## **DỰ ĐOÁN KẾT QUẢ HỌC TẬP CỦA SINH VIÊN TRÊN NỀN TẢNG HỆ THỐNG QUẢN LÝ HỌC TẬP (LMS) BẰNG PHƯƠNG PHÁP HỌC MÁY**

### LÊ QUỐC HUY, PHẠM THỊ THIẾT\*

*1Khoa Công nghệ Thông tin, Trường Đại học Công nghiệp Thành phố Hồ Chí Minh*

[*22636191.huy@student.iuh.edu.vn*](mailto:22636191.huy@student.iuh.edu.vn) *, phamthithiet@iuh.edu.vn,*

**Tóm tắt** Hiện nay xu hướng học tập online ngày càng được chú trọng đặc biệt là trên các nền tảng trực tuyến nó càng được thể hiện rõ ràng hơn trong thời kì đại dịch Covid-19 vừa qua. Việc học trực tuyến cho phép hệ thống ghi lại quá trình tương tác giữa người học và chương trình học cũng như ghi lại các dữ liệu quan trọng của người học. Thông qua việc khai thác dữ liệu trên dữ liệu sinh viên có thể giúp chúng ta hiểu rõ hơn về chương trình học và sinh viên. Đồng thời có thể hỗ trợ cho người học kịp thời đưa ra những cảnh báo quan trọng đến quá trình học tập của sinh viên. Ở bài viết này chúng tôi sẽ sử dụng các mô hình học máy để có thể đánh giá được quá trình học tập của sinh viên. Kết quả của mô hình sẽ dự đoán được điểm số của sinh viên trong tương lai, từ đó có thể đưa ra những cảnh báo cần thiết giúp sinh viên có kế hoạch học tập và cải thiện kết quả tốt hơn.

**Từ khóa.** Dự đoán kết quả học tập của sinh viên , Hệ thống quản lý học tập (LMS) , Hiệu suất học tập của sinh viên , Phân loại sinh viên, Huấn luyện mô hình học máy …

**PREDICTING STUDENT LEARNING OUTCOMES ON THE LEARNING MANAGEMENT SYSTEM (LMS) PLATFORM USING MACHINE LEARNING METHODS**

**Abstract.** Currently, the trend of online learning is increasingly emphasized, especially on online platforms, which has become more apparent during the Covid-19 pandemic. Online learning allows the system to record the interaction process between learners and the curriculum as well as capture important learner data. Through data exploration on student data, we can better understand the curriculum and students. At the same time, it can support learners by timely providing important alerts to the student learning process. In this article, we will use machine learning models to evaluate the learning process of students. The model's results will predict students' grades in the future, thereby providing necessary alerts to help students plan their studies and improve their results.

**Keyword.** Predicting students’ academic performance, Learning Management System (LMS), Students' academic performance, Classifying students, Training machine learning models….

**1. GIỚI THIỆU**

Trong những năm gần đây, bối cảnh giáo dục đã trải qua một sự chuyển đổi đáng kể, được thúc đẩy bởi những tiến bộ trong công nghệ và sự chú trọng ngày càng tăng vào các nền tảng học tập trực tuyến. Đặc biệt đáng chú ý là sự gia tăng mạnh mẽ của học tập trực tuyến, một xu hướng càng được nhấn mạnh trong bối cảnh đại dịch toàn cầu chưa từng có do Covid-19 gây ra. Sự bùng phát đòi hỏi phải nhanh chóng chuyển sang các giải pháp học tập từ xa, nêu bật tính tất yếu của Hệ thống quản lý học tập (LMS) trong việc tạo điều kiện cho việc giảng dạy không bị gián đoạn.

Nền tảng LMS đóng vai trò là khung kỹ thuật số ghi lại và tạo điều kiện thuận lợi cho sự tương tác giữa người học và nội dung giáo dục. Hơn nữa, chúng đóng một vai trò quan trọng trong việc thu thập dữ liệu có giá trị của người học, mở đường cho những hiểu biết sâu sắc dựa trên dữ liệu về cả quá trình học tập và bản thân người học. Nguồn dữ liệu dồi dào này có tiềm năng to lớn trong việc nâng cao hiểu biết của chúng ta về các chương trình giáo dục và hành vi của sinh viên, cuối cùng góp phần tạo ra các phương pháp dạy và học hiệu quả hơn.

Bằng cách thu hẹp khoảng cách giữa phân tích dữ liệu và thực tiễn giáo dục, nghiên cứu của chúng tôi nỗ lực đóng góp vào cuộc tranh luận đang diễn ra xung quanh việc tối ưu hóa trải nghiệm học tập trực tuyến. Chúng tôi tin rằng bằng cách khai thác sức mạnh của máy học kết hợp với nền tảng LMS có thể mở ra những con đường mới cho việc học tập được cá nhân hóa và thúc đẩy cải tiến liên tục về kết quả học tập của sinh viên. Thông qua việc khám phá này, chúng tôi mong muốn mở đường cho một bối cảnh giáo dục thích ứng với xu hướng phát triển hiện tại đáp ứng nhu cầu về chất lượng của người học trong thời đại kỹ thuật số. Hơn nữa, để tăng tính chất năng động của nền tảng LMS cho phép can thiệp và hỗ trợ kịp thời cho người học bằng cách xác định các vấn đề quan trọng và cung cấp phản hồi có liên quan, chúng ta cần có các mô hình dự đoán có khả năng đánh giá lộ trình học tập của học sinh và đưa ra cảnh báo kịp thời để hỗ trợ tiến độ học tập của các sinh viên.

Đối với môn học Kỹ Thuật Lập trình, việc học đòi hỏi nhiều sự kiên nhẫn và quyết tâm, điều này nói lên việc luyện tập ở nhà rất nhiều để có thể cải thiện khả năng lập trình. Bài tập tự học là một cách rất hiệu quả để sinh viên tiến hành thực hành thực tế. Ngoài ra, các câu hỏi và trả lời trong lập trình có đặc điểm riêng, những câu hỏi khác với những câu hỏi thông thường. Để kiểm tra xem bài lập trình của mình có đúng hay không, sinh viên có khả năng nộp bài tập nhiều lần trên các bộ kiểm thử thông qua hệ thống học tập trên LMS. Các kết quả sau đó được trả về ngay lập tức, giúp sinh viên để gỡ lỗi mã của họ và cải thiện nó. Môi trường này tạo ra nhiều cơ hội tự học cho sinh viên.

Trong bài báo này, một mô hình để đánh giá quá trình học tập của sinh viên được đề xuất bằng cách sử dụng các thuật toán máy học để khai thác dữ liệu người học. Mục tiêu của bài báo gồm hai phần: thứ nhất là cung cấp phân tích chuyên sâu về kết quả học tập của học sinh, thứ hai là cung cấp những kết quả có thể xảy ra nhằm tạo điều kiện cho sinh viên có thể cải thiện học tập. Thông qua mô hình này để đưa ra các dự báo kết quả học tập trong tương lai của sinh viên.

**2. CÁC NGHIÊN CỨU LIÊN QUAN**

Trong lĩnh vực học máy và dự đoán kết quả học tập của sinh viên, đã có nhiều nghiên cứu và ứng dụng được tiến hành. Các nghiên cứu trước đây đã tập trung vào việc sử dụng các phương pháp học máy để phân tích dữ liệu học tập và dự đoán kết quả học tập của sinh viên.Việc dự đoán kết quả làm bài của sinh viên trên lớp giúp người dạy xác định được nhóm người học có kết quả thấp và hướng dẫn người học cải thiện kết quả học tập tốt hơn. Một số nghiên cứu đã sử dụng các mô hình học máy cổ điển như hồi quy tuyến tính, hồi quy logistic và máy vector hỗ trợ để dự đoán điểm số học tập của sinh viên dựa trên các biến đầu vào như điểm số trước đó, tham gia lớp học, hoạt động trên nền tảng học trực tuyến, và nhiều yếu tố khác. Các nghiên cứu khác đã tìm hiểu về việc sử dụng kỹ thuật khai phá dữ liệu để hiểu hơn về quá trình học tập của sinh viên. Các phương pháp này thường bao gồm việc phân tích mẫu dữ liệu, phân tích chuỗi thời gian, và khám phá quy luật tiềm ẩn trong dữ liệu học tập. Ngoài ra, cũng có những nghiên cứu tập trung vào việc phát triển các hệ thống cảnh báo để cung cấp thông tin phản hồi và hỗ trợ cho sinh viên trong quá trình học tập. Các hệ thống này thường sử dụng các thuật toán máy học để phát hiện ra các mẫu không bình thường trong hành vi học tập của sinh viên và cung cấp các gợi ý hoặc cảnh báo phù hợp [[1]](#_bookmark0) [[2]](#_bookmark1). Tuy nhiên, mặc dù đã có nhiều nghiên cứu về việc dự đoán kết quả học tập của sinh viên, việc áp dụng các phương pháp này trên nền tảng học trực tuyến vẫn còn nhiều thách thức. Các thách thức này bao gồm sự phức tạp của dữ liệu học tập trực tuyến, sự không chắc chắn trong việc thu thập dữ liệu, và khả năng tích hợp các hệ thống dự đoán vào quá trình giảng dạy và hỗ trợ sinh viên một cách hiệu quả.

Dưới đây là một số công trình liên quan đến việc khai phá dữ liệu học tập của sinh viên [[3]](#_bookmark2). Nghiên cứu này cho thấy việc tìm ra yếu tố ảnh hướng đến kết quả của sinh viên năm đầu học tập rất quan trọng vì nó có thể ảnh hưởng đến động lực học tập của sinh viên sau này. Bài viết đã sử dụng hai thuật toán là Random Forest và Support Vector Machine, kết quả cho thấy rằng Random Forest phân loại vượt trội hơn Support Vector Machine, mô hình của bài viết này cũng có thể được áp dụng chính xác hiệu suất học tập của sinh viên theo từng khóa học.

Trong nghiên cứu [4], các tác giả áp dụng học máy để dự đoán điểm số học tập của sinh viên dựa trên các dữ liệu học tập trực tuyến. Họ sử dụng các phương pháp như hồi quy tuyến tính, máy vector hỗ trợ, …để tạo ra các mô hình dự đoán. Kết quả của bài viết này chỉ ra rằng thuật toán tốt nhất để phân loại sinh viên là Naïve Bayes có độ chính xác lên đến 98%

Nghiên cứu [5] tập trung vào dự đoán hiệu suất học tập của sinh viên sử dụng phương pháp hồi quy để dự đoán điểm và xếp loại học tập của sinh viên dựa trên hồ sơ học tập trước đó với độ chính xác 96.64%.

Trong nghiên cứu [6], này đã dự đoán hiệu suất của sinh viên trong kỳ thi cuối kỳ. Các tác giả đề nghị sử dụng K-Nearest Neighbor và Support Vector Machine, kết quả cho thầy rằng Support Vector Machine hoạt động hiệu quả hơn với độ chính xác là 96%.

Nghiên cứu [7] tập trung vào việc cải thiện tính cá nhân hóa trong hiệu suất học tập của sinh viên bằng cách sử dụng các bài kiểm tra động kèm theo một mô hình dự đoán. Đặc biệt, mục tiêu chính là tạo ra các bài kiểm tra động thích ứng để đánh giá hiệu suất học tập của sinh viên, đồng thời liên tục so sánh kết quả của việc đánh giá với mô hình dự đoán dựa trên cây quyết định. Nghiên cứu này mang lại hiểu biết sâu sắc về mức độ hiệu quả của phương pháp này trong việc cải thiện tương tác học tập động thích ứng.

Trong nghiên cứu [8], các tác giả tập trung vào dự đoán việc sinh viên sẽ bỏ học trong các khóa học trực tuyến. Họ sử dụng phương pháp cây quyết định để phân loại sinh viên thành các nhóm có nguy cơ bỏ học khác nhau dựa trên dữ liệu học tập và hành vi trước đó.

Ngoài những mô hình học tập truyền thống thì lớp học đảo ngược (Flipped classroom) là một mô hình giảng dạy đang thu hút sự chú ý trong lĩnh vực giáo dục. Khái niệm này đề cập đến việc đảo ngược của quá trình học truyền thống, trong đó các sinh viên đọc tài liệu và thông tin liên quan bài học trước khi tham gia vào các hoạt động hướng dẫn trực tiếp trong lớp học. Trong mô hình này, nội dung giảng dạy trực tuyến thường được cung cấp qua các tài liệu video, bài giảng hoặc bài đọc trước khi lớp học diễn ra. Việc này tạo điều kiện cho học viên tự học, tự điều chỉnh thời gian và tốc độ học của mình, đồng thời giúp giảng viên có thể dành thời gian trong lớp để tương tác trực tiếp với sinh viên, tập trung vào việc hỗ trợ và thúc đẩy quá trình học tập.

Một trong những ưu điểm nổi bật của lớp học đảo ngược là khả năng tăng cường sự tương tác giữa sinh viên và giảng viên. [K](#_bookmark8)ết quả của các nghiên cứu [9, 10] cho thấy việc tham gia vào lớp học đảo ngược (Flipped classroom) đã tạo ra một môi trường học tập tích cực giữa sinh viên, nó cũng kích thích sự sáng tạo và khám phá trong quá trình học. Các sinh viên trong lớp học đảo ngược cũng có xu hướng tập trung và làm việc hiệu quả hơn vào nhiệm vụ học tập. Các tác giả nhấn mạnh rằng lớp học đảo ngược không chỉ là một phương pháp giảng dạy mới mẻ mà còn là một công cụ mạnh mẽ để nâng cao hiệu suất học tập của sinh viên và thúc đẩy sự phát triển của họ trong môi trường học tập hiện đại.

Thay vì chỉ là người truyền đạt kiến thức, giáo viên trở thành người hướng dẫn và trợ giúp sinh viên hiểu sâu hơn thông qua các bài thảo luận, bài tập thực hành và dự án ngoại khóa. Hơn nữa, mô hình này cũng khuyến khích sự tích cực của sinh viên trong quá trình học, vì họ phải chuẩn bị trước khi đến lớp và trở thành người chủ động trong quá trình học tập.

1. *Hoạt động trước lớp (Pre-class ) và tại lớp (In-class) trong dạy học lập trình*

Các hoạt động trước lớp (pre-class activities) đóng vai trò quan trọng trong phương pháp giảng dạy lớp học đảo ngược. Chúng cung cấp cơ hội cho học sinh chuẩn bị kiến thức trước khi đến lớp, từ đó tối ưu hóa thời gian học tập trong lớp. Một phương pháp khác so với lớp học đảo ngược là giao các bài tập hoặc câu hỏi trước lớp để học sinh tự làm và chuẩn bị câu trả lời. Điều này khuyến khích sinh viên suy nghĩ và áp dụng kiến thức một cách sâu sắc hơn. Ưu điểm của phương pháp này là nó kích thích sự tương tác và thảo luận trong lớp, cũng như giúp học sinh xác định được những khó khăn cụ thể mà họ cần giúp đỡ. Tuy nhiên, một nhược điểm có thể là thời gian cần thiết để học sinh hoàn thành các bài tập trước lớp có thể không đồng đều và đôi khi có thể làm giảm sự đồng nhất trong quá trình học. Những phương pháp khác bao gồm giao bài tìm hiểu tự nghiên cứu, thảo luận trước lớp qua các diễn đàn trực tuyến, hoặc thậm chí là việc sử dụng các ứng dụng và trò chơi trực tuyến để học sinh tiếp cận kiến thức trước khi đến lớp. Mỗi phương pháp đều mang lại những cơ hội và thách thức riêng, và việc lựa chọn phương pháp phù hợp phụ thuộc vào mục tiêu học tập cụ thể và nhu cầu của lớp học.

Sau khi sinh viên đã chuẩn bị các pre-class activities, tại thời điểm trên lớp, giảng viên sẽ sử dụng các in-class activities để hỗ trợ quá trình học tập của họ, giúp họ củng cố các kiến thức đã tự học. Các in-class activities được thiết kế đa dạng, bao gồm các hình thức như thảo luận nhóm, trình bày trước lớp, và các hoạt động tương tác nhóm khác.

Một trong những hình thức phổ biến là thảo luận nhóm, trong đó sinh viên được phân nhóm để thảo luận và chia sẻ quan điểm về các vấn đề liên quan đến chủ đề học tập. Điều này khuyến khích sự tương tác và trao đổi ý kiến giữa các sinh viên, từ đó mở rộng góc nhìn và hiểu biết của họ về vấn đề. Ngoài ra, việc yêu cầu sinh viên trình bày trước lớp cũng là một cách hiệu quả để họ áp dụng kiến thức đã học và thể hiện sự hiểu biết của mình. Điều này không chỉ giúp họ củng cố kiến thức mà còn phát triển kỹ năng giao tiếp và tự tin trình bày ý kiến trước đám đông. Bằng cách tích hợp các hoạt động đa dạng như vậy vào lớp học, giảng viên có thể tối ưu hóa sự tham gia và hiệu quả của sinh viên trong quá trình học tập, từ đó tạo ra một môi trường học tập tích cực và động viên sự phát triển cá nhân của họ.

1. *Dự đoán kết quả học tập của sinh viên bằng phương pháp Random Forest*

Trong lĩnh vực giáo dục, việc dự đoán kết quả học tập của sinh viên đóng vai trò quan trọng trong việc xác định những học sinh có nguy cơ bị tụt lại và cung cấp hỗ trợ phù hợp cho họ. Các nghiên cứu trước đây [11] [12] [13] để dự đoán kết quả học tập của sinh viên đã sử dụng nhiều phương pháp khác nhau, bao gồm hồi quy logistic, Naive Bayes, cây quyết định và Support vector machine (SVM). Trong khi đó, phương pháp Random Forest là một lĩnh vực đang thu hút sự quan tâm của nhiều nhà nghiên cứu. Đặc biệt trong dự đoán hiệu suất học tập của sinh viên. hiệu suất học tập của sinh viên thông qua việc sử dụng Random Forest là một phương pháp quan trọng và hiệu quả trong lĩnh vực giáo dục và học máy. Random Forest là một thuật toán học máy được xây dựng trên cơ sở cây quyết định, nó hoạt động bằng cách tạo ra một "rừng" các cây quyết định và kết hợp kết quả từ các cây này để đưa ra dự đoán cuối cùng.

Trong việc dự đoán hiệu suất học tập của sinh viên, Random Forest có thể sử dụng các biến đầu vào như điểm số trước đó, tần suất tham gia lớp học, hoạt động trên nền tảng học trực tuyến, và các thông tin khác để dự đoán kết quả học tập của sinh viên trong tương lai. Phương pháp này mang lại nhiều lợi ích, bao gồm khả năng xử lý các tập dữ liệu lớn, khả năng xử lý các biến đầu vào phức tạp, và khả năng giải thích kết quả dự đoán một cách dễ dàng.

Một trong những ứng dụng quan trọng của việc dự đoán hiệu suất học tập của sinh viên bằng Random Forest là việc cung cấp thông tin hữu ích cho giáo viên và nhà trường để hỗ trợ sinh viên. Thông qua việc dự đoán, họ có thể xác định các sinh viên có nguy cơ thấp hoặc cao về việc không hoàn thành khóa học và cung cấp các biện pháp hỗ trợ phù hợp để giúp sinh viên cải thiện hiệu suất học tập của mình. Đồng thời, việc dự đoán này cũng có thể giúp cá nhân hóa quá trình học tập và cung cấp thông tin phản hồi cho sinh viên để họ có thể điều chỉnh hành vi học tập của mình để đạt được kết quả tốt nhất. Trong tổng thể, việc sử dụng Random Forest trong dự đoán hiệu suất học tập của sinh viên là một công cụ mạnh mẽ và hữu ích cho cả sinh viên và nhà trường.

## **3. PHƯƠNG ÁN ĐỀ XUẤT**

1. *Thiết kế khóa học*

Nghiên cứu của chúng tôi tập trung vào việc đánh giá sinh viên học môn Kĩ Thuật Lập Trình. Vì lập trình là một ngành học đòi hỏi sự luyện tập nhiều lần nên việc thiếu sót là không thể tránh khỏi. Việc chuẩn bị bài trước khi đến lớp sẽ giúp cho sinh viên học hỏi được nhiều kiến thức hơn. Chính vì thế chúng tôi áp dụng Flipped Classroom vào trong các giờ học thực hành của môn học. Phương pháp triển khai của chúng tôi như sau, khóa học của chúng tôi thiết kế cho sinh viên bao gồm 4 bài tập trên lớp (Inlab) và 4 bài tập ở nhà trước khi đến lớp (Prelab). Mỗi bài tập trên lớp hay bài tập ở nhà chúng tôi đều cho sinh viên làm đi làm lại nhiều lần và sẽ lấy điểm cao nhất trong các lần làm bài. Chúng tôi bắt buộc sinh viên phải làm bài tập trước khi đến lớp(Prelab) và sau đó sinh viên sẽ được học thực hành tại lớp (Inlab) các vấn đề mà sinh viên làm tại nhà để có thể củng cố lại kiến thức. Cụ thể chúng tôi sẽ thiết kế Prelab và Inlab như sau:

*PreLab:*

* 1. Tasks: Sinh viên được yêu cầu phải làm các bài tập lập trình có độ khó từ dễ đến trung bình.
  2. Resources: Các câu hỏi lập trình (hands-on program- ming?) với độ khó từ dễ tới trung bình. Những câu hỏi này đã được review bởi tập thể giảng viên.
  3. Deadline: Sinh viên cần phải hoàn thành các bài tập này trước giờ học. Kết quả làm bài của Prelab sẽ được dùng để đánh giá kết quả học tập của sinh viên.
  4. Time for completion: Mỗi Prelab sẽ được mở sau khi đã dạy xong bài Inlab trước đó. Vì vậy, sinh viên

có thể linh động trong việc quyết định thời điểm làm bài (miễn là trước buổi học kế tiếp)

# *InLab:*

1. Tasks: Trước giờ dạy, giảng viên sẽ xem nhanh qua kết quả làm bài Prelab của sinh viên thông qua phân bố điểm của mỗi câu hỏi. Sau đó, giảng viên sẽ cùng thảo luận với sinh viên về các câu hỏi Prelab dựa trên bài nộp của họ. Sau khi đã làm rõ các vấn đề, sinh viên được yêu cầu phải làm các bài tập lập trình có độ khó từ trung bình đến khó.
2. Resources: Bài nộp Prelab của sinh viên để thảo luận; Các câu hỏi lập trình với độ khó từ trung bình tới khó. Những câu hỏi này đã được review bởi tập thể giảng viên.
3. Deadline: Sinh viên cần phải hoàn thành các bài tập Inlab này trước buổi học kế tiếp. Kết quả làm bài của Inlab sẽ được dùng để đánh giá kết quả học tập của sinh viên.
4. Time for completion: Các Inlab sẽ được mở sau khi giảng viên đã thảo luận với sinh viên. Sinh viên sẽ dựa trên những thảo luận này để có thể chinh phục những câu hỏi có độ khó cao hơn.
5. *Xử lý dữ liệu*

Chúng tôi xây dựng một công cụ để có thể lấy dữ liệu của sinh viên từ hệ thống LMS về để tiến hành phân tích. Kết quả của việc lấy dữ liệu từ hệ thống về chúng tôi thu được 2 tập dataset.

* 1. new repost.csv: dataset này chứa các thông tin về : Họ và Tên, Mã Số Sinh Viên (ID number) , Tình trạng làm bài (State) , thời bắt đầu (Started on) , thời gian kết thúc (Completed), thời gian dành ra để làm bài (Time taken) , Điểm số (Grade/10) , Số lượng câu hỏi(Q1/10) , mã câu hỏi (questionID)
  2. Question bank.csv: dataset chứa các thông tin về :Bài Lab( Lab), bài về nhà, bài lên lớp hay bài tự nguyện (Pre,In,Post), Tên câu hỏi (question name) , chủ đề (topic), cấp độ ( level ), link câu hỏi(bkel-link)

Chúng tôi sẽ liên kết 2 bộ dữ liệu trên với nhau thông qua các mã câu hỏi (questionID), tuy bộ dữ liệu question của chúng tôi không có thuộc tính QuestionID nhưng ở trong giá trị của thuộc tính (bkel-link) có các giá trị của QuestionID từ đó chúng tôi có thể liên kết 2 bộ dữ liệu này lại với nhau. Sau đó chúng tôi sinh ra một thuộc tính mới tạm gọi nó là attempt (số lần làm bài), chúng tôi đếm số lần làm bài của mỗi sinh viên theo từng bài lab để sinh ra thuộc tính này . Chúng tôi tách thuộc tính "Lab" ra thành các thuộc tính riêng biệt, trong thuộc tính "Lab" có 4 giá trị tương ứng với 4 bài lab chúng tôi sẽ tiến hành tách nó ra riêng lẻ để có thể dễ dàng phân tích. Sau khi hoàn thành những giai đoạn ở trên chúng tôi sẽ có một bộ dữ liệu hoàn chỉnh bao gồm 4 bài lab của sinh viên tự học (prelab) và 4 bài lab sinh viên được giảng viên hướng dẫn tại lớp (Inlab).Tương ứng đều có các đặc trưng như : điểm số, số lần làm bài, số câu hỏi làm bài, khoảng cách giữa lần đầu và lần cuối khi nộp bài.

Cấu trúc của mỗi bài lab về nhà (Prelab) sẽ có các đặc trưng tương ứng:

* + - Prelab : Điểm của bài tập về nhà
    - Prelab-attempts : số lần làm bài tập về nhà
    - Prelab-questions : số câu hỏi làm bài tập về nhà
    - Prelab-growths : khoảng điểm giữ lần đầu với lần cuối của bài tập về nhà

Cấu trúc của mỗi bài lab tại lớp (Inlab) sẽ có các đặc trưng tương ứng:

* + - Inlab : Điểm của bài tập tại lớp
    - Inlab-attempts : số lần làm bài tập tại lớp
    - Inlab-questions : số câu hỏi làm bài tập tại lớp
    - Inlab-growths : khoảng điểm giữ lần đầu với lần cuối của bài tập tại lớp

Tóm lại trong khóa học này, chúng tôi đã tổ chức 8 bài thực hành được chia thành hai loại: 4 bài thực hành được tiến hành trực tiếp trên lớp và 4 bài thực hành tự học tại nhà. Bằng cách tổng hợp tất cả các bài thực hành này lại, chúng tôi đã tạo ra một bộ dữ liệu đầy đủ để tiến hành các phân tích và nghiên cứu.

Mối quan hệ giữa Prelab và Inlab là liên quan với nhau. Prelab (Trước buổi học), có thể cung cấp thông tin về mức độ chuẩn bị của sinh viên trước khi tham gia vào buổi học. Sinh viên có thói quen tham gia tích cực vào Prelab thường có khả năng hiểu biết sâu sắc về nội dung học và chuẩn bị tốt hơn cho các hoạt động trong Inlab. Các dữ liệu từ Prelab như việc hoàn thành bài tập, tham gia vào các cuộc thảo luận trước lớp, hoặc sự tiếp cận với tài liệu học có thể được sử dụng làm chỉ số để đo lường sự chuẩn bị của sinh viên. Inlab (Trong buổi học), hiệu suất của sinh viên trong các hoạt động Inlab có thể phản ánh mức độ hiểu biết và khả năng áp dụng kiến thức của họ trong môi trường thực tế. Các kết quả từ các bài tập, thảo luận nhóm, hoặc các bài kiểm tra trong Inlab có thể được sử dụng để đánh giá năng lực và hiệu suất học tập của sinh viên. Sự tham gia tích cực và đóng góp trong Inlab có thể phản ánh sự hứng thú và cam kết của sinh viên đối với việc học tập trên lớp. Tuy nhiên có một số sinh viên không cố gắng làm bài nên không thể dự đoán được kết quả làm bài trên dữ liệu làm bài của họ. Do đó, những người học có số lần làm bài (attempt) thấp sẽ được loại bỏ khỏi dữ liệu. Tóm lại, Prelab và Inlab có thể cung cấp thông tin quan trọng về mức độ chuẩn bị và hiệu suất học tập của sinh viên, từ đó ảnh hưởng đến khả năng dự đoán điểm số của họ trong hệ thống quản lý học tập. Bằng cách kết hợp và phân tích các dữ liệu từ cả hai phần này, giáo viên hoặc hệ thống LMS có thể đưa ra dự đoán chính xác hơn về thành tích học tập của sinh viên

1. *Chọn đặc trưng*

Prelab là dữ liệu về các bài tập mà học viên hoàn thành trước khi tham gia lớp học trực tiếp. Đây thường là nơi thể hiện sự nỗ lực và cố gắng tự học của sinh viên. Inlab là dữ liệu về kết quả làm bài của sinh viên trong quá trình tham gia lớp học trực tiếp. Hai phần dữ liệu này cung cấp cái nhìn toàn diện về quá trình học tập của học viên từ việc tự học đến việc tham gia hoạt động học tập tại lớp. Đề xuất trong nghiên cứu này là sử dụng một số thuộc tính để mô tả chi tiết hơn về quá trình làm bài Lab của học viên, bao gồm:

* 1. ***Prelab***: điểm số của người học trong bài Prelab. Một bài Prelab bao gồm nhiều câu hỏi và mỗi câu hỏi cho phép người học nộp bài nhiều lần. Điểm của người học cho một câu hỏi là điểm cao nhất trong các lần nộp bài. Prelab- result được tính bằng trung bình điểm các câu hỏi thuộc bài Prelab đó.
  2. ***Prelab-attempts (Số lần thử):***: Đây là số lần mà sinh viên đã cố gắng hoàn thành một phần của bài lab hoặc bài tập trước khi nộp bản cuối cùng. Số lần thử này có thể là một chỉ báo về mức độ cam kết và cố gắng của sinh viên đối với việc hoàn thành nhiệm vụ. Nếu một sinh viên có nhiều lần thử, có thể cho thấy họ đang gặp khó khăn và cần hỗ trợ hoặc chỉ báo về một tiến triển học tập mạnh mẽ.
  3. ***Prelab-questions (Số câu hỏi):***: Đây là số lượng câu hỏi mà sinh viên đã đạt trong quá trình làm bài lab hoặc bài tập. Số lượng câu hỏi này có thể cho thấy mức độ hiểu biết của sinh viên về nội dung cũng như mức độ tương tác và sự tò mò của họ đối với chủ đề. Một số lượng câu hỏi lớn có thể cho thấy một tinh thần nghiên cứu tích cực hoặc sự cần thiết của hỗ trợ thêm đối với sinh viên
  4. ***Prelab-growths (Sự phát triển):***Đây có thể là một chỉ số về sự phát triển trong kiến thức hoặc kỹ năng của sinh viên sau mỗi lần làm bài lab hoặc bài tập. Sự phát triển có thể được đo lường dựa trên việc làm thay đổi trong số lượng câu trả lời đúng, sự tăng lên về độ khó của các câu hỏi hoặc sự cải thiện trong thời gian làm bài.
  5. ***Prelab-timeSpent***: Đây là thời gian mà sinh viên đã chi tiêu để hoàn thành một phần của bài lab hoặc bài tập. Thời gian dành có thể là một chỉ báo về mức độ cố gắng của sinh viên cũng như mức độ khó khăn của nhiệm vụ. Một số lượng thời gian lớn có thể chỉ ra rằng sinh viên gặp khó khăn hoặc cần hỗ trợ để hoàn thành nhiệm vụ.
  6. ***Prelab-lastSubmit***:Đây là thời gian mà sinh viên đã nộp bài làm cuối cùng. Thời gian nộp cuối cùng có thể cung cấp thông tin về mức độ tự quản lý thời gian của sinh viên cũng như sự chuẩn bị và kỹ năng quản lý công việc của họ. Sinh viên thường cần hỗ trợ nếu họ thường xuyên nộp trễ.

Thành phần thứ hai sẽ là dữ liệu làm bài tại lớp học (Inlab)

1. ***Inlab***: điểm số của người học trong bài Inlab. Một bài Prelab bao gồm nhiều câu hỏi và mỗi câu hỏi cho phép người học nộp bài nhiều lần. Điểm của người học cho một câu hỏi là điểm cao nhất trong các lần nộp bài. Inlab- result được tính bằng trung bình điểm các câu hỏi thuộc bài Inlab đó.
2. ***Inlab-attempts (Số lần thử):***: Số lần mà sinh viên đã cố gắng hoàn thành một phần của bài lab hoặc bài tập trước khi nộp bản cuối cùng. Số lần thử này có thể là một chỉ báo về mức độ cam kết và cố gắng của sinh viên đối với việc hoàn thành nhiệm vụ. Nếu một sinh viên có nhiều lần thử, có thể cho thấy họ đang gặp khó khăn và cần hỗ trợ hoặc chỉ báo về một tiến triển học tập mạnh mẽ.
3. ***Inlab-questions (Số câu hỏi):***: Số lượng câu hỏi mà sinh viên đã đặt trong quá trình làm bài lab hoặc bài tập. Số lượng câu hỏi này có thể cho thấy mức độ hiểu biết của sinh viên về nội dung cũng như mức độ tương tác và sự tò mò của họ đối với chủ đề. Một số lượng câu hỏi lớn có thể cho thấy một tinh thần nghiên cứu tích cực hoặc sự cần thiết của hỗ trợ thêm đối với sinh viên
4. ***Inlab-growths (Sự phát triển):***Đây là một chỉ số về sự phát triển trong kiến thức hoặc kỹ năng của sinh viên sau mỗi lần làm bài lab hoặc bài tập. Sự phát triển có thể được đo lường dựa trên việc làm thay đổi trong số lượng câu trả lời đúng, sự tăng lên về độ khó của các câu hỏi hoặc sự cải thiện trong thời gian làm bài.
5. ***Inlab-timeSpent***: Đây là thời gian mà sinh viên đã chi tiêu để hoàn thành một phần của bài lab hoặc bài tập. Thời gian dành có thể là một chỉ báo về mức độ cố gắng của sinh viên cũng như mức độ khó khăn của nhiệm vụ. Một số lượng thời gian lớn có thể chỉ ra rằng sinh viên gặp khó khăn hoặc cần hỗ trợ để hoàn thành nhiệm vụ.
6. ***Inlab-lastSubmit***:Đây là thời gian mà sinh viên đã nộp bài làm cuối cùng. Thời gian nộp cuối cùng có thể cung cấp thông tin về mức độ tự quản lý thời gian của sinh viên cũng như sự chuẩn bị và kỹ năng quản lý công việc của họ. Sinh viên thường cần hỗ trợ nếu họ thường xuyên nộp trễ.

Để khuyến khích sự cẩn thận trong việc làm bài, chúng tôi đã giới hạn số lần thử cho mỗi câu hỏi là 5 lần. Đối với số lượng câu hỏi trong mỗi bài lab, được ký hiệu là "number Of Questions", số lần thử tối đa là 5. Thêm vào đó, chúng tôi sử dụng chỉ số "Prelab-attempts" để đo lường số lần thử của người học trong mỗi bài lab. Điều này là cần thiết vì số lần thử không hoàn toàn phản ánh số câu hỏi mà người học đã làm. Chẳng hạn, nếu một người học làm 5 câu hỏi với mỗi câu hỏi thử 2 lần, "Prelab-attempts" sẽ là 10. Tương tự, nếu người học làm 10 câu hỏi với mỗi câu hỏi thử 1 lần, "Prelab-attempts" cũng sẽ là 10. Điều này cho thấy số lần thử không phản ánh chính xác số câu hỏi đã làm trong bài lab. Do đó, yếu tố này được sử dụng để đánh giá. Trong mỗi bài lab có thể được thực hiện bằng cách so sánh số lần thử (Prelab- attempts) với số lần thử tối đa có thể có là 5, được xác định bởi "numberOfQuestions". Nếu số lần thử của một sinh viên "attempts" của 1 bài lab (có thể là Prelab hay Inlab) lớn hơn 5, điều này có thể cho thấy sinh viên đã làm nhiều câu hỏi trong bài lab. Tuy nhiên, nếu số lần thử là rất thấp, có thể người học chỉ làm một phần của các câu hỏi hoặc không hoàn thành bài lab. Tóm lại, việc đánh giá dựa trên số lần thử có thể giúp ta hiểu được mức độ hoàn thành và tương tác của sinh viên trong quá trình làm bài lab.

Mối quan hệ giữa Prelab và Inlab: Prelab phần nào thể hiện thái độ, hành vi, thói quen, sự cố gắng, chủ động của người học trong quá trình tự học. Chúng tôi chọn điểm số của người học trong Inlab (gọi là Inlab-result và được xem là nhãn của quá trình dự đoán) là yếu tố duy nhất để xác định kết quả dự đoán. Inlab sẽ có khoảng giá trị từ 0 đến 10, được tính tương tự như cách tính Prelab-result

**4. KẾT QUẢ THỰC NGHIỆM**

1. *Dữ liệu thực nghiệm*

Khóa học của chúng tôi thiết kế cho sinh viên sẽ bao gồm 4 bài inlab và 4 bài prelab. Kết quả làm bài tại lớp(Inlab) được trích xuất để dự đoán kết quả làm bài cho các lần Inlab tiếp theo ( mỗi lần sinh viên đến lớp sẽ có được một dự báo từ chúng tôi) bằng cách dựa vào dữ liệu Prelab của sinh viên tự học ở nhà hoặc tổng hợp các prelab và các Inlab đã học trước đó để dự đoán bài Inlab đến lớp gần nhất của sinh viên. Kết quả làm bài Inlab sẽ được chuyển đổi thành các nhãn theo quy tắc sau:

* + Điểm (>=0 và) < 4: học lực kém label là 0
  + Điểm >= 4 và < 5.5: học lực trung bình label là 1
  + Điểm >= 5.5 và < 7.0: học lực khá label là 2
  + Điểm >= 7.0 và < 8.5: học lực giỏi label là 3
  + Điểm >= 8.5 (và <= 10): học lực xuất sắc label là 4

Chúng tôi chọn khung điểm này làm tiêu chuẩn để dự đoán xếp loại của sinh viên, phù hợp với cách đánh giá hiện nay trong hệ thống giáo dục của Việt Nam.

Mục đích của nghiên cứu này là xây dựng một mô hình học máy để đánh giá quá trình học tập ở nhà và tại trường của sinh viên theo thang điểm đã đề cập ở trên và hệ thống sẽ đưa ra các cảnh báo kịp thời để hỗ trợ sinh viên cải thiện điểm số. Mô hình này sẽ phân tích dữ liệu về hành vi học tập của sinh viên, bao gồm việc học tập tại nhà và tương tác với giảng viên trong quá trình học tập môn học Kĩ Thuật Lập Trình. Mục tiêu cuối cùng là dự đoán tiến độ học tập của sinh viên và cung cấp gợi ý cụ thể để họ có thể cải thiện kết quả học tập của mình.

Trong nghiên cứu này, chúng tôi đã lựa chọn các mô hình như Random Forest, Logistic Regression, MLP, và KNN để giải quyết bài toán đánh giá quá trình học tập của sinh viên và đưa ra các dự đoán về kết quả học tập trong tương lai. Sau đó, chúng tôi sẽ đề xuất một mô hình có độ chính xác cao nhất để áp dụng vào thực tế.

Để đánh giá hiệu suất của mô hình, chúng tôi đã sử dụng kỹ thuật Cross-validation. Đây là một phương pháp phổ biến trong machine learning, cho phép chia bộ dữ liệu thành các phần nhỏ và thực hiện quá trình training và testing trên các phần này. Trong trường hợp này, chúng tôi chia bộ dữ liệu thành 5 phần, trong đó một phần được dành cho testing và 4 phần còn lại được sử dụng cho training. Quá trình này được lặp lại 5 lần, mỗi lần sử dụng một phần khác nhau làm test data và còn lại làm train data.

Sau mỗi lần thực hiện Cross-validation, chúng tôi ghi nhận lại độ chính xác của mô hình trên tập test và tính trung bình của các kết quả này. Qua đó, chúng tôi có thể đánh giá được hiệu suất trung bình của mô hình trên các bộ dữ liệu khác nhau. Điều này giúp chúng tôi có cái nhìn tổng quan về độ chính xác và ổn định của mô hình Random Forest trong việc dự đoán kết quả học tập của sinh viên.

Việc sử dụng kỹ thuật Cross-validation không chỉ giúp cải thiện tính khả thi và độ chính xác của mô hình, mà còn giúp kiểm tra tính tổng quát hóa của mô hình trên dữ liệu mới. Điều này là quan trọng để đảm bảo rằng mô hình có thể tổng quát hóa tốt trên dữ liệu thực tế và không bị overfitting.

1. *Đánh giá thực nghiệm*

Trong quá trình nghiên cứu, chúng tôi đã tiến hành thử nghiệm và so sánh giữa các mô hình máy học để giải quyết bài toán này. Các mô hình này bao gồm một loạt các phương pháp từ học máy cơ bản đến các mô hình phức tạp hơn. Tuy nhiên, để đảm bảo độ chính xác và hiệu suất của mô hình được tối ưu hóa, chúng tôi đã tiến hành việc lựa chọn một mô hình cuối cùng từ số lượng các mô hình đã thử nghiệm. Mô hình cuối cùng này được lựa chọn dựa trên nhiều yếu tố, bao gồm hiệu suất dự đoán, độ ổn định của kết quả, cũng như khả năng xử lý dữ liệu lớn một cách hiệu quả. Kết quả của chúng tôi chỉ ra rằng mô hình hoạt động khá tốt.

Chúng tôi thử chạy một số mô hình học máy như Random Forest, Logistic Regression, Multi-Layer Perceptron (MLP), K- nearest neigh- bors(KNN) .Các số liệu trong các bảng I,II,III,IV thể hiện hiệu suất của các mô hình khi chạy với tập dữ liệu thực nghiệm gồm các đặc trưng là của 4 bài Inlab và 4 bài Prelab :

* Prelab: Điểm của bài tập về nhà
* Prelab-attempts : số lần làm bài tập về nhà
* Prelab-questions : số câu hỏi làm bài tập về nhà
* Prelab-growths : khoảng điểm giữ lần đầu với lần cuối của bài tập về nhà
* Pre-timeSpent : thời gian mà sinh viên đã chi tiêu để hoàn thành một phần của bài lab tại nhà
* Prelab-lastSubmit :thời gian mà sinh viên đã nộp bài làm cuối cùng.
* Inlab : Điểm của bài tập tại lớp
* Inlab-attempts : số lần làm bài tập tại lớp
* Inlab-questions : số câu hỏi làm bài tập tại lớp
* Inlab-growths : khoảng điểm giữ lần đầu với lần cuối của bài tập về nhà
* Inlab-timeSpent : thời gian mà sinh viên đã chi tiêu để hoàn thành một phần của bài lab tại nhà
* Inlab-lastSubmit :thời gian mà sinh viên đã nộp bài làm cuối cùng.

Dựa vào kết cho thấy các mô hình có hiệu xuất khá tốt. Như vậy, độ chính xác khi dự đoán trên tập dữ liệu nhìn chung là tương đối.

*Bảng I : Độ chính xác khi dự đoán bài tập trên lớp (Inlab1)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Inlab1 | Accuracy | Precision | Recall | F1-core |
| Random-Forest | 0.78 | 0.62 | 0.68 | 0.68 |
| Logistic-Regression | 0.79 | 0.57 | 0.75 | 0.65 |
| MLP | 0.75 | 0.57 | 0.75 | 0.65 |
| KNN | 0.72 | 0.59 | 0.75 | 0.66 |

*Bảng II : Độ chính xác khi dự đoán bài tập trên lớp (Inlab2)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Inlab2 | Accuracy | Precision | Recall | F1-core |
| Random Forest | 0.90 | 0.84 | 0.9 | 0.87 |
| Logistic Regression | 0.9 | 0.81 | 0.90 | 0.85 |
| MLP | 0.9 | 0.81 | 0.9 | 0.85 |
| KNN | 0.88 | 0.81 | 0.9 | 0.85 |

*Bảng III : Độ chính xác khi dự đoán bài tập trên lớp (Inlab3)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Inlab3 | Accuracy | Precision | Recall | F1-core |
| Random Forest | 0.9 | 0.82 | 0.9 | 0.86 |
| Logistic Regression | 0.9 | 0.82 | 0.9 | 0.86 |
| MLP | 0.89 | 0.825 | 0.9 | 0.86 |
| KNN | 0.74 | 0.82 | 0.9 | 0.86 |

*Bảng IV : Độ chính xác khi dự đoán bài tập trên lớp (Inlab4)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Inlab4 | Accuracy | Precision | Recall | F1-core |
| Random Forest | 0.9 | 0.81 | 0.9 | 0.85 |
| Logistic Regression | 0.9 | 0.81 | 0.9 | 0.85 |
| MLP | 0.89 | .81 | 0.9 | 0.852 |
| KNN | 0.90 | 0.81 | 0.9 | 0.852 |

Trong nghiên cứu này chúng tôi đã nhận thấy rằng đặc trưng về thời gian "timeSpent"và "lastSubmit" trong cả Prelab và Inlab không đem lại hiệu quả mong đợi khi sử dụng trong mô hình phân loại. Thực tế cho thấy, thông tin về thời gian này thường không phản ánh đúng mức độ nỗ lực và chăm chỉ của sinh viên trong quá trình làm bài. Việc nhiều sinh viên chỉ mở bài mà không tiến hành làm, hoặc để bài làm treo một thời gian dài, tạo ra nhiễu loạn trong dữ liệu huấn luyện. Tuy nhiên trong quá trình thực nghiệm trên các mô hình cho thấy các đặc trưng về thời gian như "timeSpent" và "lastSubmit" của Inlab và Prelab làm cho dữ liệu huấn luyện bị nhiễu và ảnh hưởng đến hiệu suất trong việc phân loại sinh viên. Bên cạnh đó, thực tế cho thấy, thông tin về thời gian này thường không phản ánh đúng mức độ nỗ lực và chăm chỉ của sinh viên trong quá trình làm bài vì trong quá trình làm bài các sinh viên không làm một cách nghiêm túc. Rất nhiều sinh viên mở bài làm nhưng họ không làm và treo nó tận hàng giờ đồng hồ nên vì vậy đặc trưng thời gian đối với trường hợp này không hiệu quả.

Chúng tôi có thử điều chỉnh về độ thời gian trung bình mà các sinh viên cần bỏ ra để làm bài nhưng có vẻ không khả quan vì độ phức tạp của việc điều chỉnh hơn thế việc thêm các đặc trưng này vào mô hình làm cho mô hình của chúng tôi giảm độ chính xác mặc dù chúng tôi đã cố gắng điều chỉnh giá trị của nó về khoảng trung bình mà các bạn sinh viên cần bỏ ra để làm bài. Vì vậy chúng tôi đã loại bổ đặc trung "timeSpent" và "lastSubmit" ra khỏi mô hình huấn luyện của chúng tôi , đặc trưng này có lẽ chúng tôi sẽ sửa dụng lại trong bài viết tương lai khi chúng tôi có các để điều chỉnh nó lại về trạng thái thích hợp

*Bảng V : Độ chính xác khi dự đoán bài tập trên lớp (Inlab1)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Inlab1 | Accuracy | Precision | Recall | F1-core |
| Random Forest | 0.80 | 0.62 | 0.68 | 0.75 |
| Logistic Regression | 0.79 | 0.59 | 0.75 | 0.66 |
| MLP | 0.79 | 0.59 | 0.75 | 0.66 |
| KNN | 0.79 | 0.59 | 0.75 | 0.66 |

*Bảng VI : Độ chính xác khi dự đoán bài tập trên lớp (Inlab2)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Inlab2 | Accuracy | Precision | Recall | F1-core |
| Random Forest | 0.90 | 0.86 | 0.91 | 0.88 |
| Logistic Regression | 0.9 | 0.86 | 0.90 | 0.87 |
| MLP | 0.9 | 0.83 | 0.9 | 0.86 |
| KNN | 0.9 | 0.86 | 0.9 | 0.87 |

*Bảng VII : Độ chính xác khi dự đoán bài tập trên lớp (Inlab3)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Inlab3 | Accuracy | Precision | Recall | F1-core |
| Random Forest | 0.9 | 0.82 | 0.9 | 0.86 |
| Logistic Regression | 0.9 | 0.82 | 0.9 | 0.86 |
| MLP | 0.89 | 0.825 | 0.9 | 0.86 |
| KNN | 0.74 | 0.82 | 0.9 | 0.86 |

*Bảng VIII : Độ chính xác khi dự đoán bài tập trên lớp (Inlab4)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Inlab4 | Accuracy | Precision | Recall | F1-core |
| Random Forest | 0.92 | 0.81 | 0.9 | 0.85 |
| Logistic Regression | 0.9 | 0.81 | 0.9 | 0.85 |
| MLP | 0.89 | 0.81 | 0.9 | 0.852 |
| KNN | 0.90 | 0.81 | 0.9 | 0.852 |

Sau khi bỏ các đặc trưng liên quan đến thời gian cho thấy các đặc trưng được đề xuất cho dữ liệu làm bài Prelab kết hợp với các bài Inlab trước đó là khả thi để dự đoán kết quả làm bài Inlab tiếp theo của sinh viên. Chúng tôi đề xuất chọn mô hình Random Forest để tiến hành ứng dụng vào trong thực tế vì Random Forest là một phương pháp học máy dựa trên nguyên lý của Ensemble Learning, nơi nhiều mô hình được kết hợp lại để tạo ra một dự đoán cuối cùng. Đặc điểm nổi bật của Random Forest là sự kết hợp giữa việc sử dụng nhiều cây quyết định (Decision Trees) và sự ngẫu nhiên trong việc lựa chọn các đặc trưng và dữ liệu mẫu để xây dựng các cây quyết định này. Quá trình xây dựng một mô hình Random Forest bắt đầu bằng việc tạo ra một tập hợp các cây quyết định. Mỗi cây quyết định được xây dựng thông qua việc lựa chọn một tập hợp ngẫu nhiên các đặc trưng từ tập dữ liệu và một tập hợp ngẫu nhiên các mẫu từ tập dữ liệu huấn luyện. Điều này giúp mỗi cây quyết định được huấn luyện trên một phần nhỏ và đa dạng của dữ liệu. Sau khi tạo ra các cây quyết định, mô hình Random Forest kết hợp kết quả từ tất cả các cây để ra quyết định cuối cùng.

Dựa vào các kết quả thực nghiệm ở trên, mô hình Random Forest không chỉ đạt được hiệu suất ổn định mà còn mang lại các kết quả dự đoán chính xác nhất so với các mô hình khác. Sự linh hoạt và khả năng làm việc hiệu quả với dữ liệu lớn của mô hình này đã làm nổi bật nó trước các mô hình khác. Điều này thể hiện sự phù hợp của mô hình Random Forest trong việc giải quyết bài toán đánh giá và dự đoán quá trình học tập của sinh viên.

**5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

Trong bài viết này, chúng tôi đã đề xuất một phương pháp để dự đoán kết quả học thực hành môn lập trình trên lớp thông qua quá trình chuẩn bị bài của người học. Quá trình học thực hành môn lập trình được chia thành 2 giai đoạn: Prelab là giai đoạn làm bài trước khi đến lớp và Inlab là giai đoạn làm bài trong giờ học trực tiếp. Cả hai giai đoạn đều được thực hiện trên hệ thống trực tuyến và ghi nhận dữ liệu làm bài của người học. Bài viết này đề xuất sử dụng các đặc trưng như result, attempts, growths, Prelab và Inlab. Các đặc trưng này thay đổi qua các lần huấn luyện mô hình. Chúng sẽ được lần lượt đưa vào mô hình, ví dụ nếu đầu ra của chúng tôi là Inlab1, thì các đặc trưng đầu vào sẽ bao gồm toàn bộ bài Prelab1. Tương tự, nếu đầu ra của chúng tôi là Inlab2, đầu vào sẽ bao gồm Prelab1, Inlab1 và Prelab2. Quá trình này lặp lại cho đến Inlab4. Điều này giúp tăng tính linh hoạt và hiệu quả của mô hình dự đoán. Mô hình dự đoán được đề xuất sử dụng là RanDom Forest. Cách thức học tập gồm Prelab và Inlab được triển khai thực tế tại trường Đại học trong môn Fundamentals of Programming (Kĩ Thuật Lập Trình ) trong 2 học kỳ: học kỳ 2 năm 2021-2022 (SEM-212) và học kỳ 2 năm 2022-2023 (SEM-222). Dựa vào kết quả thực nghiệm trong mục IV thì mô hình Random forest có độ chính xác cáo nhất trong tất cả các lần huấn luyện, cụ thể là: Inlab1: 0.9, Inlab2: 0.9, Inlab3: 0.9, Inlab4: 0.92. Do đó, bài báo đề xuất sử dụng mô hình học máy Random Forest là một hướng tiếp cận tốt nhất để đánh giá và dự đoán quá trình học tập của sinh viên. Việc kết hợp dữ liệu từ các bài lab tự học và bài lab tại lớp để dự đoán kết quả học tập trong tương lai là một phương pháp hiệu quả và có tiềm năng. Sử dụng hệ thống phân loại điểm số dựa trên quy tắc xếp loại học lực làm đầu vào cho mô hình giúp tiêu chuẩn hóa và dễ dàng đánh giá hiệu suất của sinh viên.

Bài toán "Dự đoán Hiệu suất Học tập của Sinh viên trên Nền tảng Hệ thống Quản lý Học tập (LMS) bằng phương pháp học máy" đóng vai trò quan trọng trong việc nghiên cứu và cải thiện quá trình giảng dạy và học tập trong môi trường giáo dục đại học hiện đại. Do đó, việc nghiên cứu và áp dụng các phương pháp đánh giá và dự đoán hiệu suất học tập của sinh viên trên nền tảng LMS không chỉ mang lại lợi ích cho sinh viên mà còn giúp cải thiện chất lượng giáo dục. Vì vậy, Chúng tôi sẽ áp dụng mô hình đề xuất ở bên trên vào môi trường thực tế để phục vụ sinh viên. Bên cạnh đó, chúng tôi sẽ tích hợp thêm một số mô hình dự đoán khác để có thể đưa ra những cảnh báo phù hợp với từng học viên. Song song đó chúng tôi sẽ cố gắng cải thiện mô hình để có độ chính xác cao hơn và hơn thế chúng tôi hy vọng sẽ có hướng phát triển những nghiên cứu khác liên quan đến việc dự đoán kết quả học tập của sinh viên như :

* + Sử dụng các phương pháp học máy và khai phá dữ liệu để phân tích từ nền tảng LMS. Các thuật toán này có thể được áp dụng để dự đoán hiệu suất học tập của sinh viên dựa trên các biến như lịch sử học tập, hoạt động trên nền tảng và kết quả bài kiểm tra.
  + Xây dựng các mô hình dự đoán và đánh giá dựa trên dữ liệu để xác định các yếu tố ảnh hưởng đến hiệu suất học tập của sinh viên. Các yếu tố này có thể bao gồm thời gian dành cho học tập, tương tác với nội dung học và mức độ tham gia vào các hoạt động học tập.
  + Phát triển các hệ thống tương tác thông minh để cung cấp gợi ý học tập cá nhân hóa và phản hồi tức thì cho sinh viên trên nền tảng LMS. Những hệ thống này có thể giúp sinh viên tăng cường hiệu suất học tập và nâng cao sự cam kết của họ đối với quá trình học.
  + Phân tích sâu hơn về hành vi học tập của sinh viên trên nền tảng LMS để hiểu rõ hơn về cách họ tương tác với nội dung học và vấn đề gặp phải trong quá trình học......

Đặc biệt, hiện nay xu hướng học tập khối ngành Kĩ Thuật đặc biệt là ngành IT thường liên quan đến việc học tập tại phòng thí nghiệm tại lớp. Sinh viên sẽ được học và nghiên cứu tại lớp sau đó làm thêm ở nhà sau giờ học. Tuy nhiên phương pháp này có những hạn chế như :

* Sinh viên không có đủ thời gian để hiểu vấn đề trong lớp. Sinh viên thường cảm thấy khó hiểu những khái niệm mới và cách áp dụng chúng vào thực tế ?
* Sinh viên không có cơ hội đặt câu hỏi sau khi họ gặp lỗi trong mã của họ. Chỉ có giảng viên có thời gian để cung cấp hướng dẫn về cách giải quyết vấn đề trong giờ học, sinh viên thường gặp khó khăn trong việc tự mình giải quyết vấn đề.
* Sinh viên có thể nhút nhát và không có khả năng để giải quyết bài toán trên lớp, mà thời gian dành cho việc trên lớp không đủ để đáp ứng các câu hỏi của toán bộ sinh viên

Ngoài việc sử dụng kết quả của mô hình đề xuất để dự đoán kết quả học tập của sinh viên, chúng tôi mong muốn tiếp tục nghiên cứu và đề xuất các mô hình tốt để có thể giải quyết các câu hỏi sau:

* Có thể tăng thời lượng cho sinh viên không ? dành thời gian viết mã ở nhà để thực hành và sử dụng lớp học hoạt động giải quyết vấn đề mà sinh viên gặp phải?
* Có nên tăng lượng câu hỏi trên lớp hoặc câu hỏi bài tập về nhà cho sinh viên không ?
* Có thể dự đoán được học sinh nào đang gặp khó khăn trong một khóa học lập trình để người hướng dẫn có thể cung cấp cho họ với sự hỗ của giảng viên?

REFERENCES

1. P. Balaji, S. Alelyani, A. Qahmash, and M. Mohana, “Contributions of machine learning models towards student academic performance prediction: a systematic review,” *Applied Sciences*, vol. 11, no. 21, p. 10007, 2021.
2. H. R. Hasan, A. S. A. Rabby, M. T. Islam, and S. A. Hossain, “Machine learning algorithm for student’s performance prediction,” in *2019 10th International Conference on Computing, Communication and* *Networking Technologies (ICCCNT)*. IEEE, 2019, pp. 1–7.
3. S. K. Ghosh, F. Janan, and I. Ahmad, “Application of the classification algorithms on the prediction of student’s academic performance,” Trends in Sciences, vol. 19, no. 14, pp.5070-5070,2022.
4. M. Pojon, “Using machine learning to predict student performance,” Master’s thesis, 2017.
5. S. Hussain and M. Q. Khan, “Student-performulator: Predicting stu- dents’ academic performance at secondary and intermediate level using machine learning,” *Annals of data science*, vol. 10, no. 3, pp. 637–655, 2023.
6. H.Al-Shehri, A. Al-Qarni, L. Al-Saati, A. Batoaq, H. Badukhen, S. Alrashed, J. Alhiyafi, and S. O. Olatunji, “ Student performance prediction using support vector machine and k-nerest neighbor,” in 2017 IEEE 30th canadian coferrnce on electrical and computer engineering (CCECE). IEE, 2017, pp. 1-4
7. V. Matzavela and E. Alepis, “Decision tree learning through a predic- tive model for student academic performance in intelligent m-learning environments,” *Computers and Education: Artificial Intelligence*, vol. 2,

p. 100035, 2021.

1. S. Kotsiantis, “Educational data mining: a case study for predicting dropout-prone students,” *International Journal of Knowledge Engineer-* *ing and Soft Data Paradigms*, vol. 1, no. 2, pp. 101–111, 2009.
2. J. Bergmann and A. Sams, *Flip your classroom: Reach every student in every class every day*. International society for technology in education, 2012.
3. J. Bishop and M. A. Verleger, “The flipped classroom: A survey of the research,” in *2013 ASEE Annual Conference & Exposition*, 2013, pp. 23–1200.
4. P. Cortez and A. M. G. Silva, “Using data mining to predict secondary school student performance,” 2008.
5. M. Nachouki and M. Abou Naaj, “Predicting student performance to improve academic advising using the random forest algorithm,” *Inter- national Journal of Distance Education Technologies (IJDET)*, vol. 20, no. 1, pp. 1–17, 2022.
6. L. H. Alamri, R. S. Almuslim, M. S. Alotibi, D. K. Alkadi, I. Ul- lah Khan, and N. Aslam, “Predicting student academic performance using support vector machine and random forest,” in *Proceedings of the 2020 3rd International Conference on Education Technology Management*, 2020, pp. 1