BỘ GIÁO DỤC ĐÀO TẠO

TRƯỜNG ĐẠI HỌC ĐẠI NAM

--- 🕮 ---

ĐỒ ÁN TỐT NGHIỆP

XÂY DỰNG ỨNG DỤNG TỰ ĐỘNG GỢI Ý LỰA CHỌN CÁC HỌC PHẦN CHUYÊN NGÀNH DỰA VÀO KẾT QUẢ CÁC HỌC PHẦN CƠ SỞ

SINH VIÊN THỰC HIỆN : VŨ DUY KHƯƠNG

MÃ SINH VIÊN : 1451020130

KHOA : CÔNG NGHỆ THÔNG TIN

HÀ NỘI - 2024

BỘ GIÁO DỤC ĐÀO TẠO

TRƯỜNG ĐẠI HỌC ĐẠI NAM

VŨ DUY KHƯƠNG

XÂY DỰNG ỨNG DỤNG TỰ ĐỘNG GỢI Ý LỰA CHỌN CÁC HỌC PHẦN CHUYÊN NGÀNH DỰA VÀO KẾT QUẢ CÁC HỌC PHẦN CƠ SỞ

CHUYÊN NGÀNH : **CÔNG NGHỆ THÔNG TIN**

MÃ SỐ : **74.80.201**

NGƯỜI HƯỚNG DẪN: ThS. PHẠM THỊ TỐ NGA

HÀ NỘI - 2024

LỜI CAM ĐOAN

Em xin cam đoan đề tài: "Xây dựng ứng dụng tự động gợi ý lựa chọn các học phần chuyên ngành dựa vào kết quả các học phần cơ sở" là một công trình nghiên cứu độc lập dưới sự hướng dẫn của giáo viên hướng dẫn: Phạm Thị Tố Nga. Ngoài ra không có bất cứ sự sao chép của người khác. Đề tài, nội dung báo cáo tốt nghiệp là sản phẩm mà em đã nỗ lực nghiên cứu trong quá trình học tập tại trường cũng như được học hỏi các thầy cô, bạn bè. Các số liệu, kết quả trình bày trong báo cáo là hoàn toàn trung thực, em xin chịu hoàn toàn trách nhiệm, kỷ luật của bộ môn và nhà trường đề ra nếu như có vấn đề xảy ra.

Sinh viên thực hiện

Ký tên

LỜI NÓI ĐẦU

Trong bối cảnh phát triển nhanh chóng của công nghệ thông tin và sự phức tạp ngày càng tăng của các chương trình đào tạo đại học, việc lựa chọn các học phần chuyên ngành trở nên một thách thức không nhỏ đối với sinh viên. Việc lựa chọn đúng đắn không chỉ giúp sinh viên tiếp cận kiến thức một cách hệ thống và hiệu quả mà còn tạo nền tảng vững chắc cho sự phát triển nghề nghiệp sau này. Tuy nhiên, với sự đa dạng và phong phú của các học phần, sinh viên thường gặp khó khăn trong việc đưa ra quyết định tối ưu cho lộ trình học tập của mình.

Xuất phát từ thực tế đó, đề tài "Xây dựng ứng dụng gợi ý lựa chọn các học phần chuyên ngành dựa vào kết quả các học phần cơ sở" ra đời với mục tiêu hỗ trợ sinh viên trong quá trình lựa chọn học phần. Ứng dụng này không chỉ dựa vào kết quả học tập của các học phần cơ sở mà còn xem xét sự phù hợp với năng lực và sở thích cá nhân của từng sinh viên, từ đó đưa ra những gợi ý hợp lý và khoa học.

Bằng cách ứng dụng các thuật toán phân tích dữ liệu và trí tuệ nhân tạo, em mong muốn mang đến một công cụ hữu ích, giúp sinh viên tiết kiệm thời gian, công sức trong việc lựa chọn các học phần, đồng thời tối ưu hóa lộ trình học tập và nâng cao hiệu quả học tập. Đề tài này không chỉ góp phần cải thiện chất lượng giáo dục mà còn mở ra nhiều hướng nghiên cứu và ứng dụng mới trong lĩnh vực tư vấn học tập và quản lý giáo dục.

Em hy vọng rằng, với sự nỗ lực và tâm huyết của nhóm nghiên cứu, đề tài này sẽ mang lại những giá trị thiết thực, hỗ trợ hiệu quả cho sinh viên trong quá trình học tập và phát triển bản thân. Mong rằng ứng dụng sẽ trở thành người bạn đồng hành tin cậy của sinh viên trên con đường chinh phục tri thức và đạt được những thành công trong sự nghiệp.

LỜI CẢM ƠN

Để hoàn thành đề tài "Xây dựng ứng dụng gợi ý lựa chọn các học phần chuyên ngành dựa vào kết quả các học phần cơ sở", em đã nhận được sự hỗ trợ và giúp đỡ quý báu từ nhiều cá nhân và tập thể. Nhân dịp này, em xin bày tỏ lòng biết ơn chân thành đến tất cả những ai đã đồng hành và góp phần vào thành công của đề tài này.

Trước hết, em xin gửi lời cảm ơn sâu sắc đến cô Phạm Thị Tố Nga, giảng viên hướng dẫn, người đã tận tình chỉ bảo, hướng dẫn em trong suốt quá trình thực hiện đề tài. Sự nhiệt tình, kiến thức sâu rộng và những góp ý quý báu của cô đã giúp em vượt qua nhiều khó khăn và hoàn thiện công trình nghiên cứu này.

Em cũng xin cảm ơn nhóm nghiên cứu học máy ngành Khoa học Máy tính thuộc Khoa Công nghệ Thông tin. Sự hỗ trợ từ các thành viên trong nhóm đã cung cấp cho em những kiến thức cần thiết và những công cụ quan trọng để phát triển ứng dụng một cách hiệu quả.

Đặc biệt, em xin gửi lời tri ân đến cô Trần Thị Thanh Nhàn vì những đóng góp và sự hỗ trợ tận tình trong việc tư vấn, cung cấp tài liệu và chia sẻ kinh nghiệm quý báu trong lĩnh vực nghiên cứu.

Cuối cùng, em xin cảm ơn tập thể Khoa Công nghệ Thông tin, nơi đã tạo điều kiện thuận lợi về cơ sở vật chất cũng như môi trường học tập và nghiên cứu, giúp em có thể tập trung toàn tâm toàn lực vào việc thực hiện đề tài.

Một lần nữa, em xin chân thành cảm ơn tất cả mọi người đã giúp đỡ và hỗ trợ em trong suốt quá trình nghiên cứu. em hy vọng rằng, với những kiến thức và kinh nghiệm thu được, em có thể tiếp tục đóng góp cho sự phát triển của lĩnh vực công nghệ thông tin và giáo dục.

NHẬN XÉT

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

DANH MỤC KÝ HIỆU HOẶC CHỮ VIẾT TẮT

STT	TỪ VIẾT TẮT	VIẾT ĐẦY ĐỦ
1	BT1	Bài toán 1
2	BT2	Bài toán 2
3	CNTT	Công nghệ thông tin
4	CNPM	Chuyên ngành phần mềm
5	DC	Decision Tree
6	HT10	Hạng thang 10
7	HTNIOT	Hệ thống nhúng IOT
8	IOT	Internet of Things
9	KHMT	Khoa học máy tính
10	MDQT	Mức độ quan tâm
11	RF	Random Forest
12	SQL	Structured Query Language
13	SVM	Suport Vector Machines
14	TBC	Trung bình cộng
15	XML	eXtensible Markup Language

MỤC LỤC

DANH MỤC HÌNH ẢNH, BẢNG BIỂU

Danh mục hình ảnh

Danh mục bảng biểu

MỞ ĐẦU

1. Lý do chọn đề tài

Nghiên cứu về kỹ thuật học máy và ứng dụng trong việc hỗ trợ định hướng chuyên ngành cho sinh viên Công nghệ thông tin không chỉ là một đề tài quan trọng mà còn cực kỳ cấp thiết trong bối cảnh ngành CNTT đang phát triển mạnh mẽ ngày nay. 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 tạo ra nhu cầu về việc phân tích và sử dụng thông tin này một cách hiệu quả. Đồng thời, với sự đa dạng của các ngành và lĩnh vực trong CNTT, 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.

Học máy, một lĩnh vực trong trí tuệ nhân tạo, 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 học máy trong việc hỗ trợ định hướng chuyên ngành cho sinh viên CNTT có thể giúp phân tích sở thích, năng lực và mục tiêu cá nhân của sinh viên dựa trên dữ liệu từ học tập, dự án và các hoạt động khác. Điều này giúp sinh viên nhận biết được lĩnh vực mà họ có thể có hứng thú và tài năng, từ đó định hình lộ trình học tập và nghề nghiệp phù hợp nhất.

Ngoài ra, 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 phụ thuộc hoàn toàn 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.

Do đó, em đã chọn đề tài "Xây dựng ứng dụng tự động gợi ý lựa chọn các học phần chuyên ngành dựa vào điểm học tập các điểm cơ sở'' không chỉ đáp ứng nhu cầu thực tiễn của ngành 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 thế giới ngày nay, ngành CNTT đang trải qua sự phát triển mạnh mẽ và đa dạng hóa về cả lĩnh vực và công nghệ. Điều này đặt ra thách thức lớn đối với sinh viên khi họ phải đưa ra quyết định về chuyên ngành mà họ muốn theo đuổi. Một quyết định sai lầm có thể dẫn đến việc đánh mất cơ hội và thời gian quý báu trong học tập và sự nghiệp sau này.

Với mục tiêu giúp sinh viên tự tin và hiểu rõ hơn về sở thích, năng lực và mục tiêu cá nhân của mình, nghiên cứu này nhằm xây dựng các mô hình học máy dựa trên dữ liệu có sẵn từ kết quả học tập của cựu sinh viên. Những mô hình này sẽ phân tích và đánh giá thông tin để đề xuất các gợi ý chuyên ngành phù hợp nhất với từng sinh viên.

Bằng cách áp dụng kỹ thuật học máy và sử dụng dữ liệu thực tế từ sinh viên CNTT, đề tài này mong muốn cung cấp cho sinh viên một công cụ hỗ trợ quan trọng trong quá trình định hướng chuyên ngành, từ đó giúp sinh viên có thể phát triển sự nghiệp một cách hiệu quả và tự tin hơn trong lựa chọn con đường tương lai của mình trong ngành CNTT.

3. Phạm vi nghiên cứu

Đối tượng nghiên cứu: Gồm hai bộ dữ liệu, bộ đầu tiên là điểm của sinh viên các khoá từ K3 đến K13 của Trường Đại học Đại Nam, bộ thứ hai là khảo sát mức độ quan tâm của nhà trường đối với sinh viên và kỹ thuật học máy phân lớp dự báo chuyên ngành phù hợp cho từng sinh viên.

Phạm vi nghiên cứu: sinh viên hệ chính quy của Trường Đại học Đại Nam.

4. Phương pháp nghiên cứu

- ***Phương pháp thống kê:*** sử dụng để thu thập và phân tích dữ liệu liên quan đến bảng điểm, hiệu suất học tập, và sự phát triển theo thời gian.

- ***Phương pháp so sánh:*** bằng cách so sánh hiệu suất học tập và kiến thức cơ sở ngành của các nhóm sinh viên khác nhau (ví dụ: nhóm xuất sắc, trung bình, yếu) nhằm xác định sự khác biệt và tương quan giữa các yếu tố này.
- ***Phương pháp phân tích:*** nghiên cứu tương quan giữa các yếu tố khác nhau, như mức độ nắm vững kiến thức cơ bản và kiến thức chuyên ngành.
- ***Phương pháp mô tả:*** trình bày dữ liệu và kết quả nghiên cứu một cách chi tiết và tổ chức.

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 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.[11]

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 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.

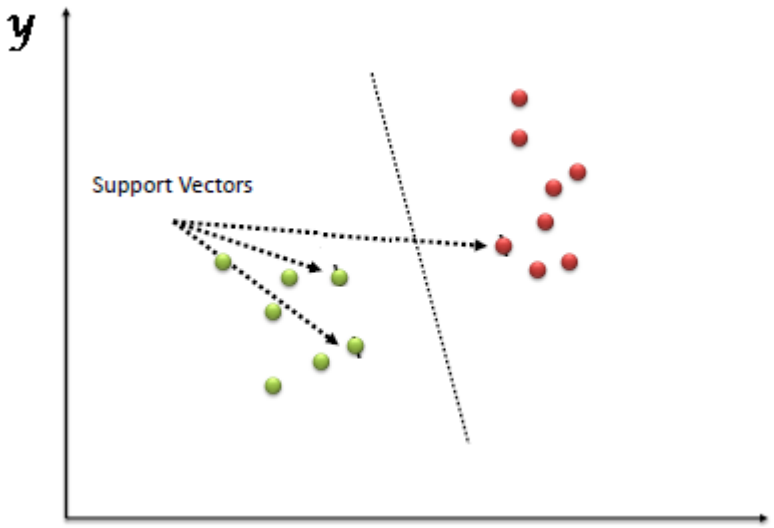
Chương 2
MỘT SỐ MÔ HÌNH HỌC MÁY

2.1. Support Vector Machine (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, 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 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.[12]

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. Mô hình biểu diễn 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 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:

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

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

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.[6]

2.1.3. Ưu và nhược điểm của mô hình

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 ta đã bàn luận ở trên.

2.2. Random Forest

2.2.1. 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

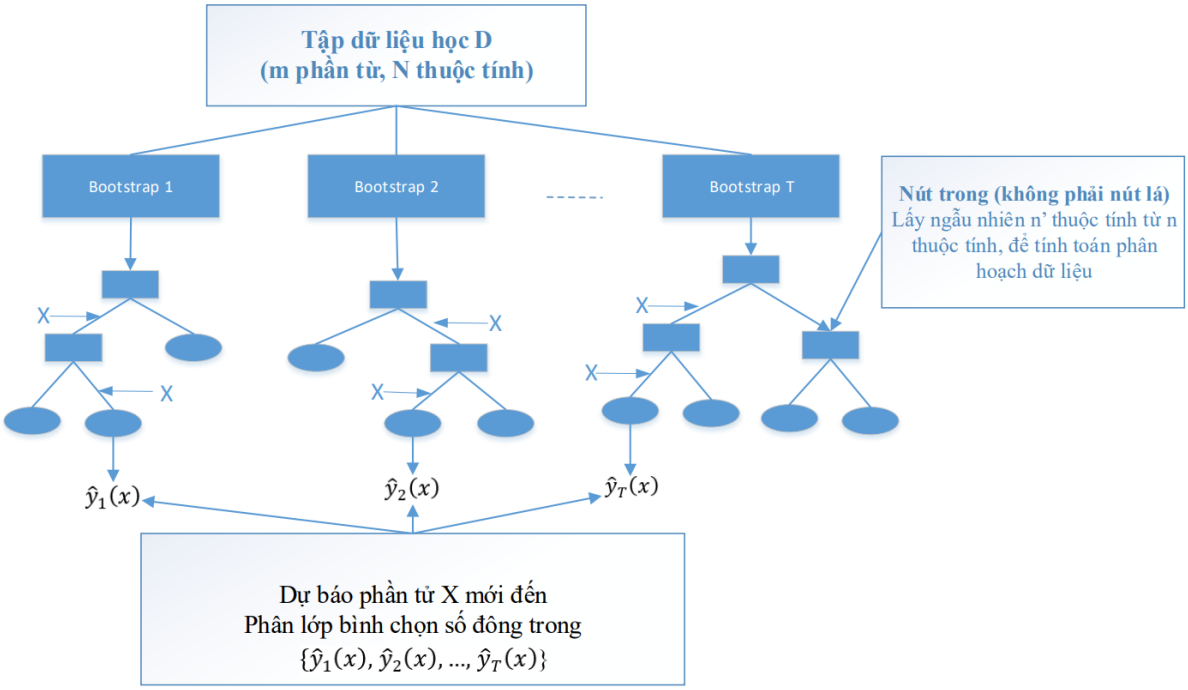
dùng luôn các mẫu out-of-bag này để kiểm thử và tính toán độ quan trọng thuộc tính của các cây CART trong rừng.[9]

2.2.2. 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 .



Hình 2. 2. Mô tả thuật toán Random Forest

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.

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. [9]

2.2.3. Ưu và nhược điểm của mô hình

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

2.3. Mô hình học máy cây quyết định (Decesion Tree)

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.[10]

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.

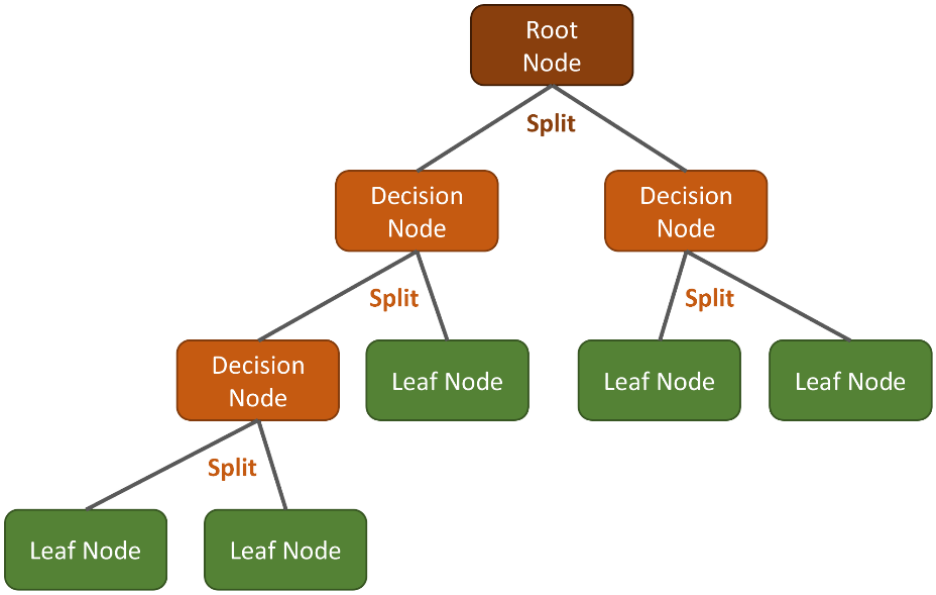
Mỗi nút của cây quyết định có thể là:

- 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. 3. Mô hình Decesion Tree

2.3.2. Các thuật toán học cây quyết định

Thuật toán ID3

ID3 (J. R. Quinlan 1993) sử dụng phương pháp tham lam tìm kiếm từ trên xuống thông qua không gian của các nhánh có thể không có backtracking. ID3 sử dụng Entropy và Information Gain để xây dựng một cây quyết định.

Qua sơ đồ, ta có thể thấy rõ ràng ràng, với phương pháp thứ nhất, ta phân loại được rõ ràng, trong khi phương pháp thứ hai, ta có một kết quả lộn xộn hơn. Và tương tự, cây quyết định sẽ thực hiện như trên khi thực hiện việc chọn các biến. [7]

Có rất nhiều hệ số khác nhau mà phương pháp cây quyết định sử dụng để phân chia. Dưới đây, hai hệ số phổ biến là ***Information Gain*** và **Gain Ratio** (ngoài ra còn hệ số Gini).

Entropy trong Cây quyết định (Decision Tree)

Entropy là thuật ngữ thuộc Nhiệt động lực học, là thước đo của sự biến đổi, hỗn loạn hoặc ngẫu nhiên. Năm 1948, Shannon đã mở rộng khái niệm Entropy sang lĩnh vực nghiên cứu, thống kê với công thức như sau:

Với một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau $x_1,x_2,\ldots,x_n$.

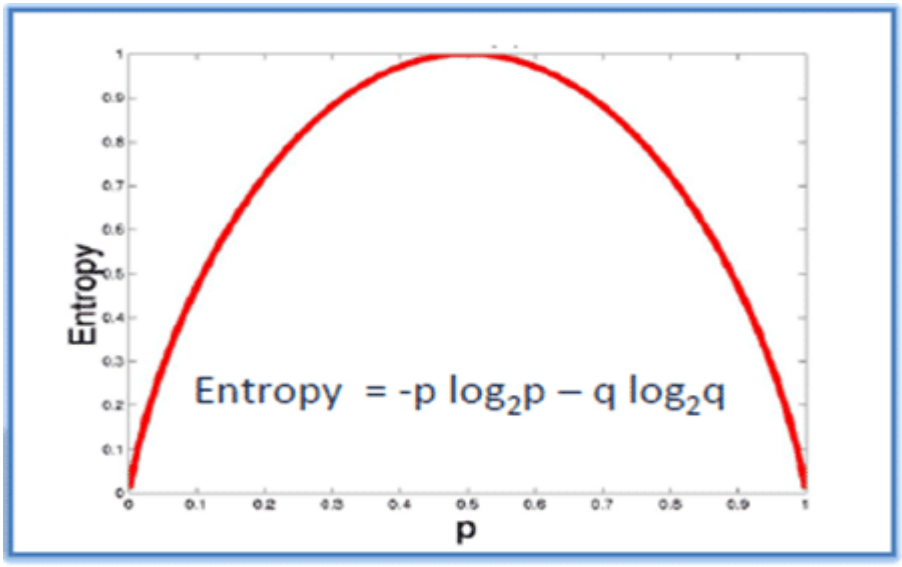
Giả sử rằng xác suất để x nhận các giá trị này là $p_i=p(x=x_i)$.

Ký hiệu phân phối là $p=(p_1 ,p_2 ,\ldots,p_n)$. Entropy của phân phối này định nghĩa là:

$$H(p)= -\sum^{n}_{n=1} p_i \log(p_i)$$

Giả sử bạn tung một đồng xu, entropy sẽ được tính như sau:

$$H = -[0.5 \ln(0.5) + 0.5 \ln(0.5)]$$



$$\text{Entropy} = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

Hình 2. 4: Hình vẽ biểu diễn sự thay đổi của hàm entropy

Hình vẽ trên biểu diễn sự thay đổi của hàm entropy. Ta có thể thấy rằng, entropy đạt tối đa khi xác suất xảy ra của hai lớp bằng nhau.

- P tinh khiết: $p_i = 0$ hoặc $p_i = 1$

- P vẩn đục: $p_i = 0.5$, khi đó hàm Entropy đạt đỉnh cao nhất

Information Gain trong Cây quyết định (Decision Tree)

Information Gain dựa trên sự giảm của hàm Entropy khi tập dữ liệu được phân chia trên một thuộc tính. Để xây dựng một cây quyết định, ta phải tìm tất cả thuộc tính trả về Infomation gain cao nhất.

Để xác định các nút trong mô hình cây quyết định, ta thực hiện tính Infomation Gain tại mỗi nút theo trình tự sau:

Bước 1: Tính toán hệ số Entropy của biến mục tiêu S có N phần tử với N_c phần tử thuộc lớp c cho trước:

$$H(S) = -\sum_{c=1}^{c}(Nc/N)\log(Nc/N)$$

Bước 2: Tính hàm số Entropy tại mỗi thuộc tính: với thuộc tính x, các điểm dữ liệu trong S được chia ra K child node S_1, S_2, …, S_K với số điểm trong mỗi child node lần lượt là m_1, m_2 ,…, m_K , ta có:

$$H(x, S) = \sum_{k=1}^{k}(mk / N) * H(Sk)$$

Bước 3: Chỉ số Gain Information được tính bằng:

$$\mathbf{G(x, S) = H(S) – H(x,S)}$$

Với ví dụ 2 trên, ta tính được hệ số Entropy như sau:

$$Entropy_{Parent} = -(0.57*ln(0.57) + 0.43*ln(0.43)) = 0.68$$

Hệ số Entropy theo phương pháp chia thứ nhất:

$$Entropy_{left} = -(.75*ln(0.75) + 0.25*ln(0.25)) = 0.56$$
$$Entropy_{right} = -(.33*ln(0.33) + 0.67*ln(0.67)) = 0.63$$

Ta có thể tính hệ số ***Information Gain*** như sau:

$$Information\ Gain = 0.68 – (4*0.56 + 3*0.63)/7 = 0.09$$

Hệ số Entropy với phương pháp chia thứ hai như sau:

$$Entropy_{left} = -(.67*ln(0.67) + 0.33*ln(0.33)) = 0.63$$

$$Entropy_{middle} = -(.5*ln(0.5) + 0.5*ln(0.5)) = 0.69$$

$$Entropy_{right} = -(.5*ln(0.5) + 0.5*ln(0.5)) = 0.69$$

Hệ số ***Information Gain***:

$$Information\ Gain = 0.68 – (3*0.63 + 2*0.69 + 2*0.69)/7 = 0.02$$

So sánh kết quả, ta thấy nếu chia theo phương pháp 1 thì ta được giá trị hệ số Information Gain lớn hơn gấp 4 lần so với phương pháp 2. Như vậy, giá trị thông tin ta thu được theo phương pháp 1 cũng nhiều hơn phương pháp 2.

Thuật toán C4.5

Thuật toán C4.5 là thuật toán cải tiến của ID3.

Trong thuật toán ID3, Information Gain được sử dụng làm độ đo. Tuy nhiên, phương pháp này lại ưu tiên những thuộc tính có số lượng lớn các giá trị mà ít xét tới những giá trị nhỏ hơn. Do vậy, để khắc phục nhược điểm trên, ta sử dụng độ đo Gain Ratio (trong thuật toán C4.5) như sau:

Đầu tiên, ta chuẩn hoá information gain với trị thông tin phân tách (split information):

$$Gain\ Ratio = \frac{Information\ Gain}{Split\ Info}$$

Trong đó: Split Info được tính như sau:

$$-\sum_{i=1}^{n} Di\ log_2 Di$$

Giả sử ta phân chia biến thành n nút cón và Di đại diện cho số lượng bản ghi thuộc nút đó. Do đó, hệ số Gain Ratio sẽ xem xét được xu hướng phân phối khi chia cây.

Áp dụng cho ví dụ trên và với cách chia thứ nhất, ta có

$$Split\ Info = – ((4/7)*log_2(4/7)) – ((3/7)*log_2(3/7)) = 0.98$$

Gain Ratio = 0.09/0.98 = 0.092

Tiêu chuẩn dừng

Trong các thuật toán Decision tree, với phương pháp chia trên, ta sẽ chia mãi các node nếu chưa tinh khiết. Như vậy, ta sẽ thu được một tree mà mọi điểm trong tập huấn luyện đều được dự đoán đúng (giả sử rằng không có hai input giống nhau nào cho output khác nhau). Khi đó, cây có thể sẽ rất phức tạp (nhiều node) với nhiều leaf node chỉ có một vài điểm dữ liệu. Như vậy, nhiều khả năng overfitting sẽ xảy ra.

Để tránh trường hừp này, ta có thể dừng cây theo một số phương pháp sau đây:

- Nếu node đó có entropy bằng 0, tức mọi điểm trong node đều thuộc một class.
- Nếu node đó có số phần tử nhỏ hơn một ngưỡng nào đó. Trong trường hợp này, ta chấp nhận có một số điểm bị phân lớp sai để tránh overfitting. Class cho leaf node này có thể được xác định dựa trên class chiếm đa số trong node.
- Nếu khoảng cách từ node đó đến root node đạt tới một giá trị nào đó. Việc hạn chế *chiều sâu của tree* này làm giảm độ phức tạp của tree và phần nào giúp tránh overfitting.
- Nếu tổng số leaf node vượt quá một ngưỡng nào đó.
- Nếu việc phân chia node đó không làm giảm entropy quá nhiều (information gain nhỏ hơn một ngưỡng nào đó).

Ngoài ID3, C4.5, còn một số thuật toán khác như:

- Thuật toán CHAID: tạo cây quyết định bằng cách sử dụng thống kê chi-square để xác định các phân tách tối ưu. Các biến mục tiêu đầu vào có thể là số (liên tục) hoặc phân loại.

- Thuật toán C&R: sử dụng phân vùng đệ quy để chia cây. Tham biến mục tiêu có thể dạng số hoặc phân loại.
- MARS
- Conditional Inference Trees. [7]

2.3.3. Ưu và nhược điểm của mô hình

Ưu điểm:

Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi bới những lợi ích:

- Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.
- Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả
- Có thể làm việc với cả dữ liệu số và dữ liệu phân loại
- Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê
- Có khả năng là việc với dữ liệu lớn

Nhược điểm:

- Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thạm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.
- Cây quyết định hay gặp vấn đề overfitting

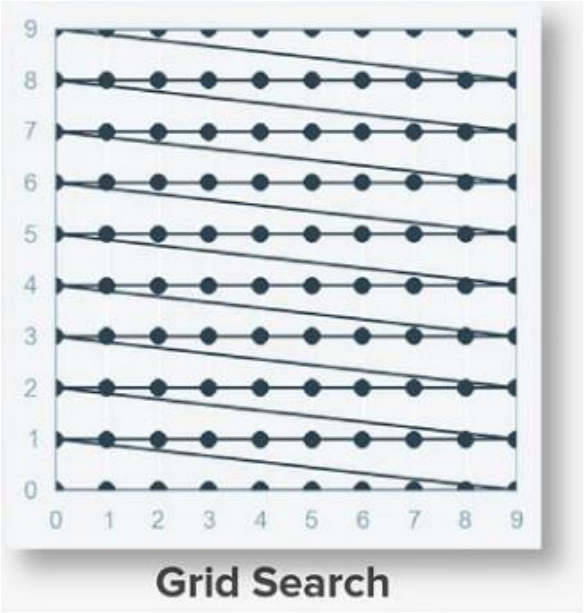
2.4. Phương pháp Grid Search

2.4.1. Khái niệm Grid Search

Grid Search là một kỹ thuật tối ưu hóa siêu tham số được sử dụng để xác định các tham số tốt nhất cho một mô hình học máy. Siêu tham số là các tham số mà giá trị của không được học trực tiếp từ dữ liệu mà cần được đặt trước khi quá trình huấn

luyện bắt đầu, ví dụ như số lượng cây trong rừng ngẫu nhiên hoặc giá trị C trong SVM.

Grid Search thực hiện tìm kiếm trên một "lưới" (grid) của các tham số có thể, bằng cách kiểm tra tất cả các kết hợp có thể của các giá trị siêu tham số và đánh giá hiệu suất của mô hình dựa trên một độ đo đánh giá nhất định (ví dụ như độ chính xác, MSE, R^2) thông qua kỹ thuật cross-validation.



Hình 2. 5. Mô hình Grid Search

2.4.2.Cách thức thực hiện Grid Search

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).

2.4.3. Ưu điểm của Grid Search

Đơn giản và trực quan: Grid Search dễ hiểu và triển khai. Nó kiểm tra tất cả các tổ hợp có thể của các siêu tham số, đảm bảo rằng không bỏ sót bất kỳ tổ hợp tiềm năng nào.

Toàn diện: Grid Search kiểm tra mọi kết hợp của các siêu tham số, giúp tìm ra tổ hợp tốt nhất cho mô hình.

Dễ dàng triển khai với các thư viện như scikit-learn: Các thư viện học máy phổ biến như scikit-learn cung cấp các công cụ để dễ dàng thực hiện Grid Search.

Tính tổng quát cao: Khi kết hợp với cross-validation, Grid Search giúp đảm bảo rằng các siêu tham số được chọn có tính tổng quát cao, phù hợp với dữ liệu chưa từng thấy.

2.4.4. Hạn chế của Grid Search

Tốn kém tài nguyên: Grid Search có thể rất tốn kém về thời gian và tài nguyên tính toán, đặc biệt khi số lượng siêu tham số và các giá trị cho mỗi tham số tăng lên. Số lượng tổ hợp tăng theo cấp số nhân.

Không hiệu quả với không gian siêu tham số lớn: Khi có nhiều siêu tham số và mỗi siêu tham số có nhiều giá trị, Grid Search có thể trở nên không khả thi.

Cố định giá trị tham số: Grid Search yêu cầu xác định trước các giá trị của các tham số để thử nghiệm. Nếu phạm vi giá trị không được chọn đúng cách, có thể dẫn đến việc bỏ sót các tổ hợp tham số tiềm năng.

Không linh hoạt: So với các phương pháp tối ưu hóa siêu tham số khác như Random Search hoặc Bayesian Optimization, Grid Search kém linh hoạt hơn và có thể bỏ lỡ các tổ hợp tham số tốt hơn nếu không được định nghĩa trong lưới ban đầu.

2.5. Phương pháp Random Search

2.5.1. Khái niệm Random Search

Random Search là một kỹ thuật tối ưu hóa siêu tham số được sử dụng trong học máy, tương tự như Grid Search nhưng với cách tiếp cận khác. Thay vì thử tất cả các tổ hợp có thể của các siêu tham số, Random Search thử ngẫu nhiên các tổ hợp trong một phạm vi xác định. Mục tiêu là tìm ra các giá trị siêu tham số tốt nhất cho mô hình mà không phải kiểm tra toàn bộ không gian tham số.

Hình 2. 6. Mô hình Random Search

2.5.2. Cách thức thực hiện Random Search

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).

2.5.3. Ưu điểm của Random Search

Tiết kiệm thời gian và tài nguyên: So với Grid Search, Random Search thường tiết kiệm hơn về thời gian và tài nguyên tính toán vì không phải kiểm tra toàn bộ không gian tham số.

Hiệu quả với không gian siêu tham số lớn: Random Search có thể khám phá không gian siêu tham số rộng lớn hơn mà không phải kiểm tra tất cả các tổ hợp có thể.

Khả năng tìm được các tổ hợp tốt hơn: Do chọn ngẫu nhiên, Random Search có khả năng tìm thấy các tổ hợp siêu tham số tối ưu mà Grid Search có thể bỏ lỡ nếu phạm vi giá trị không được chọn đúng cách.

Đơn giản và dễ triển khai: Random Search dễ dàng triển khai và có thể tích hợp với các thư viện học máy phổ biến như scikit-learn.

2.5.4. Hạn chế của Random Search

Không đảm bảo toàn diện: Vì Random Search không kiểm tra tất cả các tổ hợp, có thể bỏ lỡ một số tổ hợp siêu tham số tốt.

Kết quả phụ thuộc vào số lần thử nghiệm: Hiệu suất của Random Search phụ thuộc vào số lượng thử nghiệm (n_iter). Số lượng thử nghiệm lớn có thể cần thiết để đạt được kết quả tốt nhất, làm tăng thời gian tính toán.

Khó khăn trong việc chọn phạm vi giá trị: Xác định phạm vi giá trị phù hợp cho các siêu tham số có thể khó khăn và cần kiến thức chuyên môn về mô hình và dữ liệu.

Chương 3
PHÂN TÍCH VÀ XỬ LÝ DỮ LIỆU

3.1. Bài toán

BT1: Bài toán nhằm tạo ra một ứng dụng tự động dự đoán chuyên ngành phù hợp cho sinh viên dựa trên điểm số học tập từ các môn học cơ bản, cơ sở ngành và chuyên ngành. Quá trình chuẩn bị dữ liệu bao gồm loại bỏ dữ liệu không hợp lệ hoặc thiếu thông tin, xử lý các giá trị thiếu bằng cách điền vào giá trị phổ biến nhất, và chuyển đổi điểm số sang dạng số nguyên. Sau đó, một mô hình học máy dựa trên Naive Bayes, SVM, KNN, DF, DC,… được xây dựng và huấn luyện trên dữ liệu đã được gán nhãn, với mỗi mẫu dữ liệu đại diện cho một sinh viên với điểm số từ các môn học. Hiệu suất của mô hình được đánh giá thông qua tỷ lệ chính xác, là tỷ lệ giữa số lượng dự đoán chính xác và tổng số lượng mẫu dữ liệu. Kết quả là một hệ thống dự đoán chuyên ngành phù hợp, cung cấp một công cụ hữu ích trong quá trình lựa chọn chuyên ngành của sinh viên.

BT2: Thu thập dữ liệu từ khảo sát, sau đó làm sạch dữ liệu rồi dự đoán mức độ được quan tâm của sinh viên dựa trên các câu hỏi và dự đoán sự phụ thuộc của điểm GPA tương lai từ mức độ được quan tâm dựa trên dữ liệu khảo sát. Quá trình này bao gồm loại bỏ dữ liệu không hợp lệ, xử lý giá trị thiếu, chuyển đổi điểm số, sử dụng mô hình SVM, Hồi quy tuyến tính,… để dự đoán mức độ quan tâm với điểm GPA, và đánh giá hiệu suất mô hình thông qua tỷ lệ chính xác.

3.2. Dữ liệu nghiên cứu

3.2.1. Dữ liệu dự đoán gợi ý chuyên ngành

Bảng điểm của 11 khóa sinh viên từ khoá K3 tới K13 của Trường Đại học Đại Nam khoa Công nghệ thông tin.

3.2.2. Dữ liệu của khảo sát mức độ được quan tâm

Dữ liệu khảo sát từ goole form về mức độ được quan tâm của nhà Trường với sinh viên K14 của Trường Đại học Đại Nam khoa Công nghệ thông tin.

3.3. Thống kê các trường dữ liệu

3.3.1. Thông kê các trường dữ liệu của data dự đoán gợi ý chuyên ngành

Dữ liệu bao gồm thông tin về mười một khóa học trong lĩnh vực Công nghệ thông tin (CNTT). Số lượng dữ liệu thu thập cho từng khóa là như sau: CNTT-K3 có 35 dữ liệu, CNTT-K4 có 13 dữ liệu, CNTT-K5 có 7 dữ liệu, CNTT-K6 có 5 dữ liệu, CNTT-K7 có 14 dữ liệu, CNTT-K8 có 71 dữ liệu, CNTT-K9 có 41 dữ liệu, CNTT-K10 có 22 dữ liệu, CNTT-K11 có 15 dữ liệu, CNTT-K12 có 77 dữ liệu và CNTT-K13 có 108 dữ liệu. Số trường dữ liệu khác nhau cho mỗi khóa là: CNTT-K3 có 65 trường, CNTT-K4 có 64 trường, CNTT-K5 có 60 trường, CNTT-K6 có 58 trường, CNTT-K7 có 64 trường, CNTT-K8 có 82 trường, CNTT-K9 có 64 trường, CNTT-K10 có 61 trường, CNTT-K11 có 61 trường, CNTT-K12 có 61 trường và CNTT-K13 có 62 trường.

Tên khóa	Số lượng dữ liệu	Số trường
CNTT – K3	35	65
CNTT – K4	13	64
CNTT – K5	7	60
CNTT – K6	5	58
CNTT – K7	14	64
CNTT – K8	71	82
CNTT – K9	41	64
CNTT – K10	22	61

CNTT – K11	15	61
CNTT – K12	77	61
CNTT – K13	108	62

Bảng 3. 1. Bảng thống kê các trường dữ liệu của dữ liệu dự đoán chuyên ngành

3.3.2. Thông kê các trường dữ liệu của bộ dữ liệu khảo sát mức độ quan tâm

Dữ liệu khảo sát ghi nhận ý kiến của sinh viên về việc tiếp cận và hiểu biết về các quy định, quy chế của nhà trường, cũng như mức độ hỗ trợ từ giáo viên và cảm nhận về tài nguyên hỗ trợ từ nhà trường. Trong tổng số 12 trường dữ liệu, các thông tin thu thập bao gồm tên, mã sinh viên, cùng với các câu hỏi như mức độ tiếp cận thông tin, hiểu biết về các văn bản như hướng dẫn học tập và quy chế, sự hỗ trợ từ giáo viên, cảm nhận về tài nguyên hỗ trợ từ nhà trường, cũng như sự nhận biết và khó khăn trong việc tiếp nhận thông báo lịch thi học kỳ.

Khóa	**Số trường dữ liệu**	**Số trường**
K14	151	13

Bảng 3. 2. Bảng thống kê các trường dữ liệu khảo sát mức độ quan tâm

3.4. Tiền xử lý dữ liệu

3.4.1. Tiền xử lý dữ liệu của bộ dữ liệu dự đoán gợi ý chuyên ngành

Tiến hành gộp dữ liệu gồm 11 khóa từ K3 tới K13:

Khóa K3:

Khóa	**Tên môn tính TBC**	**Tên các môn tính TBC**	**Chú thích**
K3	**Toán**	– Giải tích 1	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
		– Hình giải tích và đại số – Giải tích 2 – Giải tích phần 3 – Toán rời rạc – Xác suất thống kê	
	Chính trị	– Những NLCBCN Mác - Lê nin (CNXH) – Những NLCBCN Mác - Lê nin (KTCT) – Những NLCBCN Mác - Lênin (Triết), Tư tưởng Hồ Chí Minh – Đường lối cách mạng của Đảng Cộng sản Việt Nam.	
	Đại cương	– Tin học đại cương – Vật lý đại cương – Pháp luật đại cương – Tin học đại cương 2 – Vật lý 2	
	Tiếng Anh	– Tiếng Anh P1 – Tiếng Anh P2 – Tiếng Anh P3 – Tiếng Anh chuyên ngành – TOEIC	
	SQL	– Lý thuyết cơ sở dữ liệu – Hệ quản trị cơ sở dữ liệu – Quản trị cơ sở dữ liệu phân tán	
	Mã nguồn mở	– Lý thuyết hệ điều hành – Hệ điều hành mã nguồn mở	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
	Đồ án tốt nghiệp	– Chuyên ngành – Cơ sở ngành – Đồ án tốt nghiệp	Nếu không có điểm Đồ án tốt nghiệp thì lấy điểm trung bình của 2 môn Chuyên ngành, Cơ sở ngành

Bảng 3. 3. Bảng tính TBC của các môn đại cương và các môn cơ sở K3

Khóa K4:

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
K4	**Toán**	– Giải tích 1 – Hình giải tích và đại số – Giải tích 2 – Giải tích phần 3 – Toán rời rạc – Xác suất thống kê	
	Chính trị	– Những NLCBCN Mác - Lê nin (CNXH) – Những NLCBCN Mác - Lê nin (KTCT), Những NLCBCN Mác - Lênin (Triết) – Tư tưởng Hồ Chí Minh – Đường lối cách mạng của Đảng Cộng sản Việt Nam.	
	Đại cương	– Tin học đại cương – Vật lý 1 – Pháp luật đại cương	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
	Tiếng Anh	– Tiếng Anh P1 – Tiếng Anh P2 – Tiếng Anh P3 – Tiếng Anh P4 – TOEIC 1 – TOEIC 2 – B1 đầu rá	
	SQL	– Lý thuyết cơ sở dữ liệu – Hệ quản trị cơ sở dữ liệu – Quản trị cơ sở dữ liệu phân tán	
	Mã nguồn mở	– Hệ điều hành mã nguồn mở	
	Đồ án tốt nghiệp	– Chuyên ngành – Cơ sở ngành	Lấy điểm chuyên ngành và cơ sở ngành / trung bình

Bảng 3. 4. Bảng tính TBC của các môn đại cương và các môn cơ sở K4

Khóa K5:

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
K5	**Toán**	– Giải tích 1 – Hình giải tích và đại số – Giải tích 2 – Toán rời rạc – Xác suất thống kê	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
	Chính trị	– Những NLCBCN Mác - Lê nin (CNXH) – Những NLCBCN Mác - Lê nin (KTCT), Những NLCBCN Mác - Lênin (Triết) – Tư tưởng Hồ Chí Minh – Đường lối cách mạng của Đảng Cộng sản Việt Nam.	
	Đại cương	– Tin học đại cương – Vật lý 1 – Pháp luật đại cương	
	Tiếng Anh	– Tiếng Anh P1 – Tiếng Anh P2 – Tiếng Anh P3 – Tiếng Anh P4 – Tiếng Anh P5	
	SQL	– Lý thuyết cơ sở dữ liệu – Hệ quản trị cơ sở dữ liệu – Quản trị cơ sở dữ liệu phân tán	
	Mã nguồn mở	– Lý thuyết hệ điều hành – Hệ điều hành mã nguồn mở	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
	Đồ án tốt nghiệp	– Chuyên ngành – Cơ sở ngành	Lấy điểm chuyên ngành và cơ sở ngành / trung bình

Bảng 3. 5. Bảng tính TBC của các môn đại cương và các môn cơ sở K5

Khóa K6:

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
K6	**Toán**	– Giải tích 1 – Hình giải tích và đại số – Giải tích 2 – Toán rời rạc – Xác suất thống kê	
	Chính trị	– Những NLCBCN Mác - Lê nin (CNXH) – Những NLCBCN Mác - Lê nin (KTCT), Những NLCBCN Mác - Lênin (Triết) – Tư tưởng Hồ Chí Minh – Đường lối cách mạng của Đảng Cộng sản Việt Nam.	
	Đại cương	– Tin học đại cương – Vật lý 1	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
		– Pháp luật đại cương	
	Tiếng Anh	– Tiếng Anh P1 – Tiếng Anh P2 – Tiếng Anh P3 – Tiếng Anh P4 – Tiếng Anh P5	
	SQL	– Lý thuyết cơ sở dữ liệu – Hệ quản trị cơ sở dữ liệu – Quản trị cơ sở dữ liệu phân tán	
	Mã nguồn mở	– Lý thuyết hệ điều hành – Hệ điều hành mã nguồn mở	
	Đồ án tốt nghiệp	– Chuyên ngành – Cơ sở ngành	Lấy điểm chuyên ngành và cơ sở ngành / trung bình

Bảng 3. 6. Bảng tính TBC của các môn đại cương và các môn cơ sở K6

Khóa K7:

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
K7	**Toán**	– Giải tích 1 – Giải tích 2 – Toán rời rạc	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
		– Xác suất thống kê	
	Chính trị	– Những NLCBCN Mác - Lê nin (CNXH) – Những NLCBCN Mác - Lê nin (KTCT), Những NLCBCN Mác - Lênin (Triết) – Tư tưởng Hồ Chí Minh – Đường lối cách mạng của Đảng Cộng sản Việt Nam.	
	Đại cương	– Tin học đại cương – Vật lý 1 – Pháp luật đại cương	
	Tiếng Anh	– Tiếng Anh P1 – Tiếng Anh P2 – Tiếng Anh P3 – Tiếng Anh P4 – Tiếng Anh P5	
	SQL	– Lý thuyết cơ sở dữ liệu – Hệ quản trị cơ sở dữ liệu – Quản trị cơ sở dữ liệu phân tán	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
	Mã nguồn mở	– Lý thuyết hệ điều hành – Hệ điều hành mã nguồn mở	
	Đồ án tốt nghiệp	– Chuyên ngành – Cơ sở ngành	Lấy điểm chuyên ngành và cơ sở ngành / trung bình

Bảng 3. 7. Bảng tính TBC của các môn đại cương và các môn cơ sở K7

Khóa K8:

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
K8	**Toán**	– Giải tích 1 – Đại số – Giải tích 2 – Toán rời rạc – Xác suất thống kê	Trong đó có 2 trường giải tích 2, 2 trường Xác xuất thống kê do khóa trước có người bị học lại và tiến hành gộp điểm 2 cột giải tích và xác xuất thống kê thành 1 trường giải tích và 1 trường xác xuất thống kê
	Chính trị	– Những NLCBCN Mác - Lê nin (CNXH)	Do có hai trường Những nguyên lý cơ bản của chủ nghĩa Mac - Lenin P2 và

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
		– Những NLCBCN Mác - Lê nin (KTCT), Những NLCBCN Mác - Lênin (Triết) – Tư tưởng Hồ Chí Minh – Đường lối cách mạng của Đảng Cộng sản Việt Nam.	Đường lối cách mạng của Đảng Cộng sản Việt Nam do khóa trước học lại đã có điểm nên tiến hành gộp điểm của 2 trường giống nhau đó thành 1 trường
	Đại cương	– Tin học đại cương – Vật lý 1 – Pháp luật đại cương	
	Tiếng Anh	– Tiếng Anh P1 – Tiếng Anh P2 – Tiếng Anh P3 – Tiếng Anh P4 – Tiếng Anh P5 – Tiếng Anh chuyên ngành – TOEIC	Có 2 trường toeic và do khóa trước học lại nên có điểm trước nên gộp lại thành một trường, và trường Tiếng Anh P5 do có một số người do khóa trước không học nên sẽ tính TBC trừ môn đó còn lại ai có điểm TAP5 thì chia trung bình cả
	SQL	– Lý thuyết cơ sở dữ liệu – Hệ quản trị cơ sở dữ liệu	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
		– Quản trị cơ sở dữ liệu phân tán	
	Mã nguồn mở	– Lý thuyết hệ điều hành – Hệ điều hành mã nguồn mở	Có thêm 1 trường Lý thuyết hệ điều hành do khóa trước học lại và do kỳ của k8 k có môn lý thuyết nên nếu ai có trường lý thuyết hệ điều hành thì cộng vào chia trung bình nếu k có thì lấy điểm Hệ điều hành mã nguồn mở
	Đồ án tốt nghiệp	– Chuyên ngành – Cơ sở ngành	Nếu không có điểm Đồ án tốt nghiệp thì lấy điểm trung bình của 2 môn Chuyên ngành, Cơ sở ngành

Bảng 3. 8. Bảng tính TBC của các môn đại cương và các môn cơ sở K8

Khóa K9:

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
K9	**Toán**	– Giải tích 1 – Hình giải tích và đại số – Giải tích 2	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
		– Toán rời rạc – Xác suất thống kê	
	Chính trị	– Những NLCBCN Mác - Lê nin (CNXH) – Những NLCBCN Mác - Lê nin (KTCT), Những NLCBCN Mác - Lênin (Triết) – Tư tưởng Hồ Chí Minh – Đường lối cách mạng của Đảng Cộng sản Việt Nam.	
	Đại cương	– Tin học đại cương – Pháp luật đại cương	
	Tiếng Anh	– Tiếng Anh P1 – Tiếng Anh P2 – Tiếng Anh P3 – Tiếng Anh P4 – Tiếng Anh chuyên ngành – TOEIC 1 – TOEIC 2	
	SQL	– Lý thuyết cơ sở dữ liệu – Hệ quản trị cơ sở dữ liệu – Quản trị cơ sở dữ liệu phân tán	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
	Mã nguồn mở	– Lý thuyết hệ điều hành – Hệ điều hành mã nguồn mở	
	Đồ án tốt nghiệp	– Chuyên ngành – Cơ sở ngành	Nếu không có điểm Đồ án tốt nghiệp thì lấy điểm trung bình của 2 môn Chuyên ngành, Cơ sở ngành

Bảng 3. 9. Bảng tính TBC của các môn đại cương và các môn cơ sở K9

Khóa K10:

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
K10	**Toán**	– Giải tích 1 – Hình giải tích và đại số – Giải tích 2 – Toán rời rạc – Xác suất thống kê	
	Chính trị	– Những NLCBCN Mác - Lê nin (CNXH) – Những NLCBCN Mác - Lê nin (KTCT), Những NLCBCN Mác - Lênin (Triết) – Tư tưởng Hồ Chí Minh	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
		– Đường lối cách mạng của Đảng Cộng sản Việt Nam.	
	Đại cương	– Tin học đại cương – Pháp luật đại cương	
	Tiếng Anh	– Tiếng Anh P1 – Tiếng Anh P2 – Tiếng Anh P3 – Tiếng Anh P4 – Tiếng Anh chuyên ngành – TOEIC 1 – TOEIC 2	
	SQL	– Lý thuyết cơ sở dữ liệu – Hệ quản trị cơ sở dữ liệu – Quản trị cơ sở dữ liệu phân tán	
	Mã nguồn mở	– Hệ điều hành mã nguồn mở	
	Đồ án tốt nghiệp	– Thực tập tốt nghiệp	Do khóa k10 k có điểm nào tốt nghiệp ngoài Thực tập tốt nghiệp, nên chọn thay cho thay cho các môn tính TBC

Bảng 3. 10. Bảng tính TBC của các môn đại cương và các môn cơ sở K10

Khóa K11:

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
K11	**Toán**	– Giải tích 1 – Hình giải tích và đại số – Giải tích 2 – Toán rời rạc – Xác suất thống kê	
	Chính trị	– Những NLCBCN Mác - Lê nin (CNXH) – Những NLCBCN Mác - Lê nin (KTCT), Những NLCBCN Mác - Lênin (Triết) – Tư tưởng Hồ Chí Minh – Đường lối cách mạng của Đảng Cộng sản Việt Nam.	
	Đại cương	– Tin học đại cương – Pháp luật đại cương	
	Tiếng Anh	– Tiếng Anh P1 – Tiếng Anh P2 – Tiếng Anh P3 – Tiếng Anh P4 – Tiếng Anh chuyên ngành	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
		– TOEIC 1 – TOEIC 2	
	SQL	– Lý thuyết cơ sở dữ liệu – Hệ quản trị cơ sở dữ liệu – Quản trị cơ sở dữ liệu phân tán	
	Mã nguồn mở	– Hệ điều hành mã nguồn mở	
	Đồ án tốt nghiệp	– Đồ án tốt nghiệp	

Bảng 3. 11. Bảng tính TBC của các môn đại cương và các môn cơ sở K11

Khóa K12:

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
K12	**Toán**	– Giải tích 1 – Hình giải tích và đại số – Giải tích 2 – Toán rời rạc – Xác suất thống kê	
	Chính trị	– Những NLCBCN Mác - Lê nin (CNXH)	

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
		– Những NLCBCN Mác - Lê nin (KTCT), Những NLCBCN Mác - Lênin (Triết) – Tư tưởng Hồ Chí Minh – Đường lối cách mạng của Đảng Cộng sản Việt Nam.	
	Đại cương	– Tin học đại cương – Pháp luật đại cương	
	Tiếng Anh	– Tiếng Anh P1 – Tiếng Anh P2 – Tiếng Anh P3 – Tiếng Anh P4 – Tiếng Anh chuyên ngành – TOEIC 1 – TOEIC 2	
	SQL	– Lý thuyết cơ sở dữ liệu – Hệ quản trị cơ sở dữ liệu – Quản trị cơ sở dữ liệu phân tán	
	Mã nguồn mở	– Cài đặt và bảo trì hệ thống máy tính	
	Đồ án tốt nghiệp	– Đồ án tốt nghiệp	

Bảng 3. 12. Bảng tính TBC của các môn đại cương và các môn cơ sở K12

Khóa K13:

Khóa	Tên môn tính TBC	Tên các môn tính TBC	Chú thích
K11	**Toán**	– Giải tích 1 – Hình giải tích và đại số – Giải tích 2 – Toán rời rạc – Xác suất thống kê	
	Chính trị	– Kinh tế chính trị Mac Lenin – Triết học Mac – Lenin – Chủ nghĩa xã hội khoa học – Tư tưởng Hồ Chí Minh	
	Đại cương	– Tin học đại cương – Pháp luật đại cương	
	Tiếng Anh	– Tiếng Anh P1 – Tiếng Anh P2 – Tiếng Anh P3 – Tiếng Anh P4 – Tiếng Anh chuyên ngành – TOEIC 1 – TOEIC 2	
	SQL	– Lý thuyết cơ sở dữ liệu – Hệ quản trị cơ sở dữ liệu	

	Mã nguồn mở	– Hệ điều hành mã nguồn mở	
	Đồ án tốt nghiệp	– Đồ án tốt nghiệp	

Bảng 3. 13. Bảng tính TBC của các môn đại cương và các môn cơ sở K13

Sau khi gộp lại thu được các môn: TBC các môn toán, chính trị, đại cương, tiếng anh, SQL, mã nguồn mở. Sau đó thu được bảng dữ liệu mới bao gồm 59 trường và có 412 dữ liệu. Trong đó K3 gồm 35 dữ liệu, K4 gồm 13 dữ liệu, K5 có 7 dữ liệu, K6 có 5 dữ liệu, K7 có 14 dữ liệu, K8 có 75 dữ liệu, K9 có 41 dữ liệu, K10 có 22 dữ liệu, K11 có 15 dữ liệu, K12 có 77 dữ liệu, K13 có 108 dữ liệu.Trong đó các trường gồm: Khóa, Số TT, Mã sinh viên, Họ và tên, Ngày sinh, Nơi sinh, Tên lớp, Điểm thưởng, Nội dung điểm thưởng, TBC HTTK, TBC HT10, Xếp loại thang 10, Số HP nợ, Số tín chỉ nợ, TBC môn toán, TBC môn chính trị, TBC môn đại cương, TBC môn tiếng anh, TBC môn SQL, TBC môn mã nguồn mở, TBC đồ án tốt nghiệp, Cấu trúc dữ liệu và giải thuật, Phương pháp tính, Các phương pháp tối ưu, Kỹ thuật đồ hoạ, Kỹ thuật vi xử lý và lập trình hệ thống, Trí tuệ nhân tạo, Công nghệ phần mềm, Đánh giá độ phức tạp của thuật toán, Mạng và Hệ điều hành mạng, Phân tích thiết kế hệ thống, Thực tập phần cứng và mạng máy tính, Xử lý ảnh, Xử lý song song, An toàn dữ liệu, Công nghệ XML, Lập trình Java, Lập trình trong môi trường Web, Quản lý dự án công nghệ thông tin, Quản trị mạng, Kiến trúc máy tính, Lập trình Dot.net, Thiết kế Web1, Lập trình hướng đối tượng trên Visual C++, Kỹ năng quản lý và khai thác tiềm năng bản thân, Cơ sở lập trình, Đồ án cấu trúc dữ liệu và giải thuật, Đồ án cơ sở lập trình, Lập trình .Net nâng cao, Bảo mật mạng máy tính, Kỹ năng giao tiếp, thuyết trình, làm việc nhóm, Đồ án hướng đối tượng, Lập trình thiết bị di động, Đồ án lập trình câu trúc dữ liệu, Học máy, IOT, Kiểm thử, Thiết kế Web 2, Khai phá dữ liệu.

STT	Tên Trường
1	Khóa
2	Số TT
3	Mã sinh viên
4	Họ và tên
5	Ngày sinh
6	Nơi sinh
7	Tên lớp
8	Điểm thưởng
9	Nội dung điểm thưởng
10	TBC HTTK
11	TBC HT10
12	Xếp loại thang 10
13	Số HP nợ
14	Số tín chỉ nợ
15	TBC môn toán
16	TBC môn chính trị
17	TBC môn đại cương
18	TBC môn tiếng anh

STT	Tên Trường
19	TBC môn SQL
20	TBC môn mã nguồn mở
21	TBC đồ án tốt nghiệp
22	Cấu trúc dữ liệu và giải thuật
23	Phương pháp tính
24	Các phương pháp tối ưu
25	Kỹ thuật đồ hoạ
26	Kỹ thuật vi xử lý và lập trình hệ thống
27	Trí tuệ nhân tạo
28	Công nghệ phần mềm
29	Đánh giá độ phức tạp của thuật toán
30	Mạng và Hệ điều hành mạng
31	Phân tích thiết kế hệ thống
32	Thực tập phần cứng và mạng máy tính
33	Xử lý ảnh
34	Xử lý song song
35	An toàn dữ liệu
36	Công nghệ XML

STT	Tên Trường
37	Lập trình Java
38	Lập trình trong môi trường Web
39	Quản lý dự án công nghệ thông tin
40	Quản trị mạng
41	Kiến trúc máy tính
42	Lập trình Dot.net
43	Thiết kế Web1
44	Lập trình hướng đối tượng trên Visual C++
45	Kỹ năng quản lý và khai thác tiềm năng bản thân
46	Cơ sở lập trình
47	Đồ án cấu trúc dữ liệu và giải thuật
48	Đồ án cơ sở lập trình
49	Lập trình .Net nâng cao
50	Bảo mật mạng máy tính
51	Kỹ năng giao tiếp, thuyết trình, làm việc nhóm
52	Đồ án hướng đối tượng
53	Lập trình thiết bị di động
54	Đồ án lập trình câu trúc dữ liệu

STT	Tên Trường
55	Học máy
56	IOT
57	Kiểm thử
58	Thiết kế Web 2
59	Khai phá dữ liệu

Bảng 3. 14. Tổng các trường sau khi gộp tính TBC

Tiếp tục tiến hành làm sạch bằng thay các điểm TB của khóa vào vào điểm khuyết và tiến hành gộp tính điểm TBC các môn chuyên ngành và môn cơ sở dựa vào các chuyên ngành đã chia và cộng điểm và chia trung bình dựa vào tín chỉ. Lấy số điểm nhân với số tín chỉ rồi cộng các môn còn lại tương tự rồi chia trung bình tổng số tín.

STT	Tên chuyên ngành	Môn chuyên ngành
1	Học phần bắt buộc cơ sở ngành	– Kỹ năng quản lý và khai thác tiềm năng bản thân – Kỹ năng giao tiếp, thuyết trình, làm việc nhóm – Cơ sở lập trình – Phương pháp tính – Đồ án cơ sở lập trình – Kiến trúc máy tính – Cấu trúc dữ liệu và giải thuật – Mạng máy tính – Lập trình hướng đối tượng

STT	Tên chuyên ngành	Môn chuyên ngành
		– Đồ án lập trình hướng đối tượng – Phân tích thiết kế hệ thống – Đánh giá độ phức tạp của thuật toán – Các phương pháp tối ưu – Công nghệ phần mềm – Trí tuệ nhân tạo – Thiết kế Web 1 – Thiết kế Web 2
2	Chuyên ngành phát triển phần mềm	– Lập trình Mobile – Quản trị dự án công nghệ thông tin – Xử lý ảnh – Lập trình Java – Lập trình .NET 1 – Lập trình . NET nâng cao – Kiểm thử – Học máy – Kỹ thuật đồ hoạ
3	Chuyên ngành hệ thống nhúng và IOT	– IOT – Quản trị mạng – Kỹ thuật vi xử lý và lập trình hệ thống
4	Chuyên ngành khoa học máy tính	– Khai phá dữ liệu – Công nghệ XML

STT	Tên chuyên ngành	Môn chuyên ngành
		– Mạng và Hệ điều hành mạng – Bảo mật mạng máy tính – An toàn dữ liệu – Xử lý song song

Bảng 3. 15. TBC các điểm chuyên ngành

Tiếp tục xử lý dữ liệu bằng các chuyển các điểm môn chuyên ngành nếu điểm Chuyên ngành phần mềm lớn chuyên ngành hệ thống nhúng và IOT và chuyên ngành hệ thống thông tin thì gán nhãn là 100, và nếu chuyên ngành nhúng và IOT lớn hơn hai chuyên ngành còn lại thì gán nhãn là 010, còn lại thì là 010.

STT	Tên chuyên ngành	Nhãn
1	Chuyên ngành phát triển phần mềm	1000
2	Chuyên ngành hệ thống nhúng và IOT	1100
3	Chuyên ngành khoa học máy tính	1110

Bảng 3. 16. Các nhãn chuyên ngành

3.4.2. Tiền xử lý dữ liệu bộ khảo sát mức độ quan tâm

Tiến hành thay đổi các câu hỏi khảo sát theo mức độ từ 1 tới 5:

STT	Câu hỏi	Câu trả lời	Quy đổi
1	**Bạn cảm thấy mức độ tiếp cận thông tin về các quy định và quy chế của nhà trường như thế nào?**	- Rất khó khăn - Khá khó khăn - Trung bình - Dễ dàng - Rất dễ dàng	- 1 - 2 - 3 - 4 - 5

STT	Câu hỏi	Câu trả lời	Quy đổi
2	**Bạn đã đọc và hiểu rõ các văn bản như hướng dẫn học tập, quy chế hành vi sinh viên, hoặc các quy định khác của nhà trường chưa?**	- Chưa đọc - Đã đọc nhưng chưa hiểu - Hiểu một phần - Hiểu tương đối - Hiểu hoàn toàn	- 1 - 2 - 3 - 4 - 5
3	**Trong quá trình học tập, bạn đã nhận được sự hỗ trợ hoặc hướng dẫn về các quy định và quy chế của trường từ giáo viên hoặc cố vấn học tập không?**	- Không bao giờ - Hiếm khi - Đôi khi - Thường xuyên - Luôn luôn	- 1 - 2 - 3 - 4 - 5
4	**Bạn cảm thấy có khó khăn khi tìm hiểu và hiểu rõ các quy định và quy chế của nhà trường không?**	- Rất khó khăn và không thể hiểu - Khá khó khăn và mất nhiều thời gian - Có chút khó khăn nhưng có thể hiểu - Tương đối dễ dàng - Rất dễ dàng và nhanh chóng	- 1 - 2 - 3 - 4 - 5

STT	Câu hỏi	Câu trả lời	Quy đổi
5	**Bạn cảm thấy như thế nào về việc nhà trường cung cấp các công cụ hoặc tài nguyên để giúp bạn tiếp cận và hiểu rõ các văn bản, quy định và quy chế của họ?**	- Rất không hữu ích - Không hữu ích - Trung bình - Hữu ích - Rất hữu ích	- 1 - 2 - 3 - 4 - 5
6	**Bạn đã từng biết đến các học bổng do nhà trường cung cấp chưa?**	- Chưa biết - Biết một ít - Biết một số - Biết hầu hết - Biết tất cả	- 1 - 2 - 3 - 4 - 5
7	**Bạn cảm thấy như thế nào về việc nhà trường thông báo lịch thi học kỳ?**	- Rất không hợp lý - Không hợp lý - Trung bình - Hợp lý - Rất hợp lý	- 1 - 2 - 3 - 4 - 5
8	**Bạn cảm thấy như thế nào về việc nhà trường cung cấp các tài liệu hướng dẫn, chuẩn bị cho kỳ thi?**	- Rất không hợp lý - Không hợp lý - Trung bình - Hợp lý - Rất hợp lý	- 1 - 2 - 3 - 4 - 5
9	**Bạn đã từng gặp khó khăn trong việc nhận biết thông**	- Luôn gặp khó khăn	- 1 - 2 - 3

	báo lịch thi học kỳ của nhà trường chưa?	- Thường xuyên gặp khó khăn - Đôi khi gặp khó khăn - Hiếm khi gặp khó khăn - Không bao giờ gặp khó khăn	- 4 - 5

Bảng 3. 17. Các câu hỏi và câu trả lời khi quy đổi

Sau đó tiến hành ghép thêm điểm GPA với từng sinh viên thu thập được và tính TBC câu hỏi:

STT	**Tên trường**
1	Mã Sinh Viên
2	Họ Tên
3	TBC Câu hỏi
4	Điểm GPA
5	CH1
6	CH2
7	CH3
8	CH4
9	CH5
10	CH6
11	CH7
12	CH8

STT	Tên trường
13	CH9

Bảng 3. 18. Gộp điểm GPA và tổng quan các trường sau khi làm sạch

3.5. Phân chia dữ liệu huấn luyện

Dữ liệu về điểm số của sinh viên được chia thành ba chuyên ngành khác nhau: Phát triển Phần mềm, Hệ thống nhúng và IoT, và Hệ thống Thông tin. 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.

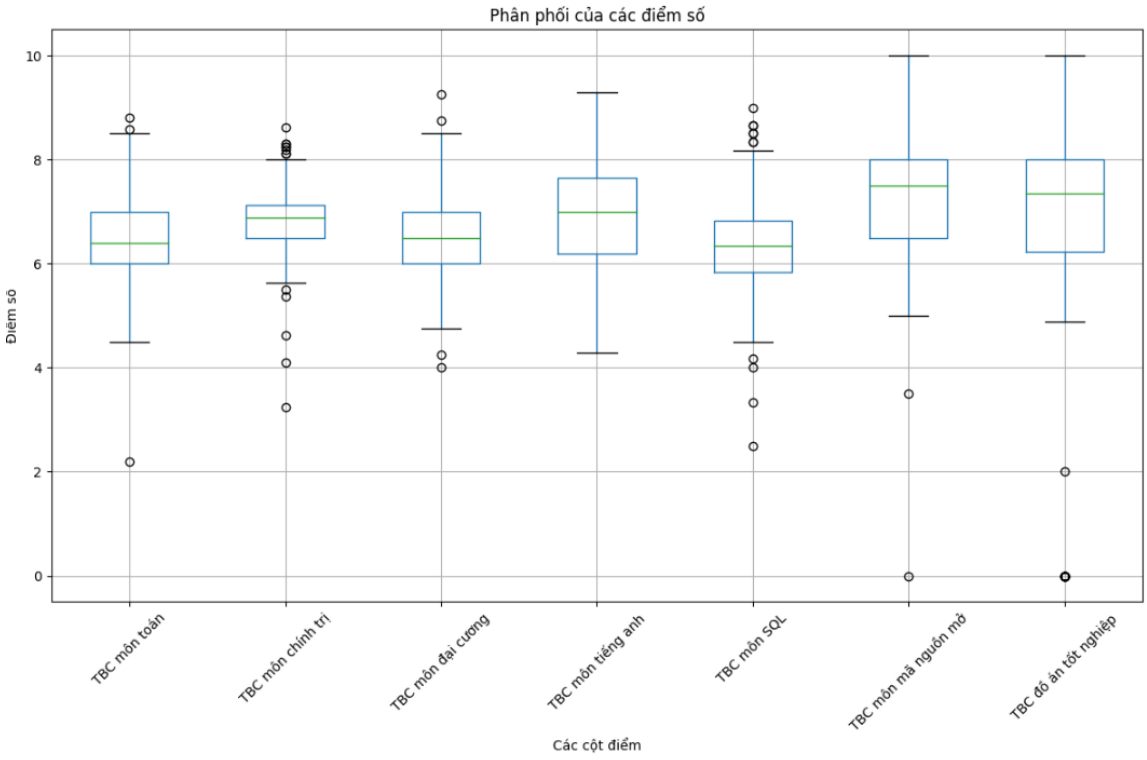
Dữ liệu khảo sát mức độ quan tâm của sinh viên được chia thành 9 câu hỏi và điểm GPA. Sau đó phân chia dữ liệu thành tập huấn luyện và tập kiểm tra.

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.

3.6. Mô hình dự đoán điểm chuyên ngành

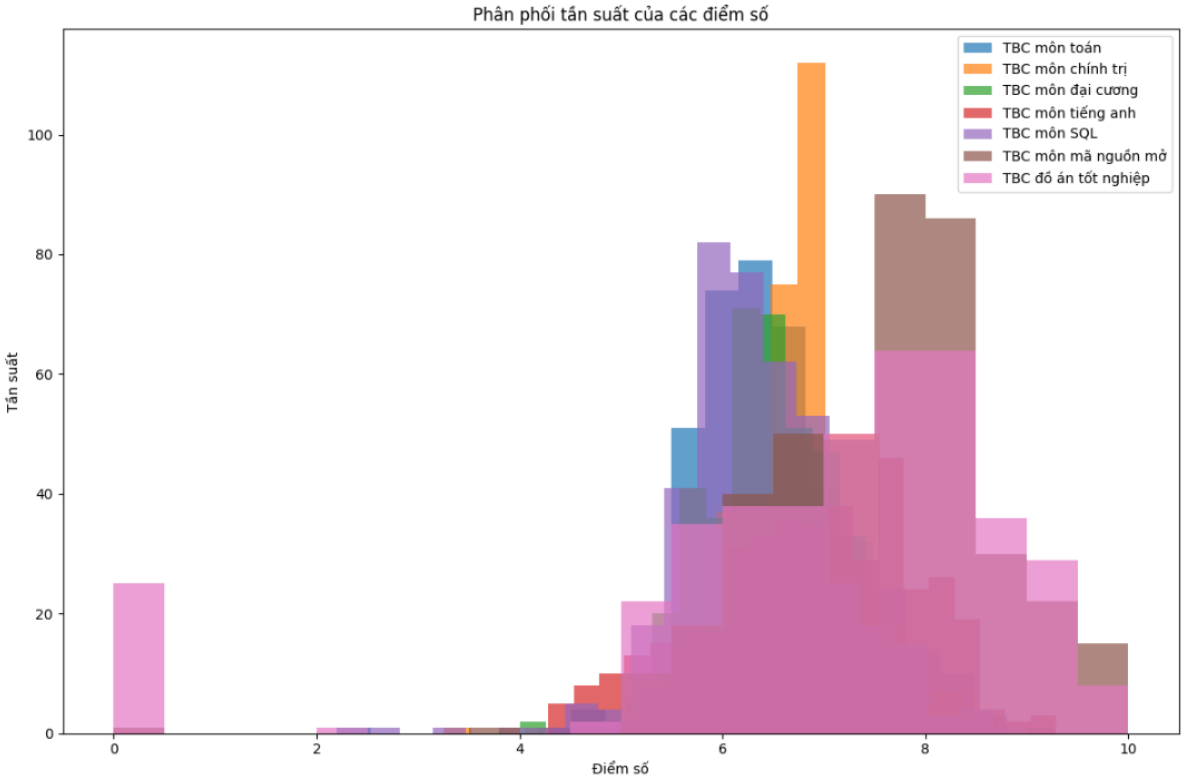
3.6.1. Trực quan hóa dữ liệu qua boxplot



Hình 3. 1. Trực quan hóa dữ liệu qua Boxplot TBC các môn học

Biểu đồ boxplot mô tả phân phối điểm số qua bảy nhóm khác nhau. Trục x biểu diễn các nhóm, mỗi nhóm có một hộp đại diện cho phạm vi tứ phân vị (50% điểm số giữa). Đường bên trong hộp là điểm trung bình của nhóm. Các 'râu' chỉ ra biến động ngoài tứ phân vị trên và dưới, thể hiện phạm vi dữ liệu. Trục y biểu diễn điểm số từ 0 đến 10, cho thấy điểm số cao nhất là 10 và thấp nhất là 0.

3.6.2. Trực quan hóa điểm qua Histogram

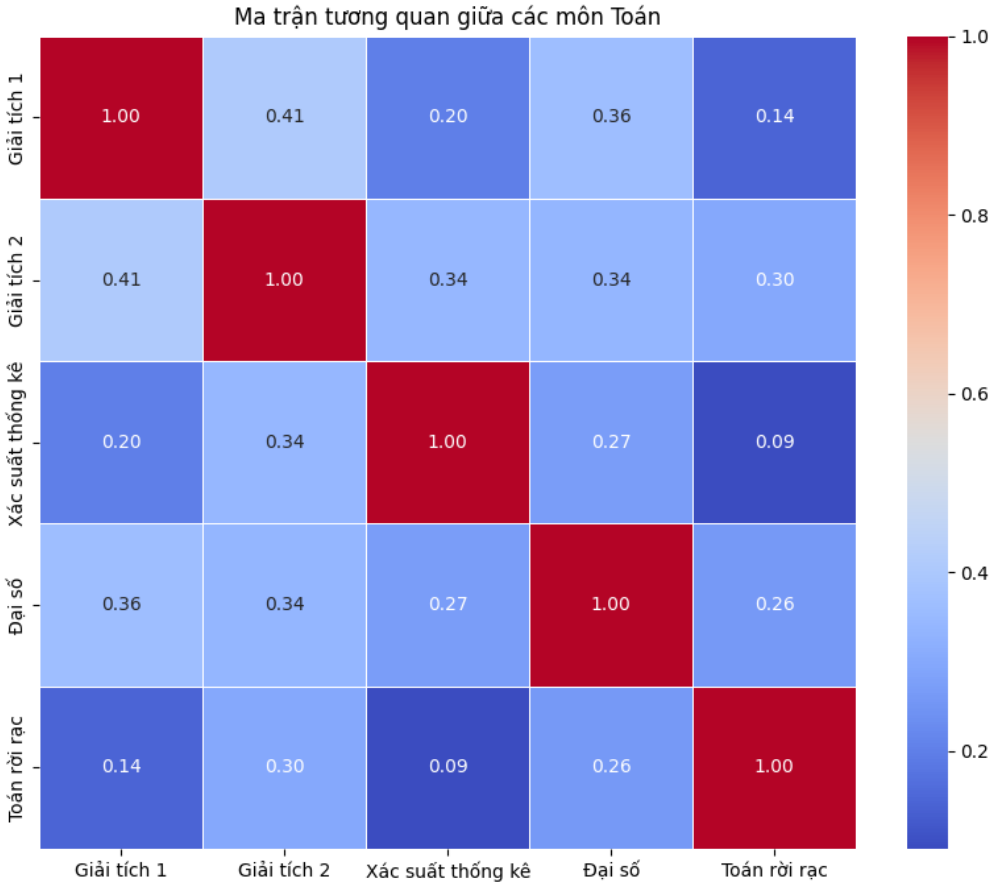


Hình 3. 2. Trực quan háo điểm qua Histogram

Biểu đồ "Phân phối tần suất của các điểm số" thể hiện phân phối tần suất các loại điểm khác nhau, từ "TBC môn toán" đến "TBC cả đời nghề nghiệp". Trục x biểu diễn "Điểm số" và trục y là "% tổng số". Mỗi loại điểm được thể hiện bằng một màu sắc khác nhau.

Phần lớn sinh viên có điểm số trong khoảng từ 4 đến 8, với tần suất giảm đáng kể dưới 4 và trên 8. Điều này cho thấy đa số sinh viên có điểm số ổn định, và chỉ một số ít có điểm số thấp hoặc cao hơn mức trung bình. Biểu đồ này giúp đánh giá hiệu suất học tập chung của sinh viên và xác định các khu vực cần cải thiện.

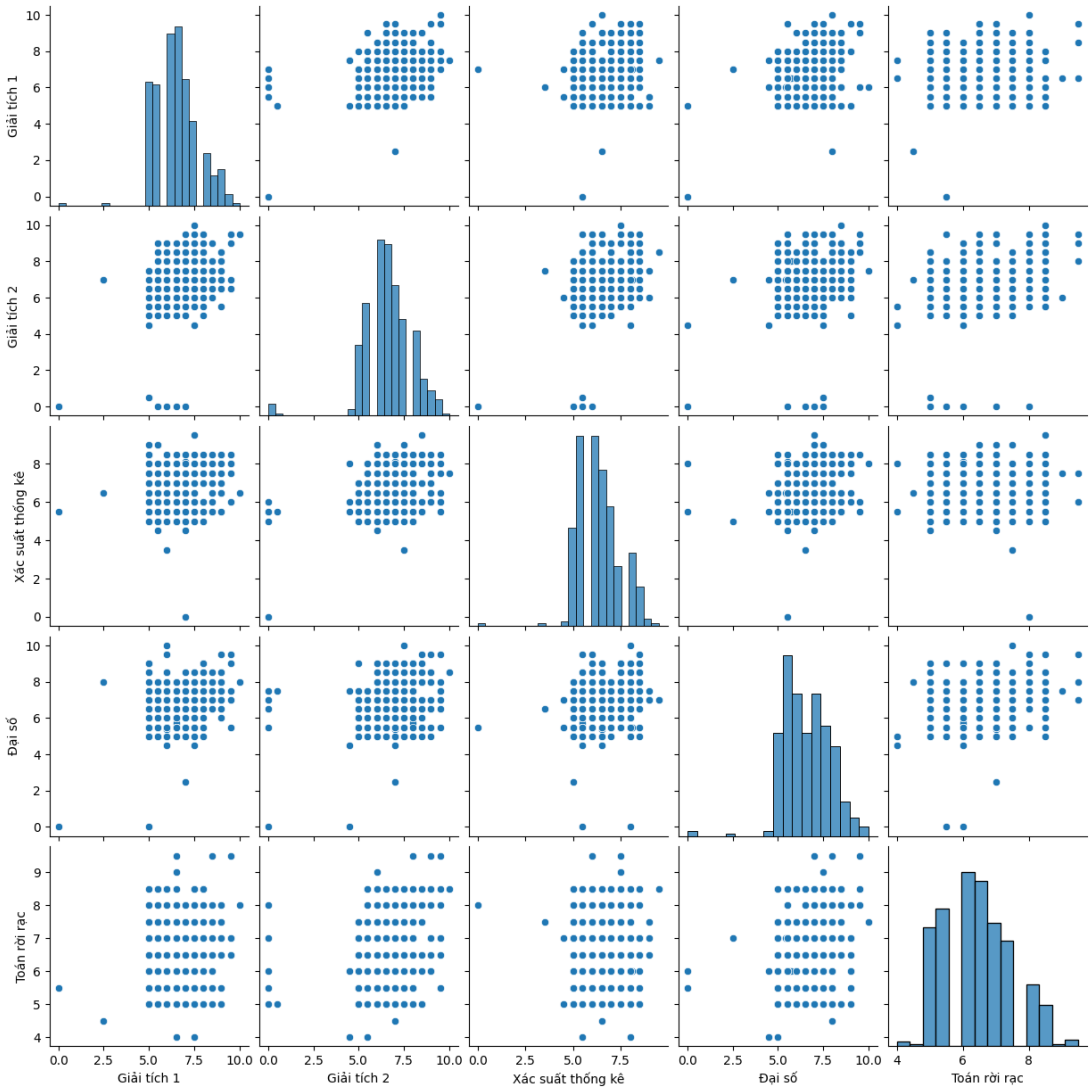
3.6.3. Xét tương quan các môn toán



Hình 3. 3. Tương quan giữa các môn toán

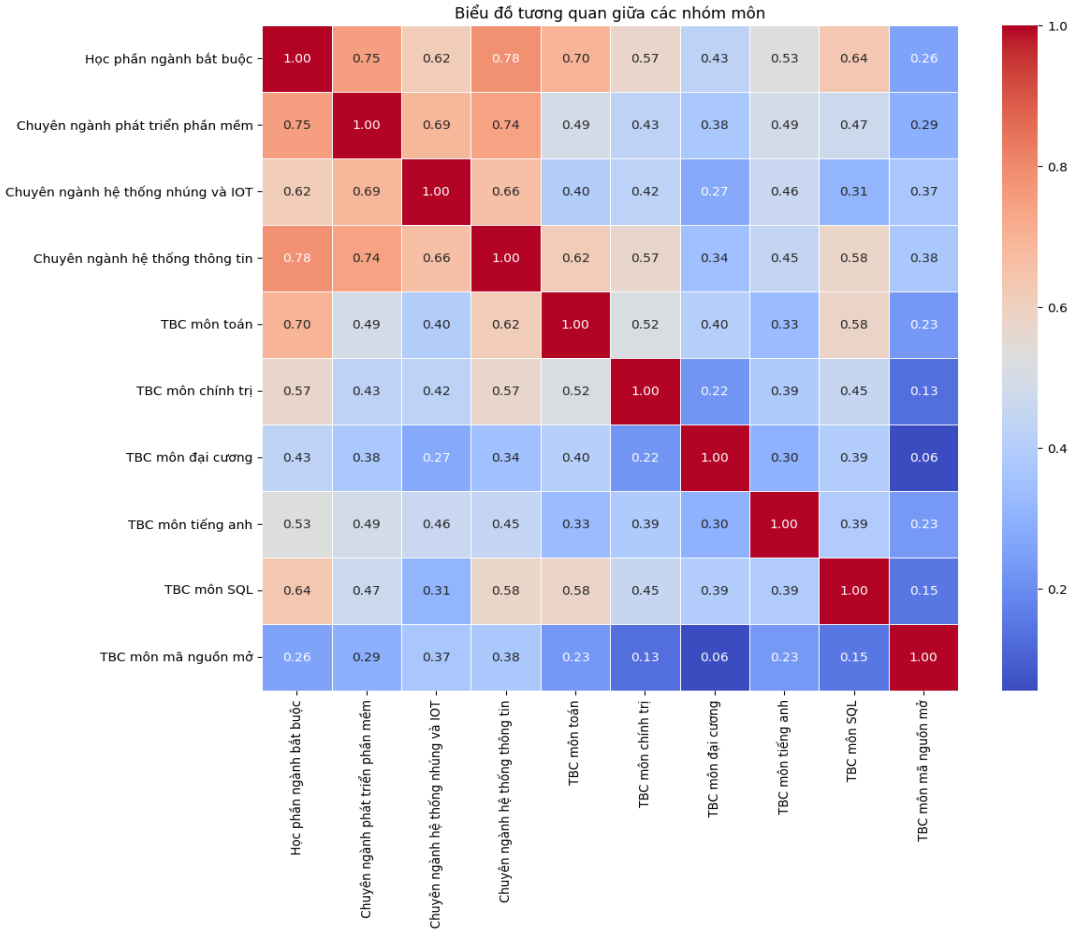
Mỗi ô trong ma trận tương quan đại diện cho hệ số tương quan giữa hai môn học tương ứng. Hệ số tương quan là một giá trị trong khoảng từ -1 đến 1, thể hiện mức độ và hướng của mối liên hệ giữa hai môn học. Màu sắc của mỗi ô cũng thể hiện mức độ và hướng của mối tương quan. Màu đỏ cho thấy mối tương quan tích cực, màu xanh cho thấy mối tương quan tiêu cực, và màu trắng cho thấy không có tương quan đáng kể..

Biểu đồ mối quan hệ giữa các môn toán:



Hình 3. 4. Mối quan hệ giữa các môn toán

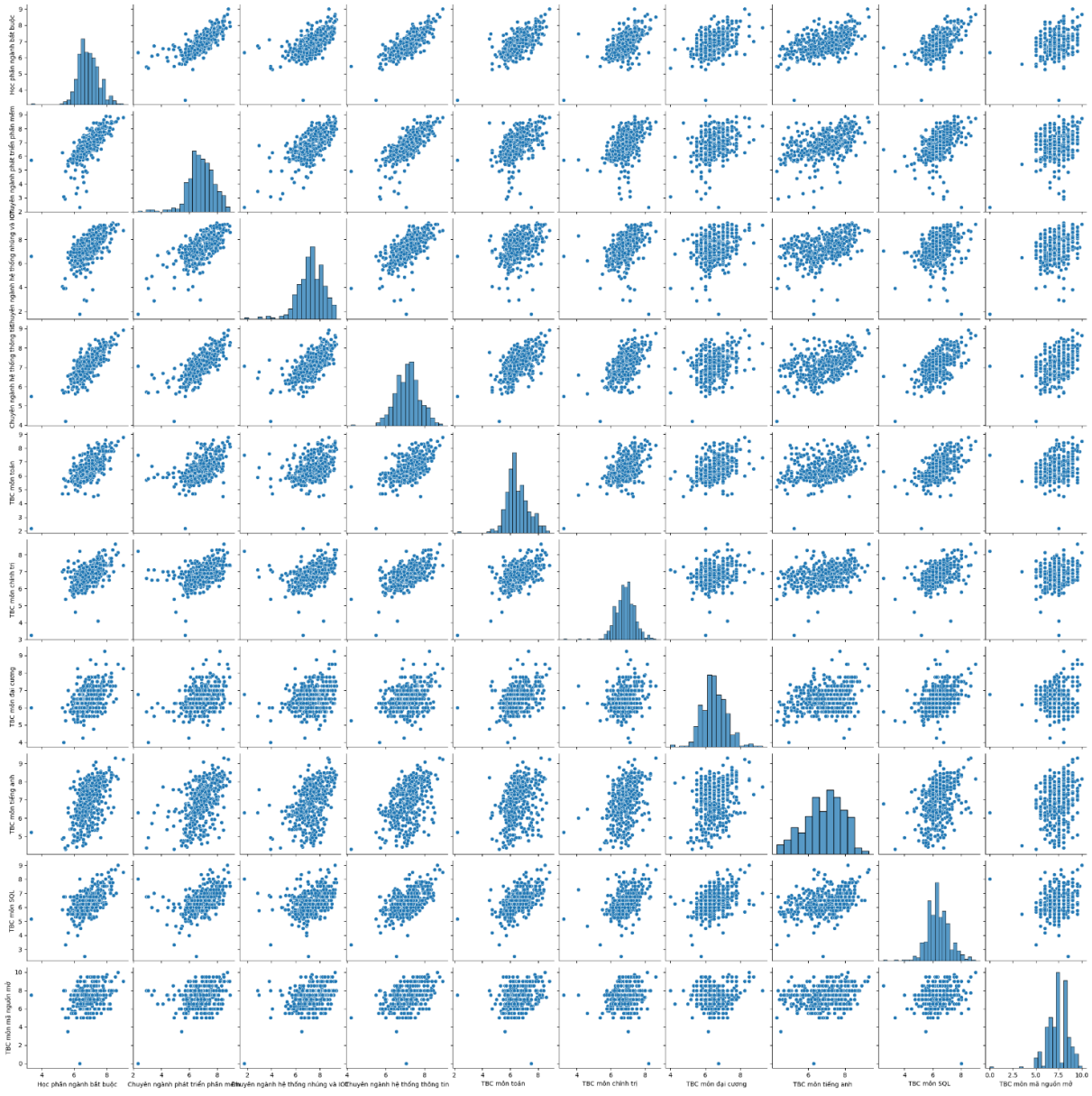
3.6.4. Tính tương quan các môn học



Hình 3. 5. Tương quan các môn học

Biểu đồ nhiệt này phân tích mối tương quan giữa các môn học. Mỗi ô đại diện cho hệ số tương quan giữa hai môn học, từ -1 đến 1. Màu đỏ thể hiện mối tương quan tích cực, xanh là tiêu cực, và trắng là không đáng kể. Ví dụ, một ô đỏ sâu tại giao điểm của "TBC môn tiếng anh" và "TBC môn mã nguồn mở" cho thấy mối tương quan tích cực mạnh.

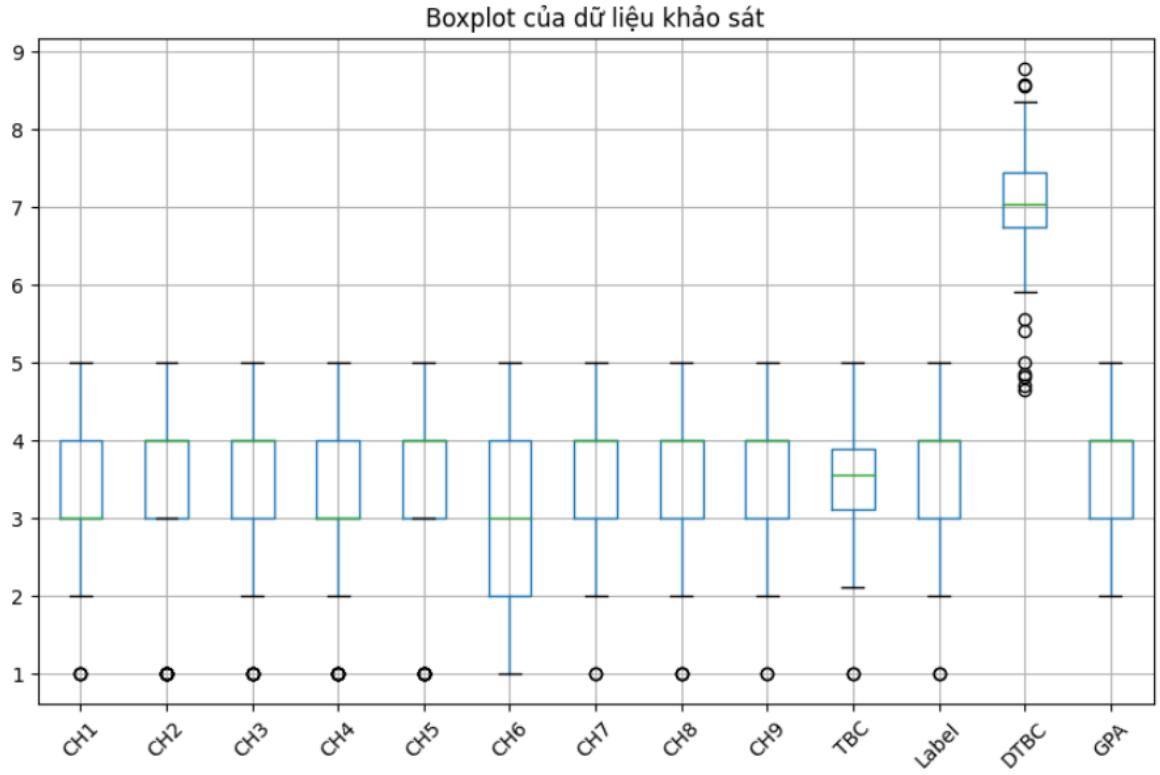
Biểu đồ mối quan hệ giữa các môn học:



Hình 3. 6. Mối quan hệ giữa các môn học

3.7. Mô hình dự đoán điểm GPA với dữ liệu khảo sát

3.7.1. Trực quan hóa dữ liệu qua Boxplot



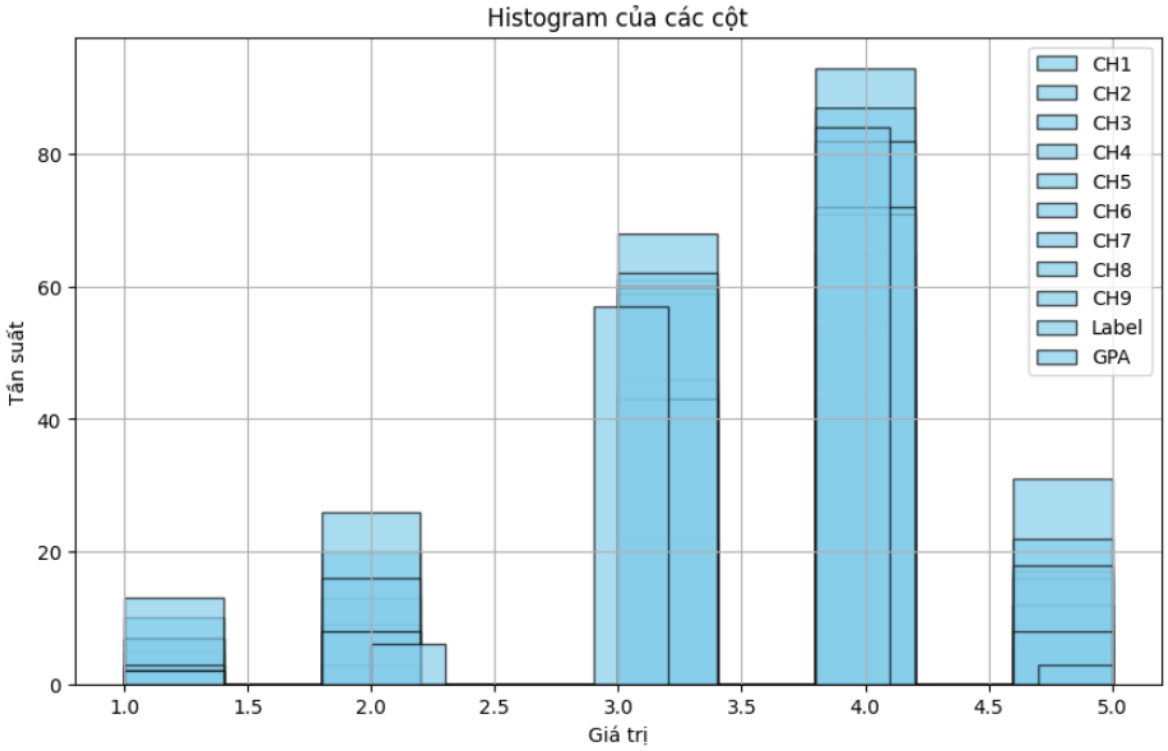
Hình 3. 7. Biểu đồ trực quan hóa dữ liệu qua Boxplot

Biểu đồ boxplot mô tả phân phối của các điểm số qua các nhóm khác nhau.

Trên trục x của biểu đồ, ta thấy bảy nhóm khác nhau được biểu diễn. Mỗi nhóm này có một hộp tương ứng trên biểu đồ, đại diện cho phạm vi tứ phân vị, hay 50% điểm số giữa của nhóm đó. Đường bên trong mỗi hộp là điểm số trung bình (median) của nhóm đó. Các ‘râu’ kéo dài từ hộp chỉ ra biến động ngoài tứ phân vị trên và dưới, cung cấp thông tin về phạm vi của dữ liệu.

Trục y của biểu đồ biểu diễn phạm vi của điểm số, từ 0 đến 10. Điều này cho thấy rằng điểm số cao nhất mà một sinh viên có thể đạt được là 10, trong khi điểm số thấp nhất là 0.

3.7.2. Trực quan hóa dữ liệu qua Histogram

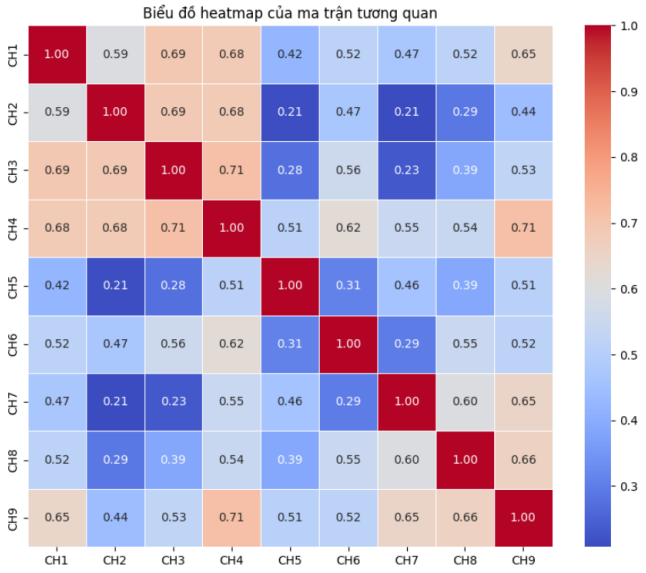


Hình 3. 8. Trực quan hóa dữ liệu của các dữ liệu khảo sát qua Histogram

Biểu đồ này mô tả sự phân bố và tần suất của dữ liệu trong các danh mục khác nhau, từ CH1 đến CH9, Label và GPA.

Cụ thể, trục ngang của biểu đồ, được gọi là “Giá trị”, biểu diễn các danh mục từ CH1 đến CH9, Label và GPA. Trục dọc, được gọi là “Tần suất”, đo lường tần suất xuất hiện của các giá trị. Các cột biểu đồ với chiều cao khác nhau cho thấy tần suất xuất hiện của từng giá trị trong mỗi danh mục.

3.7.3. Xét tính tương quan các câu hỏi

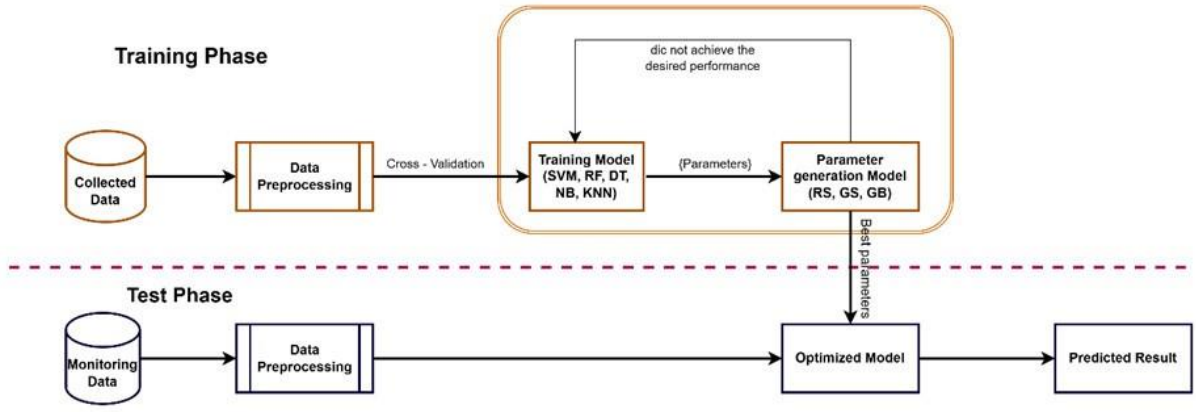


Hình 3. 9. Xét tính tương quan các câu hỏi

Biểu đồ heatmap biểu diễn một ma trận tương quan. Biểu đồ này là một lưới vuông với các ô được tô màu theo các mức độ khác nhau của màu đỏ và xanh dương, thể hiện mức độ tương quan khác nhau giữa các biến được gắn nhãn từ CH1 đến CH9 trên cả trục x và trục y. Màu đỏ biểu thị tương quan dương, trong khi màu xanh dương biểu thị tương quan âm. Mỗi ô chứa một giá trị số biểu thị mức độ mạnh của tương quan, với 1.00 là tương quan dương mạnh nhất (màu đỏ) và các giá trị gần -1 biểu thị tương quan âm mạnh hơn (màu xanh dương).

Chương 4
TỐI ƯU HÓA MÔ HÌNH KỸ THUẬT HỌC MÁY

4.1. Tổng quan về mô hình



Hình 4. 1. Các phase của mô hình

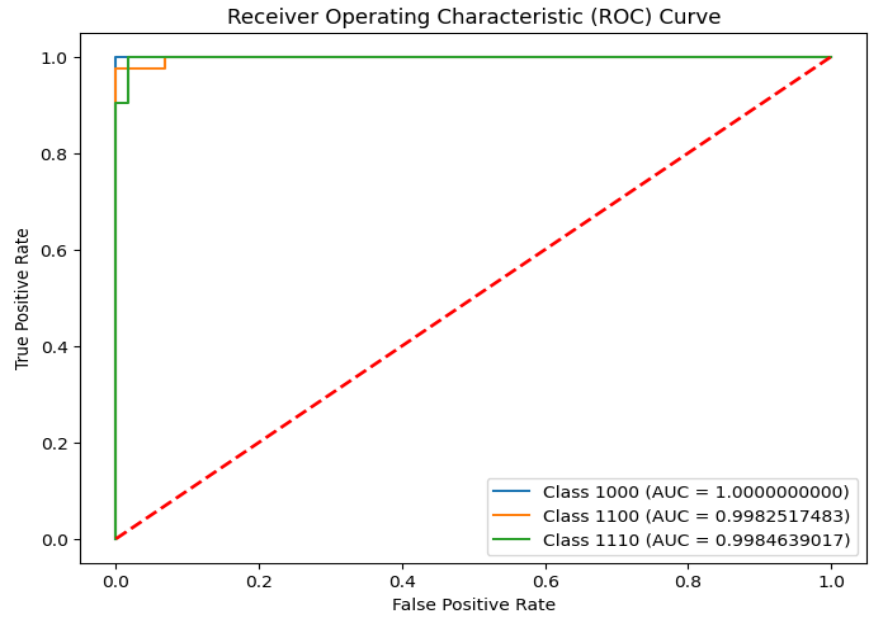
Giai đoạn Huấn luyện: Quá trình bắt đầu với việc thu thập dữ liệu từ nhiều nguồn khác nhau. Dữ liệu này sau đó được tiền xử lý để làm sạch và chuẩn bị cho quá trình huấn luyện. Tiếp theo, ta thực hiện kiểm định chéo (Cross-Validation) để đánh giá hiệu suất của mô hình trên một phần dữ liệu chưa được sử dụng trong quá trình huấn luyện. Mô hình sau đó được huấn luyện sử dụng các thuật toán như SVM, RF, DT, NB, KNN. Cuối cùng, ta sử dụng các mô hình tạo tham số như RS, GS, GB để tạo ra mô hình tối ưu.

Giai đoạn Kiểm thử: Giai đoạn này bắt đầu với việc giám sát và thu thập dữ liệu để kiểm thử mô hình. Dữ liệu này cũng được tiền xử lý để làm sạch và chuẩn bị cho quá trình kiểm thử. Dữ liệu sau khi tiền xử lý được đưa vào mô hình tối ưu để dự đoán. Mô hình tối ưu sẽ cho ra kết quả dự đoán. Nếu kết quả dự đoán không đạt được hiệu suất mong muốn, quá trình sẽ quay lại giai đoạn huấn luyện để cải thiện mô hình. Đây là một quy trình lặp đi lặp lại cho đến khi đạt được hiệu suất mong muốn.

4.2. Tối ưu hóa mô hình

4.2.1. Tối ưu hóa các mô hình của bộ dữ liệu dự đoán gợi ý chuyên ngành

Mô hình SVM



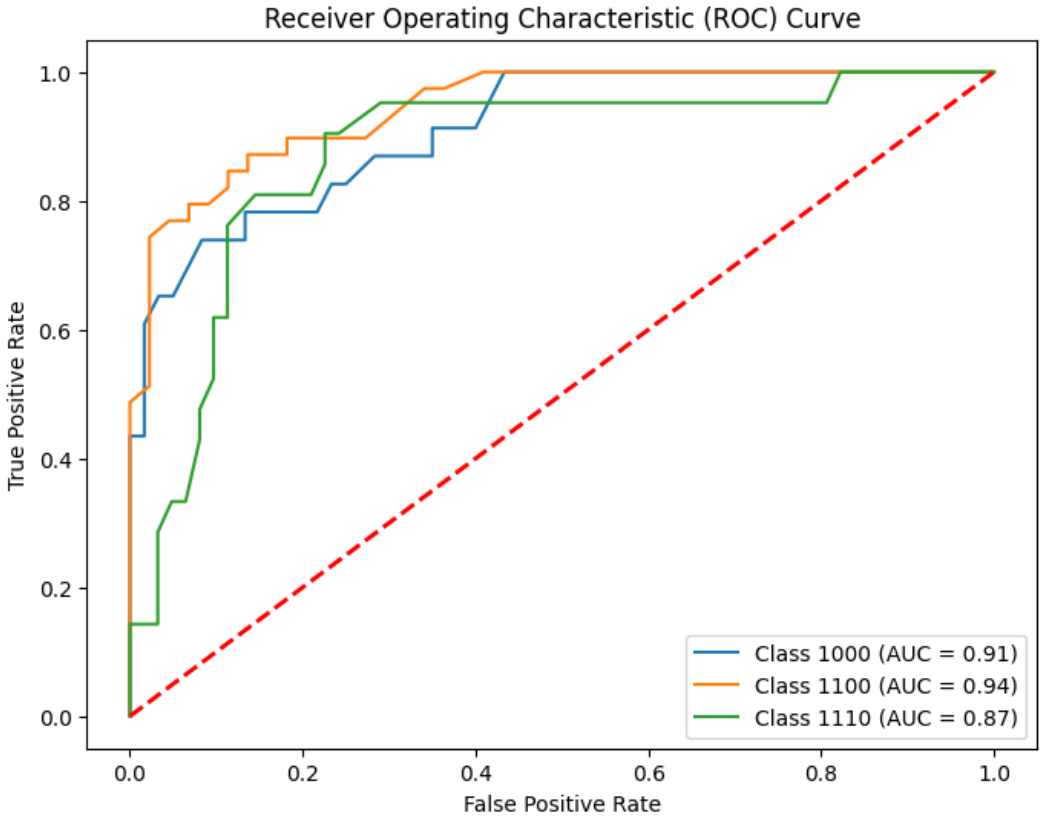
Classification Report:

	precision	recall	f1-score	support
1000	1.00	0.91	0.95	23
1100	0.93	0.97	0.95	39
1110	0.95	0.95	0.95	21

Hình 4. 2. Mô hình ROC - SVM dữ liệu dự đoán chuyên ngành

Biểu đồ Receiver Operating Characteristic (ROC) thể hiện mối quan hệ giữa Tỉ lệ Dương tính Thực (True Positive Rate - TPR) trên trục y và Tỉ lệ Dương tính Giả (False Positive Rate - FPR) trên trục x. Trong biểu đồ này, có hai đường biểu diễn cho hai lớp khác nhau: Lớp 1000 và Lớp 1110. Lớp 1000 có Diện tích dưới Đường cong (Area Under Curve - AUC) là 1.00000000000, được biểu diễn bằng đường liền màu xanh lá cây nằm sát hai trục của biểu đồ, cho thấy mô hình phân loại hoàn hảo cho lớp này. Trong khi đó, Lớp 1110 có AUC là 0.9984693017, được biểu diễn bằng đường nét đứt màu đỏ hơi lệch so với đường cong hoàn hảo của Lớp 1000, phản ánh một hiệu suất rất cao nhưng không hoàn hảo tuyệt đối. Dưới biểu đồ ROC là bảng báo cáo phân loại (classification report), cung cấp các chỉ số như độ chính xác (precision), độ nhớ (recall), điểm F1 (f1-score), và hỗ trợ (support) cho cả hai lớp trên, giúp đánh giá chi tiết hơn hiệu suất của mô hình trên từng lớp.

Mô hình random forest



Classification Report:

	precision	recall	f1-score	support
1000	0.72	0.78	0.75	23
1100	0.82	0.85	0.84	39
1110	0.72	0.62	0.67	21

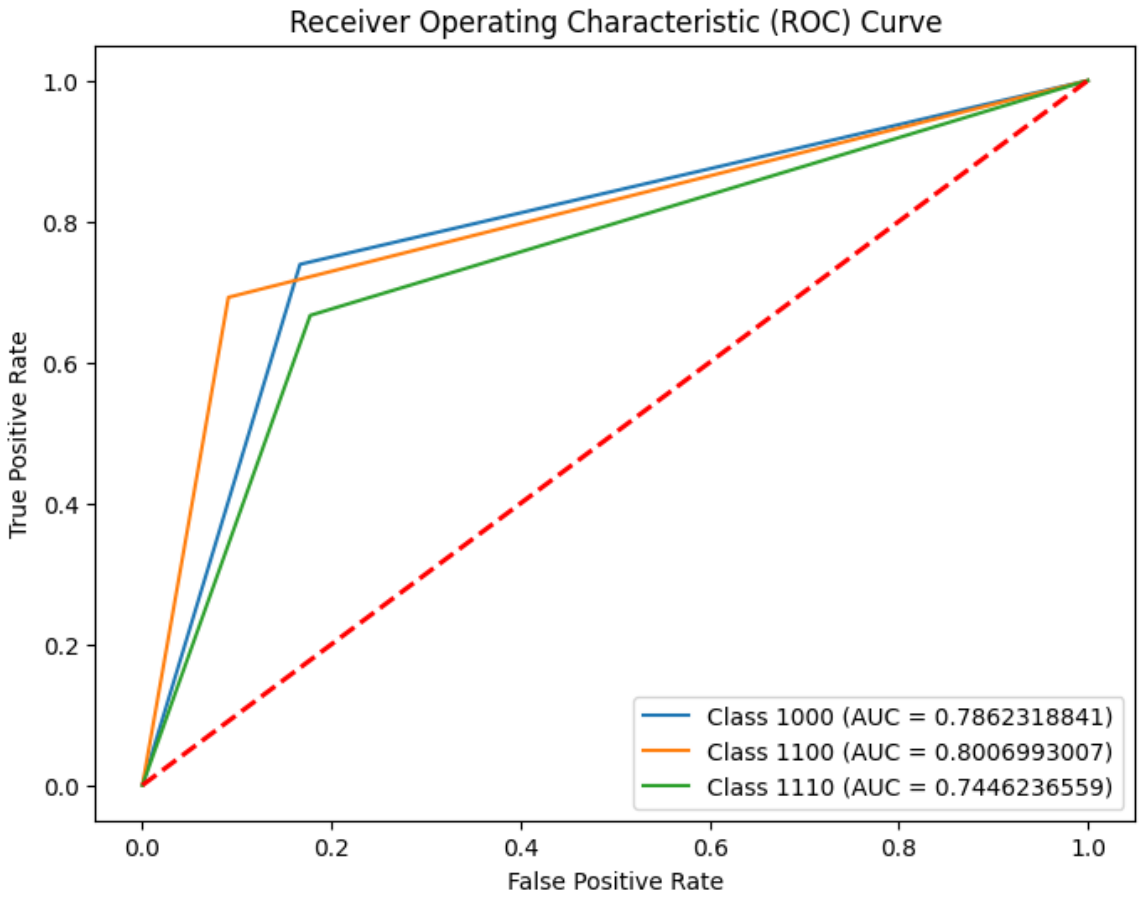
Hình 4. 3. Mô hình ROC - Random forest của dữ liệu dự đoán chuyên ngành

Ảnh này hiển thị một Đường cong Đặc trưng Hoạt động của Người nhận (ROC), là một biểu đồ đồ họa được sử dụng để thể hiện khả năng chẩn đoán của một hệ thống phân loại nhị phân khi ngưỡng phân biệt của nó thay đổi. Biểu đồ này bao gồm ba đường cong ROC cho các lớp khác nhau, mỗi lớp được biểu diễn bằng các đường màu xanh, xanh lá cây và đỏ. Đường màu xanh biểu thị cho 'Lớp 1000' với Diện tích Dưới Đường cong (AUC) là 0.91, đường màu xanh lá cây biểu thị cho 'Lớp 1100' với AUC là 0.94, và đường màu đỏ biểu thị cho 'Lớp 118' với AUC là 0.87.

Dưới biểu đồ đường cong ROC là một bảng báo cáo phân loại dưới dạng bảng, hiển thị các chỉ số như độ chính xác, độ nhớ, điểm F1 và hỗ trợ cho ba lớp được đánh dấu là 'Nhãn thực tế', 'Nhãn dự đoán' và 'Tổng cộng'. Các giá trị cụ thể là: cho Nhãn thực tế - độ chính xác: 0.82, độ nhớ: 0.78, điểm

F1: 0.75; cho Nhãn dự đoán - độ chính xác: 0.85, độ nhớ: 0.84; cho Tổng cộng - hỗ trợ: 39.

Mô hình Decesion Tree



Classification Report:

	precision	recall	f1-score	support
1000	0.63	0.74	0.68	23
1100	0.87	0.69	0.77	39
1110	0.56	0.67	0.61	21

Hình 4. 4. Mô hình ROC - Decesion Tree của data dự đoán chuyên ngành

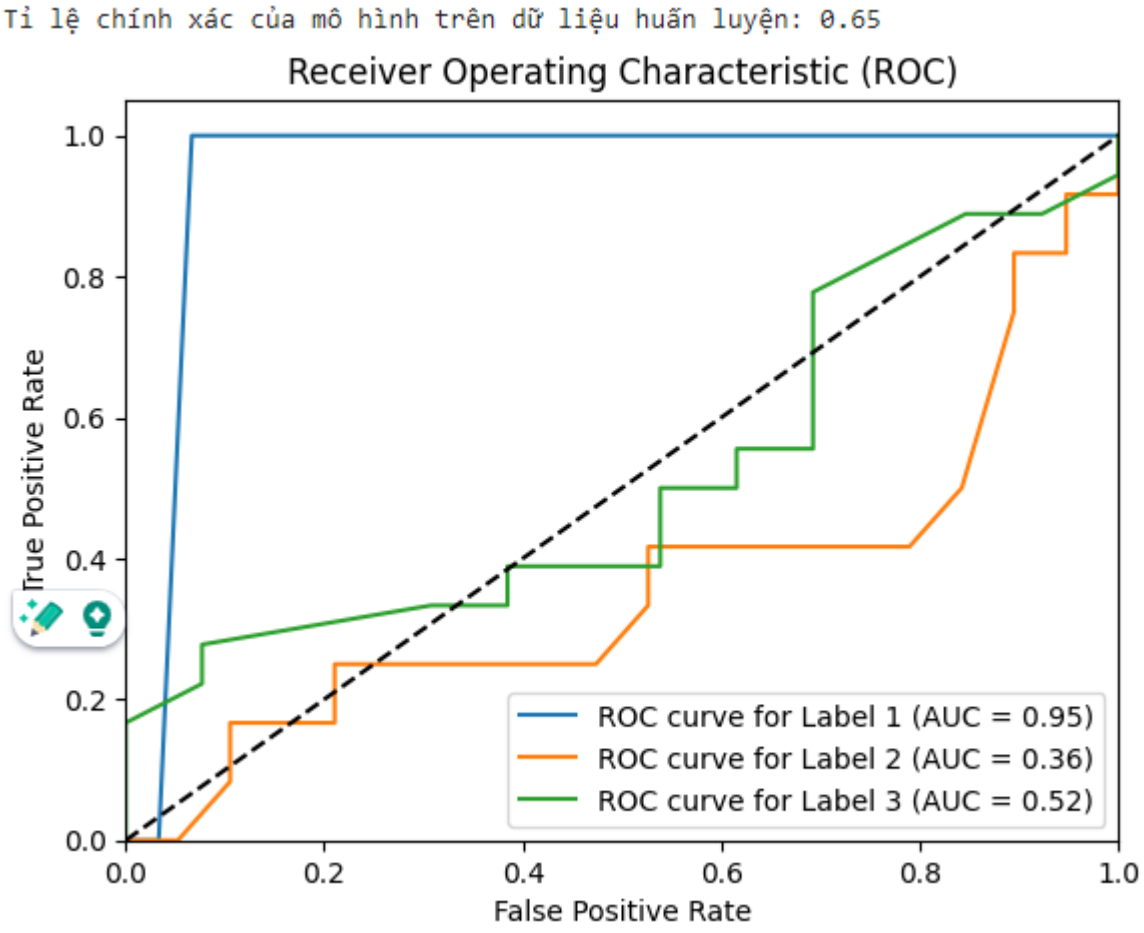
Đường cong Đặc trưng Hoạt động của Người nhận (ROC) là một công cụ quan trọng trong đánh giá hiệu suất của các mô hình phân loại. Biểu đồ ROC biểu diễn khả năng phân biệt giữa các lớp dự đoán bằng cách thay đổi ngưỡng phân biệt. Trục x biểu thị tỷ lệ dương tính giả (False Positive Rate - FPR), trong khi trục y biểu thị tỷ lệ dương tính thật (True Positive Rate - TPR).

Đường ROC là một đường cong liên tục, và mô hình tốt sẽ có diện tích dưới đường cong (AUC) gần với 1. Điều này cho thấy mô hình có khả năng

phân biệt tốt giữa các lớp. Nếu AUC gần 0.5, mô hình không có khả năng phân biệt tốt hơn ngẫu nhiên. Điều này có nghĩa là mô hình không có sự phân biệt đáng kể giữa các lớp và không thể dự đoán tốt hơn so với việc đoán ngẫu nhiên. Do đó, việc phân tích ROC và AUC là quan trọng để đánh giá hiệu suất của một mô hình phân loại.

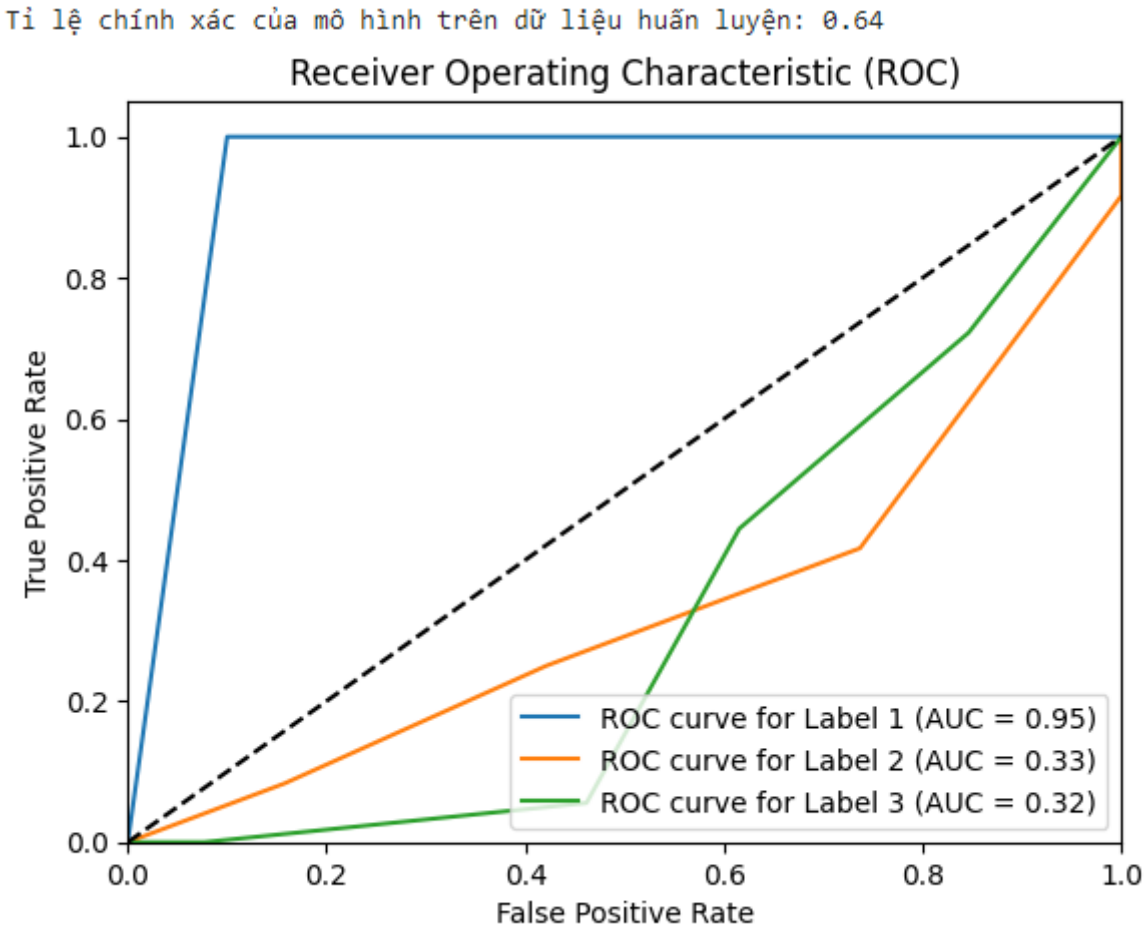
4.2.2. Tối ưu hóa các mô hình của bộ dữ liệu khảo sát mức độ quan tâm

Mô hình SVM



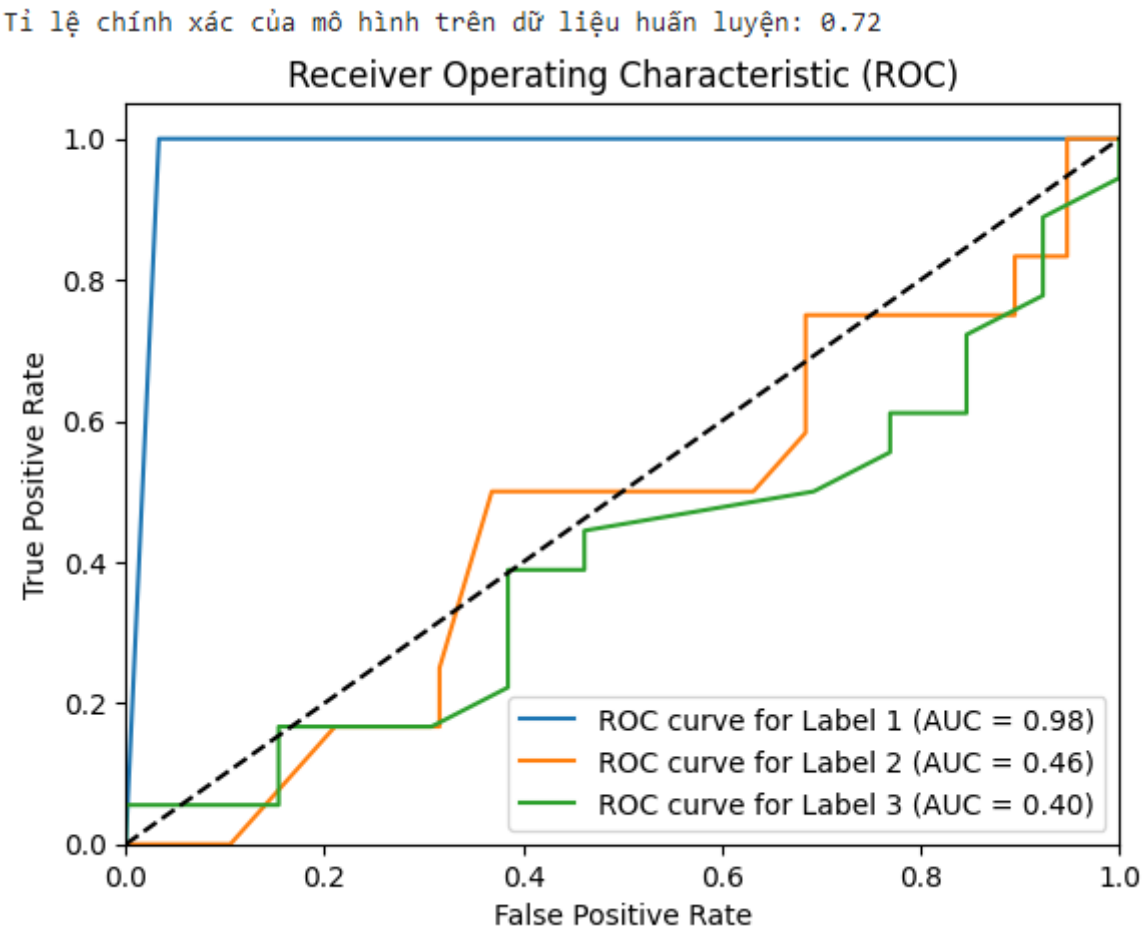
Hình 4. 5. Mô hình ROC – SVM của data khảo sát mức độ quan tâm

Mô hình KNN



Hình 4. 6. Mô hình ROC – Hồi quy KNN của data khảo sát mức độ quan tâm

Mô hình Random forest



Hình 4. 7. Mô hình ROC – Random Forest của data khảo sát mức độ quan tâm

4.3. Lựa chọn mô hình học máy phù hợp

So sánh SVM, Random Forest và Decesicon Tree

Mô hình	SVM				Random Forest				Decision Tree			
Chuyên ngành	PTPM	HTN&IOT	KHMT	TB	PTPM	HTN&IOT	KHMT	TB	PTPM	HTN&IOT	KHMT	TB
Accuracy	1.00	0.93	0.95	**0.95181**	0.72	0.82	0.72	0.771	0.63	0.87	0.56	0.698
F1 Score	0.95	0.95	0.95	**0.95186**	0.75	0.84	0.67	0.769	0.68	0.77	0.61	0.704
Runtime	0.01319				0.4539				0.0093			

Hình 4. 8. So sánh các mô hình SVM – Random Forest – Decision Tree

Dựa vào bảng so sánh hiệu suất của ba mô hình học máy, ta có thể nhận thấy các thông số cụ thể như sau: SVM có độ chính xác là 0.95 và F1 Score là 0.95, trong khi Random Forest có độ chính xác là 0.93 và F1 Score là 0.75, và Decision Tree có độ chính xác là 0.95 và F1 Score là 0.67. Tuy

nhiên, SVM có thời gian chạy lâu hơn so với hai mô hình còn lại, đặc biệt là so với Decision Tree. Nhìn chung, SVM và Decision Tree có hiệu suất tương đối tốt với độ chính xác cao và điểm F1 Score khá ổn. Tuy nhiên, quyết định chọn mô hình nào cho ứng dụng thực tế cần xem xét không chỉ độ chính xác mà còn thời gian chạy, đặc biệt khi độ chính xác và F1 Score của SVM và Decision Tree khá gần nhau.

Mô hình tối ưu tham số của 2 mô hình SVM và RF

		Prec1	Prec2	Prec3	F1	Runtime
SVM	Grid Search	1.00	0.93	1.00	0.9638	1.8413
	Random Seach	1.00	0.93	1.00	0.9638	0.9982
	GadientBoosting	0.89	0.85	0.74	0.8309	106.7181
RF	Grid Search	0.72	0.82	0.72	0.7690	0.7753
	Random Seach	0.75	0.80	0.73	0.7692	0.0967
	GadientBoosting	0.89	0.86	0.73	0.8313	47.3475

Hình 4. 9. Bảng so sánh mô hình tối ưu của SVM và RF

Hình ảnh hiển thị một bảng với các chỉ số hiệu suất cho các mô hình học máy khác nhau. Có hai loại mô hình chính là Support Vector Machine (SVM) và Random Forest (RF), mỗi mô hình được kiểm tra với ba phương pháp điều chỉnh siêu tham số khác nhau: Grid Search, Random Search và Gradient Boosting. Bảng bao gồm năm cột được ghi nhãn là Prec1, Prec2, Prec3, F1 và Runtime. Mỗi hàng đại diện cho kết quả của một sự kết hợp cụ thể giữa loại mô hình và phương pháp điều chỉnh.

Kết quả cho thấy rằng SVM với cả Grid Search và Random Search đều có độ chính xác cao (Prec1 = 1.00) và F1 Score ổn định (0.9638), nhưng SVM với Grid Search có thời gian chạy ít hơn rất nhiều so với SVM với Random Search (1.8413 giây so với 9.0982 giây). Trong khi đó, Random Forest cho thấy các kết quả khác nhau tùy thuộc vào phương pháp điều chỉnh siêu tham số được sử dụng. RF với Grid Search và Random Search có độ chính xác và F1 Score khá gần nhau, tuy nhiên, thời gian chạy của Random Forest với Random Search (0.0967 giây) nhanh hơn nhiều so với Grid Search (0.7753 giây). Cuối cùng, RF với Gradient Boosting cho

thấy độ chính xác và F1 Score tương đối cao, nhưng thời gian chạy lâu hơn nhiều so với các phương pháp điều chỉnh khác (47.3475 giây).

Chương 5
CHẠY MÔ HÌNH VÀ ĐÁNH GIÁ KẾT QUẢ

5.1. Giới thiệu thư viện sử dụng

5.1.1. Tkinter

Tkinter là thư viện GUI (Giao diện đồ họa người dùng) tiêu chuẩn của Python, được sử dụng để tạo các ứng dụng giao diện người dùng một cách nhanh chóng và dễ dàng. Tkinter là một gói tích hợp sẵn trong Python, nên không cần cài đặt bổ sung.

Cách sử dụng:

Bước 1: Khởi tạo một ứng dụng Tkinter:

```
import tkinter as tk
root = tk.Tk()
root.title("Ứng dụng Tkinter Đơn giản")
root.geometry("400x300")
```

Bước 2: Tạo các widget (các thành phần giao diện):

```
label = tk.Label(root, text="Chào mừng đến với Tkinter!")
label.pack()
button = tk.Button(root, text="Nhấn vào tôi!", command=lambda: print("Button Clicked"))
button.pack()
```

Bước 3: Chạy vòng lặp sự kiện chính:

```
root.mainloop()
```

5.1.2. Pandas

Pandas là một thư viện mạnh mẽ để phân tích và thao tác dữ liệu, đặc biệt là dữ liệu dạng bảng (tương tự như bảng tính Excel hoặc cơ sở dữ liệu SQL). Pandas cung cấp hai cấu trúc dữ liệu chính: Series và DataFrame.

Cách sử dụng Pandas

Bước 1: Cài đặt Pandas

```
pip install pandas
```

Bước 2: Đọc dữ liệu từ file CSV

```
import pandas as pd
df = pd.read_csv("data.csv")
print(df.head())
```

Bước 3: Thao tác dữ liệu

```
# Lọc dữ liệu
filtered_df = df[df['column_name'] > 10]
# Thêm cột mới
df['new_column'] = df['column1'] + df['column2']
# Nhóm và tính toán
grouped_df = df.groupby('category_column').sum()
```

Bước 4: Xuất dữ liệu ra file CSV

```
df.to_csv("output.csv", index=False)
```

5.1.3. Scikit-learn

Scikit-learn là một thư viện phổ biến cho học máy (machine learning) trong Python. Nó cung cấp các công cụ đơn giản và hiệu quả cho việc phân tích dữ liệu và các mô hình học máy.

Cách sử dụng Scikit-learn

Bước 1: Cài đặt Scikit-learn

```
pip install scikit-learn
```

Bước 2: Tạo và huấn luyện mô hình

```
from sklearn.model_selection import train_test_split

X = df.drop('target', axis=1)

y = df['target']

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

# Chọn mô hình và huấn luyện

from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(n_estimators=100, random_state=42)

model.fit(X_train, y_train)
```

Bước 3: Đánh giá mô hình

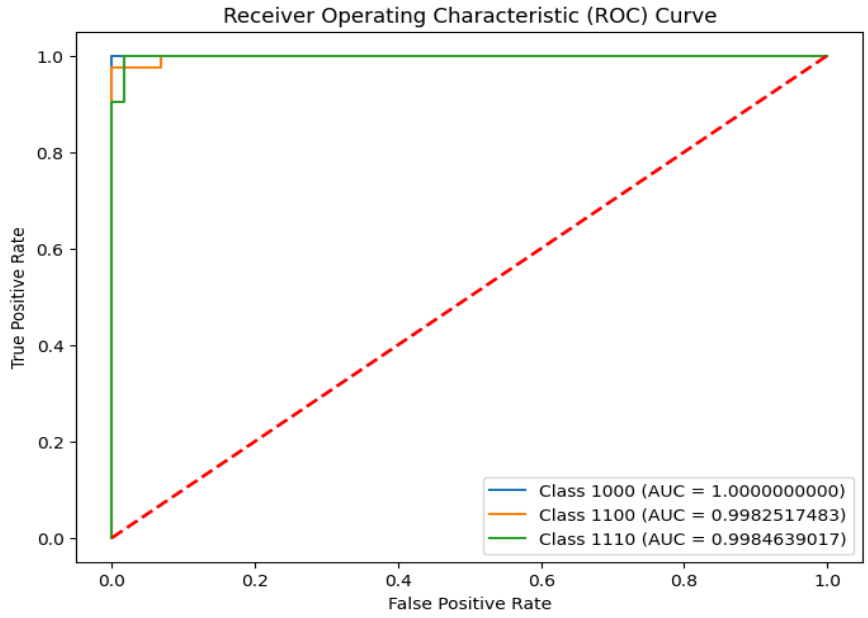
```
from sklearn.metrics import accuracy_score, classification_report

y_pred = model.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))

print("Classification Report:\n", classification_report(y_test, y_pred))
```

Bước 4: Dự đoán với dữ liệu mới

```
new_data = [[...]]  # Dữ liệu mới cần dự đoán
prediction = model.predict(new_data)
print("Prediction:", prediction)
```

5.2. Chạy mô hình dựa vào mô hình tốt nhất với bộ dữ liệu dự đoán gợi ý chuyên ngành

5.1.1. Ứng dụng mô hình qua thuật toán SVM với dữ liệu dự đoán chuyên ngành



Classification Report:

	precision	recall	f1-score	support
1000	1.00	0.91	0.95	23
1100	0.93	0.97	0.95	39
1110	0.95	0.95	0.95	21

Hình 5. 1. Mô hình tốt nhất với bộ dữ liệu dự đoán gợi ý chuyên ngành

Sau khi áp dụng các Biểu đồ Receiver Operating Characteristic (ROC) , Biểu đồ ROC – SVM cho thấy hiệu suất của hai lớp khác nhau: Lớp 1000 và Lớp 1110. Lớp 1000 có Diện tích dưới Đường cong (AUC) là 1.00000000000, biểu thị bằng đường liền màu xanh lá cây, cho thấy một mô hình phân loại hoàn hảo. Trong khi đó, Lớp 1110 có AUC là 0.9984693017, được biểu diễn bằng đường nét đứt màu đỏ, chỉ hơi lệch so với đường cong hoàn hảo của Lớp 1000, nhưng vẫn phản ánh một hiệu suất rất cao.

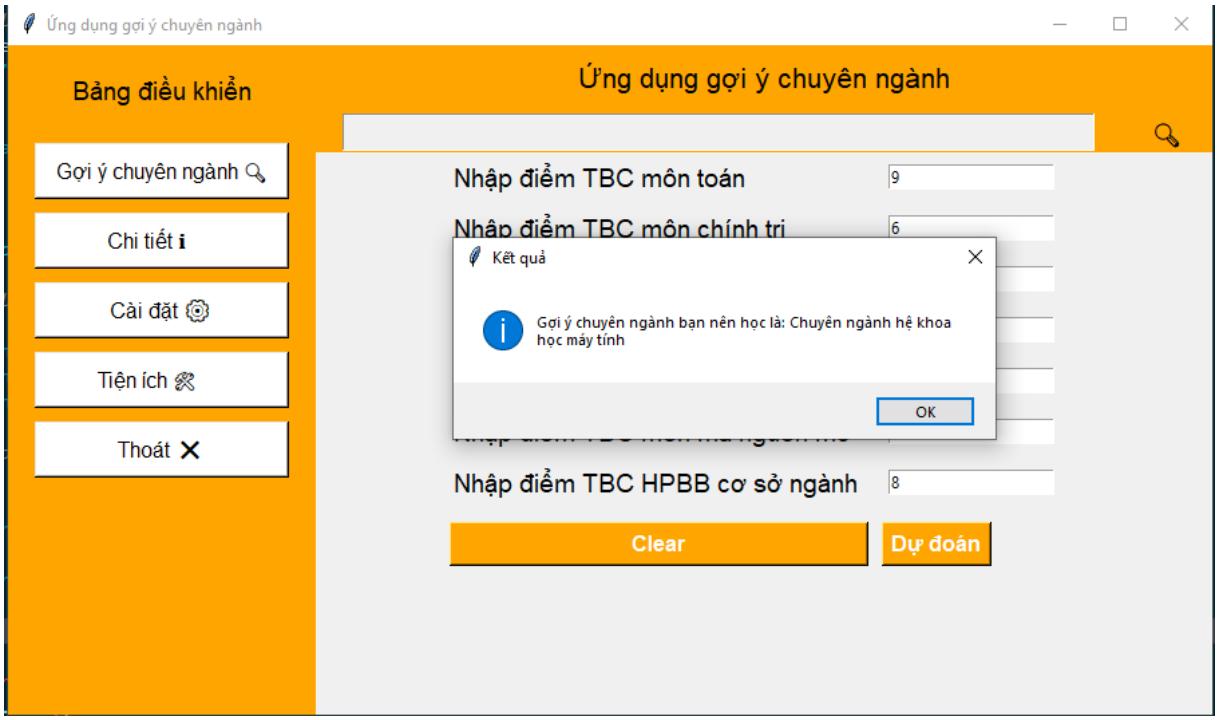
Dưới biểu đồ ROC là bảng báo cáo phân loại (classification report), cung cấp các chỉ số như độ chính xác (precision), độ nhớ (recall), điểm F1 (f1-score) và hỗ trợ (support) cho cả hai lớp trên, giúp đánh giá chi tiết hơn hiệu suất của mô hình trên từng lớp.

So với các mô hình khác như Random Forest (RF), K-Nearest Neighbors (KNN), Naive Bayes và Decision Tree (DC), mô hình này có vẻ là tốt nhất, đặc biệt là khi mô hình này có AUC gần bằng 1.0, cho thấy khả năng phân loại tốt hơn hẳn so với các mô hình khác. Tuy nhiên, để đánh giá chính xác và xác định mô hình tốt nhất, cần phải xem xét cả các chỉ số khác như precision, recall và F1-score trên bảng báo cáo phân loại.

5.1.2. Xây dựng ứng dụng tự động gợi ý lựa chọn chuyên ngành dựa vào điểm cơ sở

Sau khi sử dụng mô hình SVM với tỉ lệ chính xác cao với các mô hình còn lại tiến hành xây dựng ứng dụng tự động gợi ý chuyên ngành dựa vào điểm cơ sở.

Sử dụng thư viện Tkinker để xây dựng ứng dụng gợi ý chuyên ngành đơn giản.



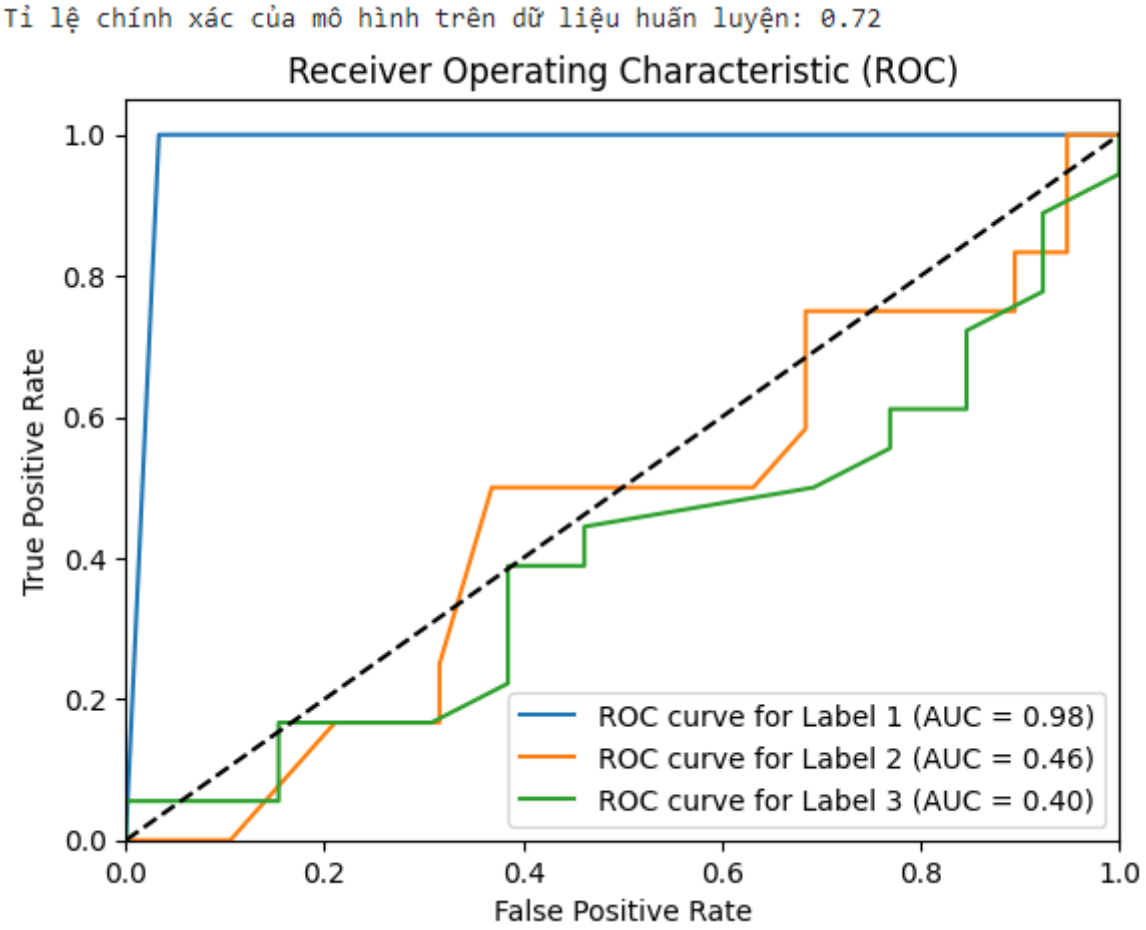
Hình 5. 2. Ứng dụng tự động gợi ý chuyên ngành

Ứng dụng gợi ý chuyên ngành sau khi nhập các đầu điểm học cơ sở sẽ tự động dự đoán và gợi ý chuyên ngành nên học.

5.2. Ứng dụng mô hình dựa vào mô hình tốt nhất với bộ dữ liệu khảo sát mức độ quan tâm

5.2.1. Ứng dụng mô hình qua thuật toán RF với dữ liệu khảo sát mức độ quan tâm

Dựa vào mô hình tốt nhất với thuật toán RF của bộ dữ liệu khảo sát thu về tỉ lệ chính xác 72%, tương đối do bộ dữ liệu khảo sát còn ít dữ liệu nên độ chính xác chỉ mang tính chất tương đối.



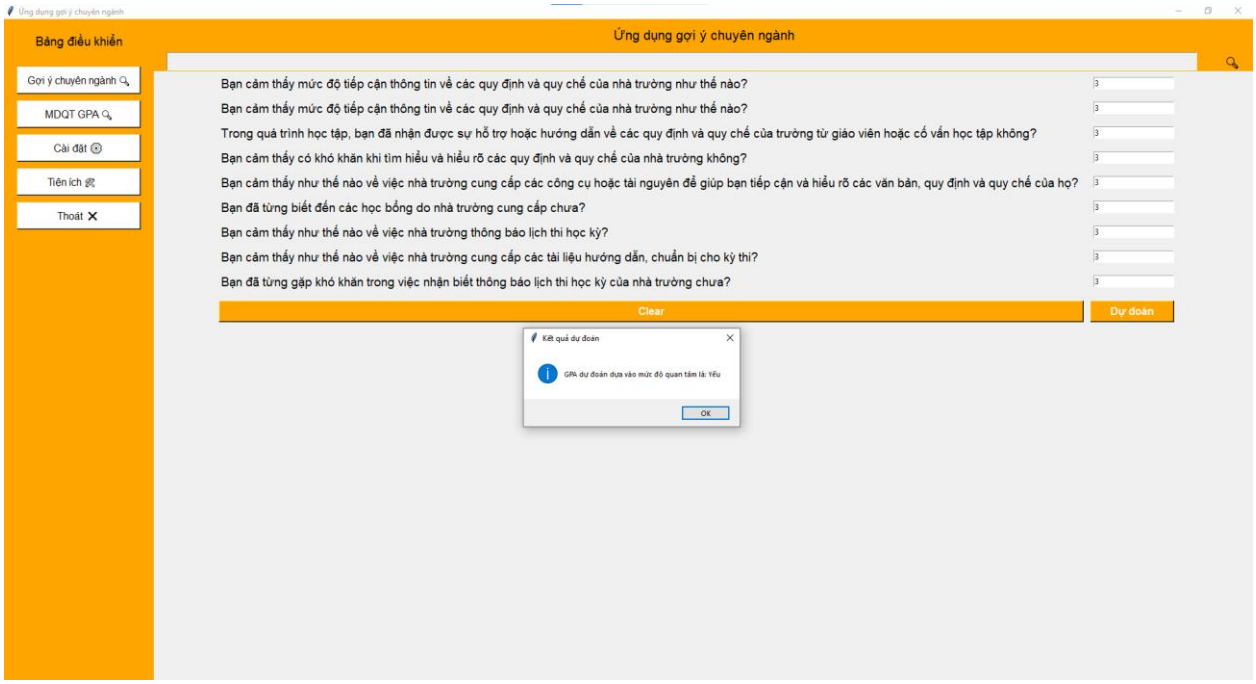
Hình 5. 3. Mô hình tốt nhất với dữ liệu khảo sát mức độ quan tâm

Trong hình ảnh của mô hình phân loại, chúng ta quan sát ba đường cong ROC được tạo ra cho ba nhãn khác nhau. Đoạn cong ROC đối với nhãn 1 có diện tích dưới đường cong (AUC) đạt mức cao, khoảng 0.98. Điều này cho thấy mô hình có khả năng phân loại tốt giữa các mẫu thuộc nhãn 1 và nhãn không phải 1. Tuy nhiên, khi quan sát các nhãn 2 và 3, chúng ta thấy AUC thấp, chỉ lần lượt là khoảng 0.46 và 0.40. Điều này cho thấy rằng mô hình gặp khó khăn trong việc phân loại các mẫu thuộc nhãn 2 và 3, có thể do dữ liệu không cung cấp đủ thông tin hoặc có sự mất cân bằng giữa các nhãn. Đánh giá tổng thể, mặc dù mô hình có khả năng phân loại tốt cho một nhãn, nhưng hiệu suất của nó vẫn còn kém đối với các nhãn khác. Điều này cần được cân nhắc để cải thiện hiệu suất phân loại của mô hình trên toàn bộ dữ liệu.

5.2.2. Xây dựng tính năng dự đoán mức độ được quan tâm

Sau khi sử dụng mô hình RF với tỉ lệ chính xác cao với các mô hình còn lại tiến hành xây dựng ứng dụng tự động gợi ý chuyên ngành dựa vào điểm cơ sở.

Sử dụng thư viện Tkinker để xây dựng ứng dụng gợi ý chuyên ngành đơn giản.



Hình 5. 4. Ứng dụng dự đoán GPA dựa vào data khảo sát

5.3. Đánh giá

Sau khi sử dụng mô hình SVM với tỉ lệ chính xác cao là 95% so với các mô hình còn lại, em đã tiến hành xây dựng một ứng dụng tự động gợi ý chuyên ngành dựa vào điểm cơ sở. Ứng dụng này được xây dựng bằng thư viện Tkinter để tạo ra một giao diện người dùng đơn giản và dễ sử dụng.

Dữ liệu dự đoán với chuyên ngành và mức độ quan tâm với GPA được sử dụng để tạo ra các gợi ý chuyên ngành cho người dùng. Mô hình SVM được chọn vì tỉ lệ chính xác cao, cho thấy khả năng phân loại tốt giữa các nhãn.

Bên cạnh đó, dựa vào mô hình tốt nhất sử dụng thuật toán Random Forest (RF) từ bộ dữ liệu khảo sát, em thu được một tỉ lệ chính xác khoảng 72%. Tuy nhiên, do bộ dữ liệu khảo sát có ít dữ liệu, nên độ chính xác chỉ mang tính tương đối.

KẾT LUẬN

Ưu điểm của dự án này là việc sử dụng mô hình SVM đã mang lại kết quả ấn tượng, với tỷ lệ chính xác cao nhất trong việc dự đoán chuyên ngành phù hợp cho sinh viên dựa trên điểm số học tập. Điều này thể hiện sự hiệu quả và tính đáng tin cậy của phương pháp học máy trong quá trình tư vấn chuyên ngành cho sinh viên.

Tính tiện ích và tính ứng dụng của ứng dụng tự động gợi ý chuyên ngành cũng là một điểm mạnh của dự án. Giao diện đơn giản và dễ sử dụng của ứng dụng, được xây dựng bằng thư viện Tkinter, giúp người dùng dễ dàng trải nghiệm và tận dụng các gợi ý chuyên ngành dựa trên điểm số của mình.

Tuy nhiên, một nhược điểm của dự án là lượng dữ liệu khảo sát thu thập không đủ lớn, ảnh hưởng đến hiệu suất của mô hình và ứng dụng. Để cải thiện và mở rộng phạm vi ứng dụng, việc tăng cường thu thập dữ liệu từ cựu sinh viên là một hướng phát triển quan trọng, giúp cải thiện độ chính xác và tính ứng dụng của hệ thống.

TÀI LIỆU THAM KHẢO

Tiếng Việt:

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

Tiếng Anh:

[2] *Predicting Students' Performance Using Machine Learning Techniques.pdf*

[3] *Predicting_Students_Final_GPA_Using_Decision_Trees.pdf*

[4] *Machine Learning Based Student Grade.pdf*

[5] *Using_Machine_Learning_Algorithms_to_Pre.pdf*

Danh mục các Website tham khảo:

[6] *https://www.matlabthayhai.com/2023/08/bai-2-thuat-toan-support-vector-machine.html*

[7] *https://trituenhantao.io/kien-thuc/decision-tree/*

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

[9] *https://www.studocu.com/vn/document/truong-dai-hoc-su-pham-ky-thuat-thanh-pho-ho-chi-minh/lap-trinh-r/random-forest-summary-lap-trinh-r/85357045*

[10] *https://one.3si.vn/vi/blog/cong-nghe-13/ml-classification-part-3-230*

[11] *https://www.linkedin.com/pulse/m%C3%A1y-h%E1%BB%8Dc-machine-learning-v%C3%A0-c%C3%A1c-m%E1%BB%91c-ph%C3%A1t-tri%E1%BB%83n-minh-giang-paul-*

[12] *https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB*