TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO BÀI TẬP LỚN HỌC PHẦN: THỰC TẬP CƠ SỞ NGÀNH

XÂY DỰNG HỆ THỐNG GỢI Ý PHIM

Giáo viên hướng dẫn: TS. Nguyễn Xuân Hoàng

Sinh viên thực hiện: Nguyễn Bá Hưởng - 2023600809
Trần Ngọc Dương - 2023600953
Nguyễn Đăng Lộc - 2023603891
Chu Thị Sương - 2021602368

Lớp, khoa: 2023DHHTTT01 - CNTT

Hà Nội - năm 2025

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO BÀI TẬP LỚN

MÔN HỌC: THỰC TẬP CƠ SỞ NGÀNH

XÂY DỰNG HỆ THỐNG GỢI Ý PHIM

Sinh viên thực hiện:

Nguyễn Bá Hưởng. Giới tính: Nam. Dân tộc: Kinh

Lớp, khoa: Hệ thống thông tin 1, Công nghệ thông tin **Năm thứ: 2/4**

Ngành học: Hệ thống thông tin

Trần Ngọc Dương. Giới tính: Nam. Dân tộc: Kinh

Lớp, khoa: Hệ thống thông tin 1, Công nghệ thông tin Năm thứ: 2/4

Ngành học: Hệ thống thông tin

Nguyễn Đăng Lộc. Giới tính: Nam. Dân tộc: Kinh

Lớp, khoa: Hệ thống thông tin 1, Công nghệ thông tin Năm thứ: 2/4

Ngành học: Hệ thống thông tin

Chu Thị Sương. Giới tính: Nữ. Dân tộc: Kinh

Lớp, khoa: Hệ thống thông tin 1, Công nghệ thông tin Năm thứ:2/4

Ngành học: Hệ thống thông tin

Hà Nội - năm 2025

LỜI CẢM ƠN

Lời đầu tiên nhóm em xin phép gửi lời cảm ơn sâu sắc tới thầy ThS. Nguyễn Xuân Hoàng. Trong quá trình học tập và thực hiện đề tài này, chúng em đã nhận được sự quan tâm giúp đỡ, hướng dẫn tận tình, tâm huyết của thầy. Những gì chúng em nhận được không chỉ dừng lại ở kiến thức môn học mà nhiều hơn thế đó là những lời khuyên, chia sẻ thực tế của thầy. Để hoàn thành được đề tài này, nhóm em xin được bày tỏ sự tri ân và lòng biết ơn sâu sắc tới giảng viên ThS. Nguyễn Xuân Hoàng là người đã trực tiếp hướng dẫn, chỉ bảo cho chúng em trong suốt quá trình nghiên cứu vừa qua để hoàn thành đề tài này.

Bên cạnh đó, chúng em xin chân thành cảm ơn nhà trường và khoa đã tạo điều kiện học tập, nghiên cứu trong suốt quá trình thực tập. Môi trường học tập năng động và định hướng thực tiễn từ chương trình đào tạo đã giúp em hiểu rõ hơn về cách vận dụng kiến thức chuyên ngành vào giải quyết các bài toán thực tế, đặc biệt là trong lĩnh vực trí tuệ nhân tạo và khoa học dữ liệu – nền tảng quan trọng để xây dựng mô hình dự đoán như đề tài đã lựa chọn.

Trong quá trình thực hiện, mặc dù đã cố gắng hết sức, nhưng do thời gian và kinh nghiệm còn hạn chế, bài báo cáo chắc chắn vẫn còn nhiều thiếu sót. Chúng em kính mong nhận những lời nhận xét, góp ý từ quý thầy và bạn đọc để có thể hoàn thiện hơn trong những nghiên cứu tiếp theo.

Chúng em xin chân thành cảm ơn!

Nhóm sinh viên thực hiện

Nhóm 11

MỤC LỤC

DANH MỤC HÌNH ẢNH

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1. Tên đề tài

Xây dựng hệ thống gợi ý phim.

1.2. Lý do chọn đề tài

Trong thời đại công nghệ số phát triển mạnh mẽ, khối lượng thông tin và dữ liệu người dùng ngày càng lớn, kéo theo nhu cầu cá nhân hóa trải nghiệm sử dụng dịch vụ. Trong đó, hệ thống gợi ý (Recommendation System) đang ngày càng trở nên phổ biến và đóng vai trò quan trọng trong nhiều lĩnh vực như thương mại điện tử, giải trí, mạng xã hội...

Xuất phát từ mong muốn tìm hiểu chuyên sâu về các kỹ thuật trí tuệ nhân tạo, học máy và xử lý dữ liệu thực tế chúng em đã chọn đề tài "Xây dựng hệ thống gợi ý phim" nhằm nghiên cứu và thực hành các kỹ thuật xây dựng hệ thống gợi ý dựa trên dữ liệu người dùng. Thông qua quá trình thực hiện đề tài, chúng em mong muốn tìm hiểu về lập trình Python, thư viện Scikit-learn, kỹ thuật cosine similarity, và phương pháp K-Nearest Neighbors (KNN).

Chính bởi những lý do trên mà đề tài "Xây dựng hệ thống gợi ý phim" là có tính thực tiễn.

1.3. Mục tiêu của đề tài

Đề tài: "Xây dựng hệ thống gợi ý phim" đáp ứng được những mục tiêu:

Nghiên cứu một số kỹ thuật học máy cơ bản như: K-Nearest Neighbors (KNN) kết hợp với phương pháp tính khoảng cách Cosine để xây dựng mô hình gợi ý.

Nắm được các kiến thức cơ bản về Python và sử dụng thành thạo các thư viện như: Pandas, Numpy, Scikit - Learn, Scipy,… cùng các kiến thức, công cụ liên quan. Quá trình phát triển được thực hiện trên nền tảng Google Colab, tận dụng khả năng tính toán trên nền tảng đám mây, hỗ trợ lập trình trực tuyến, trực quan và thuận tiện trong việc thực nghiệm, phân tích và trình bày kết quả

Hiểu được quy trình xây dựng hệ thống gợi ý gồm: tiền xử lý dữ liệu, xây dựng ma trận người dùng - phim, huấn luyện mô hình và đưa ra gợi ý.

Xây dựng một hệ thống có khả năng gợi ý phim cho người dùng trên thói quen xem và đánh giá của người dùng khác, qua đó có thể ứng dụng trong các nền tảng xem phim hoặc dịch vụ tương tự

1.4. Đối tượng và phạm vi

Đối tượng của hệ thống là người dùng có nhu cầu nhận gợi ý phim dựa trên sở thích và lịch sử xem.

Phạm vi tập trung vào xây dựng hệ thống gợi ý phim sử dụng thuật toán KNN, áp dụng trên dữ liệu đánh giá phim hoặc đặc trưng nội dung.

1.5. Kết quả dự kiến đạt được

Thông qua quá trình thực hiện đề tài *"Xây dựng hệ thống gợi ý phim"*, nhóm nghiên cứu kỳ vọng đạt được những kết quả sau:

- Xây dựng thành công một hệ thống gợi ý phim cá nhân hóa có khả năng đề xuất danh sách phim phù hợp với từng người dùng, dựa trên phân tích đánh giá và mức độ tương đồng với các người dùng khác.
- Ứng dụng hiệu quả các công cụ và thư viện lập trình phổ biến trong khoa học dữ liệu, bao gồm: Python, Pandas, NumPy, Scikit-learn, cùng với thuật toán K-Nearest Neighbors (KNN). Việc sử dụng các công cụ này giúp đảm bảo tính linh hoạt, dễ mở rộng và khả năng áp dụng vào các bài toán gợi ý tương tự.
- Nắm vững và áp dụng được quy trình chuẩn trong việc xây dựng hệ thống gợi ý, từ khâu thu thập và xử lý dữ liệu, tạo ma trận người dùng – phim, đến huấn luyện mô hình và sinh kết quả gợi ý. Điều này giúp nhóm hiểu sâu hơn về kiến trúc và các bước triển khai một hệ thống gợi ý trong thực tế.
- Triển khai hệ thống trên nền tảng Google Colab với giao diện đầu ra đơn giản, trực quan, có khả năng hiển thị danh sách phim được đề xuất cho một người dùng cụ thể, từ đó kiểm chứng hiệu quả hoạt động của mô hình.
- Nâng cao năng lực cá nhân của sinh viên trong lĩnh vực trí tuệ nhân tạo và khoa học dữ liệu, đặc biệt là trong các kỹ năng: xử lý dữ liệu, áp dụng thuật toán học máy, và xây dựng hệ thống gợi ý – những yếu tố then chốt trong các ứng dụng công nghệ hiện đại.

1.6. Phương pháp nghiên cứu

Phương pháp nghiên cứu tài liệu

- Thu thập dữ liệu và đánh giá các công trình khoa học trong và ngoài nước đã công bố liên quan đến đề tài nghiên cứu.
- Phân tích và tổng hợp các nghiên cứu có trước và thực hiện so sánh, đánh giá.
- Phân loại và hệ thống các công bố theo từng nội dung nghiên cứu của đề tài.

Phương pháp thực nghiệm

- Tiến hành thực nghiệm để đánh giá hiệu năng của giải thuật đề xuất.
- Cài đặt và thử nghiệm trên phần cứng thực tế.

1.7. Tổng kết chương

Trong chương này nhóm đã trình bày những hướng cơ bản trong nghiên cứu để tìm hiểu về đề tài. Cụ thể hơn là tìm hiểu các nội dung về lý do chọn đề tài, mục tiêu của đề tài, đối tượng và phạm vi nghiên cứu, phương pháp nghiên cứu từ đó tìm ra được hướng để tìm hiểu và hoàn thành đề tài. Đề tài cũng tạo tiền đề cho việc nghiên cứu sâu hơn các kỹ thuật đề xuất nâng cao và có khả năng mở rộng sang nhiều lĩnh vực khác như thương mại điện tử, giáo dục hay giải trí số. Ngoài ra, quá trình thực hiện đề tài còn giúp sinh viên rèn luyện kỹ năng làm việc nhóm, triển khai dự án và tư duy hệ thống – những yếu tố quan trọng trong môi trường công nghệ hiện đại.

Các chương tiếp theo sẽ trình bày chi tiết về thuật toán được lựa chọn, quy trình thu thập và xử lý dữ liệu, phương pháp thiết kế và huấn luyện mô hình của hệ thống gợi ý phim. Hy vọng rằng, những kết quả nghiên cứu của đề tài này sẽ đóng góp vào sự phát triển của lĩnh vực giải trí và mang lại những lợi ích thiết thực cho cộng đồng.

PHIẾU HỌC TẬP NHÓM 11

I. Thông tin chung

1. Tên lớp: **20242IT6055002** Khóa: ĐH K18 (2023-2027)

2. Tên nhóm: **Nhóm 11**

Họ và tên thành viên trong nhóm: **Trần Ngọc Dương** (2023600953), **Nguyễn Bá Hưởng** (2023600809), **Nguyễn Đăng Lộc** (2023603891), **Chu Thị Sương** (2021602368).

II. Nội dung học tập

1. Tên chủ đề **Xây dựng hệ thống gợi ý phim**

2. Hoạt động của sinh viên *(xác định các hoạt động chính của sinh viên trong quá trình thực hiện Tiểu luận, Bài tập lớn, Đồ án/Dự án để hình thành tri thức, kỹ năng đáp ứng mục tiêu/chuẩn đầu ra nào của học phần).*

- Xây dựng hệ thống gợi ý phim dựa trên thuật toán KNN.

3. Sản phẩm nghiên cứu *(xác định cụ thể sản phẩm của chủ đề nghiên cứu cần đạt được, ví dụ: Bản thuyết minh, bài thu hoạch, mô hình, sơ đồ, bản vẽ kỹ thuật, trang website, bài báo khoa học,...)*

- Hệ thống gợi ý phim.
- Quyển báo cáo.

III. Nhiệm vụ học tập

1. Hoàn thành Tiểu luận, Bài tập lớn, Đồ án/Dự án theo đúng thời gian quy định (từ ngày 09/03/2025 đến ngày 15/06/2025)

2. Báo cáo sản phẩm nghiên cứu theo chủ đề được giao trước giảng viên và những sinh viên khác

IV. Học liệu thực hiện Tiểu luận, Bài tập lớn, Đồ án/Dự án

1. Tài liệu học tập: Giảng viên gửi bộ tài liệu học tập trước lớp.

2. Phương tiện, nguyên liệu thực hiện Tiểu luận, Bài tập lớn, Đồ án/Dự án: Bộ tài liệu giảng viên đã gửi, tham khảo thêm các tài liệu khác.

KẾ HOẠCH THỰC HIỆN TIỂU LUẬN, BÀI TẬP LỚN, ĐỒ ÁN/DỰ ÁN

Tên lớp: **20242IT6055002** Khóa: **ĐH K18 (2023-2027)**

Tên nhóm: **Nhóm 11**

Họ và tên thành viên trong nhóm : Trần Ngọc Dương (2023600953), Nguyễn Bá Hưởng (2023600809), Nguyễn Đăng Lộc (2023603891), Chu Thị Sương (2021602368).

Tên chủ đề: **Xây dựng hệ thống gợi ý phim**

Tuần	Người thực hiện	Nội dung công việc	Phương pháp thực hiện
1	Nguyễn Bá Hưởng Trần Ngọc Dương Nguyễn Đăng Lộc Chu Thị Sương	- Lập nhóm - Tìm hiểu về các đề tài và thống nhất chọn đề tài - Bầu nhóm trưởng - Phân công vai trò cho các thành viên	- Họp nhóm qua Zoom - Bình chọn trong nhóm Zalo - Tìm hiểu đề tài qua các phương tiện thông tin
2	Nguyễn Bá Hưởng Trần Ngọc Dương Nguyễn Đăng Lộc Chu Thị Sương	- Tìm hiểu quy trình và cách thức nghiên cứu đề tài - Thống nhất quy trình thực hiện đề tài - Lên kế hoạch thực hiện đề tài	- Họp nhóm qua Zoom - Tìm hiểu qua các phương tiện thông tin
3	Nguyễn Bá Hưởng	- Phân chia công việc cho các thành viên, báo cáo tiến độ cho giảng viên, tìm hiểu thuật toán	- Sử dụng các phương tiện thông tin - Họp nhóm qua Zoom
	Trần Ngọc Dương	- Tìm hiểu dữ liệu đầu vào	
	Nguyễn Đăng Lộc	- Tìm hiểu dữ liệu đầu vào	
	Chu Thị Sương	- Tìm hiểu thuật toán	
4	Nguyễn Bá Hưởng	- Phân chia công việc cho các thành viên, báo cáo tiến	- Sử dụng google colab

		độ cho giảng viên, xây dựng mô hình thuật toán	- Họp nhóm qua Zoom - Tìm hiểu qua các phương tiện thông tin
	Trần Ngọc Dương	- Xử lý dữ liệu đầu vào	
	Nguyễn Đăng Lộc	- Xử lý dữ liệu đầu vào	
	Chu Thị Sương	- Xây dựng mô hình thuật toán	
5	Nguyễn Bá Hưởng	- Phân chia công việc cho các thành viên, báo cáo tiến độ cho giảng viên, làm các nội dung báo cáo	- Sử dụng google colab - Sử dụng Word và gg docs - Họp nhóm qua Zoom - Tìm hiểu qua các phương tiện thông tin
	Trần Ngọc Dương	- Xử lý dữ liệu đầu vào	
	Nguyễn Đăng Lộc	- Xử lý dữ liệu đầu vào	
	Chu Thị Sương	- Làm các nội dung báo cáo	
6	Nguyễn Bá Hưởng	- Phân chia công việc cho các thành viên, báo cáo tiến độ cho giảng viên, làm các nội dung báo cáo	- Sử dụng google colab - Sử dụng Word và gg docs - Họp nhóm qua Zoom - Tìm hiểu qua các phương tiện thông tin
	Trần Ngọc Dương	- Xây dựng ứng dụng	
	Nguyễn Đăng Lộc	- Xây dựng ứng dụng	
	Chu Thị Sương	- Làm các nội dung báo cáo	
7	Nguyễn Bá Hưởng	- Tập chung hoàn thiện báo cáo	- Sử dụng google colab - Sử dụng Word và gg docs - Họp nhóm qua Zoom - Tìm hiểu qua các phương tiện thông tin
	Trần Ngọc Dương		
	Nguyễn Đăng Lộc		
	Chu Thị Sương		
8	Nguyễn Bá Hưởng	- Chỉnh sửa và hoàn thiện báo cáo	- Sử dụng google colab - Sử dụng Word và gg docs
	Trần Ngọc Dương		

	Nguyễn Đăng Lộc Chu Thị Sương		- Họp nhóm qua Zoom - Tìm hiểu qua các phương tiện thông tin

Hà Nội, ngày 10 tháng 03 năm 2025

XÁC NHẬN CỦA GIẢNG VIÊN

ThS. Nguyễn Xuân Hoàng

BÁO CÁO HỌC TẬP CÁ NHÂN/NHÓM

Tên lớp: **20242IT6055002** Khóa: **ĐH K18 (2023-2027)**

Tên nhóm: Nhóm 11

Tên chủ đề: **Xây dựng hệ thống gợi ý phim**

Tuần	**Người thực hiện**	**Nội dung công việc**	**Kết quả đạt được**	**Kiến nghị với giảng viên hướng dẫn** *(Nêu những khó khăn, hỗ trợ từ phía giảng viên,...)*
1	Nguyễn Bá Hưởng Trần Ngọc Dương Nguyễn Đăng Lộc Chu Thị Sương	- Lập nhóm - Tìm hiểu về các đề tài và thống nhất chọn đề tài - Bầu nhóm trưởng - Phân công vai trò cho các thành viên	- Hoàn thành tất cả nội dung công việc - Chọn dự định thực hiện đề tài xây dựng hệ thống nhận diện khuôn mặt bằng thuật toán KNN	- Giảng viên góp ý không thể xây dựng hệ thống nhận diện khuôn mặt bằng thuật toán KNN, yêu cầu đổi đề tài.
2	Nguyễn Bá Hưởng Trần Ngọc Dương Nguyễn Đăng Lộc Chu Thị Sương	- Đổi đề tài nghiên cứu - Tìm hiểu quy trình và cách thức nghiên cứu đề tài - Thống nhất quy trình thực hiện đề tài - Lên kế hoạch thực hiện đề tài	- Hoàn thành tất cả nội dung công việc - Nhóm thống nhất chọn đề tài xây dựng hệ thống gợi ý phim bằng thuật toán KNN - Kế hoạch thực hiện đề	- Đề tài được giảng viên phê duyệt

			tài	
3	Nguyễn Bá Hưởng	- Phân chia công việc cho các thành viên, báo cáo tiến độ cho giảng viên, tìm hiểu thuật toán KNN	- Nhóm hiểu được cách thức hoạt động của thuật toán KNN - Có nguồn dữ liệu đầu vào	
	Trần Ngọc Dương	- Tìm hiểu dữ liệu đầu vào		
	Nguyễn Đăng Lộc	- Tìm hiểu dữ liệu đầu vào		
	Chu Thị Sương	- Tìm hiểu thuật toán KNN		
4	Nguyễn Bá Hưởng	- Phân chia công việc cho các thành viên, báo cáo tiến độ cho giảng viên, xây dựng mô hình thuật toán	- Nhóm xây dựng được mô hình KNN cơ bản và ví dụ ứng dụng của thuật toán - Nhóm đã xử lý được một phần dữ liệu đầu vào, giảm chiều dữ liệu – bỏ bớt một số dữ liệu không cần thiết từ nguồn dữ liệu.	
	Trần Ngọc Dương	- Xử lý dữ liệu đầu vào		
	Nguyễn Đăng Lộc	- Xử lý dữ liệu đầu vào		
	Chu Thị Sương	- Xây dựng mô hình thuật toán		
5	Nguyễn Bá Hưởng	Phân chia công việc cho các thành viên, báo cáo tiến độ cho giảng viên, làm các nội dung báo cáo	- Nhóm đã xử lý phần dữ liệu đầu vào đưa về dạng ma trận thưa để huấn luyện thuật toán nhưng có một số dữ liệu lỗi trong ma trận thưa cần xử lý	
	Trần Ngọc Dương	- Xử lý dữ liệu đầu vào		
	Nguyễn Đăng Lộc	- Xử lý dữ liệu đầu vào		
	Chu Thị Sương	- Làm các nội dung báo cáo		

			thêm. - Nhóm hoàn thiện nội dung chương 1 báo cáo : Tổng quan về học máy	
6	Nguyễn Bá Hưởng	- Phân chia công việc cho các thành viên, báo cáo tiến độ cho giảng viên, làm các nội dung tiếp theo báo cáo	- Nhóm đã hoàn thiện phần dữ liệu đầu vào. - Nhóm hoàn thiện nội dung chương 2 : Giới thiệu về thuật toán KNN và các ứng dụng	
	Trần Ngọc Dương	- Hỗ trợ hoàn thiện phần xử lý dữ liệu, làm nội dung báo cáo		
	Nguyễn Đăng Lộc	- Chỉnh sửa và hoàn thiện xử lý dữ liệu đầu vào		
	Chu Thị Sương	- Làm các nội dung tiếp theo báo cáo		
7	Nguyễn Bá Hưởng	- Phân chia công việc cho các thành viên, báo cáo tiến độ cho giảng viên, làm các nội dung tiếp theo báo cáo	- Nhóm đã hoàn thiện các nội dung công việc và ứng dụng thuật toán. - Nhóm hoàn thiện một phần nội dung chương 3 : Ứng dụng thuật toán cho gợi ý phim	
	Trần Ngọc Dương	- Xây dựng ứng dụng cho thuật toán		
	Nguyễn Đăng Lộc	- Xây dựng ứng dụng cho thuật toán		
	Chu Thị Sương	- Làm các nội dung tiếp theo báo cáo		
8	Nguyễn Bá Hưởng	- Chỉnh sửa và hoàn thiện báo cáo	- Nhóm đã chỉnh sửa và hoàn thiện báo cáo.	
	Trần Ngọc Dương			
	Nguyễn Đăng Lộc			

	Chu Thị Sương			

Hà Nội, ngày 10 tháng 03 năm 2025

XÁC NHẬN CỦA GIẢNG VIÊN

ThS. Nguyễn Xuân Hoàng

CHƯƠNG 2: TỔNG QUAN VỀ HỌC MÁY, THUẬT TOÁN KNN

2.1. Giới thiệu và các khái niệm trong học máy

2.1.1. Giới thiệu về học máy

Học máy (Machine learning) là một tập hợp con của Trí tuệ nhân tạo (AI) cho phép máy tính học từ dữ liệu và đưa ra dự đoán mà không cần được lập trình rõ ràng. Học máy dạy máy tính cách nhận dạng các mẫu và tự động đưa ra quyết định bằng cách sử dụng dữ liệu và thuật toán. Học máy về cơ bản được xây dựng dựa trên dữ liệu, đóng vai trò là nền tảng cho các mô hình đào tạo và thử nghiệm.

Trong bối cảnh hiện nay, khi lượng dữ liệu ngày càng tăng mạnh mẽ và các ứng dụng công nghệ ngày càng trở nên phức tạp, học máy đã trở thành công cụ hỗ trợ đắc lực trong nhiều lĩnh vực như nhận dạng hình ảnh, phân tích ngôn ngữ tự nhiên, hệ thống gợi ý (như gợi ý phim, sản phẩm) và các ứng dụng chẩn đoán trong y tế. Điều này không chỉ giúp tăng cường hiệu quả công việc mà còn mở ra những hướng đi mới trong nghiên cứu và phát triển công nghệ.

2.1.2. Các khái niệm cơ bản trong học máy

- Dữ liệu (Data): Dữ liệu là yếu tố đầu vào quyết định chất lượng và hiệu quả của mô hình học máy. Dữ liệu có thể ở nhiều dạng như số, văn bản, hình ảnh, video,… Với mục tiêu khai thác thông tin có giá trị, việc thu thập dữ liệu cần đảm bảo tính đầy đủ, chính xác và có sự đại diện của tập hợp đối tượng nghiên cứu.
- Mô hình học máy: Mô hình là kết quả biểu diễn toán học của quá trình học từ dữ liệu. Mô hình được xây dựng dựa trên bộ dữ liệu huấn luyện thông qua quá trình tối ưu hóa nhằm giảm thiểu sai số dự đoán. Có nhiều loại mô hình khác nhau tùy thuộc vào bài toán như:
    + Mô hình hồi quy (Regression Models): Dùng để dự đoán giá trị liên tục (ví dụ: dự đoán giá nhà, dự báo lượng tiêu thụ).
    + Mô hình phân loại (Classification Models): Dùng để phân loại dữ liệu vào các nhóm hoặc lớp cụ thể (ví dụ: phân loại email spam và không spam, xác định thể loại phim).

- Thuật toán học máy: Thuật toán là quy trình hay tập hợp các bước thực hiện nhằm xây dựng mô hình từ dữ liệu. Một số thuật toán phổ biến bao gồm:
    + K-Nearest Neighbors (KNN)
    + Decision Trees
    + Support Vector Machines (SVM)
    + Neural Networks
- Huấn luyện (Training): Quá trình huấn luyện là giai đoạn mà mô hình học máy "học" từ bộ dữ liệu. Trong quá trình này, các tham số của mô hình được điều chỉnh nhằm tối thiểu hóa sai số (loss function) giữa giá trị dự đoán và giá trị thực tế.
- Đánh giá (Evaluation): Sau khi huấn luyện, mô hình cần được đánh giá để kiểm tra khả năng dự đoán trên dữ liệu mà mô hình chưa từng tiếp xúc (tập kiểm tra hoặc dữ liệu xác thực). Các chỉ số đánh giá phụ thuộc vào loại bài toán, ví dụ đối với phân loại có thể sử dụng độ chính xác (accuracy), precision, recall, F1-score; đối với hồi quy thì có thể sử dụng Mean Squared Error (MSE) hay Root Mean Square Error (RMSE).
- Phân chia dữ liệu (Data Splitting): Để đảm bảo mô hình không bị overfitting (quá khớp) hoặc underfitting (không khớp), bộ dữ liệu thường được chia thành hai hay ba phần: tập huấn luyện (training set), tập kiểm tra (test set) và đôi khi tập xác thực (validation set).
    + Overfitting (Quá khớp): Xảy ra khi mô hình học quá kỹ các chi tiết, nhiễu của dữ liệu huấn luyện, dẫn đến hiệu năng dự đoán kém trên dữ liệu mới. Các biện pháp khắc phục bao gồm sử dụng regularization, pruning, hoặc tăng cường dữ liệu.
    + Underfitting (Không khớp): Xảy ra khi mô hình quá đơn giản và không nắm bắt được các đặc trưng cần thiết trong dữ liệu, dẫn đến hiệu suất học tập và dự đoán thấp. Giải pháp là lựa chọn mô hình phức tạp hơn hoặc cải thiện các tính năng đầu vào.
- Tiền xử lý dữ liệu: là bước quan trọng nhằm chuyển đổi dữ liệu thô thành dạng có thể sử dụng được cho mô hình học máy:
    + Xử lý giá trị thiếu: Áp dụng các kỹ thuật như thay thế giá trị trung bình, trung vị hoặc sử dụng mô hình dự đoán.

+ Chuẩn hóa và chuẩn hóa lại dữ liệu: Giúp đưa các đặc trưng về cùng một thang đo để giảm thiểu ảnh hưởng của khoảng cách tới kết quả cuối cùng.
+ Chuyển đổi dữ liệu: Áp dụng các phương pháp giảm chiều dữ liệu (PCA, t-SNE) hoặc tạo ra các biến mới (feature engineering) nhằm nâng cao hiệu quả của mô hình.

2.1.3. Phân loại học máy

- Học có giám sát (Supervised Learning):
  + Học có giám sát, còn được gọi là học máy có giám sát, được xác định bằng cách sử dụng các bộ dữ liệu được gắn nhãn để huấn luyện các thuật toán nhằm phân loại dữ liệu hoặc dự đoán kết quả một cách chính xác.
  + Khi dữ liệu đầu vào được đưa vào mô hình, mô hình sẽ điều chỉnh trọng số của nó cho đến khi nó được thiết lập phù hợp. Điều này xảy ra như một phần của quá trình xác thực chéo để đảm bảo rằng mô hình tránh bị quá khớp hoặc dưới khớp.
  + Một số ví dụ về các thuật toán học máy được sử dụng trong học có giám sát bao gồm Neural Networks, Naïve Bayes, hồi quy tuyến tính, hồi quy logistic, K lân cận (K-nearest neighbor hay KNN), cây quyết định (decision tree), rừng ngẫu nhiên (random forest) và máy vectơ hỗ trợ (SVM)
- Học không giám sát (Unsupervised Learning):
  + Học không giám sát, còn được gọi là học máy không giám sát, sử dụng thuật toán học máy để phân tích và phân cụm các tập dữ liệu không được gắn nhãn (các tập hợp con được gọi là cụm). Các thuật toán này khám phá các mẫu hoặc nhóm dữ liệu ẩn mà không cần sự can thiệp của con người.
  + Một số ví dụ về các thuật toán học máy được sử dụng trong học không giám sát bao gồm phân tích thành phần chính (PCA) – được sử dụng để giảm số lượng đặc trưng trong mô hình thông qua quá trình giảm số chiều của tập dữ liệu, Neural Networks, phân cụm k-mean và phương pháp phân cụm xác suất.
- Học giám sát một phần (Semi-supervised Learning):

+ Học giám sát một phần cung cấp một phương tiện hài hòa giữa học có giám sát và không giám sát. Trong quá trình huấn luyện, nó sử dụng tập dữ liệu được gắn nhãn nhỏ hơn để hướng dẫn phân loại và trích xuất đặc trưng từ tập dữ liệu lớn hơn, không được gắn nhãn.
+ Học giám sát một phần có thể giải quyết vấn đề không có đủ dữ liệu được gắn nhãn cho thuật toán học có giám sát. Nó cũng hữu ích nếu việc gắn nhãn cho dữ liệu quá tốn kém.

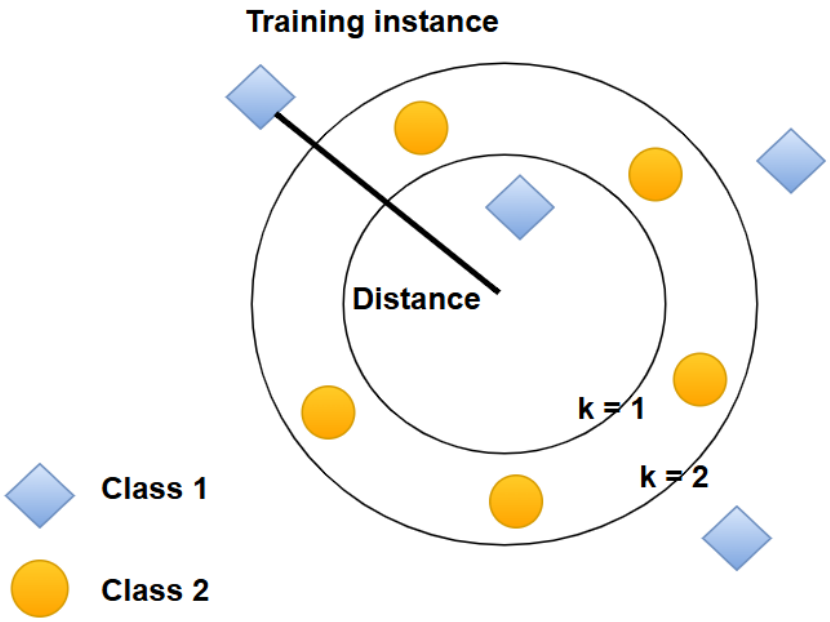
- Học tăng cường (Reinforcement Learning):
  + Học máy tăng cường là một mô hình học máy tương tự như học có giám sát, nhưng thuật toán không được huấn luyện bằng dữ liệu mẫu. Mô hình này học bằng cách sử dụng phương pháp thử và sai. Cách tiếp cận này học thông qua một mô hình thích ứng bằng cách tự huấn luyện một tập hợp các hành động thử nghiệm nhất định và quan sát các phản ứng đối với trạng thái môi trường.
  + Học máy tăng cường liên quan đến việc khám phá chuỗi hành động hoặc hành vi thích ứng của một tác nhân thông minh (tác nhân RL) trong một môi trường nhất định với động cơ tối đa hóa phần thưởng tích lũy. Hành động của tác nhân thông minh gây ra sự thay đổi có thể quan sát được về trạng thái của môi trường.

2.2. Thuật toán K-nearest neighbor

2.2.1. Định nghĩa

K - nearest neighbor (KNN) là một trong những thuật toán học có giám sát đơn giản nhất trong khai phá dữ liệu và học máy. Ý tưởng của thuật toán này là nó không học một điều gì từ tập dữ liệu học, mọi tính toán được thực hiện khi nó cần dự đoán nhãn của dữ liệu mới.

Lớp (nhãn) của một đối tượng dữ liệu mới có thể dự đoán từ các lớp (nhãn) của k hàng xóm gần nó nhất.



Hình 2.1: Nguyên lý hoạt động của thuật toán KNN

2.2.2. Quy trình làm việc của thuật toán KNN

Bước 1: Xác định tham số **k** = số hàng xóm gần nhất

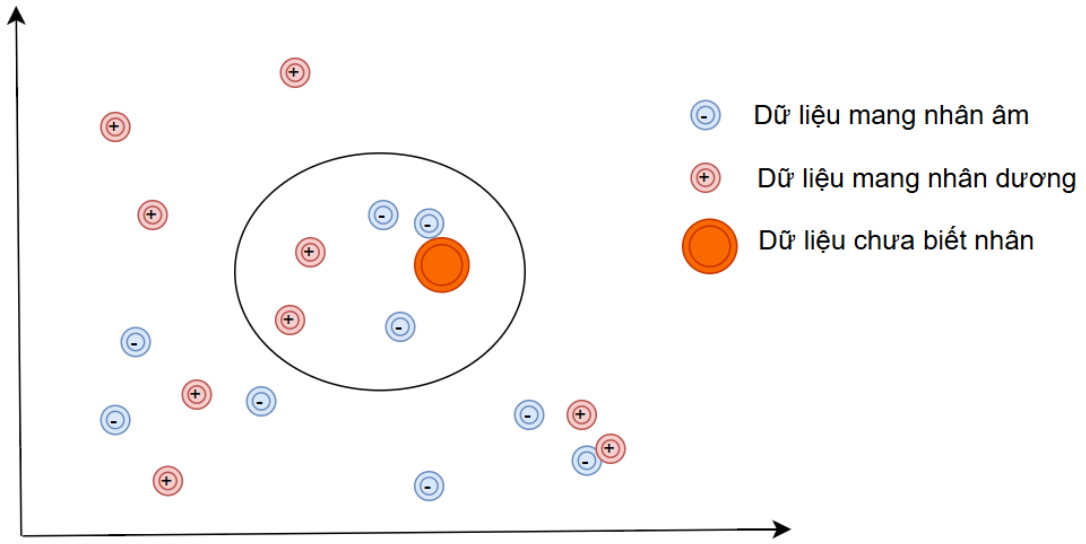
Bước 2 : Tính khoảng cách đối tượng cần phân lớp với tất cả các đối tượng trong training data

Bước 3: Sắp xếp khoảng cách theo thứ tự tăng dần và xác định **k** láng giềng gần nhất với đối tượng cần phân lớp

Bước 4: Lấy tất cả các lớp của **k** hàng xóm gần nhất

Bước 5: Dựa vào phần lớn lớp của **k** để xác định lớp cho đối tượng cần phân lớp.

2.2.3. Ví dụ minh họa



Hình 2.2: Ví dụ minh họa thuật toán KNN

Giả sử ta có D là tập các dữ liệu đã được phân loại thành 2 nhãn (+) và (-) được biểu diễn trên trục tọa độ như hình vẽ và một điểm dữ liệu mới A chưa biết nhãn. Vậy làm cách nào để chúng ta có thể xác định được nhãn của A là (+) hay (-)?

Có thể thấy đơn giản nhất là so sánh tất cả các đặc điểm của dữ liệu A với tất cả dữ liệu học đã được gắn nhãn và xem nó giống cái nào nhất, nếu dữ liệu (đặc điểm) của A gần giống với dữ liệu mang nhãn (+) thì điểm A mang nhãn (+), nếu dữ liệu A giống với dữ liệu nhãn (-) thì nó mang nhãn (-), trông có vẻ đơn giản nhưng đó là những gì mà KNN làm.

Trong trường hợp của KNN, thực tế nó không so sánh dữ liệu mới (không được phân lớp) với tất cả các dữ liệu khác, thực tế nó thực hiện một phép tính toán để đo khoảng cách giữa dữ liệu mới với tất cả các điểm trong tập dữ liệu học D để thực hiện phân lớp. Phép tính khoảng cách giữa 2 điểm có thể là Euclidian, Manhattan, trọng số, Minkowski,...

2.2.4. Ưu điểm, nhược điểm của thuật toán

- Ưu điểm:
    + Thuật toán đơn giản và dễ hiểu

+ Dễ cài đặt và sử dụng
+ Độ phức tạp tính toán nhỏ
+ Thích ứng tốt với dữ liệu mới, dự đoán dễ dàng do khả năng lưu dữ liệu huấn luyện và tính toán ngay khi dự đoán

- Nhược điểm:
  + Dễ đưa ra kết quả không chính xác khi **k** không hợp lý
  + Khi số chiều dữ liệu tăng, mọi điểm trở nên cách đều nhau hơn, dẫn đến giảm khả năng phân biệt và độ chính xác.
  + Mất nhiều thời gian tính toán và dự đoán khi tập dữ liệu lớn vì thuật toán phải lưu trữ toàn bộ tập dữ liệu huấn luyện và mỗi lần dự đoán với một mẫu mới

2.3. Khoảng cách trong không gian vector

Và đó chính là lý do mà khái niệm norm ra đời. Có nhiều loại norm khác nhau mà bạn sẽ thấy ở dưới đây:

Để xác định khoảng cách giữa hai vector y và z, người ta thường áp dụng một hàm số lên vector hiệu **x = y - z**. Một hàm số được dùng để đo khoảng cách các vector cần có một vài tính chất đặc biệt

2.3.1. Định nghĩa

Trong không gian một chiều, việc đo khoảng cách giữa hai điểm là rất quen thuộc : lấy trị tuyệt đối của hiệu giữa hai giá trị đó. Trong không gian hai chiều, tức mặt phẳng, chúng ta thường dùng khoảng cách Euclid để đo khoảng cách giữa hai điểm.

Việc đo khoảng cách giữa hai điểm dữ liệu nhiều chiều, tức hai vector, là rất cần thiết trong Machine Learning. Chúng ta cần đánh giá xem điểm nào là điểm gần nhất của một điểm khác; chúng ta cũng cần đánh giá xem độ chính xác của việc ước lượng và trong rất nhiều ví dụ khác nữa.

2.3.2. Một số norm thường dùng

Giả sử các vector $\mathbf{x} = [x_1, x_2,...,x_n]$, $\mathbf{y} = [y_1, y_2,...,y_n]$.

Nhận thấy khoảng cách Euclid chính là một norm, norm này thường được gọi là norm 2:

$$||\mathbf{x}||_2 = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} \qquad (1)$$

Với p là một số không nhỏ hơn 1 bất kỳ, hàm số sau đây:

$$||\mathbf{x}||_p = (|x_1|^p + |x_2|^p + ... + |x_n|^p)^{\frac{1}{p}} \quad (2)$$

Được chứng minh thoả mãn ba điều kiện trên, và được gọi là norm p.

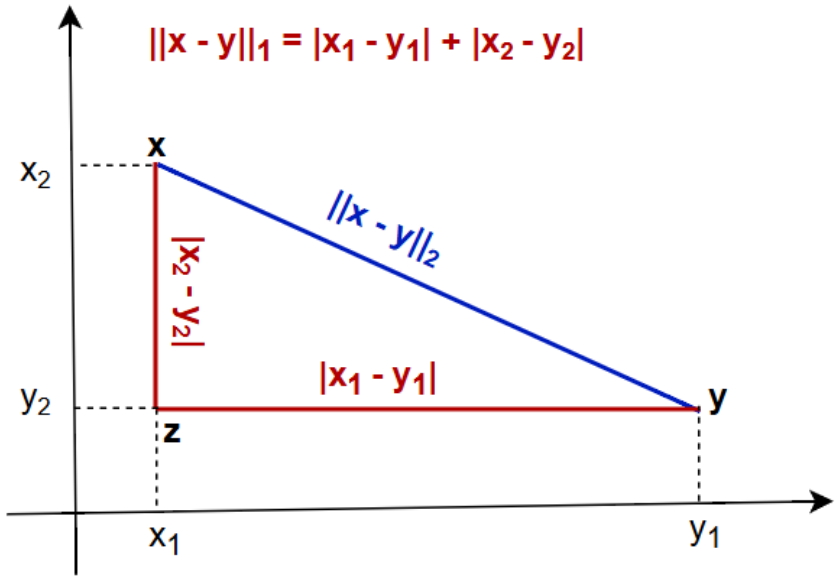
Nhận thấy rằng khi $\mathbf{p} \to 0$ thì biểu thức bên trên trở thành số các phần tử khác 0 của x. Hàm số (2) khi $\mathbf{p} = 0$ được gọi là giả chuẩn (pseudo-norm). Nó không phải là một norm vì nó không thoả mãn điều kiện 2 và 3 của norm. Giả-chuẩn nayd, thường được ký hiệu $||\mathbf{x}||_0$, khá quan trọng trong ML vì trong nhiều bài toán, chúng ta cần có ràng buộc ""sparse", tức số lượng thành phần "active" của **x** nhỏ.

Có một vài giá trị của p thường được dùng:

- Khi p = 2 chúng ta có norm2 như ở trên.
- Khi p = 1 chúng ta có:

$$||\mathbf{X}||_1 = |x_1| + |x_2| + |x_3| + ... + |x_n| \quad (3)$$

Là tổng các giá trị tuyệt đối của từng phần tử của **x**. Norm 1 thườn được dùng như xấp xỉ của norm 0 trong các bài toán có ràng buộc. Dưới đây là một ví dụ so sánh norm 1 và norm 2 trong không gian hai chiều:



Hình 2.3: Norm 1 và norm 2 trong không gian hai chiều

Norm 2 (màu xanh) chính là đường chim bay nối giữa vector **x** và vector **y**. Khoảng cách norm 1 giữa hai điểm này (màu đỏ) có thể diễn giải như là một đường đi từ x đến y trong một thành phố mà thành phố được tạo hình bàn cờ, chúng ta chỉ có thể đi theo dọc bàn cờ chứ không thể đi theo đường thẳng.

Khi $p \to \infty$, ta có norm p chính là trị tuyệt đối của phần tử lớn nhất của vector đó:

$$||x||\infty = \max |x_i| \qquad i = 1,2,..n \qquad (4)$$

2.4. Tổng kết chương

Chương 2 giới thiệu tổng quan về học máy, bao gồm các thành phần chính như dữ liệu, mô hình, thuật toán, quá trình huấn luyện và đánh giá. Bốn loại học máy cơ bản – có giám sát, không giám sát, giám sát một phần và học tăng cường – được trình bày cùng các ví dụ ứng dụng thực tiễn.

Ngoài ra, chương còn tập trung vào thuật toán K-nearest neighbor (KNN), cách hoạt động của nó và cách đo khoảng cách trong không gian vector bằng các chuẩn (norm). Những kiến thức này giúp người học hiểu rõ hơn về các phương pháp phân loại trong học máy.

CHƯƠNG 3: ỨNG DỤNG THUẬT TOÁN

3.1. Bài toán gợi ý phim

3.1.1. Giới thiệu bài toán

Hệ thống gợi ý phim là một dạng ứng dụng của hệ thống gợi ý, nhằm mục tiêu cá nhân hóa trải nghiệm người dùng bằng cách đưa ra các đề xuất phim phù hợp với sở thích, hành vi hoặc lịch sử tương tác của họ. Đối với bài toán này, việc ứng dụng thuật toán K-Nearest Neighbors (KNN) được triển khai theo hướng lọc cộng tác dựa trên người dùng hoặc dựa trên sản phẩm

3.1.2. Dữ liệu đánh giá phim

Dữ liệu đầu vào thường bao gồm:

- Người dùng (User): Mỗi người dùng được gắn với một ID duy nhất.
- Phim (Item): Mỗi bộ phim cũng có một ID hoặc tên riêng.
- Đánh giá (Rating): Người dùng đánh giá phim theo thang điểm phản ánh mức độ yêu thích.

3.1.3. Mục tiêu của hệ thống gợi ý phim:

Bài toán đặt ra : Với mỗi người dùng, dự đoán các bộ phim mà họ có thể yêu thích nhưng chưa xem, từ đó tạo danh sách gợi ý cá nhân hóa.

3.2. Tiền xử lý dữ liệu

3.2.1. Dữ liệu

Để xây dựng ứng dụng gợi ý phim bằng Python ta sử dụng bộ dữ liệu về phim của Fabian Daniel trên Kaggle. Thư viện Python và bộ dữ liệu được mở cho mục đích học tập. Trong đó ta chỉ sử dụng bộ dữ liệu ratings_small.csv, movies.csv.

movies.csv: nơi chứa thông tin về tất cả các bộ phim có trong cơ sở dữ liệu

	movieId	title	genres	popularity	runtime	vote_average
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	21.946943	81.0	7.7
1	2	Jumanji (1995)	Adventure|Children|Fantasy	17.015539	104.0	6.9
2	3	Grumpier Old Men (1995)	Comedy|Romance	11.712900	101.0	6.5
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance	3.859495	127.0	6.1
4	5	Father of the Bride Part II (1995)	Comedy	8.387519	106.0	5.7
5	6	Heat (1995)	Action|Crime|Thriller	17.924927	170.0	7.7
6	7	Sabrina (1995)	Comedy|Romance	6.677277	127.0	6.2
7	8	Tom and Huck (1995)	Adventure|Children	2.561161	97.0	5.4
8	9	Sudden Death (1995)	Action	5.231580	106.0	5.5
9	10	GoldenEye (1995)	Action|Adventure|Thriller	14.686036	130.0	6.6

ratings_small.csv: chứa đánh giá của người dùng về các bộ phim họ đã xem

index	userId	movieId	rating	timestamp
0	1	31	2.5	1260759144
1	1	1029	3.0	1260759179
2	1	1061	3.0	1260759182
3	1	1129	2.0	1260759185
4	1	1172	4.0	1260759205
5	1	1263	2.0	1260759151
6	1	1287	2.0	1260759187
7	1	1293	2.0	1260759148
8	1	1339	3.5	1260759125
9	1	1343	2.0	1260759131
10	1	1371	2.5	1260759135
11	1	1405	1.0	1260759203
12	1	1953	4.0	1260759191
13	1	2105	4.0	1260759139
14	1	2150	3.0	1260759194
15	1	2193	2.0	1260759198
16	1	2294	2.0	1260759108
17	1	2455	2.5	1260759113
18	1	2968	1.0	1260759200
19	1	3671	3.0	1260759117
20	2	10	4.0	835355493
21	2	17	5.0	835355681
22	2	39	5.0	835355604

3.2.2. Tiền xử lý dữ liệu

Nạp dữ liệu: kết nối với GG drive, sử dụng pandas để đọc dữ liệu từ CSDL

```
from google.colab import drive

drive.mount('/content/drive')

import pandas as pd

movies = pd.read_csv('/content/drive/MyDrive/movies.csv')

rating= pd.read_csv('/content/drive/MyDrive/ratings_small.csv')
```

movies_pop là bảng đã loại bỏ các thuộc tính không sử dụng trong bảng movies: 'genres', 'popularity', 'runtime', 'vote_average'.

```
movies_pop = movies.drop(['genres', 'popularity', 'runtime', 'vote_average'], axis=1, errors='ignore')
```

	movieId	title
0	1	Toy Story (1995)
1	2	Jumanji (1995)
2	3	Grumpier Old Men (1995)
3	4	Waiting to Exhale (1995)
4	5	Father of the Bride Part II (1995)
...	...	...
87580	292731	The Monroy Affaire (2022)
87581	292737	Shelter in Solitude (2023)
87582	292753	Orca (2023)
87583	292755	The Angry Breed (1968)
87584	292757	Race to the Summit (2023)

87585 rows × 2 columns

Tạo userid_200 là danh sách userId của 200 người dùng đầu tiên (không cùng userId), ratingby_200 là bảng con của rating, chỉ chứa các dòng mà userId nằm trong danh sách userid_200.

```
userid_200 = rating['userId'].unique()[:200]

ratingby_200 = rating[rating['userId'].isin(userid_200)]
```

	userId	movieId	rating	timestamp
0	1	31	2.5	1260759144
1	1	1029	3.0	1260759179
2	1	1061	3.0	1260759182
3	1	1129	2.0	1260759185
4	1	1172	4.0	1260759205
...	...	...	...	...
27673	200	106920	3.5	1438025417
27674	200	111659	4.0	1438026414
27675	200	129354	4.0	1457720823
27676	200	134130	4.5	1457123787
27677	200	136592	1.5	1438020227

27678 rows × 4 columns

tạo movieIds là danh sách được sắp xếp theo movieId ratingby_200 ra theo chiều tăng dần và mỗi movieId phải là duy nhất

```
import numpy as np
movieIds = np.sort(ratingby_200['movieId'].unique())
```

```
array([     1,      2,      3, ..., 161155, 161594, 162376])
```

Tạo user_to_index là một từ điển dạng như {user_id1: 0, user_id2: 1, ...}.với userId lấy từ userId_200. tương tự với movie_to_index

```
user_to_index = {int(user_id): idx for idx, user_id in enumerate(userid_200)}
movie_to_index = {int(movie_id): idx for idx, movie_id in enumerate(movieIds)}
user_to_index
```

```
{1: 0,
 2: 1,
 3: 2,
 4: 3,
 5: 4,
 6: 5,
 7: 6,
 8: 7,
 9: 8,
 10: 9,
```

```
movie_to_index
```

```
{1: 0,
 2: 1,
 3: 2,
 4: 3,
 5: 4,
 6: 5,
 7: 6,
 8: 7,
 9: 8,
 10: 9,
```

3.3. Huấn luyện

Sử dụng KNN của thư viện sklearn với độ tương đồng cosine, duyệt toàn bộ dữ liệu brute, huấn luyện với dữ liệu là user_movie_matrix

```
from sklearn.neighbors import NearestNeighbors
model_knn = NearestNeighbors(metric='cosine', algorithm='brute')
model_knn.fit(user_movie_matrix)
```

Chọn ngẫu nhiên 1 người dùng, lấy sở thích xem phim của người đó, tính distances: chứa khoảng cách cosine giữa người này và 6 người giống nhất, indices: chứa vị trí (chỉ số hàng) của 6 người tương tự đó trong ma trận.

```
query_index = np.random.choice(user_movie_matrix.shape[0]) query_vector = user_movie_matrix[query_index]
distances, indices = model_knn.kneighbors(query_vector, n_neighbors=6)
```

Lấy ID người dùng mà ta đang muốn gợi ý phim, danh sách các phim mà người dùng đã xem, ID người dùng tương tự, Lọc ra các phim mà người dùng tương tự đã xem nhưng người dùng chính chưa xem., Lấy phim được đánh giá cao nhất trong số đó, Lấy tên phim từ bảng thông tin phim. In ra gợi ý. Chọn người dùng cần gợi ý. Tìm các người dùng tương tự. Gợi ý phim mà họ thích nhưng người dùng chính chưa xem.

```
query_user_id = userid_200[query_index] print(f" Recommendations for userId
{query_user_id}:\n")
seen_movies = ratingby_200[ratingby_200['userId'] ==
query_user_id]['movieId'].values for i in range(1, len(distances.flatten())):
   similar_user_idx = indices.flatten()[i]
   similar_user_id = userid_200[similar_user_idx]    similar_ratings =
ratingby_200[ratingby_200['userId'] == similar_user_id]     unseen_movies =
similar_ratings[~similar_ratings['movieId'].isin(seen_movies)]
   if not unseen_movies.empty:        top_movie =
unseen_movies.sort_values(by='rating', ascending=False).iloc[0]        movie_id =
top_movie['movieId']
      title = movies_pop[movies_pop['movieId'] == movie_id]['title'].values
      print(f"{i}: {title[0]}")    else:
      print(f"{i}: Không có phim phù hợp để gợi ý từ userId {similar_user_id}")
```

```
 Recommendations for userId 183:

1: Blade Runner (1982)
2: Taxi Driver (1976)
3: Dumb & Dumber (Dumb and Dumber) (1994)
4: Misérables, Les (1995)
5: So I Married an Axe Murderer (1993)
```

3.4. Tổng kết chương

Chương 3 đã trình bày chi tiết quá trình xây dựng hệ thống gợi ý phim sử dụng thuật toán K-Nearest Neighbors (KNN), từ việc phân tích bài toán, chuẩn bị dữ liệu đến triển khai mô hình và đưa ra gợi ý cá nhân hóa cho người dùng.

Đầu tiên, bài toán gợi ý phim được phân tích theo hướng lọc cộng tác (Collaborative Filtering), trong đó mục tiêu chính là dự đoán sở thích tiềm năng của người dùng dựa

trên những người dùng có hành vi tương tự. Dữ liệu đánh giá được thể hiện qua ma trận người dùng – phim, đóng vai trò làm đầu vào cho mô hình huấn luyện.

Tiếp theo, quá trình tiền xử lý dữ liệu bao gồm: nạp dữ liệu, lọc người dùng, sắp xếp danh sách phim, và xây dựng các chỉ số ánh xạ người dùng/phim thành dạng số nguyên phục vụ cho huấn luyện mô hình. Việc lựa chọn một tập con gồm 200 người dùng giúp đơn giản hóa việc thử nghiệm nhưng vẫn đảm bảo tính đa dạng và đầy đủ cho mô hình đánh giá.

Sau đó, hệ thống được triển khai bằng thư viện Scikit - learn với thuật toán KNN, sử dụng độ đo Cosine similarity để tính toán mức độ tương đồng giữa các người dùng. Mô hình tiến hành huấn luyện trên ma trận người dùng – phim đã chuẩn hóa, và đưa ra các gợi ý phim dựa trên hành vi đánh giá của những người dùng tương tự. Kết quả gợi ý được hiển thị rõ ràng, minh họa hiệu quả của hệ thống khi chọn một người dùng ngẫu nhiên.

Tổng thể, chương 3 không chỉ làm rõ quy trình xây dựng một hệ thống gợi ý phim cơ bản mà còn minh chứng khả năng ứng dụng thực tế của thuật toán KNN trong việc cá nhân hóa nội dung, góp phần nâng cao trải nghiệm người dùng. Những nội dung này sẽ là cơ sở cho chương tiếp theo – đánh giá và thảo luận kết quả mô hình.

KẾT LUẬN

Trong báo cáo này, nhóm chúng em đã triển khai thành công quy trình xây dựng một hệ thống gợi ý phim cá nhân hóa dựa trên thuật toán K-Nearest Neighbors (KNN). Từ việc thu thập và tiền xử lý dữ liệu người dùng - phim, thiết lập ma trận đánh giá, đến việc áp dụng KNN với độ đo Cosine similarity và lựa chọn tham số k tối ưu, từng bước đã được thực hiện đầy đủ và có hệ thống. Thí nghiệm với tập dữ liệu Movie đã cho thấy khả năng phân biệt sở thích của người dùng và đề xuất những bộ phim phù hợp dựa trên hành vi của nhóm người dùng tương tự.

Bên cạnh những ưu điểm về tính đơn giản, dễ triển khai và khả năng cập nhật nhanh khi có dữ liệu mới, chúng em cũng nhận thấy một số hạn chế cần khắc phục trong giai đoạn tiếp theo: chi phí tính toán và bộ nhớ tăng lên khi mở rộng tập dữ liệu, vấn đề cold-start với người dùng hoặc phim mới, cũng như độ thưa (sparsity) của ma trận đánh giá có thể ảnh hưởng tới chất lượng gợi ý.

Để nâng cao hiệu suất và độ chính xác, nhóm đề xuất một số hướng nghiên cứu tiếp theo: ứng dụng kỹ thuật giảm chiều (PCA, SVD) trước khi tính KNN, triển khai các phương pháp tìm kiếm lân cận gần đúng (Approximate Nearest Neighbors) để gia tăng tốc độ truy vấn, kết hợp mô hình hybrid giữa KNN và các phương pháp phân rã ma trận hoặc deep learning, đồng thời bổ sung cơ chế cold-start thông qua thông tin ngữ nghĩa (thể loại, đạo diễn) và dữ liệu hồ sơ người dùng.

Cuối cùng, chúng em xin chân thành cảm ơn thầy hướng dẫn, nhà trường và các bạn trong nhóm đã hỗ trợ, đóng góp ý kiến trong suốt quá trình thực hiện đề tài. Mọi góp ý của quý thầy cô và bạn đọc sẽ là động lực để nhóm hoàn thiện và phát triển giải pháp cho đề tài càng hiệu quả hơn.

TÀI LIỆU THAM KHẢO

https://www.geeksforgeeks.org/machine-learning/

https://www.geeksforgeeks.org/k-nearest-neighbours/

BÁO CÁO HỌP NHÓM

Tên lớp: **20242IT6055002** Khóa: **ĐH K18 (2023-2027)**

Tên nhóm: Nhóm 11

Tên chủ đề: **Xây dựng hệ thống gợi ý phim**

STT	Buổi họp	Thành viên tham dự	Nội dung buổi họp	Minh chứng
1	1	Buổi họp đủ tất cả thành viên	- Nhóm thảo luận về đề tài và thống nhất chọn đề tài xây dựng hệ thống nhận diện khuôn mặt - Nhóm bầu bạn Nguyễn Bá Hưởng làm trưởng nhóm. - Trưởng nhóm giao nhiệm vụ cho các thành viên về tìm hiểu về quy trình thực hiện đề tài và những phương tiện cần thiết hỗ trợ quá trình thực hiện đề tài.	https://youtu.be/Hy9kFBZ405U
2	2	Buổi họp đủ tất cả thành viên	- Nhóm thống nhất chọn lại đề tài theo gợi ý của giảng viên: hệ thống gợi ý phim bằng thuật toán KNN	https://youtu.be/tRsc5kink50

			- Nhóm thống nhất quy trình và kế hoạch thực hiện đề tài. - Trưởng nhóm giao nhiệm vụ cho các thành viên thực hiện trong tuần tiếp theo.	
3	3	Buổi họp đủ tất cả thành viên	- Nhóm thảo luận về các phần nội dung chính của thuật toán cần tìm hiểu để ứng dụng cho đề tài. - Trưởng nhóm giao nhiệm vụ cho các thành viên thực hiện trong tuần tiếp theo.	https://youtu.be/p-gqZ949Yh0
4	4	Buổi họp đủ tất cả thành viên	- Các thành viên lần lượt trình bày kết quả thực hiện được giao: thuật toán KNN cơ bản, xử lý dữ liệu đầu vào. - Nhóm hỗ nhau các nội dung kiến thức. - Trưởng nhóm giao nhiệm vụ cho	https://youtu.be/6-gCLMUc7l4

			các thành viên thực hiện trong tuần tiếp theo.	
5	5	Buổi họp đủ tất cả thành viên	- Các thành viên lần lượt trình bày kết quả thực hiện được giao: nội dung báo cáo chương 1, phần xử lý dữ liệu đầu vào. - Nhóm thảo luận sôi nổi, giải đáp thắc mắc và hỗ nhau các phần nội dung. - Trưởng nhóm giao nhiệm vụ cho các thành viên thực hiện trong tuần tiếp theo.	https://youtu.be/CNLmIF7urlc
6	6	Buổi họp đủ tất cả thành viên	- Các thành viên lần lượt trình bày kết quả thực hiện được giao: nội dung báo cáo chương 2, chỉnh sửa phần xử lý dữ liệu. - Nhóm thảo luận hỗ trợ nhau các phần nội dung.	https://youtu.be/dJhapBUjp3s

			- Trưởng nhóm giao nhiệm vụ cho các thành viên thực hiện trong tuần tiếp theo.	
7	7	Buổi họp đủ tất cả thành viên	- Các thành viên lần lượt trình bày kết quả thực hiện được giao: nội dung báo cáo chương 3. - Trưởng nhóm giao nhiệm vụ cho các thành viên thực hiện trong tuần tiếp theo.	https://youtu.be/UxBdUfon1To
8	8	Buổi họp đủ tất cả thành viên	- Các thành viên lần lượt trình bày kết quả thực hiện được giao: nội dung chỉnh sửa chương 3, phần phiếu học tập và nội dung chương giới thiệu đề tài - Trưởng nhóm giao nhiệm vụ cho các thành viên hoàn thiện báo cáo.	https://youtu.be/EQgfc1n_y8g

TRƯỞNG NHÓM

Nguyễn Bá Hưởng