dụng GridSearchCV để tìm các siêu tham số tối ưu như n_estimators, max_depth, min_samples_split và min_samples_leaf. Việc chủ động điều chỉnh và thử nghiệm giúp em hiểu rõ hơn về cách tinh chỉnh mô hình để đạt hiệu quả cao, tránh tình trạng mô hình ghi nhớ quá kỹ dữ liệu huấn luyện.

Việc triển khai và so sánh hai mô hình không chỉ giúp em rèn luyện kỹ năng lập trình và phân tích đánh giá, mà còn giúp em hình thành tư duy lựa chọn thuật toán phù hợp với từng loại bài toán cụ thể. Đây là một trong những kỹ năng mà em cảm nhận rõ sự tiến bộ của bản thân sau khi hoàn thành học phần.

V. KỸ NĂNG ÁP DỤNG MÔ HÌNH KHUYẾN NGHỊ THEO LUẬT

Một trong những kỹ thuật mà em được tiếp cận trong học phần là xây dựng hệ thống khuyến nghị dựa trên **luật kết hợp (association rules)**, với các thuật toán tiêu biểu như **Apriori** và **FP-Growth**. Đây là phương pháp mang tính khai phá dữ liệu, cho phép phát hiện các mối quan hệ tiềm ẩn giữa các mục thường xuyên xuất hiện cùng nhau trong tập dữ liệu giao dịch. Em nhận thấy kỹ thuật này đặc biệt hữu ích trong các hệ thống đề xuất sản phẩm, nơi hành vi mua hàng trong quá khứ có thể gợi ý cho hành động tiếp theo.

Trong phần thực hành, em đã sử dụng thư viện mlxtend để xây dựng mô hình trên tập dữ liệu giao dịch. Sau khi mã hóa dữ liệu ở dạng one-hot, em áp dụng hàm apriori để khai thác các tập mục phổ biến và sử dụng association_rules nhằm trích xuất các luật có độ tin cậy (confidence) và chỉ số lift cao. Việc trực tiếp thao tác với dữ liệu và điều chỉnh các

ngưỡng **support** và **confidence** cho phép em hiểu rõ hơn tác động của các thông số này đến chất lượng luật được tạo ra.

Thông qua quá trình thực hành, em cảm nhận kỹ năng này đã phát triển rõ rệt. Từ chỗ còn bỡ ngỡ với khái niệm luật kết hợp, em dần thành thạo hơn trong việc lựa chọn tiêu chí lọc, phân tích và diễn giải các luật sinh ra để đưa vào hệ thống khuyến nghị. Đây cũng là một trong những phần em cảm thấy hứng thú nhất vì giúp em nhìn rõ hơn vai trò của khai phá dữ liệu trong việc tạo ra giá trị ứng dụng thực tiễn.

VI. KẾT LUẬN

Qua quá trình tham gia học phần *Hệ thống khuyến nghị*, em không chỉ nắm vững những kiến thức lý thuyết cốt lõi mà còn từng bước hoàn thiện kỹ năng thực hành thông qua các bài tập và dự án cuối kỳ. Các kỹ năng như tiền xử lý dữ liệu, xây dựng mô hình lọc cộng tác, áp dụng các thuật toán học máy và khai phá luật kết hợp đều được em tiếp cận một cách hệ thống và trải nghiệm trực tiếp trong môi trường thực hành.

Mặc dù vẫn còn nhiều kiến thức cần tiếp tục trau dồi, song nền tảng lý thuyết và kỹ năng thực hành mà em tích lũy được trong học phần này chính là hành trang quý báu, giúp em tự tin hơn trên con đường phát triển chuyên môn trong lĩnh vực khoa học dữ liệu và trí tuệ nhân tạo.