

XỬ LÝ DỮ LIỆU SINH VIÊN THÔNG QUA ỨNG DỤNG KỸ THUẬT HỌC MÁY ĐỂ HỖ TRỢ CÔNG TÁC TUYỂN SINH

Nguyễn Thị Kim Sơn, Nguyễn Xuân Hải, Tô Hồng Đức,
Phạm Tuấn Anh, Đỗ Thị Thu Trang

Trường Đại học Thủ đô Hà Nội

Tóm tắt: Bài báo trình bày kết quả nghiên cứu về việc xây dựng tập dữ liệu sinh viên và kết quả ứng dụng kỹ thuật học máy để lập chương trình dự báo cho loại tốt nghiệp sinh viên, dự báo các yếu tố trong tổ hợp tuyển sinh ảnh hưởng tới kết quả học tập của sinh viên. Để giải quyết các bài toán trên, chúng tôi tiến hành nghiên cứu trên bộ dữ liệu của ngành Giáo dục tiểu học của trường Đại học Thủ đô Hà Nội (dữ liệu trong 5 năm từ 2016 đến 2020). Các kỹ thuật học máy được sử dụng bao gồm kỹ thuật Logistic Regression (để dự báo kết quả tốt nghiệp của sinh viên) và một kỹ thuật cải tiến của kỹ thuật Linear discriminant analysis (để dự báo nhân tố quan trọng ảnh hưởng tới kết quả học tập của sinh viên) - kỹ thuật Discriminative Feature Selection. Từ đó nhóm tác giả đưa ra những đề xuất khuyến nghị về xu hướng tuyển sinh trình độ đại học hệ chính quy một số khuyến nghị về việc tổ chức đào tạo và chiến lược tuyển sinh cho trường Đại học Thủ đô Hà Nội.

Từ khóa: Học máy, machine learning, khoa học dữ liệu, khoa học giáo dục, dự báo kết quả học tập, bài toán tuyển sinh.

Nhận bài ngày 7.6.2021; gửi phản biện, chỉnh sửa, duyệt đăng ngày 22.7.2021

Liên hệ tác giả: Nguyễn Thị Kim Sơn; Email: ntkson@daihocthudo.edu.vn

1. MỞ ĐẦU

Với sự phát triển của khoa học, dữ liệu của con người được thu thập và lưu trữ thông qua các hoạt động kinh tế, xã hội, các hoạt động nghiên cứu khoa học ngày càng nhiều, chúng được lưu trữ trong các hệ thống máy tính có dung lượng lên đến hàng terabyte thậm chí là petabyte. Tuy nhiên, rất khó để con người hiểu và sử dụng đầy đủ những dữ liệu này. Đứng trước thực tế này, các hướng nghiên cứu mới trong phát hiện tri thức và khai phá dữ liệu đã hình thành và phát triển nhanh chóng trong 20 năm qua. Học máy và khai phá dữ liệu lớn là một lĩnh vực phát triển rất nhanh chóng, là lĩnh vực giao thoa giữa nhiều lĩnh vực liên quan như: công nghệ cơ sở dữ liệu, thống kê, học máy, thuật toán học và các lĩnh vực liên quan khác nhằm trích rút ra những tri thức hữu ích từ những tập dữ liệu rất lớn. Công tác tuyển sinh có ý nghĩa và vai trò quan trọng, được coi như là “điều kiện tiên quyết” đảm bảo cho toàn bộ quá trình tổ chức đào tạo cũng như các hoạt động khác trong trường đại học

được diễn ra. Với sự cạnh tranh khốc liệt của các trường đại học trong nước về công tác tuyển sinh, việc ứng dụng CNTT nói chung, công nghệ học máy và bài toán dự báo nói riêng được coi là một trong các giải pháp hữu hiệu nhằm đạt được kết quả tuyển sinh mong muốn. Từ việc phân tích thực trạng công tác tuyển sinh tại trường Đại học Thủ đô Hà Nội, đặc biệt là những thuận lợi và khó khăn trong việc tổ chức thực hiện quy trình tuyển sinh và dự báo xu hướng tuyển sinh; trên nền tảng chuyển đổi số và khoa học công nghệ của Nhà trường chưa phải là ưu thế, chúng tôi đưa ra sự cần thiết phải ứng dụng công nghệ trong việc hỗ trợ ra quyết định, giúp các nhà quản lý và chuyên viên phụ trách chuyên môn lập kế hoạch và phương hướng tuyển sinh. Trong bài báo này, chúng tôi đề cập đến việc ứng dụng của công nghệ học máy, bài toán dự báo, hỗ trợ ra quyết định,... làm cơ sở cho việc thực hiện nghiên cứu xu hướng tuyển sinh trình độ đại học hệ chính quy của Trường Đại học Thủ đô Hà Nội giai đoạn 2021-2025.

2. NỘI DUNG

2.1. Thực trạng công tác tuyển sinh của Trường Đại học Thủ đô Hà Nội

Trường Đại học Thủ đô Hà Nội là trường đại học công lập trực thuộc Ủy ban nhân dân thành phố Hà Nội được thành lập theo Quyết định số 2402/QĐ-TTg ngày 31/12/2014 của Thủ tướng Chính Phủ. Trường Đại học Thủ đô Hà Nội là cơ sở giáo dục đào tạo đại học trong hệ thống giáo dục quốc dân, tổ chức đào tạo nguồn nhân lực đa ngành, đa lĩnh vực trình độ cao đẳng, đại học và trên đại học; tổ chức các hoạt động giáo dục nghề nghiệp theo nhu cầu xã hội và theo quy định của pháp luật; là đơn vị sự nghiệp công lập trực thuộc UBND thành phố Hà Nội, chịu sự quản lý trực tiếp, toàn diện của UBND thành phố Hà Nội và sự quản lý chuyên môn của Bộ GD&ĐT; là đơn vị sự nghiệp công lập tự bảo đảm một phần kinh phí hoạt động, có tư cách pháp nhân, hoạt động theo nguồn kinh phí sự nghiệp do nhà nước cấp và có một số hoạt động nghiên cứu, tư vấn, dịch vụ có thu theo quy định của nhà nước, được mở tài khoản tại khoa bạc nhà nước và ngân hàng, có con dấu riêng và giao dịch theo quy định của pháp luật.

Trường Đại học thủ đô Hà Nội tổ chức đào tạo, giáo dục, bồi dưỡng nguồn nhân lực đa ngành, đa lĩnh vực có trình độ cao đẳng, đại học và cao học. Số lượng giảng viên cơ hữu của Nhà trường tham gia giảng dạy ở các trình độ đào tạo là 287, trong đó có 01 Giáo sư- Tiến sĩ, 09 Phó Giáo sư – Tiến sĩ, 68 Tiến sĩ. Với quy mô tuyển sinh và đào tạo trình độ cao đẳng, đại học chính quy của Nhà trường trong thời gian 04 năm vừa qua đạt được rất khả quan thu hút rất nhiều sinh viên học tập cũng như đào tạo tại Nhà trường. Việc tuyển sinh các ngành đạt chỉ tiêu và số lượng nguyện vọng vào các ngành trọng điểm tăng theo từng năm. Nhìn chung, kết quả tuyển sinh theo từng năm, thay đổi tùy từng ngành và khó dự báo được quy luật. Việc tuyển sinh không đủ chỉ tiêu cũng phụ thuộc nhiều vào các yếu tố bên ngoài, như xu hướng trong nước và thế giới về ngành nghề đào tạo, chất lượng đầu ra và nghề nghiệp kỳ vọng của học sinh và phụ huynh học sinh phổ thông, cách tính điểm chuẩn và chính sách của Bộ giáo dục và đào tạo,... Tuy nhiên, một yếu tố không nhỏ tác động tới kết quả tuyển sinh chính là cách thức tính điểm cho các tổ hợp tuyển sinh và lựa chọn mức điểm chuẩn cho

từng ngành đào tạo. Đây là một công việc khó khăn đối với cán bộ tham gia công tác tuyển sinh cũng như nhà quản lý khi phải ra quyết định.

Trường Đại học Thủ đô Hà Nội chưa thực sự có thương hiệu lớn, cơ sở vật chất chưa đáp ứng cho đảm bảo việc giảng dạy và học tập, xuống cấp,... Đặc biệt hai năm tuyển sinh liên tiếp là 2019 và 2020 đều diễn ra trong bối cảnh bệnh dịch Covid-19 do đó Nhà trường đã phải áp dụng nhiều phương thức tuyển sinh trong đó có những phương thức hoàn toàn mới mẻ không chỉ riêng với Trường Đại học Thủ đô Hà Nội mà còn mới mẻ với toàn xã hội. Nhiều quy định của Bộ Giáo dục và Đào tạo còn chồng chéo và chưa rõ ràng. Các văn bản quy định về phương thức, thời gian, đối tượng tham gia công tác xét tuyển của Bộ GD&ĐT thường xuyên thay đổi. Nhiều tình huống phát sinh ngoài các văn bản hướng dẫn, quy định của Bộ GD&ĐT. Hệ thống công nghệ thông tin được là một trong những hạn chế lớn nhất của Nhà trường, trong đó là vai trò đặc biệt của công nghệ thông tin đối với công tác tuyển sinh do hệ thống hạ tầng phần mềm đã quá cũ, lỗi thời, không thể cập nhật, đã có những sự cố đáng tiếc xảy ra như sập nguồn trong thời gian đăng kí nguyện vọng, không lưu trữ hoặc không cung cấp minh chứng cho thí sinh để xác nhận đăng kí nguyện vọng, nhập học, không đăng kí đúng thời hạn do quá tải,... Bên cạnh đó, sự chủ động của Nhà trường trong công tác tuyển sinh cũng còn nhiều hạn chế, đặc biệt là việc đưa ra các định hướng, xây dựng chiến lược, kế hoạch tuyển sinh cho những năm qua, chưa dự báo được công tác này có tính chất “dài hơi” và chủ động.

Một trong những công tác phục vụ đánh giá hoạt động tuyển sinh đó là việc tiến hành thu thập, tổng hợp, phân tích dữ liệu của các bên liên quan đến giáo dục là một quá trình mất nhiều thời gian, tốn kém và không phải lúc nào cũng có mức độ chính xác cao, nhất là trong thực trạng việc chuyển đổi số và lưu trữ dữ liệu dạng số trong lĩnh vực giáo dục ở Việt Nam còn chưa được tiến hành bài bản và triệt để, trong đó có trường Đại học Thủ đô Hà Nội. Việc quản lý thông tin tuyển sinh, trong đó có điểm thi trung học phổ thông quốc gia, do Bộ giáo dục và đào tạo quản lý. Trong khi đó các trường Đại học chỉ quản lý dữ liệu điểm thi của các môn học trong quá trình đào tạo và điểm tốt nghiệp. Thực tế hiện nay, mặc dù các trường đại học đều đã quan tâm đến việc xây dựng tập dữ liệu sinh viên nhưng dữ liệu được xây dựng, quản lý bởi nhiều bộ phận khác nhau (phòng đào tạo, phòng công tác học sinh sinh viên, các khoa đào tạo,...). Các dữ liệu vì thế khó đồng bộ, việc kết nối để khai thác và sử dụng dữ liệu của các bên liên quan cũng gặp nhiều khó khăn.

Mặt khác, các dữ liệu khác của sinh viên như: Văn hóa, truyền thống gia đình, kinh tế, nguyện vọng cá nhân, định hướng nghề nghiệp, kết quả học tập phổ thông, kế hoạch học tập, chương trình đào tạo, đội ngũ giảng viên, cơ sở vật chất của cơ sở giáo dục, việc tham gia các tổ chức xã hội, đoàn thể, yếu tố nhân chủng học, yếu tố văn hóa, kinh tế, tâm lý học,... cần được nghiên cứu xây dựng cần đảm bảo tập dữ liệu gồm nhiều trường thông tin, nhiều tham số đại diện, ảnh hưởng qua lại lẫn nhau trực tiếp ảnh hưởng tới quá trình đào tạo của SV và kết quả học tập của sinh viên đó. Các yếu tố này cần được phân tích, xử lý, đưa ra cách thức cải thiện để làm cho dữ liệu giáo dục có ý nghĩa hơn đối với sinh viên, giảng viên và các bên liên quan khác. Đặc biệt trong bối cảnh Trường Đại học Thủ đô Hà Nội đang

từng bước mở rộng và phát triển các ngành nghề đào tạo, việc làm tốt công tác tuyển sinh, phân tích tình hình ra quyết định tuyển sinh chính xác là rất cần thiết.

2.2. Sự cần thiết phải ứng dụng khai phá dữ liệu hỗ trợ công tác tuyển sinh

Trên thế giới cũng như ở Việt Nam, khoa học dữ liệu được hiểu là một ngành cho phép kết hợp rất hữu hiệu sự phát triển vượt bậc của thế giới công nghệ cao, trong đó trước hết phải kể đến Công nghệ thông tin (CNTT) và Internet với trình độ tổ chức và cơ cấu quản trị hoạt động, kinh doanh của các tổ chức. Khoa học và thực tiễn quản lý – quản trị ngày càng trở nên hiện đại hơn dựa vào hạ tầng CNTT, trình độ ứng dụng tin học vào quản trị tài nguyên dữ liệu lớn. Cách mạng công nghệ 4.0 (CMCN 4.0) được định hình không tách rời với dữ liệu và phân tích dữ liệu, đặt ra thách thức với các tổ chức khi phải xử lý tốt dữ liệu để có thể giảm thiểu rủi ro thất bại. Đối với hoạt động tuyển sinh đại học cũng như vậy, dữ liệu lớn được ví là tài sản, là phần cốt lõi của các chiến lược hoạt động, mang lại giá trị rộng lớn giúp tăng tính cạnh tranh. Đặc điểm của các dữ liệu trong thời đại kỹ thuật số là khối lượng rất lớn, cấu trúc phức tạp, biến đổi nhanh, vì vậy các kỹ thuật ngày phải càng được phát triển để đáp ứng với nhu cầu phân tích dữ liệu mới. Nguồn dữ liệu khổng lồ nếu không biết kiểm soát, tổ chức khai thác, xử lý, phân tích sẽ nhanh chóng trở nên không thể kiểm soát được. Đối với hoạt động tuyển sinh, thông tin quan trọng này có được và sử dụng được chủ yếu thông qua phân tích dữ liệu. Nhu cầu thiết yếu này tạo ra nhu cầu cần có một hệ thống phân tích dữ liệu và dự báo xu thế tuyển sinh nhằm đáp ứng nhu cầu của trong khu vực và trên thế giới. Nhiệm vụ của hệ thống phân tích dữ liệu, dự báo xu hướng là sử dụng thuần thục các công cụ được học để chuyển hóa dữ liệu thành thông tin và tiếp tục chuyển hóa thông tin thành nghiệp vụ, chiến lược, giúp nhà quản lý ra các quyết định quan trọng trong công tác quản lý đào tạo.

Nhu cầu cần phải có một hệ thống phân tích dữ liệu và dự báo cần có đủ năng lực tiếp cận với cơ sở dữ liệu lớn một cách chủ động, am hiểu và tự tin giải quyết các vấn đề phức tạp tồn tại khách quan về công tác tuyển sinh, báo cáo, kiểm soát thực trạng nhu cầu của xã hội; phân tích tìm hiểu nguyên nhân các vấn đề; giám sát các công việc đang được triển khai; dự báo các xu hướng,...

Phát triển nguồn nhân lực là một trong những yếu tố quan trọng quyết định sự thành công trong phát triển kinh tế - xã hội của quốc gia, do vậy tất cả các nước trên thế giới đều quan tâm đến phát triển nguồn nhân lực. Hơn lúc nào hết, khi nước ta đã trở thành thành viên WTO, đòi hỏi phải có một nguồn nhân lực chất lượng cao đáp ứng quá trình hội nhập. Kinh nghiệm nhiều nước cho thấy, ví dụ như Trung Quốc, sau 5 năm gia nhập WTO, kinh tế phát triển gần gấp đôi, nhưng kèm theo đó là việc thiếu nhân lực trầm trọng nhất là nguồn nhân lực có trình độ tay nghề cao. Tình trạng này có trầm trọng hay không, có thể vượt qua được hay không là phụ thuộc rất nhiều vào công tác dự báo nhu cầu nguồn lao động việc làm như thế nào. Để thực hiện tốt công tác dự báo cung ứng nguồn nhân lực thì đòi hỏi các trường Đại học cần có hệ thống dự báo xu hướng tuyển sinh để đáp ứng nhu cầu thị trường. Xuất phát từ quản lý thông tin cơ bản của hồ sơ học tập sinh viên, việc sử dụng công nghệ thông tin hỗ trợ khai thác dữ liệu từ đó hình thành cơ sở để xây dựng mô hình tổ chức, tuyển

sinh, bố trí các chuyên ngành, chỉ tiêu tuyển sinh phù hợp. Bên cạnh đó, hệ thống dự báo tuyển sinh còn có thể tìm ra giải pháp khắc phục những mặt còn hạn chế, lúng túng trong công tác quản lý đào tạo, giải quyết một cách hài hòa mối quan hệ giữa đào tạo các chuyên ngành và tuyển sinh. Từ đó xây dựng các chương trình đào tạo có chuyên môn cao đáp ứng nhu cầu của thị trường lao động và xã hội.

Tuy nhiên để thực hiện công tác này không phải là một công việc dễ dàng, đơn giản. Hơn nữa, những chuyên viên làm công tác tuyển sinh, đào tạo mặc dù có nhiều năm kinh nghiệm nhưng lại có rất ít kiến thức về công nghệ thông tin nên rất khó đưa ra một mô hình dự báo hợp lý cho công tác tuyển sinh. Nhu cầu nguồn nhân lực ngoài thị trường lao động không ngừng đổi mới do vậy cần tiến hành các nghiên cứu để nâng cao chất lượng công tác dự báo nhu cầu lao động của thị trường, từ đó giúp nhà trường đưa ra những dự báo trong chính sách tuyển sinh trong những năm tiếp theo. Quản lý thông tin của sinh viên, phát hiện những thông tin tiềm ẩn hỗ trợ dự báo công tác tuyển sinh không chỉ là vấn đề bức thiết hiện nay mà còn là vấn đề khó khả thi nếu không có sự hỗ trợ đắc lực của công nghệ.

Từ những phân tích nêu trên, việc nghiên cứu dự báo xu hướng tuyển sinh và ứng dụng công nghệ trong việc hỗ trợ ra quyết định, giúp các nhà quản lý và chuyên viên phụ trách chuyên môn lập kế hoạch và phương hướng tuyển sinh hợp lý là một nhu cầu bức thiết của Nhà trường. Thông qua việc nghiên cứu xu hướng tuyển sinh trình độ đại học hệ chính quy của Trường Đại học Thủ đô Hà Nội giai đoạn 2021-2025 thông qua ứng dụng công nghệ Học máy (Machine Learning), chúng tôi sẽ xây dựng bộ dữ liệu tuyển sinh của trường Đại học Thủ đô Hà Nội, dựa vào các kỹ thuật học máy, để lập trình xử lý dữ liệu, đưa ra các dự báo hỗ trợ ra quyết định cho công tác tuyển sinh của Nhà trường.

2.3. Ứng dụng kỹ thuật học máy trong xử lý dữ liệu tuyển sinh và dữ liệu học tập của sinh viên

Có thể nói dữ liệu đặc biệt trong bối cảnh cuộc cách mạng công nghiệp 4.0, dữ liệu đã được nâng cấp lên thành các dữ liệu lớn (Big Data), tuy nhiên đây không phải là nguồn tài nguyên vô tận, nếu như có các dữ liệu cần thiết nhưng không có công cụ hoặc phương pháp phân tích thì các dữ liệu này đều sẽ không phát huy hết được các giá trị. Việc phân tích dữ liệu, đặc biệt trong hoạt động đào tạo của nhà trường, mà cụ thể ở đây là công tác xử lý dữ liệu tuyển và học tập của sinh viên là hết sức cần thiết. Ngày nay, bối cảnh giáo dục có sự thay đổi rất lớn khi điều kiện học tập của người học được nâng lên với sự đầu tư cả ở cấp độ quốc gia, nhà trường và người học. Công nghệ đã trở thành một phần tư liệu sản xuất của quá trình giáo dục. Bên cạnh đó, nhu cầu học tập cá nhân được chú trọng [1,3]. Do đó, các nghiên cứu giáo dục học đang được chuyển hướng đến nghiên cứu sâu hành vi người học để thiết lập các chương trình học cá nhân; đồng thời khai phá dữ liệu lớn những người học để sớm chẩn đoán và định hướng lại quá trình học tập của người học nói riêng, quản lý/ điều hành quá trình giáo dục nói chung [2,4,6].

Trên thế giới, các ứng dụng về kỹ thuật học máy cũng đã được triển khai nhằm đưa ra các xu hướng dự báo các dữ liệu phục vụ cho tuyển sinh cũng như đánh giá tính khả thi của

lĩnh vực đào tạo, trong đó phải kể đến các ứng dụng học máy dự báo kết quả dữ liệu học tập của sinh viên. Tác giả Kotsiantis (2012) mô tả lĩnh vực của học máy trong lĩnh vực giáo dục và đào tạo, đây là nghiên cứu thể hiện rõ nét nhất vai trò của học máy đối với hoạt động giáo dục và đào tạo, trong đó có thể áp dụng trong hoạt động tuyển sinh, trong nghiên cứu này các dữ liệu đặc trưng của sinh viên và dữ liệu điểm được khai thác dưới dạng tập dữ liệu cho phương pháp học máy hồi quy được sử dụng để dự đoán khả năng học tập trong tương lai của sinh viên [8]Error! Reference source not found..

Bên cạnh đó, các thông tin cần thiết xử lý dữ liệu của sinh viên có thể cần đề cập đến đó là các dữ liệu đầu vào phục vụ cho quá trình tuyển sinh như kết quả học tập bậc trung học phổ thông, có thể được nghiên cứu để đưa ra dự báo đánh giá về tiềm năng của các học sinh có khả năng lựa chọn ngành nghề phù hợp với kết quả học tập tại các trường đại học. Với sự phát triển của công nghệ AI, có thể nói các nghiên cứu về học máy là cần thiết cũng như cần được ứng dụng ở mức độ cao hơn, cần có sự nâng cấp từ bước thu thập và khai phá dữ liệu đơn thuần hiện nay ở công tác tuyển sinh trở thành bước dự báo và nhận diện các dữ liệu phục vụ cho công tác tuyển sinh của Nhà trường.

Từ đó có thể khẳng định, việc ứng dụng kỹ thuật học máy trong xử lý dữ liệu là vô cùng tiềm năng, đặc biệt ở tính ứng dụng trong việc xử lý dữ liệu đối với hoạt động tuyển sinh góp phần nâng cao chất lượng về hiệu quả của hoạt động giáo dục cũng như hoạt động quản lý, hoạt động quản trị nhà trường trong bối cảnh hiện nay. Bài toán mà chúng tôi giải quyết trong bài báo này có hai bài toán con chính đó là (1) dự đoán xếp loại học tập: từ dữ liệu học tập năm 1 (năm 2) của sinh viên cùng với dữ liệu tuyển sinh, chúng tôi xây dựng chương trình ứng dụng học máy để dự đoán loại tốt nghiệp của sinh viên và (2) xác định thuộc tính quan trọng: Từ dữ liệu học tập và dữ liệu tuyển sinh của sinh viên, chúng tôi xác định môn nào là quan trọng nhất trong số các môn của tập dữ liệu đầu vào. Từ đó chúng tôi đưa ra những tổ hợp tuyển sinh với trọng số tuyển sinh hợp lý để có thể lựa chọn được những sinh viên có đầu ra tốt nghiệp đạt kỳ vọng.

Dữ liệu được sử dụng cho bài báo này là dữ liệu tuyển sinh và dữ liệu đào tạo của sinh viên ngành Giáo dục Tiểu học. Ngành Giáo dục Tiểu học là một trong những ngành truyền thống của Nhà trường, bắt đầu tuyển sinh từ năm 1959, ngành Giáo dục Tiểu học của Trường Đại học Thủ đô Hà Nội đã cung cấp lượng giáo viên tiểu học lớn cho các trường tiểu học trên địa bàn Thành phố Hà Nội và các địa bàn lân cận. Trong bài báo này, chúng tôi xét vấn đề tuyển sinh trình độ đại học, hệ chính quy, ngành Giáo dục Tiểu học của Trường Đại học Thủ đô Hà Nội. Dữ liệu bao gồm điểm tuyển sinh đầu vào, điểm các học phần năm 1, năm 2, năm 3, năm 4, điểm thực tập và điểm tốt nghiệp, tình hình tài chính có thể được trích xuất từ phần mềm quản lý đào tạo để phân tích dữ liệu. Bộ dữ liệu bao gồm 2426 mẫu quan sát của sinh viên ngành Giáo dục tiểu học bao gồm cả hệ đại học và hệ cao đẳng trong những năm gần đây. Riêng dữ liệu cho hệ đại học được lấy từ khóa D2016 đến khóa D2020 và 44 biến thuộc tính. Mỗi mẫu quan sát được thể hiện bằng 1 hàng. Các biến được sử dụng trong phân tích dữ liệu được thể hiện trong bảng 1 dưới đây.

Bảng 1. Các biến thuộc tính

Mã Học phần	Tên học phần
M1	Môn 1 (Toán)
M2	Môn 2 (Văn)
M3	Môn 3 (Tiếng anh)
TĐ	Tổng điểm xét tuyển
20TRA002	Giáo dục quốc phòng an ninh 1
20TRA003	Những nguyên lý cơ bản của CN Mác Lenin 1
30PRI002	Rèn kỹ năng sử dụng tiếng Việt ở Tiểu học
30PRI003	Văn học và dạy tác phẩm văn học ở Tiểu học
30PRI072	Hoạt động nghệ thuật ở trường TH 1
30PRI073	Rèn kỹ năng sử dụng Tiếng Việt
30PRI120	Cơ sở lý thuyết Toán 1
30PRI301	Giáo dục nghệ thuật ở trường TH 1
30TRA002	Pháp luật đại cương
30TRA045	Giáo dục thể chất 1
30TRA054	Tâm lý học
30TRA121	Triết học Mác-Lênin
30TRA126	GDQP_ AN 1
TA1	Tiếng Anh (đầu vào)
20TRA004	Tin học
20TRA006	Giáo dục quốc phòng an ninh 2
20TRA007	Tiếng Anh 1
20TRA008	Những nguyên lý cơ bản của CN Mác Lenin 2
20TRA013	Logic học
20TRA026	Rèn luyện phẩm chất người giáo viên
30PRI004	Cơ sở đại số ở tiểu học
30PRI006	Sinh lý trẻ lứa tuổi tiểu học
30PRI074	Văn học
30TRA001	Quản lý hành chính nhà nước và quản lý ngành
30TRA023	Tiếng Anh
30TRA024	Tiếng Trung Quốc
30TRA025	Tiếng Hàn Quốc
30TRA031	Giáo dục học
30TRA032	Nghệ vụ sư phạm 1
30TRA110	Phát triển năng lực thông tin trong kỉ nguyên số
30TRA122	Kinh tế chính trị Mác-Lênin
30TRA127	GDQP_ AN 2
30TRA128	GDQP_ AN 3
20TRA009	Hà Nội học

20TRA010	Tiếng Việt thực hành
20TRA011	Cơ sở văn hóa Việt Nam
20TRA012	Lịch sử văn minh thế giới
20TRA014	Giáo dục vì sự phát triển bền vững
20TRA015	Dân số môi trường và phòng chống ma túy
20TRA016	Biển và hải đảo Việt Nam
20TRA017	Giáo dục quốc phòng an ninh 3
20TRA018	Tiếng Anh 2
20TRA906	Toán xác suất thống kê
30CIV006	Kinh tế học đại cương
30CIV075	Những vấn đề của thời đại ngày nay
30PRI016	Cơ sở tự nhiên xã hội
30PRI020	PPDH tiếng việt ở tiểu học 1
30PRI049	rèn luyện năng lực dạy học
30PRI075	Cơ sở lý thuyết Toán 2
30PRI079	Hướng dẫn làm đồ dùng dạy học truyền thống
30PRI181	Tâm lí học và giáo dục học tiểu học 1
30PRI210	Giáo dục đạo đức ở tiểu học
30PRI257	Công tác chủ nhiệm ở TH
30PRI303	Phương tiện dạy học ở tiểu học
30TRA003	Giáo dục thể chất 3
30TRA033	Nghệ thuật vụ sư phạm 2
30TRA039	Thực hành Âm nhạc
30TRA040	Thực hành Mỹ thuật
30TRA046	Giáo dục thể chất 2
30TRA058	Kỹ năng quản lý tài chính cá nhân
30TRA070	Phương pháp nghiên cứu khoa học
30TRA080	Kĩ năng quản lí tài chính cá nhân
30TRA111	Âm nhạc và cảm thụ âm nhạc
30TRA112	Mĩ thuật và cảm thụ mỹ thuật
30TRA123	Chủ nghĩa xã hội khoa học
30TRA129	GDQP_AN 4
20TRA019	Đường lối Cách mạng của Đảng cộng sản Việt Nam
20TRA031	Rèn luyện năng lực giáo dục
20TRA040	Tiếng Anh tăng cường
30PRI007	Tiếng Việt
30PRI009	PPDH Tiếng Việt 1
30PRI021	PPDH tiếng việt ở trường tiểu học 2
30PRI023	PPDH TNXH 1

30PRI036	tìm hiểu lịch sử - địa lý hà nội phục vụ dạy TH
30PRI037	Ngữ nghĩa - ngữ dụng và UD trong dạy TV ở tiểu học
30PRI041	đàn phím điện tử và ứng dụng trong dạy học tiểu học
30PRI055	Thực tập sư phạm 1
30PRI077	PPDH Toán Tiểu học 1
30PRI078	Giáo dục sức khỏe và thể chất ở tiểu học 1
30PRI080	PPDH Toán Tiểu học 2
30PRI081	UDCNTT và TT trong dạy học tiểu học
30PRI082	Giáo dục sức khỏe và thể chất ở tiểu học 2
30PRI190	Mỹ thuật và PPDH mỹ thuật ở tiểu học
30TRA034	Nghịệp vụ sư phạm 3
30TRA124	Tư tưởng Hồ Chí Minh

Bài toán dự báo xếp loại học tập và bài toán xác định thuộc tính quan trọng được thực hiện theo các bước sau: Bước 1: Thu thập dữ liệu sinh viên/Bước 2: Tiền xử lý dữ liệu/Bước 3: Tách dữ liệu thu thập được thành hai tập, một tập là huấn luyện, tập còn lại là tập kiểm thử/Bước 4: Huấn luyện/Bước 5: Sử dụng mô hình dự báo để (a) dự báo kết quả học tập của sinh viên và (b) xác định thuộc tính quan trọng./Bước 6: Đánh giá kết quả, độ chính xác của mô hình dự báo.

Tập dữ liệu kết quả học tập của sinh viên thu được từ trường Đại học Thủ đô ở trên là không tách được tuyến tính nên một số phương pháp học máy như hồi quy tuyến tính, PLA,... không áp dụng được. Bài toán phân loại kết quả học tập ở trên là bài toán đa lớp dẫn đến hồi quy logistics với hàm sigmoid sẽ không hiệu quả. Do đó chúng tôi đề xuất áp dụng hồi quy logistics với hàm softmax cho bài toán con thứ nhất.

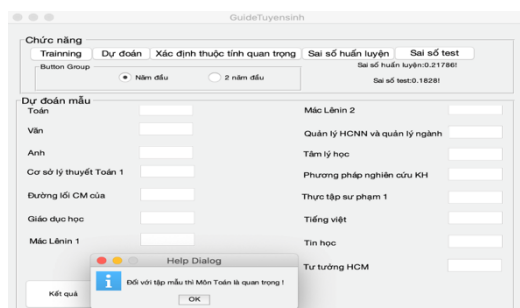
Bài toán con thứ hai đó là xác định thuộc tính quan trọng (cụ thể là xác định môn nào là quan trọng nhất trong số các môn của tập dữ liệu đầu vào). Việc xác định thuộc tính quan trọng là một bài toán tối ưu NP-Hard trên không gian rời rạc. Do đó, việc nghiên cứu tìm thuật toán lựa chọn đặc trưng có độ phức tạp thấp hơn là một vấn đề quan trọng khi giải bài toán con thứ hai này. Một cách thích hợp để giải bài toán con thứ hai với độ phức tạp tính toán chấp nhận được đó là quy bài toán tối ưu trên không gian rời rạc về bài toán tối ưu trên không gian liên tục như trong [5].

Phân tích hồi quy logistic được sử dụng để kiểm tra mã tốt nghiệp trên mỗi sinh viên, dựa trên 44 trường dữ liệu đầu vào, số điểm thi trung học phổ thông quốc gia, số điểm thi học phần của các môn học trong hai năm đầu. Bài toán được tiếp cận với 3 mức độ của các dữ liệu đầu vào: Mức độ 1 dựa trên 16 thuộc tính gồm điểm xét tuyển đại học và 12 môn học của năm thứ nhất đại học, Mức độ 2 là phân tích dựa trên điểm số của 29 môn học năm thứ nhất và thứ hai đại học, cùng với điểm thi tốt nghiệp trung học phổ thông quốc gia.

Cả hai trường hợp đều cho thấy mối quan hệ tăng dần độ chính xác với một sự giảm dần độ lệch chuẩn trong kết quả dự báo (sai số từ 21% đến 15% trên tập huấn luyện và kiểm thử). Điều này có thể giải thích tại sao các trường đại học nên chuyển hướng tích cực sang việc

dạy học hai giai đoạn. Dựa trên quy mô đào tạo đại cương và điểm số tuyển sinh, điểm học năm thứ nhất và thứ hai đại học, chúng ta có thể dự báo được những sở trường của sinh viên phù hợp với ngành nghề đào tạo nào của trường Đại học Thủ đô Hà Nội. Việc xác định thuộc tính quan trọng nhất được xây dựng dựa trên thuật toán Linear discriminant analysis (LDA) với nhiều trường dữ liệu và thư viện con được xây dựng. Hình 1 mô tả kết quả huấn luyện cho mẫu huấn luyện của khóa Đại học ngành Giáo dục tiểu học, trường Đại học Thủ đô Hà Nội. Theo kết quả huấn luyện này, môn Toán sẽ là thuộc tính quan trọng nhất ảnh hưởng tới kết quả đầu ra của tập dữ liệu nếu chúng ta xét ở mức độ 1, dựa trên dữ liệu tuyển sinh đại học và kết quả học tập của năm đầu đại học.

Nghiên cứu trong tương lai có thể được tiến hành để sử dụng thêm các biện pháp đánh giá kết quả học tập nhằm xác định kết quả học tập trong tương lai của sinh viên. Mô hình hồi quy logistic có thể được cải thiện để đưa ra dự đoán tốt hơn về kết quả học tập của sinh viên. Sẽ rất thú vị nếu so sánh hiệu suất của các mô hình phân loại trên các tập dữ liệu khác, hoặc thậm chí cải thiện mô hình để tăng độ chính xác dự đoán của mô hình hiện tại. Nghiên cứu trong tương lai cũng có thể bao gồm đánh giá các đặc điểm hành vi của sinh viên, các yếu tố nhân học, yếu tố cá nhân và yếu tố lịch sử của quá trình học tập của sinh viên, cũng như thái độ học tập và các yếu tố kinh tế xã hội khác khi chúng liên quan đến kết quả học tập của sinh viên dựa trên các cách thức đánh giá. Có thể tiến hành phân tích bổ sung bằng cách sử dụng các bộ phân loại khác nhau trên cùng một tập dữ liệu, bao gồm nhận thức đa lớp và mạng nơ-ron nhân tạo.



Hình 1. Xác định môn quan trọng

Việc xác định kết quả dự báo cho nhân tố ảnh hưởng chính đến kết quả tốt nghiệp của sinh viên trong các môn của tổ hợp tuyển sinh sẽ hỗ trợ những nhà quản lý ra quyết định trong việc chọn các tổ hợp tuyển sinh phù hợp với định hướng chuẩn đầu ra của ngành. Đồng thời xác định hệ số cho mỗi môn trong tổ hợp tuyển sinh phù hợp sao cho lựa chọn được những sinh viên có lực học phù hợp nhất với yêu cầu của ngành nghề đào tạo. Mặt khác, các dữ liệu khác của sinh viên như: văn hóa, truyền thống gia đình, kinh tế, nguyện vọng cá nhân, định hướng nghề nghiệp, kết quả học tập phổ thông, kế hoạch học tập, chương trình đào tạo, đội ngũ giảng viên, cơ sở vật chất của cơ sở giáo dục, việc tham gia các tổ chức xã hội, đoàn thể, yếu tố nhân chủng học, yếu tố văn hóa, kinh tế, tâm lý học,... cần được nghiên cứu xây dựng cần đảm bảo tập dữ liệu gồm nhiều trường thông tin, nhiều tham số đại diện, ảnh hưởng qua lại lẫn nhau trực tiếp ảnh hưởng tới quá trình đào tạo của sinh viên và kết quả

học tập của sinh viên đó. Các yếu tố này cần được phân tích, xử lý, đưa ra cách thức cải thiện để làm cho dữ liệu giáo dục có ý nghĩa hơn đối với sinh viên, giảng viên và các bên liên quan khác.

2.4. Dự báo xu hướng tuyển sinh trình độ đại học hệ chính quy của Trường Đại học Thủ đô Hà Nội giai đoạn 2021-2025

Từ những phân tích nêu trên, xu hướng tất yếu của bài toán tuyển sinh cũng như quản lý đào tạo tại các trường đại học nói chung, trường đại học Thủ đô Hà Nội nói riêng, sẽ phải dựa trên công nghệ, lượng hóa dữ liệu để ra những quyết định đúng đắn, hợp lý trong giải quyết các vấn đề tuyển sinh. Ngoài ra, việc quản lý đào tạo đối với sinh viên, đặc biệt là sinh viên chính quy, cần được dữ liệu hóa và liên kết với dữ liệu tuyển sinh, để có được những thông tin định hướng trong quá trình tuyển sinh những năm tiếp theo. Xu hướng tuyển sinh sẽ diễn ra dưới nhiều hình thức tuyển sinh khác nhau, nhưng có trọng số cho những tổ hợp định hướng, những môn quan trọng. Đồng thời những yếu tố xung quanh học sinh như: yếu tố văn hóa, tình hình kinh tế gia đình, sở thích cá nhân... sẽ được quan tâm đúng mức trong các yếu tố quyết định chọn trường của sinh viên. Việc này sẽ được kiểm soát ở các trường đại học khi dữ liệu được đồng bộ từ Bộ Giáo dục và Đào tạo đến các cơ sở giáo dục đại học và các trường phổ thông.

2.5. Một số khuyến nghị về tổ chức đào tạo và chiến lược tuyển sinh

2.5.1. Khuyến nghị đối với Trường Đại học Thủ đô Hà Nội

Xây dựng bộ dữ liệu sinh viên, thực hiện chuyển đổi số toàn diện trong quản lý đào tạo. Cần thay đổi về quan niệm, cách phân loại, cách thức và quy trình xây dựng bộ dữ liệu sinh viên. Tập dữ liệu sinh viên cần bao gồm nhiều trường thông tin, đảm bảo sự kết nối để có thể khai thác, sử dụng trong nhiều lĩnh vực hoạt động của các chủ thể quản lý khác nhau trong nhà trường (xây dựng và ban hành những chính sách, quyết định của lãnh đạo nhà trường; hoạt động quản lý của các phòng ban chức năng, hoạt động giảng dạy của giảng viên, hoạt động học tập của sinh viên,...). Thống nhất đầu mối quản lý dữ liệu, chia sẻ account đến các đơn vị liên quan (phòng Quản lý đào tạo và Công tác học sinh sinh viên là đơn vị quản lý, các khoa đào tạo là đơn vị thành viên cấu thành dữ liệu, quản lý thứ cấp, các đơn vị khác hỗ trợ cung cấp dữ liệu và cùng khai thác dữ liệu,...). Lưu trữ dữ liệu số trong công tác tuyển sinh một cách hệ thống và đầy đủ, tập hợp dữ liệu của nhiều năm để phục vụ việc tư vấn tuyển sinh và dự báo xu hướng tuyển sinh. Sử dụng công nghệ học máy để đưa ra các dự báo phục vụ ra quyết định trong công tác tuyển sinh của Nhà trường đối với những năm tuyển sinh tiếp theo trên cơ sở đảm bảo đầy đủ dữ liệu tuyển sinh trong thời gian ít nhất là 03 năm liên tục và kết quả học tập của toàn bộ sinh viên toàn trường trong năm thứ nhất và năm thứ hai tính đến thời điểm hiện tại. Hệ thống quản lý phần mềm trong tuyển sinh và kết quả học tập của sinh viên Nhà trường cần đáp ứng phù hợp với nhu cầu dữ liệu cho công nghệ học máy trong bài toán dự báo. Sử dụng kết quả ứng dụng công nghệ học máy vào bài toán tuyển sinh các ngành đào tạo của Trường Đại học Thủ đô Hà Nội trong việc xây dựng đề án tuyển sinh, trong đó dự báo và xác định số lượng thí sinh đăng ký xét tuyển, nhập học và theo học

tại Nhà trường đối với từng ngành đào tạo, sử dụng trong truyền thông cho công tác tuyển sinh, đào tạo của Nhà trường. Xây dựng đội ngũ kỹ thuật (công nghệ thông tin) có đủ khả năng, kỹ năng sử dụng công nghệ học máy và bài toán dự báo áp dụng trong công tác tuyển sinh của Nhà trường.

2.5.2. Khuyến nghị đối với Bộ Giáo dục và Đào tạo

Điều chỉnh kỹ thuật trong quy trình tổ chức thi. Cụ thể, về đề thi, nội dung đề thi nằm trong chương trình cấp THPT, chủ yếu là chương trình lớp 12, đảm bảo ngưỡng cơ bản để xét tốt nghiệp THPT và có độ phân hóa phù hợp để làm cơ sở cho tuyển sinh. Xây dựng bộ dữ liệu chung về hồ sơ học tập của học sinh, trong đó có lưu trữ các dữ liệu liên quan, ảnh hưởng tới kết quả học tập của học sinh cũng như tác động tới yếu tố chọn ngành nghề, chọn trường đại học khi tốt nghiệp trung học phổ thông như: văn hóa, truyền thống gia đình, kinh tế, nguyện vọng cá nhân, định hướng nghề nghiệp, kết quả học tập phổ thông, kế hoạch học tập, chương trình đào tạo, đội ngũ giảng viên, cơ sở vật chất, các khóa sinh viên, kết quả học tập tại đại học (đánh giá giữa kỳ, cuối kỳ, đánh giá quá trình,...); việc tham gia các tổ chức xã hội, đoàn thể, yếu tố nhân chủng học, yếu tố văn hóa, kinh tế, tâm lý học,... Chia sẻ bộ dữ liệu chung cho các trường Đại học, các trường chuyên nghiệp để có phân tích, định hướng trong vấn đề tuyển sinh của từng trường và giải quyết bài toán cân bằng cung-cầu, đảm bảo đáp ứng nhu cầu xã hội. Bộ Giáo dục và Đào tạo cũng cần tăng cường công tác chỉ đạo các sở Giáo dục và Đào tạo trong công tác tổ chức dạy học, kiểm tra đánh giá kết quả học tập ở trường phổ thông, hạn chế tình trạng học sinh có kết quả học tập tốt nhưng điểm thi tốt nghiệp trung học phổ thông lại thấp, điều này khiến các trường đại học, cao đẳng không đánh giá được chính xác năng lực của thí sinh khi xét tuyển. Tiếp tục giao các trường đại học chủ động xây dựng và công bố đề án tuyển sinh đảm bảo nguyên tắc tự chủ; theo đó, ngoài phương thức sử dụng kết quả kỳ thi trung học phổ thông quốc gia làm cơ sở tuyển sinh, có thể sử dụng các phương thức khác để tuyển sinh.

3. KẾT LUẬN

Ứng dụng công nghệ học máy, khai phá dữ liệu lớn trong phân tích dữ liệu sinh viên, bài báo đã đề cập đến phương thức lượng hóa hỗ trợ ra quyết định trong vấn đề tuyển sinh đại học hệ chính quy của Trường Đại học Thủ đô Hà Nội. Dựa trên quy mô đào tạo đại cương và điểm số tuyển sinh, điểm học năm thứ nhất và thứ hai đại học, thông qua ứng dụng kỹ thuật học máy, bài báo đưa ra kết quả dự báo về những nhân tố quan trọng trong tổ hợp tuyển sinh ảnh hưởng tới chất lượng học tập và kết quả đầu ra của sinh viên. Từ đó dự báo xu hướng và phương thức tuyển sinh của trường Đại học Thủ đô Hà Nội nói riêng, các trường đại học nói chung (khi có dữ liệu phù hợp).

TÀI LIỆU THAM KHẢO

1. An, N. T. T., Thành, N. V., Oanh, Đ. T. K., & Thứ, N. T. N. (2016), “Những nhân tố ảnh hưởng kết quả học tập của sinh viên năm I-II Trường Đại học Kỹ thuật - Công nghệ Cần Thơ”, *Tạp Chí Khoa Học Trường Đại Học Cần Thơ*, tr.46 - 82.

2. Sang, L. H., Điện, T. T., Nghe, N. T., & Hải, N. T. (2020), “Dự báo kết quả học tập bằng kỹ thuật học sâu với mạng nơ-ron đa tầng”, *Can Tho University Journal of Science*, 56(3) (June), tr.20-28.
3. Fei, M and Yeung, D-Y. (2015), “Temporal Models for Predicting Student Dropout in Massive Open Online Courses”, trong *Kỷ yếu Hội thảo 2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 256–263, DOI: <https://doi.org/10.1109/ICDMW.2015.174>.
4. Gray, G, McGuinness, C and Owende, P. (2014), *An application of classification models to predict learner progression in tertiary education*, trong *Kỷ yếu Hội thảo 2014 4th IEEE International Advance Computing Conference (IACC)*, 549–554. DOI: <https://doi.org/10.1109/IAdCC.2014.6779384>.
5. Masaeli, M., Fung, G. & Dy, J. G. (2010), “From transformation-based dimensionality reduction to feature selection”, trong *Kỷ yếu Hội thảo ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, tr. 751–758.
6. Shahiri, AM, Husain, W and Rashid, NA. (2015), “A Review on Predicting Student’s Performance Using Data Mining Techniques”, *Procedia Computer Science*, 72: 414–422. DOI: <https://doi.org/10.1016/j.procs.2015.12.157>.
- 8., S. B. Kotsiantis (2012), “Use of machine learning techniques for educational proposes: A decision support system for forecasting students’ grades”, *Artificial Intelligence Review*, 37(4), 331–344, DOI: <https://doi.org/10.1007/s10462-011-9234-x>.

APPLYING MACHINE LEARNING TECHNIQUES IN PROCESSING STUDENT DATA TO ASSIST UNIVERSITY ADMISSION

Abstract: The article presents research results on building a student data set and the results of applying machine learning techniques to make a prediction program for the type of student graduation, predicting factors in the student enrollment mix that affects student learning outcomes. To solve the above problems, we conducted a research on the primary education data set of Hanoi Metropolitan University (data for 5 years from 2016 to 2020). Machine learning techniques used include Logistic Regression technique (to predict student graduation results) and an improved technique of Linear discriminant analysis technique (to predict important factors affecting student learning outcomes) - Discriminative Feature Selection technique. From there, the authors make recommendations on the trend of enrollment at university level, some recommendations on training organization and enrollment strategy for Hanoi Metropolitan University.

Keywords: Machine learning, machine learning techniques, data science, education science, learning outcomes prediction, enrollment problem.