BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC ĐẠI NAM

ĐỒ ÁN TỐT NGHIỆP

NGHIÊN CỨU THUẬT TOÁN HỒI QUY ĐỂ DỰ ĐOÁN ĐIỂM TRUNG BÌNH LỚP 12 DỰA VÀO ĐIỂM LỚP 10, LỚP 11

SINH VIÊN THỰC HIỆN : VŨ CÔNG HOAN

MÃ SINH VIÊN : 1451020093

KHOA : CÔNG NGHỆ THÔNG TIN

HÀ NỘI - 2024

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC ĐẠI NAM

VŨ CÔNG HOAN

NGHIÊN CỨU THUẬT TOÁN HỒI QUY ĐỂ DỰ ĐOÁN ĐIỂM TRUNG BÌNH LỚP 12 DỰA VÀO ĐIỂM LỚP 10, LỚP 11

CHUYÊN NGÀNH : **CÔNG NGHỆ THÔNG TIN**

MÃ SỐ : **74.80.201**

NGƯỜI HƯỚNG DẪN : ThS. TRẦN THU TRANG

HÀ NỘI - 2024

LỜI CAM ĐOAN

Em tên là Vũ Công Hoan cam đoan rằng đồ án thực tập về đề tài "Tìm hiểu một số thuật toán hồi quy để dự đoán điểm trung bình lớp 12 dựa vào điểm lớp 10, 11" được trình bày dưới đây là công trình nghiên cứu của bản thân em dưới sự hướng dẫn của cô Trần Thu Trang và không sao chép từ bất kỳ nguồn tài liệu nào khác.

Em cam đoan rằng trong quá trình thực hiện đồ án này, em đã tuân thủ các quy định về nghiên cứu khoa học và trích dẫn tài liệu. Mọi thông tin, số liệu và kết quả được trình bày trong báo cáo đều được thu thập và xử lý một cách trung thực và đáng tin cậy.

Em xin chịu trách nhiệm về tính chính xác và độ tin cậy của đồ án này. Đồ án được trình bày dưới dạng một tài liệu tham khảo cho mục đích học tập và nghiên cứu

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành đến cô Trần Thu Trang đã hỗ trợ và giúp đỡ em trong quá trình thực hiện đề tài "Tìm hiểu một số thuật toán hồi quy để dự đoán điểm trung bình lớp 12 dựa vào điểm lớp 10, 11". Sự chỉ dẫn, hướng dẫn và kiến thức chuyên môn mà cô đã chia sẻ với em đã là nguồn cảm hứng và động lực để em tiến hành nghiên cứu. Cô đã dành thời gian và tận tâm hỗ trợ em trong quá trình nắm bắt và hiểu rõ về các thuật toán hồi quy. Em xin chân thành cảm ơn cô.

Em xin chân thành cảm ơn!

Hà Nội, tháng năm 2024

Sinh viên

LỜI NÓI ĐẦU

Trong bối cảnh ngày nay, việc áp dụng các phương pháp máy học và thuật toán hồi quy để dự đoán điểm trung bình của học sinh là một lĩnh vực nghiên cứu đầy tiềm năng và có ứng dụng rộng rãi trong hệ thống giáo dục. Đặc biệt, việc dự đoán điểm trung bình của học sinh lớp 12 dựa trên điểm các năm học trước đó, như điểm trung bình của lớp 10 và lớp 11, không chỉ giúp các nhà quản lý giáo dục và giáo viên đánh giá được tiềm năng học tập của học sinh mà còn hỗ trợ quá trình đưa ra quyết định về việc cung cấp hỗ trợ học tập đúng đắn.

Chính vì lý do này, đề tài này đã được lựa chọn để thực hiện với mục tiêu tìm hiểu và áp dụng một số thuật toán hồi quy để dự đoán điểm trung bình của học sinh lớp 12 dựa vào dữ liệu về điểm số của các năm học trước đó. Phạm vi nghiên cứu của đề tài sẽ tập trung vào việc phân tích, so sánh và đánh giá hiệu suất của các thuật toán hồi quy được áp dụng trong bài toán dự đoán điểm trung bình lớp 12 từ dữ liệu điểm lớp 10 và lớp 11.

Phương pháp nghiên cứu sẽ bao gồm việc thu thập dữ liệu, tiền xử lý dữ liệu, lựa chọn và huấn luyện mô hình, đánh giá hiệu suất của mô hình và so sánh các kết quả thu được. Mục tiêu cuối cùng của đề tài là cung cấp thông tin hữu ích và khả năng ứng dụng thực tiễn cho các nhà quản lý giáo dục và các chuyên gia trong việc dự đoán và đánh giá tiềm năng học tập của học sinh dựa trên dữ liệu điểm số từ các năm học trước đó.

NHẬN XÉT

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

DANH MỤC KÝ HIỆU

STT	Ký hiệu chữ viết tắt	Chữ viết tắt đầy đủ
1	ML	Machine Learning
2	HK	Học kỳ
3	CMND	Chứng minh nhân dân
4	CN	Cả năm
5	GDCD	Giáo dục công dân

DANH MỤC HÌNH ẢNH

Hình 2. 1: Thuật toán SVM 21
Hình 2. 2: Mô hình Decision Trees 26
Hình 2. 3: Thuật toán rừng ngẫu nhiên 29
Hình 3. 1: Dữ liệu ban đầu 41
Hình 4. 1: Trực quan hóa dữ liệu 48
Hình 4. 2: Dữ liệu khối tự nhiên 49
Hình 4. 3: Dữ liệu khối xã hội 50
Hình 4. 4: Kết quả mô hình Linear Regression 51
Hình 4. 5: Kết quả mô hình Random Forest 51
Hình 4. 6: Kết quả mô hình SVM 52
Hình 4. 7: Kết quả mô hình Decision Trees 52
Hình 4. 8: Mô hình tối ưu hóa tham số GridSearch của Linear Regression 53
Hình 4. 9: Mô hình tối ưu hóa tham số GridSearch của Decision Tree 54
Hình 4. 10: Mô hình tối ưu hóa tham số RandomSearch của Linear Regression 55
Hình 4. 11: Mô hình tối ưu hóa tham số RandomSearch của Decision Tree 55
Hình 5. 1: Ứng dụng mô hình Linear Regression 58
Hình 5. 2: Ứng dụng dự đoán điểm 59

DANH MỤC BẢNG BIỂU

Bảng 3. 1: Bảng phân tích dữ liệu 36
Bảng 3. 2: Bảng các trường loại bỏ 42
Bảng 3. 3: Bảng xử lý dữ liệu trống 42
Bảng 3. 4: Bảng thống kê dữ liệu 42
Bảng 4. 1: Bảng so sánh 2 mô hình tối ưu bằng GridSearch 54
Bảng 4. 2: Bảng so sánh 2 mô hình tối ưu bằng RandomSearch 55

MỤC LỤC

MỞ ĐẦU 13

1. Tính cấp thiết của đề tài 13

2. Mục đích nghiên cứu 14

3. Phạm vi nghiên cứu 14

4. Phương pháp nghiên cứu 14

Chương 1 CƠ SỞ LÝ THUYẾT VÀ KỸ THUẬT VỀ HỌC MÁY 15

1.1. Giới thiệu về học máy 15

1.2. Lịch sử phát triển của học máy 15

1.3. Phân loại về học máy 17

1.3.1. Dựa trên cách học 17

1.3.2. Dựa trên mục tiêu 18

1.3.3. Dựa trên cách tiếp cận 18

1.3.4. Dựa trên mục đích ứng dụng 19

1.4. Ứng dụng học máy trong định hướng chuyên ngành cho sinh viên CNTT 20

Chương 2 MỘT SỐ MÔ HÌNH HỌC MÁY 21

2.1. Mô hình SVM 21

2.1.1. Khái niệm 21

2.1.2. Thuật toán Support Vector Machine 21

2.1.3. Ưu và nhược điểm 22

2.2. Mô hình Linear Regression 23

2.2.1. Khái niệm 23

2.2.2. Thuật toán Linear Regression 23

2.2.3. Ưu và nhược điểm 24

2.3. Mô hình Decision Trees 25
2.3.1. Khái niệm 25
2.3.2. Thuật toán hồi quy của Decision Trees 26
2.3.3. Ưu và nhược điểm 27
2.4. Mô hình Random Forest 27
2.4.1. Khái niệm 27
2.4.2. Thuật toán Rừng ngẫu nhiên 29
2.4.3. Ưu và nhược điểm 30
2.5. Thuật toán tối ưu Grid Search 31
2.5.1. Khái niệm 31
2.5.2. Thuật toán 31
2.5.3. Ưu và nhược điểm 32
2.6. Thuật toán tối ưu Random Search 32
2.6.1. Khái niệm 32
2.6.2. Thuật toán 32
2.6.3. Ưu và nhược điểm 33
Chương 3 PHÂN TÍCH VÀ XỬ LÝ DỮ LIỆU 35
3.1. Bài toán 35
3.2. Dữ liệu nghiên cứu 36
3.3. Thống kê các trường dữ liệu 36
3.3.1. Dữ liệu ban đầu 36
3.3.2. Tiền xử lý dữ liệu 42
3.4. Phân chia dữ liệu huấn luyện 46
Chương 4 TỐI ƯU HÓA MÔ HÌNH KỸ THUẬT HỌC MÁY 48

4.1. Tính tương quan của các điểm 48
4.2. Chạy các mô hình học máy hồi quy 50
4.2.1. Mô hình Linear Regression 50
4.2.2. Mô hình Random Forest 51
4.2.3. Mô hình SVM 52
4.2.4. Mô hình Decision Trees 52
4.3. So sánh 4 mô hình 53
4.4. Mô hình tối ưu tham số 53
4.4.1. Tối ưu bằng Grid Search 53
4.4.2. Tối ưu bằng Random Search 55
Chương 5 CHẠY MÔ HÌNH VÀ ĐÁNH GIÁ KẾT QUẢ 57
5.1. Chạy mô hình dựa vào mô hình tốt nhất 57
5.1.1. Ứng dụng mô hình Linear Regression 57
5.1.2. Ứng dụng dự đoán điểm trung bình lớp 12 dựa vào điểm các môn lớp 10, 11 58
5.2. Đánh giá kết quả chạy 59
KẾT LUẬN 61
TÀI LIỆU THAM KHẢO 63

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Đề tài của đồ án tốt nghiệp tập trung vào nghiên cứu và ứng dụng thuật toán hồi quy để dự đoán điểm trung bình của học sinh lớp 12 dựa trên các điểm số thu thập được từ lớp 10 và lớp 11. Trong bối cảnh ngành Công nghệ thông tin đang phát triển mạnh mẽ, việc hỗ trợ định hướng chuyên ngành cho sinh viên Công nghệ thông tin là một vấn đề quan trọng và cấp thiết.

Sự tiến bộ của công nghệ thông tin và sự phổ biến của máy tính đã tạo ra một lượng lớn dữ liệu. Từ đó, việc phân tích và sử dụng thông tin này một cách hiệu quả trở thành một nhu cầu quan trọng. Đồng thời, với sự đa dạng của các ngành và lĩnh vực trong Công nghệ thông tin, việc định hướng chuyên ngành đúng đắn không chỉ giúp sinh viên xác định được hướng đi của mình mà còn giúp họ phát triển sự nghiệp sau này một cách hiệu quả và phù hợp với thị trường lao động.

Trong lĩnh vực trí tuệ nhân tạo, học máy đóng vai trò quan trọng trong việc xử lý dữ liệu và tạo ra các mô hình dự đoán từ dữ liệu đó. Áp dụng thuật toán hồi quy trong việc dự đoán điểm trung bình của học sinh lớp 12 dựa trên điểm số từ lớp 10 và lớp 11 có thể giúp phân tích sở thích, năng lực và mục tiêu cá nhân của học sinh.

Việc áp dụng kỹ thuật học máy trong hỗ trợ định hướng chuyên ngành cũng giúp tối ưu hóa quá trình tư vấn cho sinh viên. Thay vì chỉ dựa vào kinh nghiệm và cảm nhận của các cố vấn học thuật, việc sử dụng dữ liệu và mô hình học máy có thể cung cấp thông tin phản hồi chính xác và phản ánh đa chiều về sở thích và khả năng của sinh viên. Điều này giúp tăng cường sự hiệu quả và tính cá nhân hóa trong quá trình tư vấn, từ đó nâng cao khả năng thành công của sinh viên sau này.

Đề tài "Nghiên cứu thuật toán hồi quy để dự đoán điểm trung bình lớp 12 dựa vào điểm lớp 10, 11" không chỉ đáp ứng nhu cầu thực tiễn của ngành Công nghệ thông tin mà còn mang lại nhiều lợi ích lớn cho sinh viên và các cơ sở giáo dục. Đây là một đề tài cực kỳ cấp thiết và tiềm năng trong lĩnh vực giáo dục và công nghệ thông tin hiện nay.

2. Mục đích nghiên cứu

Trong thực tế, việc dự đoán điểm trung bình của sinh viên lớp 12 dựa trên điểm lớp 10 và 11 có thể hỗ trợ học sinh và giáo viên trong quá trình định hướng học tập và lựa chọn con đường tương lai. Điều này đặt ra mục tiêu cho nghiên cứu này, đó là xây dựng một mô hình học máy sử dụng thuật toán hồi quy để dự đoán điểm trung bình lớp 12 dựa vào điểm lớp 10 và 11.

Mục tiêu của nghiên cứu là tạo ra một công cụ dự đoán hiệu quả và chính xác, giúp học sinh và giáo viên có cái nhìn rõ ràng hơn về tiềm năng và khả năng học tập của học sinh. Kết quả dự đoán có thể được sử dụng để định hướng tổ hợp, lựa chọn trường đại học, và xác định các biện pháp hỗ trợ học tập phù hợp.

3. Phạm vi nghiên cứu

Đối tượng nghiên cứu: dữ liệu điểm của trường đại học Quốc gia, thu thập trong kì thi đánh giá năng lực năm 2022-2023.

Phạm vi nghiên cứu: Nghiên cứu tập trung vào việc áp dụng thuật toán hồi quy trên dữ liệu điểm của học sinh trong kì thi đánh giá.

4. Phương pháp nghiên cứu

Phương pháp nghiên cứu sẽ sử dụng các kỹ thuật thu thập và phân tích dữ liệu thống kê. Dữ liệu điểm số từ lớp 10 và 11 của sinh viên sẽ được thu thập và tiền xử lý để chuẩn bị cho quá trình huấn luyện và đánh giá mô hình hồi quy. Các phương pháp thống kê và phân tích sẽ được áp dụng để đánh giá mức độ dự đoán và tương quan giữa điểm trung bình lớp 12 và điểm lớp 10, 11. Kết quả nghiên cứu sẽ được trình bày một cách chi tiết và tổ chức trong báo cáo đồ án tốt nghiệp.

Chương 1
CƠ SỞ LÝ THUYẾT VÀ KỸ THUẬT VỀ HỌC MÁY

1.1. Giới thiệu về học máy

Học máy là một lĩnh vực của trí tuệ nhân tạo (AI) tập trung vào việc xây dựng và phát triển các thuật toán mà máy tính có thể "học" từ dữ liệu và trải nghiệm để tự động cải thiện hiệu suất mà không cần được lập trình một cách cụ thể. Điều này có nghĩa là máy tính có khả năng nhận biết các mẫu trong dữ liệu và dự đoán, phân loại hoặc ra quyết định mà không cần sự can thiệp trực tiếp từ con người.

Trong học máy, dữ liệu là yếu tố chính để huấn luyện các mô hình. Các thuật toán học máy được thiết kế để phát hiện các mẫu, tổ chức thông tin và tạo ra các dự đoán hoặc quyết định dựa trên dữ liệu này. Các ứng dụng của học máy rất đa dạng, từ dự đoán thị trường tài chính, phát hiện gian lận tín dụng, nhận diện khuôn mặt, đến tự động lái xe và nhiều lĩnh vực khác.

Một số phương pháp phổ biến trong học máy bao gồm học có giám sát (supervised learning), học không giám sát (unsupervised learning), và học tăng cường (reinforcement learning). Mỗi phương pháp này đều có các ưu điểm và hạn chế riêng, và chúng được áp dụng trong các tình huống khác nhau tùy thuộc vào loại dữ liệu và mục tiêu cụ thể của vấn đề.

1.2. Lịch sử phát triển của học máy

Lịch sử phát triển của học máy có những bước đột phá quan trọng cùng các mốc thời gian:

- 1950 - Nhà bác học Alan Turing đã tạo ra "Turing Test (phép thử Turing)" để xác định xem liệu một máy tính có trí thông minh thực sự hay không. Để vượt qua bài kiểm tra đó, một máy tính phải có khả năng đánh lừa một con người tin là con người.
- 1952 - Arthur Samuel đã viết ra chương trình học máy (computer learning) đầu tiên. Chương trình này là trò chơi cờ đam, và hãng máy tính IBM đã cải tiến trò chơi này để có thể tự học và tổ chức những nước đi trong chiến lược để giành chiến thắng.
- 1957 - Frank Rosenblatt đã thiết kế mạng nơron (neural network) đầu tiên cho máy tính, trong đó mô phỏng quá trình suy nghĩ của bộ não con người.
- 1967 - Thuật toán "nearest neighbor" đã được viết, cho phép các máy tính bắt đầu sử dụng những mẫu nhận dạng (pattern recognition) rất cơ bản. Được sử dụng để vẽ ra lộ trình

cho một người bán hàng có thể bắt đầu đi từ một thành phố ngẫu nhiên nhưng đảm bảo anh ta sẽ đi qua tất cả các thành phố khác theo một quãng đường ngắn nhất.

- 1979 - Sinh viên tại trường đại học Stanford đã phát minh ra giỏ hàng "Stanford Cart" có thể điều hướng để tránh các chướng ngại vật trong một căn phòng.
- 1981 - Gerald Dejong giới thiệu về khái niệm Explanation Based Learning (EBL), trong đó một máy tính phân tích dữ liệu huấn luyện và tạo ra một quy tắc chung để có thể làm theo bằng cách loại bỏ đi những dữ liệu không quan trọng.
- 1985 - Terry Sejnowski đã phát minh ra NetTalk, có thể học cách phát âm các từ giống như cách một đứa trẻ tập nói.
- 1990s - Machine Learning đã dịch chuyển từ cách tiếp cận hướng kiến thức (knowledge-driven) sang cách tiếp cận hướng dữ liệu (data-driven). Các nhà khoa học bắt đầu tạo ra các chương trình cho máy tính để phân tích một lượng lớn dữ liệu và rút ra các kết luận - hay là "học" từ các kết quả đó.
- 1997 - Deep Blue của hãng IBM đã đánh bại nhà vô địch cờ vua thế giới.
- 2006 - Geoffrey Hinton đã đưa ra một thuật ngữ "deep learning" để giải thích các thuật toán mới cho phép máy tính "nhìn thấy" và phân biệt các đối tượng và văn bản trong các hình ảnh và video.
- 2010 - Microsoft Kinect có thể theo dõi 20 hành vi của con người ở một tốc độ 30 lần mỗi giây, cho phép con người tương tác với máy tính thông qua các hành động và cử chỉ.
- 2011 - Máy tính Watson của hãng IBM đã đánh bại các đối thủ là con người tại Jeopardy.
- 2011 - Google Brain đã được phát triển, và mạng deep nơron (deep neural network) có thể học để phát hiện và phân loại nhiều đối tượng theo cách mà một con mèo thực hiện.
- 2012 - X Lab của Google phát triển một thuật toán machine learning có khả năng tự động duyệt qua các video trên YouTube để xác định xem video nào có chứa những con mèo.
- 2014 - Facebook phát triển DeepFace, một phần mềm thuật toán có thể nhận dạng hoặc xác minh các cá nhân dựa vào hình ảnh ở mức độ giống như con người có thể.
- 2015 - Amazon ra mắt nền tảng machine learning riêng của mình.
- 2015 - Microsoft tạo ra Distributed Machine Learning Toolkit, trong đó cho phép phân phối hiệu quả các vấn đề machine learning trên nhiều máy tính.

- 2015 - Hơn 3.000 nhà nghiên cứu AI và Robotics, được sự ủng hộ bởi những nhà khoa học nổi tiếng như Stephen Hawking, Elon Musk và Steve Wozniak (và nhiều người khác), đã ký vào một bức thư ngỏ để cảnh báo về sự nguy hiểm của vũ khí tự động trong việc lựa chọn và tham gia vào các mục tiêu mà không có sự can thiệp của con người.
- 2016 - Thuật toán trí tuệ nhân tạo của Google đã đánh bại nhà vô địch trò chơi Cờ Vây, được cho là trò chơi phức tạp nhất thế giới (khó hơn trò chơi cờ vua rất nhiều). Thuật toán AlphaGo được phát triển bởi Google DeepMind đã giành chiến thắng 4/5 trước nhà vô địch Cờ Vây.

Sự tiến bộ trong việc thu thập dữ liệu, công nghệ tính toán và sự hiểu biết sâu sắc về các thuật toán học máy đã làm cho lĩnh vực này trở thành một trong những lĩnh vực nghiên cứu và ứng dụng quan trọng nhất trong thời đại hiện đại.

1.3. Phân loại về học máy

1.3.1. Dựa trên cách học

Học có giám sát (Supervised Learning)

Mỗi mẫu dữ liệu trong tập huấn luyện đi kèm với một nhãn. Mô hình học từ cặp dữ liệu (đặc trưng) và nhãn tương ứng để dự đoán nhãn cho các dữ liệu mới.

Ví dụ: Dự đoán nếu một email là spam (nhãn 1) hoặc không phải spam (nhãn 0) dựa trên nội dung và tiêu đề của email.

Học không giám sát (Unsupervised Learning)

Dữ liệu không có nhãn. Mô hình phải tự tìm hiểu cấu trúc hoặc thông tin ẩn trong dữ liệu mà không có sự hướng dẫn từ các nhãn.

Ví dụ: Phân nhóm khách hàng thành các nhóm có tính chất tương đồng, phát hiện biên trong dữ liệu giao dịch tài chính để phát hiện gian lận.

Học bán giám sát (Semi-supervised Learning)

Một phần của dữ liệu được gán nhãn và phần còn lại không. Mô hình học từ cả dữ liệu có nhãn và không nhãn để cải thiện hiệu suất.

Ví dụ: Dự đoán rating cho sản phẩm dựa trên các đánh giá có sẵn (có nhãn) cùng với thông tin về sản phẩm (không nhãn).

Học tăng cường (Reinforcement Learning)

Mô hình tương tác với một môi trường và nhận phản hồi dựa trên hành động. Mục tiêu là tối ưu hóa phần thưởng (reward) thông qua các hành động.

Ví dụ: Huấn luyện một robot để tự động lái xe, trong đó mô hình phải học cách tương tác với môi trường (đường) và tối ưu hóa phần thưởng (đến đích một cách an toàn và nhanh chóng).

1.3.2. Dựa trên mục tiêu

Học phân loại (Classification)

Dự đoán nhãn hoặc lớp của dữ liệu mới dựa trên các thông tin đã học từ tập dữ liệu huấn luyện.

Ví dụ: Gmail xác định xem một email có phải là spam hay không; các hãng tín dụng xác định xem một khách hàng có khả năng thanh toán nợ hay không. Ba ví dụ phía trên được chia vào loại này

Học hồi quy (Regression)

Dự đoán giá trị liên tục cho các biến mục tiêu.

Ví dụ: Dự đoán giá nhà dựa trên diện tích và số phòng, dự đoán doanh số bán hàng dựa trên quảng cáo và giá cả.

Học gom cụm (Clustering):

Phân nhóm các dữ liệu không có nhãn vào các nhóm có tính chất tương tự nhau.

Ví dụ: Phân nhóm khách hàng dựa trên hành vi mua hàng, phân loại văn bản vào các chủ đề tương tự nhau.

Học giảm chiều dữ liệu (Dimensionality Reduction)

Giảm số chiều của dữ liệu bằng cách giữ lại thông tin quan trọng nhất.

Ví dụ: Sử dụng phương pháp như Principal Component Analysis (PCA) để giảm số chiều của dữ liệu mà vẫn giữ lại các đặc trưng quan trọng nhất.

1.3.3. Dựa trên cách tiếp cận

Học tập cơ sở (Instance-based Learning)

Mô hình dự đoán dựa trên các trường hợp tương tự đã được lưu trữ trong tập dữ liệu huấn luyện. Không tạo ra một mô hình tổng quát mà thay vào đó lưu trữ trực tiếp các trường hợp đã được học.

Khi cần dự đoán, mô hình tìm các trường hợp tương tự nhất trong tập dữ liệu huấn luyện và sử dụng nhãn của chúng để dự đoán cho dữ liệu mới.

Ví dụ: K-nearest neighbors (KNN) là một phương pháp học tập cơ sở, trong đó nhãn của một điểm dữ liệu mới được dự đoán dựa trên các điểm dữ liệu láng giềng gần nhất trong không gian đặc trưng.

Học dựa trên mô hình (Model-based Learning)

Mô hình được xây dựng dựa trên một cấu trúc được xác định trước và được điều chỉnh thông qua việc huấn luyện trên dữ liệu. Mô hình này thường là một biểu diễn toán học hoặc thống kê của quan hệ giữa các đặc trưng và nhãn trong tập dữ liệu huấn luyện.

Khi cần dự đoán, mô hình sử dụng các tham số đã học từ dữ liệu huấn luyện để dự đoán kết quả cho dữ liệu mới.

Ví dụ: Các mô hình như Linear Regression, Logistic Regression, Decision Trees, Neural Networks là các phương pháp học dựa trên mô hình.

1.3.4. Dựa trên mục đích ứng dụng

Học máy cơ bản (Basic Machine Learning)

- Mục tiêu: Áp dụng các phương pháp cơ bản trong học máy cho các bài toán phổ biến như phân loại, hồi quy.
- Phương pháp và mô hình được sử dụng thường là những phương pháp truyền thống như Linear Regression, Logistic Regression, Naive Bayes, Decision Trees, và Support Vector Machines.
- Thích hợp cho các bài toán đơn giản và dữ liệu có cấu trúc tương đối đơn giản.
- Ví dụ: Dự đoán giá nhà dựa trên diện tích, phân loại email là spam hoặc không phải spam.

Học máy nâng cao (Advanced Machine Learning)

- Mục tiêu: Sử dụng các phương pháp và mô hình phức tạp hơn để giải quyết các bài toán phức tạp hơn và tinh vi hơn.

- Bao gồm các kỹ thuật như học sâu (deep learning), học tăng cường (reinforcement learning), và các phương pháp tiên tiến khác như học chuyển giao (transfer learning), học đa nhiệm (multi-task learning).

- Thích hợp cho các bài toán có tính phức tạp cao, dữ liệu lớn và có cấu trúc phức tạp.

- Đòi hỏi tài nguyên tính toán và dữ liệu huấn luyện lớn.
- Ví dụ: Nhận diện hình ảnh, dịch máy, tự động lái xe, xử lý ngôn ngữ tự nhiên.

1.4. Ứng dụng học máy trong định hướng chuyên ngành cho sinh viên CNTT

Ứng dụng học máy trong định hướng chuyên ngành cho sinh viên Công nghệ thông tin (CNTT) là một công cụ mạnh mẽ giúp họ hiểu rõ hơn về các lĩnh vực và cơ hội nghề nghiệp trong ngành. Một trong những ứng dụng quan trọng nhất của học máy là trong việc phân tích dữ liệu cá nhân của sinh viên, bao gồm kỹ năng, sở thích, và mục tiêu nghề nghiệp. Dựa trên dữ liệu này, các hệ thống tư vấn sự nghiệp có thể được xây dựng để đề xuất những lựa chọn chuyên ngành phù hợp nhất với từng cá nhân.

Ngoài ra, học máy cũng được áp dụng để dự đoán xu hướng công nghệ trong tương lai. Sinh viên CNTT có thể sử dụng các công nghệ học máy để phân tích dữ liệu về các lĩnh vực công nghệ đang phát triển, từ đó định hình quyết định về việc lựa chọn chuyên ngành. Việc này giúp sinh viên hiểu rõ hơn về những lĩnh vực đang nổi bật và có tiềm năng phát triển trong tương lai, từ đó tạo ra cơ hội nghề nghiệp trong ngành CNTT.

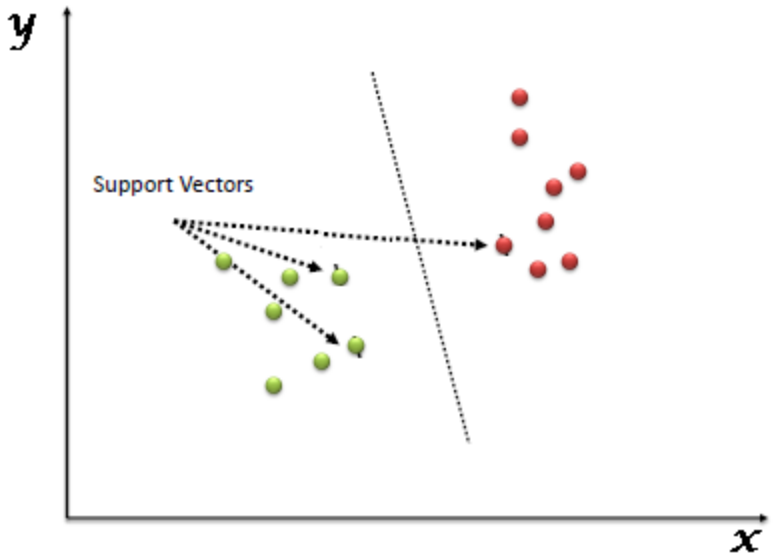
Chương 2
MỘT SỐ MÔ HÌNH HỌC MÁY

2.1. Mô hình SVM

2.1.1. Khái niệm

SVM là một thuật toán giám sát, có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại. Trong thuật toán này, chúng ta vẽ đồi thị dữ liệu là các điểm trong n chiều (ở đây n là số lượng các tính năng bạn có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay" (hyper-plane) phân chia các lớp. Hyper-plane chỉ hiểu đơn giản là 1 đường thẳng có thể phân chia các lớp ra thành hai phần riêng biệt.

Support Vectors hiểu một cách đơn giản là các đối tượng trên đồ thị tọa độ quan sát, Support Vector Machine là một biên giới để chia hai lớp tốt nhất.



Hình 2. 1: Thuật toán SVM

2.1.2. Thuật toán Support Vector Machine

Support Vector Machine (SVM) là một thuật toán học máy được sử dụng chủ yếu trong các bài toán phân loại và hồi quy. Mục tiêu của SVM là tìm ra một siêu phẳng trong không gian n chiều (n là số lượng biến đầu vào) sao cho siêu phẳng này tối đa hoá khoảng cách giữa các điểm dữ liệu thuộc các lớp khác nhau.

Mô hình toán học:

Giả sử chúng ta có tập dữ liệu huấn luyện gồm các điểm dữ liệu (x_1,y_1), (x_2,y_2)….. (x_n,y_n), trong đó x_i là vector đặc trưng của mẫu và y_i là nhãn của mẫu (-1 hoặc 1 trong bài toán phân loại nhị phân).

Mục tiêu của SVM là tìm ra siêu phẳng phân chia (decision boundary) tốt nhất giữa các lớp dữ liệu. Siêu phẳng này được mô tả bởi phương trình:

$f(x)=w^Tx+b=0$

Trong đó:

- w là vector trọng số của siêu phẳng.
- b là hệ số bias.
- x là vector đặc trưng của mẫu.

Siêu phẳng này chia không gian thành hai phần, mỗi phần chứa các điểm dữ liệu của một lớp. Các điểm dữ liệu nằm gần siêu phẳng và được gọi là các vector hỗ trợ (support vectors).

Mục tiêu của SVM là tìm ra siêu phẳng sao cho khoảng cách từ các điểm dữ liệu đến siêu phẳng là lớn nhất. Điều này có thể được biểu diễn bằng việc tối thiểu hóa độ lớn của vector trọng số $||w||$ trong khi giữ cho tất cả các điểm dữ liệu nằm đúng phía của siêu phẳng:

minimize $\frac{1}{2}||w||^2$

subject to $y_i(w^Tx_i + b)\geq 1$, for all $i=1,2,...,n$

Trong đó y_i là nhãn của mẫu x_i

Vấn đề này có thể được giải quyết bằng các phương pháp tối ưu hóa convex như phương pháp Gradient Descent hoặc bằng các phương pháp tối ưu hóa convex cơ bản khác.

2.1.3. Ưu và nhược điểm

Ưu điểm

- Xử lý trên không gian số chiều cao: SVM là một công cụ tính toán hiệu quả trong không gian chiều cao, trong đó đặc biệt áp dụng cho các bài toán phân loại văn bản và phân tích quan điểm nơi chiều có thể cực kỳ lớn.

- Tiết kiệm bộ nhớ: Do chỉ có một tập hợp con của các điểm được sử dụng trong quá trình huấn luyện và ra quyết định thực tế cho các điểm dữ liệu mới nên chỉ có những điểm cần thiết mới được lưu trữ trong bộ nhớ khi ra quyết định.
- Tính linh hoạt - phân lớp thường là phi tuyến tính. Khả năng áp dụng Kernel mới cho phép linh động giữa các phương pháp tuyến tính và phi tuyến tính từ đó khiến cho hiệu suất phân loại lớn hơn.

Nhược điểm:

- Bài toán số chiều cao: Trong trường hợp số lượng thuộc tính (p) của tập dữ liệu lớn hơn rất nhiều so với số lượng dữ liệu (n) thì SVM cho kết quả khá tồi.
- Chưa thể hiện rõ tính xác suất: Việc phân lớp của SVM chỉ là việc cố gắng tách các đối tượng vào hai lớp được phân tách bởi siêu phẳng SVM. Điều này chưa giải thích được xác suất xuất hiện của một thành viên trong một nhóm là như thế nào. Tuy nhiên hiệu quả của việc phân lớp có thể được xác định dựa vào khái niệm margin từ điểm dữ liệu mới đến siêu phẳng phân lớp mà chúng ta đã bàn luận ở trên.

2.2. Mô hình Linear Regression

2.2.1. Khái niệm

Linear Regression là một phương pháp trong học máy dùng để mô hình hóa mối quan hệ tuyến tính giữa biến đầu vào và biến đầu ra. Ý tưởng chính là xây dựng một đường thẳng (trong trường hợp một chiều) hoặc một siêu phẳng (trong trường hợp nhiều chiều) sao cho tổng bình phương của sai số giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất.[2]

2.2.2. Thuật toán Linear Regression

Linear Regression (Hồi quy tuyến tính) là một mô hình thống kê được sử dụng để mô hình hóa mối quan hệ tuyến tính giữa một biến độc lập (x) và một biến phụ thuộc (y). Mô hình tuyến tính giả định rằng có một mối quan hệ tuyến tính giữa các biến này, có thể được biểu diễn bằng một đường thẳng.

Đường thẳng tuyến tính được biểu diễn bởi phương trình:

$$y=mx+b$$

Trong đó:

y là biến phụ thuộc (đầu ra).

x là biến độc lập (đầu vào).

m là hệ số của đường thẳng (độ dốc).

b là hệ số ức chế (chặn).

Mục tiêu của tính năng khôi phục tuyến tính là giá trị của m và b sao cho đường thẳng tạo ra dự đoán gần nhất với dữ liệu thực tế. Điều này thường được thực hiện bằng cách giảm thiểu hóa tổng bình phương của độ chênh lệch giữa giá trị dự đoán và giá trị thực tế. Quá trình này được gọi là phương pháp bình phương tối thiểu (Phương pháp bình phương tối thiểu).

Hàm Mất mát (Loss Function):

$$L(m, b) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - (mx_i + b))^2$$

Trong đó:

n là số lượng điểm dữ liệu.

y_i và x_i là giá trị thực tế và giá trị dự đoán của điểm dữ liệu thứ i.

Tối thiểu hóa Hàm Mất mát:

Mục tiêu là tìm m và b sao cho giá trị của hàm mất mát là nhỏ nhất. Điều này có thể được thực hiện thông qua đạo hàm riêng và giải hệ phương trình đạo hàm bằng không.

2.2.3. Ưu và nhược điểm

Ưu điểm:

- Dễ hiểu và triển khai: Linear Regression là một phương pháp đơn giản, dễ hiểu và dễ triển khai.
- Hiệu quả với dữ liệu tuyến tính: Khi dữ liệu phù hợp với mô hình tuyến tính, Linear Regression cho kết quả hiệu quả.

Nhược điểm:

- Giả định về tuyến tính: Mô hình giả định rằng mối quan hệ giữa biến đầu vào và đầu

ra là tuyến tính, điều này có thể không phù hợp cho các vấn đề có mối quan hệ phức tạp hơn.

- Nhạy cảm với nhiễu: Linear Regression có thể bị ảnh hưởng bởi dữ liệu nhiễu, làm giảm độ chính xác của mô hình.

2.3. Mô hình Decision Trees

2.3.1. Khái niệm

Cây quyết định là một kiểu mô hình dự báo (predictive model), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng.

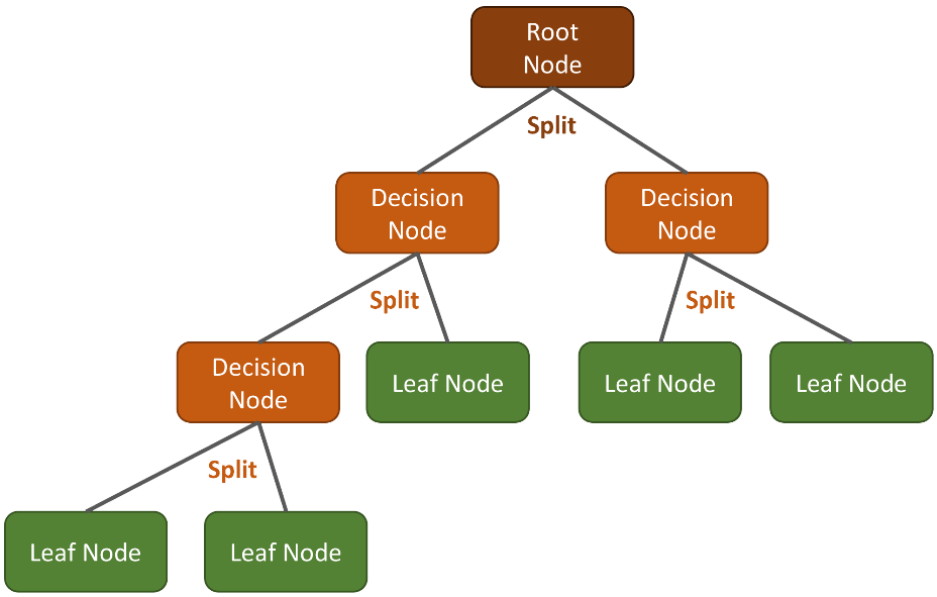
Cây quyết định có cấu trúc hình cây và là một sự tượng trưng của một phương thức quyết định cho việc xác định lớp các sự kiện đã cho. Mỗi nút của cây chỉ ra một tên lớp hoặc một phép thử cụ thể, phép thử này chia không gian các dữ liệu tại nút đó thành các kết quả có thể đạt được của phép thử. Mỗi tập con được chia ra là không gian con của các dữ liệu được tương ứng với vấn đề con của sự phân loại. Sự phân chia này thông qua một cây con tương ứng. Quá trình xây dựng cây quyết định có thể xem như là một chiến thuật chia để trị cho sự phân loại đối tượng. Một cây quyết định có thể mô tả bằng các khái niệm nút và đường nối các nút trong cây.

Mỗi nút của cây quyết định có thể là:

- Nút lá (leaf node) hay còn gọi là nút trả lời (answer node), biểu thị cho một lớp các trường hợp (bản ghi), nhãn là tên của lớp.

- Nút không phải là lá (non-leaf node) hay còn gọi là nút trong (inner node), nút này xác định một phép thử thuộc tính (attribute test), nhãn của nút này có tên của thuộc tính và sẽ có một nhánh (hay đường đi) nối nút này đến cây con (subtree) ứng với mỗi kết quả có thể có của phép thử. Nhãn của nhánh này chính là giá trị của thuộc tính đó. Nút không phải lá nằm trên cùng là nút gốc (root node).

Một cây quyết định sử dụng để phân loại dữ liệu bằng cách bắt đầu đi từ nút gốc của cây và đi xuyên qua cây theo các nhánh cho tới khi gặp nút lá, khi đó ta sẽ được lớp của dữ kiện đang xét.



Hình 2. 2: Mô hình Decision Trees

2.3.2. Thuật toán hồi quy của Decision Trees

Cây quyết định hồi quy tạo ra các quy tắc (rules) để dự đoán giá trị đầu ra dựa trên giá trị của các đặc trưng đầu vào. Mỗi lá của cây đại diện cho một phân khúc của không gian đầu vào và có một giá trị dự đoán.

Phương trình cơ bản của cây quyết định hồi quy có thể được biểu diễn như sau:

$$\hat{y} = f(x)$$

Trong đó:

$\hat{y}$ là giá trị dự đoán hồi quy.

x là vector đặc trưng đầu vào.

Quy trình xây dựng cây quyết định hồi quy

Chia Phân Khúc:

Cây quyết định bắt đầu với một nút gốc, đại diện cho toàn bộ không gian đầu vào.

Nút được chia thành hai (hoặc nhiều hơn) nút con dựa trên giá trị của một đặc trưng.

Lựa Chọn Đặc Trưng và Ngưỡng Chia:

Lựa chọn đặc trưng tại mỗi nút dựa trên tiêu chí như Mean Squared Error (MSE) hoặc Mean Absolute Error (MAE).

Chọn ngưỡng chia tối ưu để tối thiểu hóa lỗi dự đoán.

Tạo Các Nút Con:

Mỗi nút con tiếp tục quá trình chia bằng cách lựa chọn đặc trưng và ngưỡng chia tối ưu cho không gian đầu vào của nó.

Quá trình này được lặp lại cho đến khi một điều kiện dừng được đạt được, chẳng hạn như độ sâu tối đa hoặc số lượng mẫu tối thiểu trong mỗi lá.

Xác Định Giá Trị Dự Đoán tại Lá:

Khi một lá được tạo, giá trị dự đoán tại lá được xác định, thường là trung bình của các giá trị đầu ra trong lá đó.

2.3.3. Ưu và nhược điểm

Ưu Điểm:

Dễ diễn giải và hiểu.

Có thể xử lý cả dữ liệu số và dữ liệu phân loại.

Không yêu cầu chuẩn bị dữ liệu đặc biệt như chuẩn hóa.

Nhược Điểm:

Dễ bị overfitting nếu không kiểm soát độ sâu hoặc số lượng lá.

Khả năng không ổn định khi dữ liệu thay đổi nhỏ.

Có thể tạo ra cây quá phức tạp với nhiều lá khi không kiểm soát.

2.4. Mô hình Random Forest

2.4.1. Khái niệm

Random Forest (rừng ngẫu nhiên) là phương pháp học tập thể (ensemble) để phân loại, hồi quy được phát triển bởi Leo Breiman tại đại học California, Berkeley. Breiman cũng

đồng thời là đồng tác giả của phương pháp CART . Random Forest (RF) là phương pháp cải tiến của phương pháp tổng hợp bootstrap (bagging). RF sử dụng 2 bước ngẫu nhiên, một là ngẫu nhiên theo mẫu (sample) dùng phương pháp bootstrap có hoàn lại (with replacement), hai là lấy ngẫu nhiên một lượng thuộc tính từ tập thuộc tính ban đầu. Các tập dữ liệu con (sub-dataset) được tạo ra từ 2 lần ngẫu nhiên này có tính đa dạng cao, ít liên quan đến nhau, giúp giảm lỗi phương sai (variance). Các cây CART được xây dựng từ tập các tập dữ liệu con này tạo thành rừng. Khi tổng hợp kết quả, RF dùng phương pháp bỏ phiếu (voting) cho bài toán phân loại và lấy giá trị trung bình (average) cho bài toán hồi quy. Việc kết hợp các mô hình CART này để cho kết quả cuối cùng nên RF được gọi là phương pháp học tập thể.

Đối với bài toán phân loại, cây CART sử dụng công thức Gini như là một hàm điều kiện để tính toán điểm tách nút của cây. Số lượng cây là không hạn chế, các cây trong RF được xây dựng với chiều cao tối đa.

Trong những năm gần đây, RF được sử dụng khá phổ biến bởi những điểm vượt trội so với các thuật toán khác: xử lý được với dữ liệu có số lượng các thuộc tính lớn, có khả năng ước lượng được độ quan trọng của các thuộc tính, thường có độ chính xác cao trong phân loại (hoặc hồi quy), quá trình học nhanh. Trong RF, mỗi cây chỉ chọn một tập nhỏ các thuộc tính trong quá trình xây dựng (bước ngẫu nhiên thứ 2), cơ chế này làm cho RF thực thi với tập dữ liệu có số lượng thuộc tính lớn trong thời gian chấp nhận được khi tính toán. Người dùng có thể đặt mặc định số lượng các thuộc tính để xây dựng cây trong rừng, thông thường giá trị mặc định tối ưu là $\sqrt{p}$ cho bài toán phân loại và p/3 với các bài toán hồi quy (p là số lượng tất cả các thuộc tính của tập dữ liệu ban đầu). Số lượng các cây trong rừng cần được đặt đủ lớn để đảm bảo tất cả các thuộc tính đều được sử dụng một số lần. Thông thường là 500 cây cho bài toán phân loại, 1000 cây cho bài toán hồi quy. Do sử dụng phương pháp bootstrap lấy mẫu ngẫu nhiên có hoàn lại nên các tập dữ liệu con có khoảng 2/3 các mẫu không trùng nhau dùng để xây dựng cây, các mẫu ngày được gọi là in-bag. Khoảng 1/3 số mẫu còn lại gọi là out-of-bag, do không tham gia vào việc xây dựng cây nên RF dùng luôn các mẫu out-of-bag này để kiểm thử và tính toán độ quan trọng thuộc tính của các cây CART trong rừng.

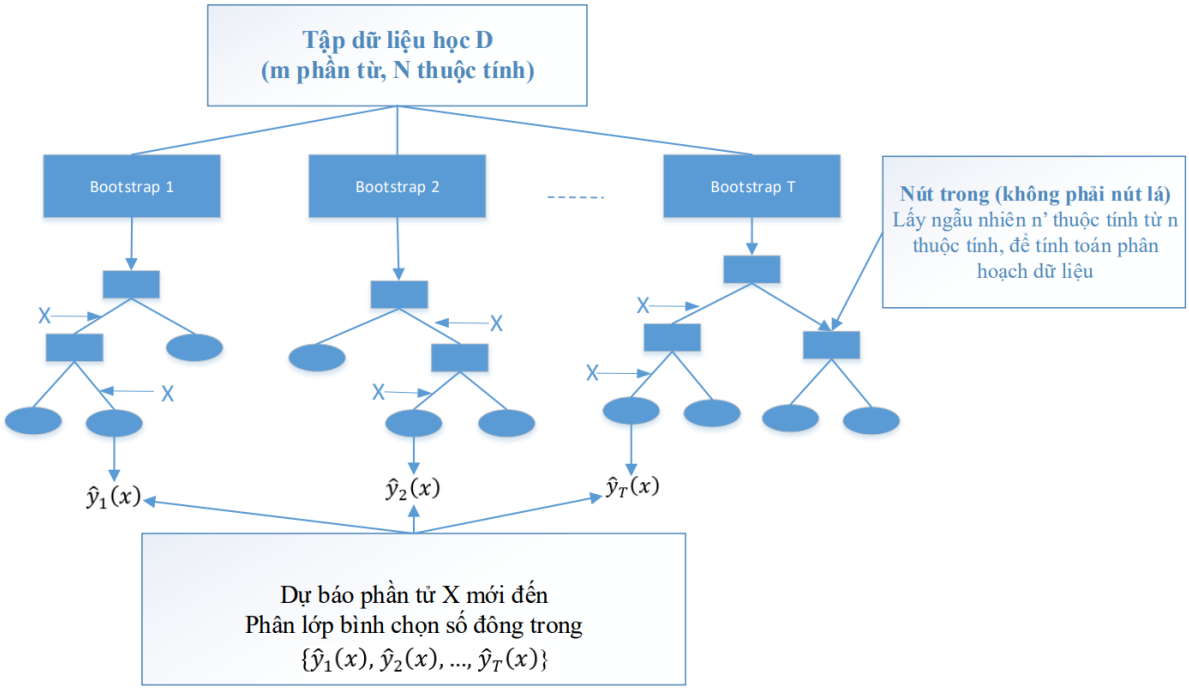
2.4.2. Thuật toán Rừng ngẫu nhiên

Mô tả thuật toán Tóm tắt thuật toán Random Forest cho phân loại dữ liệu:

Bước 1: Từ tập dữ liệu huấn luyện D, ta tạo dữ liệu ngẫu nhiên (mẫu bootstrap).

Bước 2: Sử dụng các tập con dữ liệu lấy mẫu ngẫu nhiên D_1, D_2,…, D_k xây dựng nên các cây T_1, T_2,…,T_k .

Bước 3: Kết hợp các cây: sử dụng chiến lược bình chọn theo số đông với bài toán phân loại hoặc lấy trung bình các giá trị dự đoán từ các cây với bài toán hồi quy.



Hình 2. 3: Thuật toán rừng ngẫu nhiên

Quá trình học của Random Forest bao gồm việc sử dụng ngẫu nhiên giá trị đầu vào, hoặc kết hợp các giá trị đó tại mỗi node trong quá trình dựng từng cây quyết định. Trong đó Random Forest có một số thuộc tính mạnh như:

(1) Độ chính xác của RF tương đối cao.

(2) Thuật toán giải quyết tốt các bài toán có nhiều dữ liệu nhiễu.

(3) Thuật toán chạy nhanh hơn so với bagging.

(4) Có những sự ước lượng nội tại như độ chính xác của mô hình dự đoán hoặc độ mạnh và liên quan giữa các thuộc tính.

(5) Dễ dàng thực hiện song song.

(6) Tuy nhiên để đạt được các tính chất mạnh trên, thời gian thực thi của thuật toán khá lâu và phải sử dụng nhiều tài nguyên của hệ thống

Tính chất thứ 4 được quan tâm rất nhiều và là tính chất được sử dụng để giải quyết bài toán trích chọn thuộc tính. Sau khi thực hiện học sẽ thu được một danh sách các thuộc được xếp hạng dựa theo một trong hai tiêu chí. Tiêu chí thứ nhất là thu được sau quá trình kiểm tra độ chính xác sử dụng các mẫu out-of-bag. Tiêu chí thứ hai là mức độ dầy đặc tại các node khi phân chia thuộc thuộc tính, và được tính trung bình trên tất cả các cây.

Qua những tìm hiểu trên về giải thuật RF ta có nhận xét rằng RF là một phương pháp phân loại tốt do:

(1) Trong RF các phương sai (variance) được giảm thiểu do kết quả của RF được tổng hợp thông qua nhiều bộ học (learner).

(2) Việc chọn ngẫu nhiên tại mỗi bước trong RF sẽ làm giảm mối tương quan (correlation) giữa các bộ phận lớp trong việc tổng hợp các kết quả.

2.4.3. Ưu và nhược điểm

Ưu điểm: Random forests được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình này. Thuật toán không bị vấn đề overfitting. Lý do chính là mất trung bình của tất cả các dự đoán, trong đó hủy bỏ những thành kiến. Thuật toán có thể được sử dụng trong cả hai vấn đề phân loại và hồi quy. Random forests cũng có thể xử lý các giá trị còn thiếu. Có hai cách để xử lý các giá trị này: sử dụng các giá trị trung bình để thay thế các biến liên tục và tính toán mức trung bình gần kề của các giá trị bị thiếu. Bạn có thể nhận được tầm quan trọng của tính năng tương đối, giúp chọn các tính năng đóng góp nhiều nhất cho trình phân loại.

Nhược điểm: Random forests chậm tạo dự đoán bởi vì có nhiều cây quyết định. Bất cứ khi nào đưa ra dự đoán, tất cả các cây trong rừng phải đưa ra dự đoán cho cùng một đầu vào cho trước và sau đó thực hiện bỏ phiếu trên đó. Toàn bộ quá trình này tốn thời gian. Mô hình khó hiểu hơn so với cây quyết định, nơi bạn có thể dễ dàng đưa ra quyết định bằng cách đi theo đường dẫn trong cây.

2.5. Thuật toán tối ưu Grid Search

2.5.1. Khái niệm

Grid Search là một phương pháp tìm kiếm siêu tham số (hyperparameter) trong các mô hình học máy. Phương pháp này thực hiện bằng cách thử tất cả các tổ hợp có thể của các giá trị siêu tham số được xác định trước, từ đó tìm ra tổ hợp tối ưu nhất dựa trên một tiêu chí đánh giá như độ chính xác hoặc F1 score.

2.5.2. Thuật toán

Grid Search tạo ra một lưới các giá trị cho mỗi siêu tham số được xác định và sau đó kiểm tra hiệu suất của mô hình với mỗi tổ hợp giá trị. Mỗi tổ hợp sẽ được đánh giá bằng cách sử dụng một phương pháp cross-validation để đánh giá mô hình và tiêu chí đánh giá được chọn trước.

Xác định mô hình và các siêu tham số cần tối ưu hóa:

Chọn mô hình học máy mà bạn muốn tối ưu hóa (ví dụ: SVM, Random Forest, v.v.). Liệt kê các siêu tham số của mô hình mà bạn muốn tối ưu hóa (ví dụ: C và gamma trong SVM, số lượng cây trong Random Forest).

Xác định phạm vi giá trị cho các siêu tham số:

Đặt các giá trị có thể cho mỗi siêu tham số. Các giá trị này tạo thành một "lưới" các tổ hợp tham số để thử nghiệm.

Thiết lập Grid Search với cross-validation:

Sử dụng cross-validation để đánh giá hiệu suất của mô hình cho mỗi tổ hợp tham số trên lưới.

Chia dữ liệu thành nhiều tập con và kiểm tra mô hình trên các tập con khác nhau để đảm bảo tính tổng quát.

Chạy Grid Search:

Thử nghiệm tất cả các tổ hợp siêu tham số có thể và huấn luyện mô hình tương ứng.

Đánh giá hiệu suất của mỗi mô hình và ghi lại kết quả.

Chọn tổ hợp siêu tham số tốt nhất:

Chọn tổ hợp tham số có hiệu suất tốt nhất dựa trên độ đo đánh giá đã chọn (ví dụ: độ chính xác, F1-score, MSE).

2.5.3. Ưu và nhược điểm

Ưu Điểm:

- Hoàn chỉnh: Grid Search đảm bảo mọi tổ hợp siêu tham số đều được kiểm tra, giúp đảm bảo tìm ra tổ hợp tối ưu nhất.
- Dễ dàng sử dụng: Phương pháp này dễ dàng triển khai và hiểu, đặc biệt là cho những người mới bắt đầu trong lĩnh vực học máy.

Nhược Điểm:

- Chi phí tính toán cao: Grid Search có thể trở nên tốn kém tính toán khi số lượng siêu tham số và giá trị được kiểm tra tăng lên.
- Không linh hoạt: Grid Search có thể bỏ qua các giá trị siêu tham số quan trọng nếu chúng không được xác định trong lưới giá trị được chọn.

2.6. Thuật toán tối ưu Random Search

2.6.1. Khái niệm

Random Search là một phương pháp tìm kiếm siêu tham số bằng cách chọn ngẫu nhiên các giá trị cho các siêu tham số được xác định trước. Thay vì thử tất cả các tổ hợp có thể như Grid Search, Random Search sẽ chọn một số lượng lớn các tổ hợp siêu tham số ngẫu nhiên để đánh giá.

2.6.2. Thuật toán

Random Search chọn ngẫu nhiên các giá trị cho mỗi siêu tham số từ phân phối xác định trước. Sau đó, mỗi tổ hợp giá trị được kiểm tra bằng cách sử dụng một phương pháp cross-validation và tiêu chí đánh giá được chọn trước.

Xác định mô hình và các siêu tham số cần tối ưu hóa:

- Chọn mô hình học máy mà bạn muốn tối ưu hóa (ví dụ: SVM, Random Forest, v.v.).

- Liệt kê các siêu tham số của mô hình mà bạn muốn tối ưu hóa.

- Xác định phạm vi giá trị cho các siêu tham số:

- Đặt các giá trị có thể hoặc phân phối xác suất cho mỗi siêu tham số. Random Search sẽ chọn ngẫu nhiên từ các giá trị này.

Thiết lập Random Search với cross-validation:

- Sử dụng cross-validation để đánh giá hiệu suất của mô hình cho mỗi tổ hợp tham số được chọn ngẫu nhiên.

- Chia dữ liệu thành nhiều tập con và kiểm tra mô hình trên các tập con khác nhau để đảm bảo tính tổng quát.

Chạy Random Search:

- Thử nghiệm một số lượng nhất định các tổ hợp tham số ngẫu nhiên và huấn luyện mô hình tương ứng.

- Đánh giá hiệu suất của mỗi mô hình và ghi lại kết quả.

- Chọn tổ hợp siêu tham số tốt nhất:

- Chọn tổ hợp tham số có hiệu suất tốt nhất dựa trên độ đo đánh giá đã chọn (ví dụ: độ chính xác, F1-score, MSE).

2.6.3. Ưu và nhược điểm

Ưu Điểm:

- Hiệu quả về mặt tính toán: Random Search thường hiệu quả hơn so với Grid Search vì không cần phải thử tất cả các tổ hợp giá trị siêu tham số.
- Linh hoạt: Phương pháp này linh hoạt hơn vì có thể chọn từ một phân phối xác định thay vì ràng buộc bởi một lưới giá trị nhất định.

Nhược Điểm:

- Không đảm bảo tối ưu: Do không kiểm tra tất cả các tổ hợp giá trị siêu tham số,

Random Search không đảm bảo tìm ra tổ hợp tối ưu nhất.

- Cần nhiều lượt thử: Đôi khi có thể cần nhiều lượt thử hơn để tìm ra tổ hợp tốt nhất so với Grid Search.

Chương 3
PHÂN TÍCH VÀ XỬ LÝ DỮ LIỆU

3.1. Bài toán

Trong một trường học gồm các học sinh đang học lớp 12. Để giúp quản lý lớp học và hỗ trợ học sinh, nhà trường quyết định thực hiện một nghiên cứu để dự đoán điểm trung bình của học sinh lớp 12 dựa vào điểm số từ lớp 10 và lớp 11 của họ. Mục tiêu là xây dựng một mô hình dự đoán chính xác điểm trung bình của học sinh từ các năm học trước.

Mục Tiêu:

Xây dựng một mô hình hồi quy có khả năng dự đoán điểm trung bình của học sinh lớp 12 dựa vào điểm lớp 10 và lớp 11 của họ.

Phân Tích:

Tiến Hành Tiền Xử Lý Dữ Liệu: Xử lý dữ liệu thiếu, chuẩn hóa dữ liệu nếu cần, và chia dữ liệu thành tập huấn luyện và tập kiểm tra.

Xây Dựng Mô Hình:

Sử dụng các thuật toán hồi quy như Linear Regression, Decision Tree Regression, SVM hoặc Random Forest Regression để xây dựng mô hình dự đoán điểm trung bình của học sinh lớp 12.

Đánh Giá Mô Hình:

Đánh giá hiệu suất của mô hình bằng các độ đo như Mean Squared Error (MSE), Mean Absolute Error (MAE) trên tập kiểm tra.

Kết Quả Dự Kiến:

Một mô hình hồi quy có khả năng dự đoán điểm trung bình của học sinh lớp 12 dựa vào điểm lớp 10 và lớp 11 của họ. Mô hình này có thể được sử dụng để đưa ra dự đoán về điểm trung bình của các học sinh trong tương lai và hỗ trợ quyết định trong việc cung cấp hỗ trợ học tập và phát triển cá nhân cho học sinh.

3.2. Dữ liệu nghiên cứu

Dữ liệu điểm của trường đại học Quốc gia, thu thập trong kì thi đánh giá năng lực năm 2022-2023.

3.3. Thống kê các trường dữ liệu

3.3.1. Dữ liệu ban đầu

Bộ dữ liệu: điểm HB.xlsx gồm 1555 dữ liệu tương ứng với 1555 học sinh

Mỗi một học sinh sẽ có 116 thông tin tương ứng với 116 trường

Bảng 3. 1: Bảng phân tích dữ liệu

Trường	Mô tả
Mã ĐTN	Mã trường của học sinh
Tên ĐTN	Tên trường của học sinh
Số CMND	Số chứng minh nhân dân của học sinh
Họ và tên	Họ tên học sinh
Ngày sinh	Ngày sinh của học sinh
Giới tính	Giới tính của học sinh
10.Điểm tổng kết HK I	Điểm tổng kết học kì 1 năm lớp 10
10.Điểm tổng kết HK II	Điểm tổng kết học kì 2 năm lớp 10
10.Điểm tổng kết CN	Điểm tổng kết cả năm năm lớp 10
10.Học lực HK I	Học lực học kì 1 năm lớp 10
10.Học lực HK II	Học lực học kì 2 năm lớp 10
10.Học lực CN	Học lực cả năm năm lớp 10
10.Hạnh kiểm HK I	Hạnh kiểm học kì 1 năm lớp 10
10.Hạnh kiểm HK II	Hạnh kiểm học kì 2 năm lớp 10
10.Hạnh kiểm CN	Hạnh kiểm cả năm năm lớp 10

10.Toán HK I	Điểm toán học kì 1 năm lớp 10
10.Toán HK II	Điểm toán học kì 2 năm lớp 10
10.Toán CN	Điểm toán cả năm năm lớp 10
10.Văn HK I	Điểm văn học kì 1 năm lớp 10
10.Văn HK II	Điểm văn học kì 2 năm lớp 10
10.Văn CN	Điểm văn cả năm năm lớp 10
10.Vật lí HK I	Điểm vật lý học kì 1 năm lớp 10
10.Vật lí HK II	Điểm vật lý học kì 2 năm lớp 10
10.Vật lí CN	Điểm vật lý cả năm năm lớp 10
10.Hóa học HK I	Điểm hóa học học kì 1 năm lớp 10
10.Hóa học HK II	Điểm hóa học học kì 2 năm lớp 10
10.Hóa học CN	Điểm hóa học cả năm năm lớp 10
10.Sinh học HK I	Điểm sinh học học kì 1 năm lớp 10
10.Sinh học HK II	Điểm sinh học học kì 2 năm lớp 10
10.Sinh học CN	Điểm sinh học cả năm năm lớp 10
10.Lịch sử HK I	Điểm lịch sử học kì 1 năm lớp 10
10.Lịch sử HK II	Điểm lịch sử học kì 2 năm lớp 10
10.Lịch sử CN	Điểm lịch sử cả năm năm lớp 10
10.Địa lí HK I	Điểm địa lý học kì 1 năm lớp 10
10.Địa lí HK II	Điểm địa lý học kì 2 năm lớp 10
10.Địa lí CN	Điểm địa lý cả năm năm lớp 10
10.GDCD HK I	Điểm giáo dục công dân học kì 1 năm lớp 10
10.GDCD HK II	Điểm giáo dục công dân học kì 2 năm lớp 10

10.GDCD CN	Điểm giáo dục công dân cả năm năm lớp 10
10.Ngoại ngữ HK I	Điểm ngoại ngữ học kì 1 năm lớp 10
10.Ngoại ngữ HK II	Điểm ngoại ngữ học kì 2 năm lớp 10
10.Ngoại ngữ CN	Điểm ngoại ngữ cả năm năm lớp 10
10.Môn ngoại ngữ	Tên của môn ngoại ngữ
11.Điểm tổng kết HK I	Điểm tổng kết học kì 1 năm lớp 11
11.Điểm tổng kết HK II	Điểm tổng kết học kì 2 năm lớp 11
11.Điểm tổng kết CN	Điểm tổng kết cả năm năm lớp 11
11.Học lực HK I	Học lực học kì 1 năm lớp 11
11.Học lực HK II	Học lực học kì 2 năm lớp 11
11.Học lực CN	Học lực cả năm năm lớp 11
11.Hạnh kiểm HK I	Hạnh kiểm học kì 1 năm lớp 11
11.Hạnh kiểm HK II	Hạnh kiểm học kì 2 năm lớp 11
11.Hạnh kiểm CN	Hạnh kiểm cả năm năm lớp 11
11.Toán HK I	Điểm toán học kì 1 năm lớp 11
11.Toán HK II	Điểm toán học kì 2 năm lớp 11
11.Toán CN	Điểm toán cả năm năm lớp 11
11.Văn HK I	Điểm văn học kì 1 năm lớp 11
11.Văn HK II	Điểm văn học kì 2 năm lớp 11
11.Văn CN	Điểm văn cả năm năm lớp 11
11.Vật lí HK I	Điểm vật lý học kì 1 năm lớp 11
11.Vật lí HK II	Điểm vật lý học kì 2 năm lớp 11
11.Vật lí CN	Điểm vật lý cả năm năm lớp 11

11.Hóa học HK I	Điểm hóa học học kì 1 năm lớp 11
11.Hóa học HK II	Điểm hóa học học kì 2 năm lớp 11
11.Hóa học CN	Điểm hóa học cả năm năm lớp 11
11.Sinh học HK I	Điểm sinh học học kì 1 năm lớp 11
11.Sinh học HK II	Điểm sinh học học kì 2 năm lớp 11
11.Sinh học CN	Điểm sinh học cả năm năm lớp 11
11.Lịch sử HK I	Điểm lịch sử học kì 1 năm lớp 11
11.Lịch sử HK II	Điểm lịch sử học kì 2 năm lớp 11
11.Lịch sử CN	Điểm lịch sử cả năm năm lớp 11
11.Địa lí HK I	Điểm địa lý học kì 1 năm lớp 11
11.Địa lí HK II	Điểm địa lý học kì 2 năm lớp 11
11.Địa lí CN	Điểm địa lý cả năm năm lớp 11
11.GDCD HK I	Điểm giáo dục công dân học kì 1 năm lớp 11
11.GDCD HK II	Điểm giáo dục công dân học kì 2 năm lớp 11
11.GDCD CN	Điểm giáo dục công dân cả năm năm lớp 11
11.Ngoại ngữ HK I	Điểm ngoại ngữ học kì 1 năm lớp 11
11.Ngoại ngữ HK II	Điểm ngoại ngữ học kì 2 năm lớp 11
11.Ngoại ngữ CN	Điểm ngoại ngữ cả năm năm lớp 11
11.Môn ngoại ngữ	Tên của môn ngoại ngữ
12.Điểm tổng kết HK I	Điểm tổng kết học kì 1 năm lớp 12
12.Điểm tổng kết HK II	Điểm tổng kết học kì 2 năm lớp 12
12.Điểm tổng kết CN	Điểm tổng kết cả năm năm lớp 12
12.Học lực HK I	Học lực học kì 1 năm lớp 12

12.Học lực HK II	Học lực học kì 2 năm lớp 12
12.Học lực CN	Học lực cả năm năm lớp 12
12.Hạnh kiểm HK I	Hạnh kiểm học kì 1 năm lớp 12
12.Hạnh kiểm HK II	Hạnh kiểm học kì 2 năm lớp 12
12.Hạnh kiểm CN	Hạnh kiểm cả năm năm lớp 12
12.Toán HK I	Điểm toán học kì 1 năm lớp 12
12.Toán HK II	Điểm toán học kì 2 năm lớp 12
12.Toán CN	Điểm toán cả năm năm lớp 12
12.Văn HK I	Điểm văn học kì 1 năm lớp 12
12.Văn HK II	Điểm văn học kì 2 năm lớp 12
12.Văn CN	Điểm văn cả năm năm lớp 12
12.Vật lí HK I	Điểm vật lý học kì 1 năm lớp 12
12.Vật lí HK II	Điểm vật lý học kì 2 năm lớp 12
12.Vật lí CN	Điểm vật lý cả năm năm lớp 12
12.Hóa học HK I	Điểm hóa học học kì 1 năm lớp 12
12.Hóa học HK II	Điểm hóa học học kì 2 năm lớp 12
12.Hóa học CN	Điểm hóa học cả năm năm lớp 12
12.Sinh học HK I	Điểm sinh học học kì 1 năm lớp 12
12.Sinh học HK II	Điểm sinh học học kì 2 năm lớp 12
12.Sinh học CN	Điểm sinh học cả năm năm lớp 12
12.Lịch sử HK I	Điểm lịch sử học kì 1 năm lớp 12
12.Lịch sử HK II	Điểm lịch sử học kì 2 năm lớp 12
12.Lịch sử CN	Điểm lịch sử cả năm năm lớp 12

12.Địa lí HK I	Điểm địa lý học kì 1 năm lớp 12
12.Địa lí HK II	Điểm địa lý học kì 2 năm lớp 12
12.Địa lí CN	Điểm địa lý cả năm năm lớp 12
12.GDCD HK I	Điểm giáo dục công dân học kì 1 năm lớp 12
12.GDCD HK II	Điểm giáo dục công dân học kì 2 năm lớp 12
12.GDCD CN	Điểm giáo dục công dân cả năm năm lớp 12
12.Ngoại ngữ HK I	Điểm ngoại ngữ học kì 1 năm lớp 12
12.Ngoại ngữ HK II	Điểm ngoại ngữ học kì 2 năm lớp 12
12.Ngoại ngữ CN	Điểm ngoại ngữ cả năm năm lớp 12

Tất cả các điểm đều từ 0 đến 10 và không có dữ liệu nhiễu (là những số lớn hơn 10 hoặc những số âm,…)

10.Toán HK II	10.Toán CN	10.Văn HK I	10.Văn HK II	10.Văn CN	10.Vật lí HK	10.Vật lí HK II	10.Vật lí CN	10.Hóa học HK I	10.Hóa học HK II	10.Hóa học CN	10.Sinh học HK	10.Sinh học HK I	10.Sinh học CN	10.Lịch sử HK	10.Lịch sử HK I	10.Lịch sử CN	10.Địa lí HK	10.Địa lí HK I	10.Địa lí CN
9.6	9.5	6.8	7.3	7.1	9.4	8.7	8.9	8.8	8.9	8.9	8.6	8.6	8.6	8.5	8.7	8.6	8.2	7.3	7.6
6.3	6.4	7	7.3	7.1	6.2	5.4	5.7	7.5	6.7	7	6.6	6.3	6.4	6.6	6.7	6.7	6.5	7.8	7.4
9.5	9.2	7.8	7.3	7.9	8.1	8.3	8.2	8.6	8.7	8.7	7.9	9.4	8.9	7.6	8.7	8.3	8	8.6	8.4
8.1	8.2	6.3	7.3	6.6	8.7	9	8.9	8.1	8	8	7.6	6.3	6.7	6.7	7	6.9	7.2	7.8	7.6
6.7	6.8	7.4	7.3	7.3	9	9.5	9.3	6.8	7.6	7.3	7.9	7.5	7.6	8.5	8.5	8.5	7.7	7.1	7.3
8.7	8.7	8.1	7.3	8.3	7.2	6.5	6.7	7.8	7.3	7.5	7.1	9.1	8.4	7.4	8.3	8	7.3	8.1	7.8
7.8	7.6	6.2	7.3	6.3	5.6	6.4	6.1	5.3	6.5	6.1	7.4	6.7	6.9	5.9	7.2	6.8	6.3	7.3	7
8	8	6.7	7.3	6.7	6.9	8.1	7.7	7.9	7.6	7.7	7.6	7.9	7.8	6.6	7.5	7.2	7.2	7.7	7.5
6.3	6	7.2	7.3	7.1	6.7	5.7	6	6.8	5.5	5.9	6.9	7.2	7.1	7.6	7.6	7.6	6.7	7.8	7.4
9.2	9.1	8.2	7.3	8.2	9.4	9.4	9.4	8.5	7.9	8.1	9	9.4	9.3	8	9.1	8.7	8.5	8	8.2
7.3	7.3	7	7.3	7.1	6.1	8.1	7.4	6.4	6	6.1	5.9	7.1	6.7	7.3	6.7	6.9	7.1	7.6	7.4
7.7	7.7	7.4	7.3	7.5	6.8	7.8	7.5	8	7.4	7.6	7.9	7.2	7.4	7.6	7.8	7.7	8.2	6.7	7.2
9.1	8.8	8	7.3	8.3	6.6	7.6	7.3	7.4	7.4	7.4	7.3	8.7	8.2	7.3	7.6	7.5	7.3	8.5	8.1
8.5	8.7	6.5	7.3	6.8	9	8.7	8.8	7.1	6.5	6.7	6.8	7.9	7.5	8.1	8.6	8.4	6.9	8.8	8.2
7	7	7.4	7.3	7.2	6.9	7.8	7.5	7	7.9	7.6	7.1	5.5	6	7.9	7.2	7.4	7	7.3	7.2
9.3	9.2	8	7.3	8.1	8.8	8.9	8.9	8.9	8.8	8.8	7.9	8.2	8.1	8.3	8	8.1	7.3	9.2	8.6
8.6	8.5	8	7.3	8.3	7.9	7.9	7.9	7.5	8.2	8	6.9	7.8	7.5	8.1	7.9	8	8.4	9.2	8.9
6.8	7	7.4	7.3	7.5	7.1	8.9	8.3	5.7	6.4	6.2	7.5	8.2	8	7.6	7.6	7.6	7.4	8	7.8
9.8	9.7	7.7	7.3	7.8	8.5	8.4	8.4	9.3	7.9	8.4	8.7	8.4	8.5	8.4	8.3	8.3	8.3	8.6	8.5
8.8	8.5	6.5	7.3	6.8	8.9	8.8	8.8	7.4	9	8.5	7.9	8.1	8	8	7.9	7.9	7.9	8.1	8
7.3	7.7	8.1	7.3	8	8	7.7	7.8	9.6	8.6	8.9	8.2	8.2	8.2	8.6	7.8	8.1	7.7	8.9	8.5
9	9	7	7.3	7.4	8.8	7.9	8.2	7.2	7.9	7.7	7	8.3	7.9	7	7.3	7.2	6.8	8.6	8
6.4	6.8	6.8	7.3	6.7	7.2	7.5	7.4	7.1	7.2	7.2	6.6	7.8	7.4	7.7	6.8	7.1	7.2	7.3	7.3
8.5	8.4	8	7.3	8.2	6.6	8.1	7.6	7.5	6.2	6.6	6.9	7.7	7.4	7	7.9	7.6	6.6	7.9	7.5
5.7	5.8	5.9	7.3	6.6	6.2	7.4	7	4.6	5.3	5.1	6.6	7.5	7.2	6.1	8.7	7.8	6.3	7.3	7
7.5	7.6	7	7.3	7.2	7.4	8.3	8	6.2	5.9	6	7.2	8.4	8	6.9	7.9	7.6	7	7.9	7.6
7.3	7.5	6.5	7.3	6.8	8.8	8.3	8.5	8.3	7	7.4	8.8	7.9	8.2	7.6	8.6	8.3	8.2	5.9	6.7
8.7	8.4	7.6	7.3	7.7	8.4	8.8	8.7	7.4	9	8.5	7.8	9.3	8.8	8.3	8.7	8.6	8.4	8.4	8.4
7.5	7.1	7.3	7.3	7.4	6.6	7.1	6.9	7.2	8	7.7	7	8.3	7.9	6.2	8.3	7.6	7.9	6.9	7.2
8.3	8.1	7.5	7.3	8	7.5	7.8	7.7	7.2	6.9	7	6.8	9.6	8.7	7.2	8.1	7.8	7.4	8.9	8.4
6.9	6.7	7.3	7.3	7.4	7	7.2	7.1	7.9	7.4	7.6	5.2	7.3	6.6	7.6	7.9	7.8	6.8	7.1	7
8.5	8.7	5.7	7.3	5.9	8.4	8.6	8.5	7.6	7.9	7.8	6.6	7.6	7.3	5.8	6.9	6.5	6.9	6.8	6.8
6.3	6.3	5.8	7.3	6.3	5.9	6.4	6.2	6.5	5.7	6	7.4	7.1	7.2	7.2	8.4	8	7.6	7.8	7.7
7	7.1	6.7	7.3	6.8	7.7	7.1	7.3	7	7.1	7.1	5.5	6.1	5.9	6.4	5.9	6.1	6.3	6.9	6.7
6.9	6.8	6.3	7.3	6.5	8.2	7.1	7.5	8	8.3	8.2	6	7.9	7.3	7.1	6.4	6.6	6.1	6.5	6.4
5	5	5.9	7.3	6.6	7.9	7.1	7.4	4.3	5.9	5.4	5.5	6.6	6.2	6.1	6.4	6.3	5.6	7.5	6.9
8.9	8.7	7	7.3	7	8.8	9	8.9	9.1	9.1	9.1	9.4	8.5	8.8	8.5	8.2	8.3	8.9	8	8.3
8.3	8.2	6.8	7.3	6.7	8.7	9	8.9	7.7	8.3	8.1	7.5	7.8	7.7	7	8.2	7.8	7.1	8.2	7.8
8.7	8.6	7.6	7.3	7.6	8.8	8.1	8.3	9.3	9.3	9.3	9	8.5	8.7	8.5	8.2	8.3	8.5	8.2	8.3
8.9	8.5	6.8	7.3	6.9	8.7	9	8.9	7.5	8.7	8.3	7.4	8.5	8.1	7.9	8.3	8.2	7.8	8.3	8.1
9.1	8.8	7	7.3	7.1	9	8.9	8.9	8.8	8.7	8.7	6.6	9.3	8.4	6.7	8.2	7.7	8.4	8.6	8.5

Hình 3. 1: Dữ liệu ban đầu

3.3.2. Tiền xử lý dữ liệu

Tiến hành loại bỏ một số trường dữ liệu không cần thiết:

Bảng 3. 2: Bảng các trường loại bỏ

STT	Tên Trường
1	Mã ĐTN
2	Tên ĐTN
3	Số CMND
4	Họ và tên
5	Ngày sinh

Xử lý dữ liệu bị trống:

Bảng 3. 3: Bảng xử lý dữ liệu trống

Dữ liệu bị trống	Dữ liệu thay thế
Điểm thuộc học kì I	Điểm trung bình học kì I
Điểm thuộc học kì II	Điểm trung bình học kì II
Điểm cả năm	Điểm trung bình cả năm

Thống kê kiểu dữ liệu của các trường và tổng dữ liệu của từng trường:

Bảng 3. 4: Bảng thống kê dữ liệu

STT	Tên trường	Kiểu dữ liệu	Tổng
0	Giới tính	object	2
1	10.Điểm tổng kết HK I	float64	37
2	10.Điểm tổng kết HK II	float64	35
3	10.Điểm tổng kết CN	float64	35
4	10.Học lực HK I	object	4

5	10.Học lực HK II	object	4
6	10.Học lực CN	object	4
7	10.Hạnh kiểm HK I	object	4
8	10.Hạnh kiểm HK II	object	3
9	10.Hạnh kiểm CN	object	3
10	10.Toán HK I	float64	53
11	10.Toán HK II	float64	56
12	10.Toán CN	float64	52
13	10.Văn HK I	float64	46
14	10.Văn HK II	float64	45
15	10.Văn CN	float64	42
16	10.Vật lí HK I	float64	59
17	10.Vật lí HK II	float64	55
18	10.Vật lí CN	float64	54
19	10.Hóa học HK I	float64	57
20	10.Hóa học HK II	float64	55
21	10.Hóa học CN	float64	53
22	10.Sinh học HK I	float64	56
23	10.Sinh học HK II	float64	49
24	10.Sinh học CN	float64	46
25	10.Lịch sử HK I	float64	51
26	10.Lịch sử HK II	float64	52
27	10.Lịch sử CN	float64	47
28	10.Địa lí HK I	float64	52
29	10.Địa lí HK II	float64	47
30	10.Địa lí CN	float64	44
31	10.GDCD HK I	float64	54
32	10.GDCD HK II	float64	44
33	10.GDCD CN	float64	44

34	10.Ngoại ngữ HK I	float64	57
35	10.Ngoại ngữ HK II	float64	55
36	10.Ngoại ngữ CN	float64	52
37	10.Môn ngoại ngữ	object	3
38	11.Điểm tổng kết HK I	float64	41
39	11.Điểm tổng kết HK II	float64	33
40	11.Điểm tổng kết CN	float64	36
41	11.Học lực HK I	object	4
42	11.Học lực HK II	object	3
43	11.Học lực CN	object	3
44	11.Hạnh kiểm HK I	object	3
45	11.Hạnh kiểm HK II	object	3
46	11.Hạnh kiểm CN	object	3
47	11.Toán HK I	float64	49
48	11.Toán HK II	float64	46
49	11.Toán CN	float64	45
50	11.Văn HK I	float64	45
51	11.Văn HK II	float64	42
52	11.Văn CN	float64	41
53	11.Vật lí HK I	float64	61
54	11.Vật lí HK II	float64	52
55	11.Vật lí CN	float64	52
56	11.Hóa học HK I	float64	59
57	11.Hóa học HK II	float64	54
58	11.Hóa học CN	float64	53
59	11.Sinh học HK I	float64	51
60	11.Sinh học HK II	float64	47
61	11.Sinh học CN	float64	45
62	11.Lịch sử HK I	float64	50

63	11.Lịch sử HK II	float64	48
64	11.Lịch sử CN	float64	45
65	11.Địa lí HK I	float64	48
66	11.Địa lí HK II	float64	45
67	11.Địa lí CN	float64	44
68	11.GDCD HK I	float64	50
69	11.GDCD HK II	float64	37
70	11.GDCD CN	float64	37
71	11.Ngoại ngữ HK I	float64	54
72	11.Ngoại ngữ HK II	float64	52
73	11.Ngoại ngữ CN	float64	50
74	11.Môn ngoại ngữ	object	3
75	12.Điểm tổng kết HK I	float64	29
76	12.Điểm tổng kết HK II	float64	29
77	12.Điểm tổng kết CN	float64	29
78	12.Học lực HK I	object	3
79	12.Học lực HK II	object	3
80	12.Học lực CN	object	2
81	12.Hạnh kiểm HK I	object	3
82	12.Hạnh kiểm HK II	object	2
83	12.Hạnh kiểm CN	object	2
84	12.Toán HK I	float64	39
85	12.Toán HK II	float64	37
86	12.Toán CN	float64	36
87	12.Văn HK I	float64	43
88	12.Văn HK II	float64	43
89	12.Văn CN	float64	42
90	12.Vật lí HK I	float64	49
91	12.Vật lí HK II	float64	46

92	12.Vật lí CN	float64	43
93	12.Hóa học HK I	float64	45
94	12.Hóa học HK II	float64	44
95	12.Hóa học CN	float64	41
96	12.Sinh học HK I	float64	45
97	12.Sinh học HK II	float64	41
98	12.Sinh học CN	float64	41
99	12.Lịch sử HK I	float64	43
100	12.Lịch sử HK II	float64	44
101	12.Lịch sử CN	float64	38
102	12.Địa lí HK I	float64	41
103	12.Địa lí HK II	float64	40
104	12.Địa lí CN	float64	39
105	12.GDCD HK I	float64	40
106	12.GDCD HK II	float64	38
107	12.GDCD CN	float64	35
108	12.Ngoại ngữ HK I	float64	49
109	12.Ngoại ngữ HK II	float64	46
110	12.Ngoại ngữ CN	float64	43
111	12.Môn ngoại ngữ	object	3

3.4. Phân chia dữ liệu huấn luyện

Dữ liệu về điểm số của học sinh được lấy của 3 khối là 10, 11, 12. Trước khi bắt đầu quá trình huấn luyện và đánh giá mô hình, bước quan trọng nhất là phân chia dữ liệu thành tập huấn luyện và tập kiểm tra. Quyết định phân chia dữ liệu này quan trọng để đảm bảo rằng mô hình được đánh giá trên dữ liệu mà nó chưa từng thấy trước đó, giúp đánh giá tính tổng quát của mô hình.

Phương pháp phân chia dữ liệu 20-80 đã được sử dụng. Điều này có nghĩa là 20% của dữ liệu được dành cho tập kiểm tra, trong khi 80% còn lại được sử dụng cho tập huấn luyện.

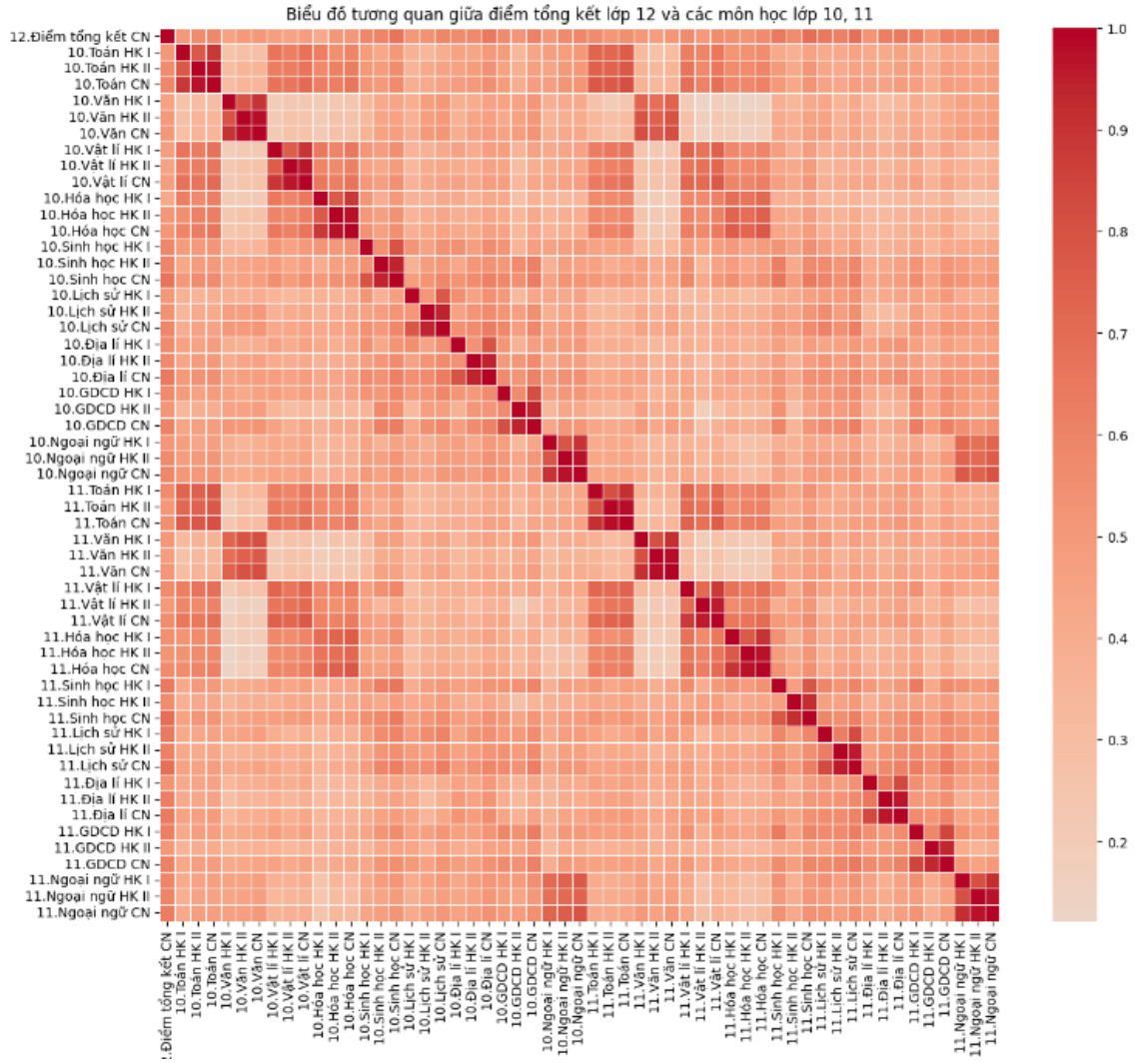
Việc này đảm bảo rằng một phần lớn dữ liệu được sử dụng để huấn luyện mô hình, trong khi một phần nhỏ dành cho việc đánh giá mô hình trên dữ liệu không nhìn thấy trước đó.

Sau khi dữ liệu đã được chia thành các tập huấn luyện và kiểm tra, mô hình hồi quy tuyến tính, SVM, Random forest, Decision Tree, được huấn luyện trên tập huấn luyện và sau đó được sử dụng để dự đoán điểm số cho tập kiểm tra. Kết quả của việc dự đoán này được sử dụng để đánh giá hiệu suất của mô hình trên dữ liệu mới, không nhìn thấy trước đó, thông qua các độ đo như độ chính xác.

Chương 4
TỐI ƯU HÓA MÔ HÌNH KỸ THUẬT HỌC MÁY

4.1. Tính tương quan của các điểm

Trong trường hợp có từ 3 đến 4 biến định lượng trở lên, thay vì biểu diễn trực quan tất cả dữ liệu của các biến định lượng lên biểu đồ, ta có thể tính toán hệ số tương quan Pearson giữa từng cặp biến và trực quan số liệu này lên biểu đồ. Phương pháp này được gọi là trực quan bằng biểu đồ tương quan Correlogram. Biểu đồ sử dụng một dải màu để chỉ ra các giá trị hệ số tương quan trong khoảng từ [-1;1].[1]



Hình 4. 1: Trực quan hóa dữ liệu

Mỗi ô trong ma trận tương quan đại diện cho hệ số tương quan giữa hai môn học tương ứng. Hệ số tương quan là một giá trị trong khoảng từ -1 đến 1, thể hiện mức độ và hướng của mối liên hệ giữa hai môn học. Một hệ số tương quan gần với 1 cho thấy một mối liên hệ mạnh mẽ và tích cực: khi điểm số của một môn học tăng, điểm số của môn học khác cũng tăng. Một hệ số tương quan gần với -1 cho thấy một mối liên hệ mạnh mẽ và tiêu cực: khi điểm số của một môn học tăng, điểm số của môn học khác giảm. Một hệ số tương quan gần với 0 cho thấy rằng không có mối liên hệ rõ ràng giữa hai môn học.

Màu sắc của mỗi ô trong ma trận tương quan cũng thể hiện mức độ và hướng của mối tương quan. Màu đỏ càng đậm cho thấy mối tương quan mạnh mẽ(ví dụ điểm toán cả năm lớp 10), màu càng nhạt cho thấy mối tương quan không đáng kể(ví dụ điểm hóa kì 1 lớp 10).

Kết Luận: Dựa vào độ tương quan của dữ liệu và tính thực tế là năm lớp 12 các học sinh sẽ chia ra thành 2 khối là tự nhiên và xã hội để xét tốt nghiệp và thi đại học nên em sẽ chia dữ liệu thành 2 phần tương ứng:

Dự đoán điểm trung bình của năm lớp 12 dựa vào điểm các môn thuộc khối tự nhiên(Toán, Văn, Anh, Vật lí, Hóa Học, Sinh Học)(gồm 1555 dữ liệu)

Giới tính	10.Toán CN	10.Văn CN	10.Vật lí CN	10.Hóa học CN	10.Sinh học CN	10.Ngoại ngữ CN	11.Toán CN	11.Văn CN	11.Vật lí CN	11.Hóa học CN	11.Sinh học CN	11.Ngoại ngữ CN
Nam	9.5	7.1	8.9	8.9	8.6	7.5	9.7	7.7	9.4	9.6	9.1	7.6
Nam	6.4	7.1	5.7	7	6.4	7	7.9	7.3	7.1	6.2	7.1	7.4
Nam	9.2	7.9	8.2	8.7	8.9	7.9	8.9	8	8.4	8.8	8.3	8.8
Nam	8.2	6.6	8.9	8	6.7	7.8	8.5	6.9	8.3	6.9	7.7	8.6
Nam	6.8	7.3	9.3	7.3	7.6	9.1	7.5	7.4	6.6	6.5	7.9	9.3
Nam	8.7	8.3	6.7	7.5	8.4	7.9	8.4	8.5	6.8	7.5	8.3	8.7
Nam	7.6	6.3	6.1	6.1	6.9	5.9	8.3	6.5	8.4	8.2	7.6	6.2
Nam	8	6.7	7.7	7.7	7.8	6.9	8.1	7	6.8	7.5	7.9	7.6
Nữ	6	7.1	6	5.9	7.1	6.4	5.6	7.4	5.9	5.3	7.7	7.7
Nữ	9.1	8.2	9.4	8.1	9.3	9.3	9.5	8	9.1	7.9	9.1	9.3
Nữ	7.3	7.1	7.4	6.1	6.7	6.2	7.1	7.8	5.5	7.6	8	8.5
Nữ	7.7	7.5	7.5	7.6	7.4	8.3	8.9	7.4	8	8.1	7.9	8.2
Nữ	8.8	8.3	7.3	7.4	8.2	7.5	8.9	8.5	7.6	7	8.1	7.7
Nữ	8.7	6.8	8.8	6.7	7.5	5.9	9	8.1	9.1	8.3	8.5	7.6
Nữ	7	7.2	7.5	7.6	6	7.9	8.1	7.1	5.3	6.3	7.2	8.4
Nữ	9.2	8.1	8.9	8.8	8.1	8.1	9.5	8.3	8.6	9.3	8.5	8.5
Nữ	8.5	8.3	7.9	8	7.5	9	8.3	7.9	8.9	8.4	8.2	8
Nữ	7	7.5	8.3	6.2	8	8.3	8.2	6.8	8.9	7.4	8.6	8.3
Nữ	9.7	7.8	8.4	8.4	8.5	9.3	9.3	8.4	8.5	7.7	9	9.9
Nam	8.5	6.8	8.8	8.5	8	8	8.2	7.3	8.5	8.4	7.8	7.5
Nữ	7.7	8	7.8	8.9	8.2	8	8.4	7.7	7.2	7.1	8.4	8.2
Nam	9	7.4	8.2	7.7	7.9	7.8	9.1	7.5	8.9	8.4	7.7	8.6
Nam	6.8	6.7	7.4	7.2	7.4	7.6	8.3	7.4	8.2	8.2	7.4	6.9
Nữ	8.4	8.2	7.6	6.6	7.4	6	8.2	8.6	6.7	6.1	7.2	7.5
Nữ	5.8	6.6	7	5.1	7.2	6.3	5.9	7.2	6.1	5.9	7.2	8.2
Nữ	7.6	7.2	8	6	8	6.2	7.6	8.5	8	7.3	8.3	7
Nữ	7.5	6.8	8.5	7.4	8.2	8.3	8.7	7.1	8.5	8.6	9.1	8.1
Nam	8.4	7.7	8.7	8.5	8.8	7.8	8.9	8.4	8.1	8.5	8.6	8.1

Hình 4. 2: Dữ liệu khối tự nhiên

Dự đoán điểm trung bình của năm lớp 12 dựa vào điểm các môn thuộc khối xã hội(Toán, Văn, Anh, Sử, Địa lí, GDCD)(gồm 1555 dữ liệu)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Giới tính	10.Toán C	10.Văn CN	10.GDCD CN	10.Lịch sử CN	10.Địa lí CN	10.Ngoại r	11.Toán C	11.Văn CN	11.Lịch sử CN	11.Địa lí CN	11.GDCD CN	11.Ngoại ngữ CN			
2	Nam	9.5	7.1	9.1	8.6	7.6	7.5	9.7	7.7	8.9	8.9	9.1	7.6			
3	Nam	6.4	7.1	7.7	6.7	7.4	7	7.9	7.3	7.7	6.9	7.5	7.4			
4	Nam	9.2	7.9	8.9	8.3	8.4	7.9	8.9	8	8.6	8.4	8.6	8.8			
5	Nam	8.2	6.6	7.4	6.9	7.6	7.8	8.5	6.9	7.6	8	7.8	8.6			
6	Nam	6.8	7.3	8.6	8.5	7.3	9.1	7.5	7.4	8.4	7.7	8.5	9.3			
7	Nam	8.7	8.3	8.9	8	7.8	7.9	8.4	8.5	8.8	8.4	8.7	8.7			
8	Nam	7.6	6.3	6.6	6.8	7	5.9	8.3	6.5	6.2	7.5	8	6.2			
9	Nam	8	6.7	8.4	7.2	7.5	6.9	8.1	7	7.5	7.7	7.2	7.6			
10	Nữ	6	7.1	7.1	7.6	7.4	6.4	5.6	7.4	7.4	7.3	7.4	7.7			
11	Nữ	9.1	8.2	9	8.7	8.2	9.3	9.5	8	9.3	8.7	9.2	9.3			
12	Nữ	7.3	7.1	7.6	6.9	7.4	6.2	7.1	7.8	6.4	6.8	7	8.5			
13	Nữ	7.7	7.5	7.6	7.7	7.2	8.3	8.9	7.4	7.1	8.1	8.2	8.2			
14	Nữ	8.8	8.3	8	7.5	8.1	7.5	8.9	8.5	7.6	8.7	7.9	7.7			
15	Nữ	8.7	6.8	8.5	8.4	8.2	5.9	9	8.1	7.9	8.7	8.5	7.6			
16	Nữ	7	7.2	7.9	7.4	7.2	7.9	8.1	7.1	7.1	6.6	7.3	8.4			
17	Nữ	9.2	8.1	8.7	8.1	8.6	8.1	9.5	8.3	8.9	9.5	8.7	8.5			
18	Nữ	8.5	8.3	7.9	8	8.9	9	8.3	7.9	7.6	8.6	8.1	8			
19	Nữ	7	7.5	7.1	7.6	7.8	8.3	8.2	6.8	8.2	8	7.9	8.3			
20	Nữ	9.7	7.8	8.5	8.3	8.5	9.3	9.3	8.4	8.5	8.9	8.7	9.9			
21	Nam	8.5	6.8	8.6	7.9	8	8	8.2	7.3	8.3	7.6	8.6	7.5			
22	Nữ	7.7	8	8.3	8.1	8.5	8	8.4	7.7	8.8	8	8.4	8.2			
23	Nam	9	7.4	8.2	7.2	8	7.8	9.1	7.5	8.2	8.7	8.5	8.6			
24	Nam	6.8	6.7	7.8	7.1	7.3	7.6	8.3	7.4	6.9	7.7	7.2	6.9			
25	Nữ	8.4	8.2	7.8	7.6	7.5	6	8.2	8.6	6.9	9.3	7.6	7.5			
26	Nữ	5.8	6.6	7.3	7.8	7	6.3	5.9	7.2	7.4	6.6	6.7	8.2			
27	Nữ	7.6	7.2	7.8	7.6	7.6	6.2	7.6	8.5	8.3	8.6	9	7			
28	Nữ	7.5	6.8	8	8.3	6.7	8.3	8.7	7.1	8.1	8.3	8.4	8.1			
29	Nam	8.4	7.7	8.3	8.6	8.4	7.8	8.9	8.4	8.5	8.4	8.9	8.1			

Hình 4. 3: Dữ liệu khối xã hội

4.2. Chạy các mô hình học máy hồi quy

Các mô hình đều được chạy trên colab với 2 tập biến là:

Biến độc lập:

```
X = data[['10.Toán CN', '10.Văn CN', '10.Vật lí CN', '10.Hóa học CN', '10.Sinh học CN', '10.Ngoại ngữ CN', '11.Toán CN', '11.Văn CN', '11.Vật lí CN', '11.Hóa học CN', '11.Sinh học CN', '11.Ngoại ngữ CN', '10.Lịch sử CN', '10.Địa lí CN', '10.GDCD CN', '11.Lịch sử CN', '11.Địa lí CN', '11.GDCD CN']]
```

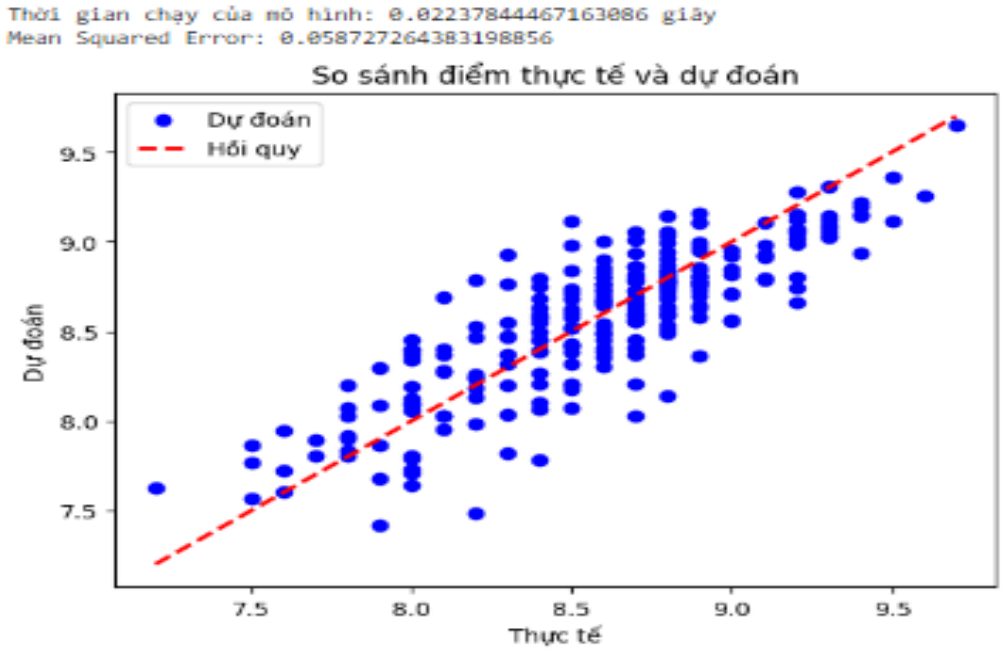
Biến phụ thuộc:

```
y = data['12.Điểm tổng kết CN']
```

Phân chia dữ liệu huấn luyện theo tỉ lệ 20%- 80%

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

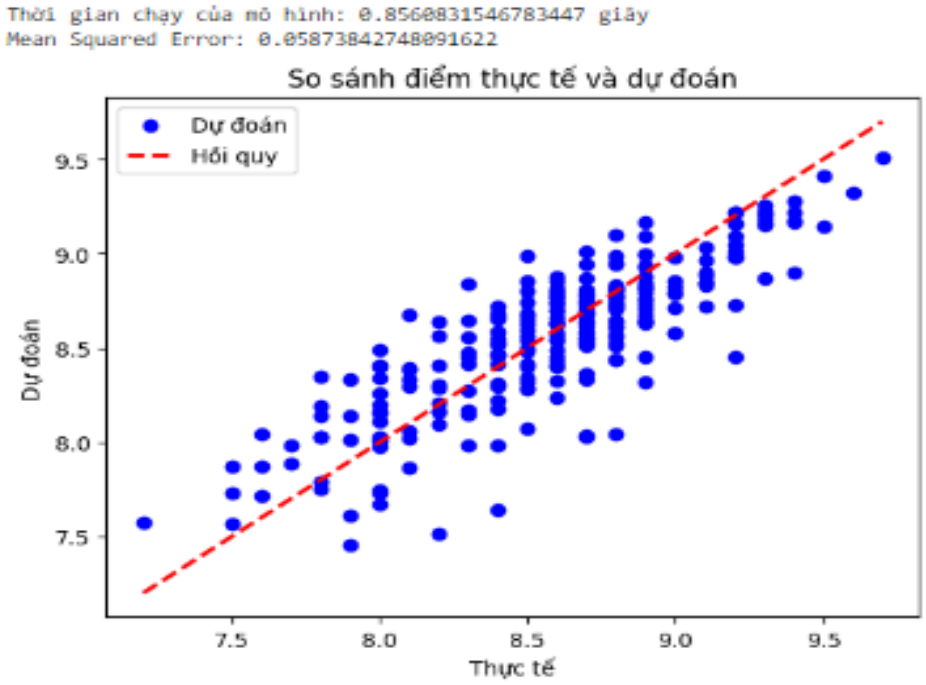
4.2.1. Mô hình Linear Regression



Hình 4. 4: Kết quả mô hình Linear Regression

Như vậy MSE của mô hình là 0,06 và thời gian chạy là 0,022 giây

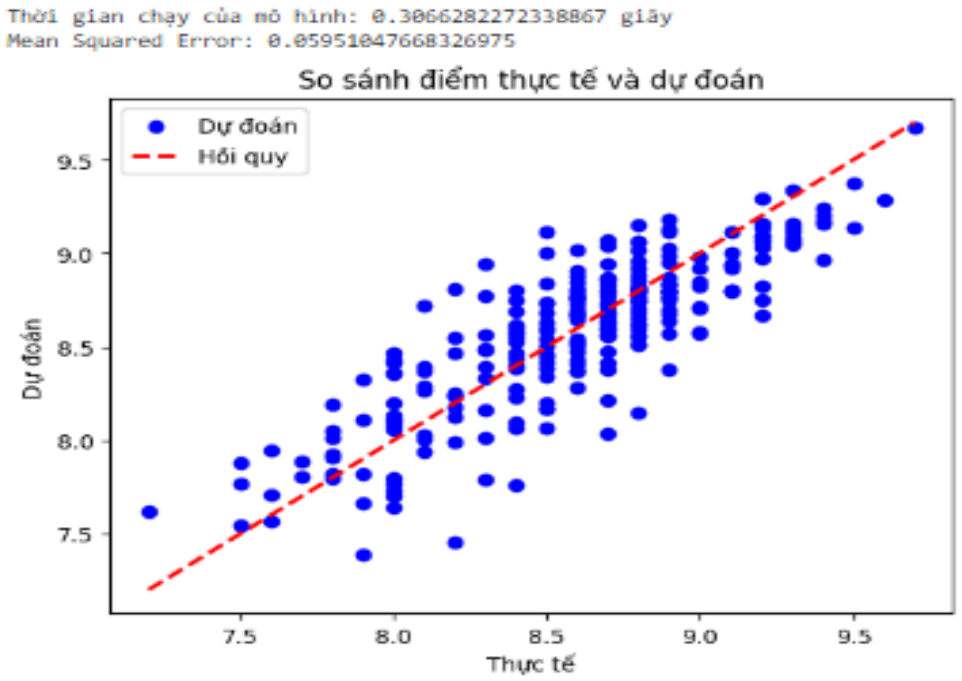
4.2.2. Mô hình Random Forest



Hình 4. 5: Kết quả mô hình Random Forest

Như vậy MSE của mô hình là 0,06 và thời gian chạy là 0,856 giây

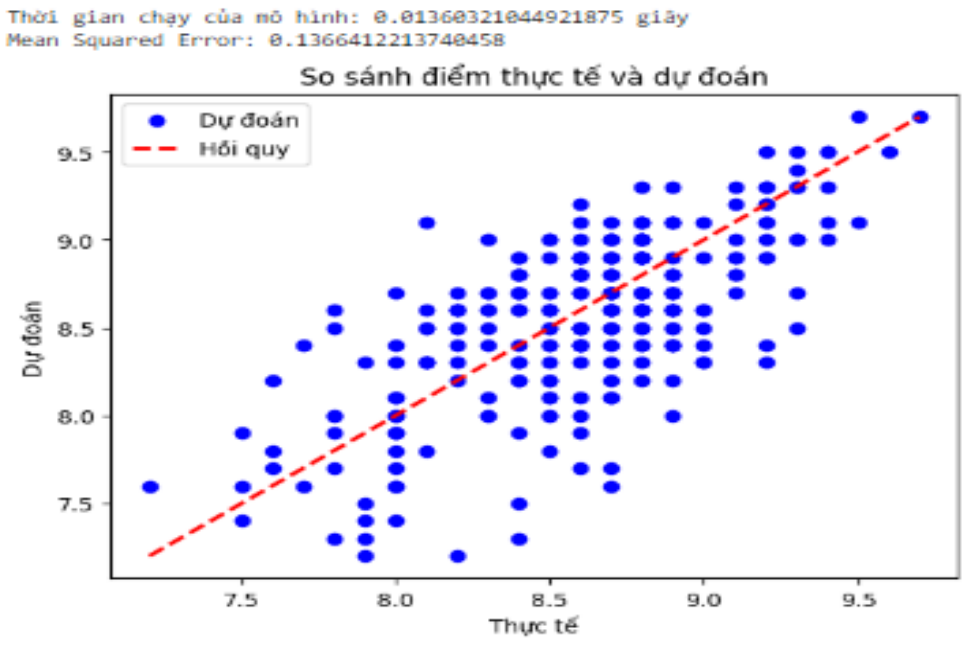
4.2.3. Mô hình SVM



Hình 4. 6: Kết quả mô hình SVM

Như vậy MSE của mô hình là 0,06 và thời gian chạy là 0,306 giây

4.2.4. Mô hình Decision Trees



Hình 4. 7: Kết quả mô hình Decision Trees

Như vậy MSE của mô hình là 0,136 và thời gian chạy là 0,0136 giây

4.3. So sánh 4 mô hình

Cả bốn mô hình Linear Regression, Random Forest, Decision Trees và SVM đều cho kết quả tương đối tốt trong việc dự đoán điểm trung bình lớp 12 từ điểm lớp 10 và lớp 11. Sai số của các mô hình:

Linear Regression: 0,06 và 0,022 giây thời gian chạy.

Random Forest: 0,06 và 0,85 giây thời gian chạy.

SVM: 0,06 và 0,4 giây thời gian chạy.

Decision Trees: 0,136 và 0,0136 giây thời gian chạy.

Như vậy:

Linear Regression và Decision Trees có thời gian chạy đáng kể ngắn hơn so với Random Forest và SVM. Thời gian chạy của Linear Regression là 0.022 giây và Decision Trees là 0,0136 giây trong khi Random Forest mất 0,85 giây, SVM mất 0,4 giây. Nhưng về sai số thì Decision Trees lại lớn hơn hẳn so với các mô hình còn lại; vì vậy cho thấy Linear Regression có hiệu suất tính toán tốt và thời gian chạy trong trường hợp này.

Lựa chọn mô hình:

Dựa trên kết quả độ chính xác và thời gian chạy, ta sẽ tiến hành tối ưu 2 mô hình là Linear Regression và Decision Trees để đưa ra lựa chọn tốt nhất.

4.4. Mô hình tối ưu tham số

4.4.1. Tối ưu bằng Grid Search

```
Thời gian chạy của mô hình: 0.1174917221069336 giây
Best parameters found:  {'regressor__fit_intercept': True}
Mean Squared Error: 0.06
Mean Absolute Error: 0.19
```

Kết quả mô hình Linear Regression

	MSE	MAE
Linear Regression	0.06	0.19

Hình 4. 8: Mô hình tối ưu hóa tham số GridSearch của Linear Regression

```
Thời gian chạy của mô hình: 4.7449305057525635 giây
Best parameters found:  {'regressor__max_depth': 10, 'regressor__min_samples_leaf': 4, 'regressor__min_samples_split': 10}
Mean Squared Error: 0.1
Mean Absolute Error: 0.25
```

Kết quả mô hình Decision Tree Regression

	MSE	MAE
Decision Tree Regression	0.1	0.25

Hình 4. 9: Mô hình tối ưu hóa tham số GridSearch của Decision Tree

Bảng 4. 1: Bảng so sánh 2 mô hình tối ưu bằng GridSearch

Mô hình	**Linear Regression**	**Decision Tree**
Thời gian chạy	0,117 giây	4,744 giây
MSE	0,06	0,1
MAE	0,19	0,25

- Thời gian chạy của mô hình:

Đây là thời gian (tính bằng giây) mà mô hình cần để hoàn thành việc huấn luyện bằng cách sử dụng GridSearchCV. Thời gian này bao gồm việc thử nghiệm tất cả các kết hợp siêu tham số và huấn luyện mô hình với từng kết hợp để tìm ra bộ siêu tham số tốt nhất.

- Best parameters found:

Đây là bộ siêu tham số tối ưu mà GridSearchCV tìm ra được.

- Mean Squared Error (MSE):

MSE là giá trị trung bình của bình phương sai số giữa giá trị dự đoán và giá trị thực tế. Giá trị này càng nhỏ thì mô hình càng chính xác.

- Mean Absolute Error (MAE):

MAE là giá trị trung bình của sai số tuyệt đối giữa giá trị dự đoán và giá trị thực tế. MAE cũng được sử dụng để đánh giá độ chính xác của mô hình, với giá trị càng nhỏ càng tốt.

4.4.2. Tối ưu bằng Random Search

```
Thời gian chạy của mô hình: 0.12218655212402344 giây
Best parameters found:  {'regressor__fit_intercept': True}
Mean Squared Error: 0.06
Mean Absolute Error: 0.19
```

Kết quả mô hình Linear Regression

	MSE	MAE
Linear Regression	0.06	0.19

Hình 4. 10: Mô hình tối ưu hóa tham số RandomSearch của Linear Regression

```
Thời gian chạy của mô hình: 0.8413636684417725 giây
Best parameters found:  {'regressor__min_samples_split': 2, 'regressor__min_samples_leaf': 4, 'regressor__max_features': 'log2', 'regressor__max_depth': 50}
Mean Squared Error: 0.09
Mean Absolute Error: 0.24
```

Kết quả mô hình Decision Tree Regression

	MSE	MAE
Decision Tree Regression	0.09	0.24

Hình 4. 11: Mô hình tối ưu hóa tham số RandomSearch của Decision Tree

Bảng 4. 2: Bảng so sánh 2 mô hình tối ưu bằng RandomSearch

Mô hình	**Linear Regression**	**Decision Tree**
Thời gian chạy	0,122 giây	0,841 giây
MSE	0,06	0,09
MAE	0,19	0,24

- Thời gian chạy của mô hình:

Đây là thời gian (tính bằng giây) mà mô hình cần để hoàn thành việc huấn luyện bằng cách sử dụng RandomSearchCV. Thời gian này bao gồm việc thử nghiệm tất cả

các kết hợp siêu tham số và huấn luyện mô hình với từng kết hợp để tìm ra bộ siêu tham số tốt nhất.

- Best parameters found:

Đây là bộ siêu tham số tối ưu mà RandomSearchCV tìm ra được.

- Mean Squared Error (MSE):

MSE là giá trị trung bình của bình phương sai số giữa giá trị dự đoán và giá trị thực tế. Giá trị này càng nhỏ thì mô hình càng chính xác.

- Mean Absolute Error (MAE):

MAE là giá trị trung bình của sai số tuyệt đối giữa giá trị dự đoán và giá trị thực tế. MAE cũng được sử dụng để đánh giá độ chính xác của mô hình, với giá trị càng nhỏ càng tốt.

Kết luận: Mô hình Linear Regression được tối ưu bằng GridSearch(0,117 giây và MSE = 0,06) là kết quả tốt nhất; vì vậy em lựa chọn mô hình Linear Regression được tối ưu tham số bằng GridSearch để thiết kế ứng dụng dự đoán điểm.

Chương 5
CHẠY MÔ HÌNH VÀ ĐÁNH GIÁ KẾT QUẢ

5.1. Chạy mô hình dựa vào mô hình tốt nhất

5.1.1. Ứng dụng mô hình Linear Regression

```
import time
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
import numpy as np

# Đường dẫn đến tệp Excel trên Google Drive của bạn
file_path = '/content/drive/MyDrive/điểm HB.xlsx'

# Đọc dữ liệu từ tệp Excel
data = pd.read_excel(file_path)

# Loại bỏ các dòng chứa giá trị NaN
data.dropna(inplace=True)

# Xác định biến độc lập và phụ thuộc
X = data[['10.Toán CN', '10.Văn CN', '10.Vật lí CN', '10.Hóa học CN', '10.Sinh học CN', '10.Ngoại ngữ CN',
          '11.Toán CN', '11.Văn CN', '11.Vật lí CN', '11.Hóa học CN', '11.Sinh học CN', '11.Ngoại ngữ CN',
          '10.Lịch sử CN', '10.Địa lí CN', '10.GDCD CN', '11.Lịch sử CN', '11.Địa lí CN', '11.GDCD CN']]
y = data['12.Điểm tổng kết CN']

# Phân chia dữ liệu thành tập huấn luyện và tập kiểm tra
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Tạo pipeline kết hợp chuẩn hóa và mô hình hồi quy tuyến tính
pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('regressor', LinearRegression())
])

# Định nghĩa các siêu tham số cần tìm kiếm
param_grid = {
    'regressor__fit_intercept': [True, False]
}

# Thực hiện Grid Search
grid_search = GridSearchCV(pipeline, param_grid, cv=5, scoring='neg_mean_squared_error')
start_time = time.time()
grid_search.fit(X_train, y_train)
end_time = time.time()
```

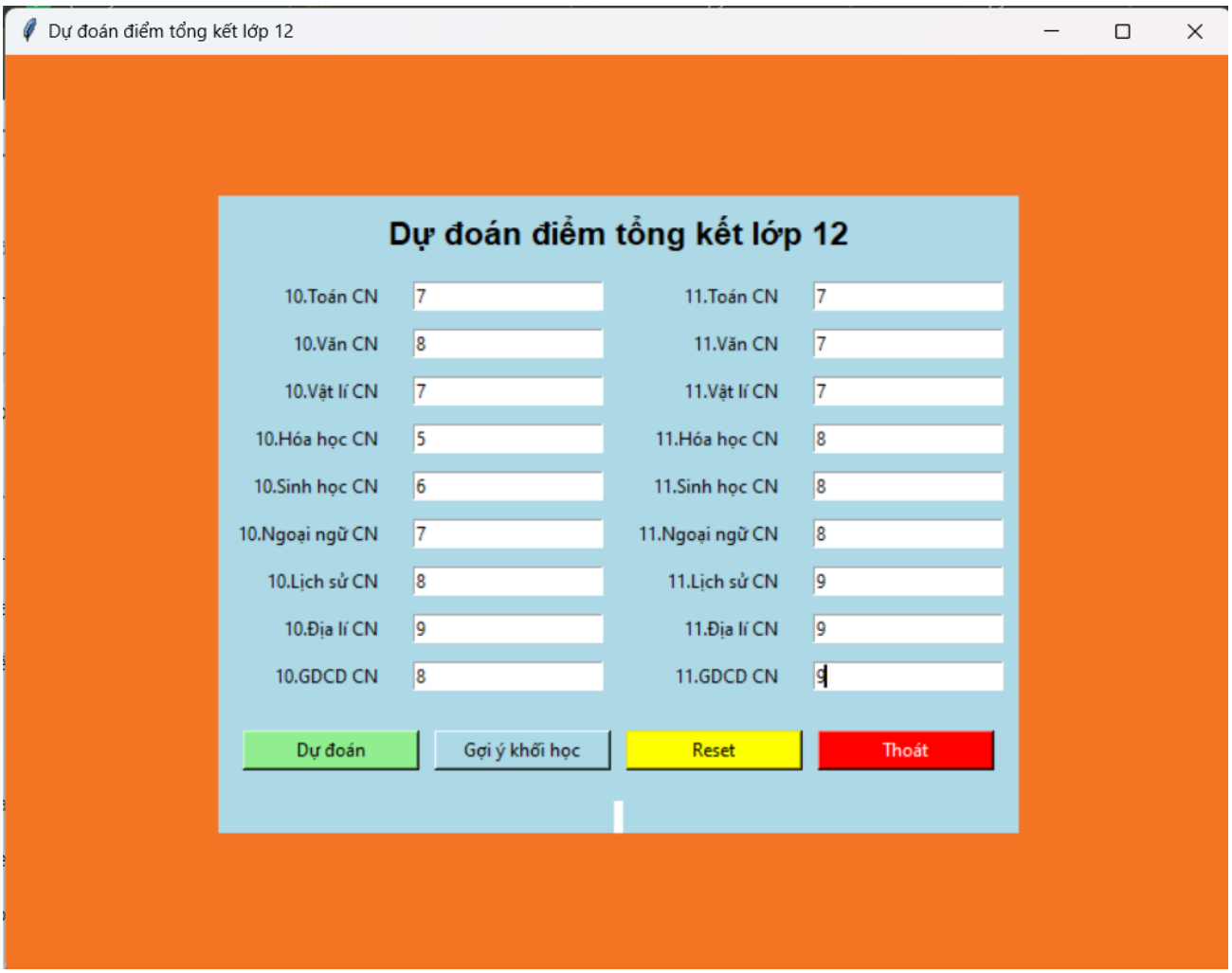
```
# In ra thời gian chạy của mô hình
execution_time = end_time - start_time
print("Thời gian chạy của mô hình:", execution_time, "giây")

# In ra các siêu tham số tốt nhất
print("Best parameters found: ", grid_search.best_params_)

# Dự đoán điểm lớp 12 cho tập kiểm tra
y_pred = grid_search.predict(X_test)

# Đánh giá mô hình bằng các chỉ số phù hợp với hồi quy
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)

# Làm tròn các kết quả đến 2 chữ số thập phân
mse = round(mse, 2)
mae = round(mae, 2)

print("Mean Squared Error:", mse)
print("Mean Absolute Error:", mae)

# Tạo DataFrame để hiển thị các chỉ số
results_df = pd.DataFrame({
    'MSE': [mse],
    'MAE': [mae]
}, index=['Linear Regression'])

# Hiển thị bảng kết quả
fig, ax = plt.subplots(figsize=(6, 1.5))  # Tạo figure và axis cho bảng
ax.axis('tight')
ax.axis('off')
table = ax.table(cellText=results_df.values, colLabels=results_df.columns, rowLabels=results_df.index, cellLoc='center', loc='center')

# Tùy chỉnh hiển thị bảng
table.auto_set_font_size(False)
table.set_fontsize(12)
table.scale(1.2, 1.2)

plt.title("Kết quả mô hình Linear Regression", fontsize=14)
plt.show()
```

Hình 5. 1: Ứng dụng mô hình Linear Regression

5.1.2. Ứng dụng dự đoán điểm trung bình lớp 12 dựa vào điểm các môn lớp 10, 11

Sau khi sử dụng mô hình Linear Regression với tỉ lệ chính xác cao với các mô hình còn lại tiến hành xây dựng ứng dụng dự đoán điểm lớp 12 dựa vào điểm các môn lớp 10,11.

Sử dụng thư viện Tkinker để xây dựng ứng dụng dự đoán điểm lớp 12 dựa vào điểm các môn lớp 10,11. Ngoài ra dựa vào điểm mà bạn nhập sẽ được ứng dụng xử lý để đưa ra lựa chọn và khối mà bạn nên lựa chọn khi vào lớp 12.



Hình 5. 2: Ứng dụng dự đoán điểm

5.2. Đánh giá kết quả chạy

Thời gian chạy:

Thời gian chạy mô hình khá nhanh, chỉ 0.117 giây, cho thấy mô hình có tốc độ xử lý dữ liệu tương đối tốt.

Các thông số tối ưu hóa:

Mô hình đã tìm được các tham số tối ưu, trong đó "regressor__fit_intercept" được xác định là True. Điều này cho thấy mô hình sẽ tính toán cả hệ số (intercept) trong phương trình hồi quy.

Chỉ số độ chính xác:

Mean Squared Error (MSE): 0.06 - Giá trị này khá thấp, cho thấy mô hình dự báo khá chính xác.

Mean Absolute Error (MAE): 0.19 - Giá trị MAE thấp, có nghĩa là sai số tuyệt đối trung bình chỉ khoảng 0.19 điểm.

Tổng hợp lại, với thời gian chạy nhanh, các thông số tối ưu hóa hợp lý, và các chỉ số độ chính xác như MSE, MAE ở mức khá tốt, ta có thể kết luận rằng mô hình hồi quy đã được xây dựng và tối ưu hóa khá hiệu quả, có thể dự báo điểm vào đồ án tốt nghiệp với độ tin cậy khá cao.

KẾT LUẬN

Ưu Điểm:

Tiềm năng cải thiện độ chính xác và hiệu suất: Mặc dù mô hình chưa được triển khai trên quy mô lớn, nhưng việc tiếp tục cải thiện và optimizing mô hình trên các tập dữ liệu lớn hơn và phức tạp hơn sẽ giúp nâng cao độ chính xác và hiệu suất của mô hình, đáp ứng tốt hơn các yêu cầu thực tế.

Khả năng tổng quát hóa: Mở rộng và đa dạng hóa tập dữ liệu huấn luyện sẽ giúp cải thiện khả năng tổng quát hóa của mô hình, tăng cường khả năng áp dụng rộng rãi cho các bối cảnh khác nhau.

Giá trị ứng dụng: Việc nghiên cứu và cải thiện các mô hình dự đoán điểm trung bình lớp 12 có vai trò quan trọng trong việc cung cấp các công cụ hỗ trợ quan trọng cho việc đánh giá và dự đoán hiệu suất học tập của học sinh. Điều này sẽ góp phần cải thiện chất lượng giáo dục và hỗ trợ tốt hơn cho học sinh.

Tính sáng tạo và tiềm năng phát triển: Mặc dù mô hình chưa được triển khai rộng rãi, nhưng việc nghiên cứu và phát triển các mô hình dự đoán như vậy cho thấy sự sáng tạo và tiềm năng của nhóm nghiên cứu. Đây là nền tảng tốt để tiếp tục cải thiện và mở rộng ứng dụng trong tương lai.

Nhược Điểm:

Dự án chưa được triển khai trên quy mô lớn hoặc thực tế, do đó, độ chính xác và hiệu suất của mô hình có thể cần được cải thiện để đáp ứng yêu cầu thực tế.

Tập dữ liệu huấn luyện có thể cần được mở rộng và đa dạng hơn để cải thiện khả năng tổng quát hóa của mô hình.

Việc nghiên cứu và cải thiện các mô hình dự đoán điểm trung bình lớp 12 sẽ đóng vai trò quan trọng trong việc cung cấp các công cụ hỗ trợ quan trọng cho việc đánh giá và dự đoán hiệu suất học tập của học sinh.

Phương hướng phát triển:

Mở rộng tập dữ liệu:

Tăng cường tập dữ liệu huấn luyện bằng cách thu thập thêm dữ liệu từ các nguồn khác nhau hoặc mở rộng tập dữ liệu hiện có để bao gồm các năm học mới và các địa điểm khác nhau. Điều này sẽ giúp cải thiện khả năng tổng quát hóa của mô hình và làm cho nó phù hợp hơn với các tình huống thực tế.

Đánh giá thêm các yếu tố khác:

Xem xét thêm các yếu tố khác ngoài điểm lớp 10 và 11, chẳng hạn như sự nghiệp học tập trước đó, hoạt động ngoại khóa, hoặc các yếu tố gia đình và xã hội để cải thiện dự đoán hiệu suất học tập.

Triển khai hệ thống:

Nếu có khả năng, triển khai hệ thống dự đoán vào môi trường thực tế để đánh giá hiệu suất và ứng dụng thực tế của nó. Làm trang web giúp dự đoán điểm trung bình lớp 12 dựa vào điểm lớp 10, 11.

TÀI LIỆU THAM KHẢO

Tiếng Việt:

[1]. Giáo trình FIT 4103 (book)_Machine Learning Vũ Hữu Tiệp.pdf

[2]. *Nguyễn Văn Tuân*, Mô hình hồi qui và khám phá khoa học, Nhà xuất bản TPHCM, 2022.

Danh mục các Website tham khảo:

[1]. https://luatminhkhue.vn/bieu-do-tuong-quan-la-gi.aspx

[2]. https://aws.amazon.com/vi/what-is/linear-regression/

[3]. https://machinelearningcoban.com/2017/04/09/smv/

[4]. https://scikit-learn.org/stable/modules/tree.html

[5]. https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html