Phương Pháp Dự Báo Khả Năng Tốt Nghiệp Của Sinh Viên Dựa Trên Thuật Toán Phân Lớp

Methods for Student Graduation Prediction using Classification Algorithms

Đặng Minh Tiến, Lê Công Diễn, Nguyễn Đồng Đoan Thục, Võ Duy Tân

Khoa Công nghệ Thông tin, Đại học Nông Lâm TP.HCM

Tóm tắt. Một trong những ưu điểm của chương trình đào tạo tín chỉ là sinh viên có thể chủ động hơn với kế hoạch học tập của mình để có thể tốt nghiệp trước thời hạn hoặc đúng hạn. Do đó, việc dự đoán được khả năng tốt nghiệp của sinh viên không chỉ mang tính nghiên cứu khoa học mà còn mang tính thực tiễn rất cao. Trong bài báo này, chúng tôi trình bày một số phương pháp phân lớp áp dụng vào bài toán dự báo khả năng tốt nghiệp của sinh viên như Cây quyết định, Naïve Bayes, KNN. Các thuật toán này sử dụng đầu vào là điểm số các môn đã được học tính đến thời điểm hiện tại để kiểm tra xem liệu sinh viên có khả năng tốt nghiệp đúng hạn hay không. Kết quả thực nghiệm với dữ liệu 454 sinh viên từ các khoá 2014, 2015, 2016 Khoa Công nghệ Thông tin cho thấy thuật toán Naïve Bayes cho ra kết quả có độ chính xác cao hơn thuật toán cây quyết định và KNN.

Abstract. One of the advantages of the credit system is that students can be more proactive with their study plans to graduate before or on time. Therefore, predicting the graduation ability of students is not only scientific but also very practical. In this paper, we present a number of classification methods applied to the problem of predicting students' graduation such as Decision Tree, Naïve Bayes, KNN. These algorithms use as input the scores of the subjects studied so far to test whether a student is likely to graduate on time or not. Experimental results with data of 454 students from 2014, 2015, 2016 Faculty of Information Technology showed that Naïve Bayes algorithm gave higher accuracy than decision tree algorithm and KNN.

Keywords: Classification, Student Graduation Prediction, Decision Tree, Naive Bayes, KNN

1. GIỚI THIỆU

Ngày nay, hầu hết các trường đại học đều chuyển sang đào tạo tín chỉ. Một trong những ưu điểm của chương trình đào tạo tín chỉ là sinh viên có thể chủ động hơn với kế hoạch

học tập của mình để có thể tốt nghiệp trước thời hạn hoặc đúng hạn. Tuy nhiên, việc nghiên cứu các môn học có ảnh hưởng đến tương lai sau này của sinh viên cũng là một thách thức lớn. Do đó phần lớn sinh viên không biết cách chủ động sắp xếp kế hoạch học tập dẫn đến trường hợp không thể ra trường đúng hạn. Vì thế, việc dự đoán được khả năng tốt nghiệp của sinh viên không chỉ mang tính nghiên cứu khoa học mà còn mang tin thực tiễn rất cao. Một mặt giúp sinh viên biết được khả năng học tập của mình để có thể phấn đấu hơn, mặt khác giúp cho các đơn vị đào tạo cũng như các cố vấn học tập nắm đươc tình hình của sinh viên để có thể đưa ra các giải pháp hỗ trợ sinh viên trong việc cải thiện kết quả học tập. Cho đến nay, đã có rất nhiều nghiên cứu liên quan đến việc áp dụng các kỹ thuật khai khá dữ liệu và bài toán dự đoán khả năng tốt nghiệp của sinh viên từ nước ngoài như (Tekin 2014; Lesinski et al. 2016; Alyahyan et al. 2020); trong nước như (Đỗ et al. 2014; Nguyễn 2016; Nguyễn et al. 2019).

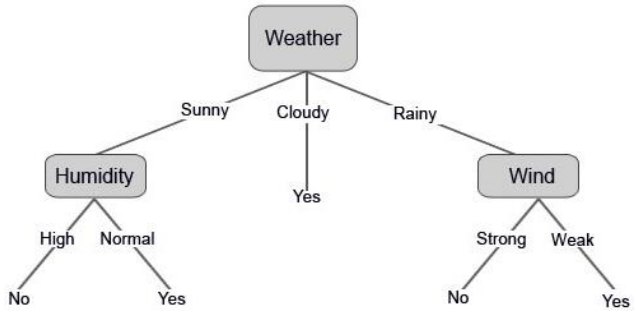
Phân lớp là một trong những lĩnh vực nghiên cứu chính của khai phá dữ liệu. Công nghệ này đã và đang phát triển mạnh mẽ, được áp dụng rộng rãi và các lĩnh vực thương mại, y tế (Jothi et al. 2015), giáo dục (Mohamad et al. 2013) ... Trong các thuật toán tiêu biểu của mô hình phân lớp, Naive Bayes được coi là một phương pháp phân lớp hiệu quả và có độ chính xác cao nhất và thuật toán này chính là nhân tố trung tâm và quan trọng nhất để giải quyết bài toán. Kết quả này được rút ra từ việc nghiên cứu các thuật toán phân lớp: Cây quyết định, Naive Bayes, KNN, từ đó tập trung phân tích, đánh giá và so sánh các thuật toán này với nhau (Ashraf et al. 2018). Với các chiến lược lựa chọn thuộc tính phát triển, cách thức phân bố dữ liệu, cách thức cắt tỉa dữ liệu và một số đặc trưng khác nhau dẫn đến kết quả cho ra có sự khác biệt rõ rệt và thuật toán Naive Bayes được lựa chọn do có tỉ lệ dự đoán cũng như hiểu quả cao nhất đồng thời thời gian dự đoán cũng không quá lâu. Trong công trình (Nguyễn et al. 2019) nhóm tác giả đã áp dụng 2 thuật toán khai phá dữ liệu là Logistic Regression và Naïve Bayes để đưa ra dự báo tình trạng học tập của sinh viên các khoá tiếp theo. Kết quả thực nghiệm cho thấy Naïve Bayes cho ra kết quả dự báo tốt hơn so với Logistic Regression. Các thuật toán này sử dụng đầu vào là điểm số các môn đã được học tính đến thời điểm hiện tại để kiểm tra xem liệu sinh viên có khả năng tốt nghiệp đúng hạn hay không. Mô hình phân lớp được xây dựng với 454 sinh viên từ khoá 2014, 2015, 2016 làm dữ liệu đầu vào đùng để dự đoán cho các sinh viên khoá sau.

2. PHƯƠNG PHÁP

Trong phần này, chúng tôi trình bày một số kỹ thuật sử dụng trong bài toán dự đoán khả năng tốt nghiệp của sinh viên.

2.1 Thuật toán cây quyết định

Cây quyết định là một kiểu mô hình dự báo (*predictive model*) thuộc loại Supervised Learning – học có giám sát (Tan et al. 2006). Thuật toán được xây dựng giống hình dạng một cái cây có ngọn cây, thân cây và các lá được nối kết bằng các cành cây, mỗi thành phần đều có ý nghĩa của riêng nó như là một yếu tố tác động lên quyết định cuối cùng. Mục đích của cây quyêt định là để thể hiện kết cấu, kiến trúc của một hệ thống ra quyết định, hay nói cách khác là cách con người tư duy, logic ra sao để đi đến quyết định cuối cùng. Tóm lại, với đầu vào là dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các dữ liệu chưa biết. Ví dụ:



Hình 1. Cây quyết định cho dữ liệu thời tiết

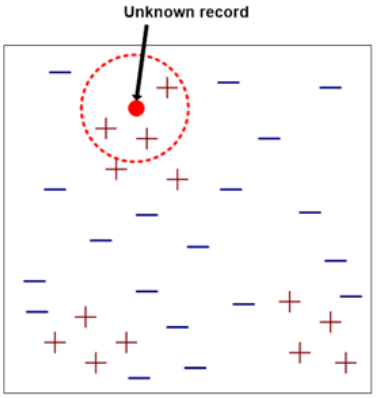
Ứng dụng: Cây quyết định có thể được dùng để áp dụng vào các bài toán phân loại như phân loại báo, phân loại các giao dịch trong ngân hàng, trong y học có thể dùng cây quyết định để phân loại và dự đoán các tế bào ung thư là lành tính hay ác tính,...

2.2 Thuật toán KNN

KNN tên viết tắt của K – Nearest Neighbor hay còn gọi là K láng giềng gần nhất, là 1 phương pháp học có giám sát (Supervised Learning), tức là dựa trên biến mục tiêu đã xác định trước đó, thuật toán sẽ xem xét dữ liệu đã được phân loại trước đó để học và tìm ra những record có thể ảnh hưởng đến biến mục tiêu (Tan et al. 2006).

KNN dựa trên giả định là những thứ tương tự hay có tính chất gần giống nhau sẽ nằm ở vị trí gần nhau, với giả định như vậy, KNN được xây dựng trên công thức toán học để phục vụ có việc tính khoảng cách giữ 2 record với nhau để đưa ra sự tương quan giữa chúng.

KNN còn được gọi là phương pháp học "Lazy" vì tính đơn giản của nó, nghĩa là không có quá trình training. Điều này nghĩa là KNN không có giai đoạn xây dựng model nhưng giai đoạn Testing sẽ chậm hơn do chúng ta cần tính toán nhiều lần dựa vào giá trị K.



Hình 2. Ví dụ thuật toán KNN

Trong trường hợp xấu nhất, tức dữ liệu phân bố không đồng đều hay nói cách khác là dữ liệu bị nhiễu, thời gian quét dữ liệu của KNN sẽ tốn nhiều hơn và kết quả sẽ có độ chính xác thấp. Ngoài ra KNN không cần dựa trên các tham số khác nhau để tiến hành phân loại dữ liệu cũng như không đưa ra bất kỳ kết luận cụ thể nào giữa biến đầu vào và biến mục tiêu, mà chỉ dựa trên khoảng cách giữa cách dữ liệu cần phân loại và dữ liệu đã phân loại trước đó. Đây là ưu điểm lớn nhất của KNN khi so với các thuật toán khác vì hầu hết các dữ liệu thực tế đều không thực sự tuân theo bất kỳ một giả định cụ thể nào cả. Ứng dụng: KNN có thể được sử dụng hiệu quả trong việc phát hiện các ngoại lệ như phát hiện gia lận thẻ tín dụng, phân loại các loài hoa dựa trên màu sắc hoặc mùi hương, ...

2.3 Thuật toán Naive Bayes

Thuật toán Naive Bayes là một thuật toán thuộc loại học có giám sát (Supervised Learning) dựa trên định lý Bayes về lý thuyết xác suất để đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các dữ liệu được quan sát và thống kê (Tan et al. 2006). Được ứng dụng nhiều trong các lĩnh vực Machine Learning dùng để đưa ra các dự đoán có chính xác cao và nhanh chóng dựa trên các tính toán số học bằng định lý Bayes đối với các dữ liệu đã được phân lớp trước đó. Thuật toán hoạt động hoàn toàn dựa trên giả định độc lập, tức là không cần thiết phải xây dựng model dự đoán, điều này cho phép tiết kiệm được thời gian hơn so với các thuật toán khác. Tuy vậy giải định độc lập cũng là nhược điểm của thuật toán này do hầu hết các trường hợp thực tế trong đó các thuộc tính trong các đối tượng thường phụ thuộc lẫn nhau. Bên cạnh đó thuật toán này cho phép kết hợp tri thức tiền nhiệm (prior knowledge) và dữ liệu đã quan sát được (observed data), điều này cho phép cải thiện kết quả dự đoán khi có sự chênh lệch số lượng giữa các phân loại (hay dữ liệu bị nghiêng về một phía). Ứng dụng: Dự đoán theo thời gian thực (Real time Prediction): Naive Bayes chạy khá nhanh nên nó thích hợp ứng dụng nhiều vào các ứng dụng chạy thời gian thực như hệ thống cảnh báo phát hiện sự cố. Phân loại văn bản/

Spam mail: thuật toán này cũng rất thích hợp cho các hệ thống phân loại văn bản hay ngôn ngữ tự nhiên vì tính chính xác của nó lớn hơn sao với các thuật toán khác. Ngoài ra hệ thống chống thư rác cũng rất ưa chuộng thuật toán này.

2.4. Đánh giá mô hình phân lớp:

Mô hình phân lớp được đánh giá dựa vào các tham số trong confusion matrix (Han et al. 2006) dưới đây:

Bảng 1. Confusion matrix

Actual class\Predicted class	C_1	¬ C_1
C_1	**True Positives (TP)**	**False Negatives (FN)**
¬ C_1	**False Positives (FP)**	**True Negatives (TN)**

Trong đó:

Actual class là lớp của dữ liệu từ tập test và Predicted class là lớp mà mô hình dự đoán cho dữ liệu từ tập test.

- **Accuracy**: Được dùng để đo độ chính xác của mô hình phân lớp.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Precision**: Precision được định nghĩa là tỉ lệ số điểm Positive mô hình dự đoán đúng trên tổng số điểm mô hình dự đoán là Positive. Hay nói cách khác, đây là giá trị thể hiện khả năng phát hiện tất cả các Positive. Precision càng cao, tức là số điểm mô hình dự đoán là positive đều là positive càng nhiều. Precision = 1, tức là tất cả số điểm mô hình dự doán là Positive đều đúng, hay không có điểm nào có nhãn là Negative mà mô hình dự đoán nhầm là Positive.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**: Recall được định nghĩa là tỉ lệ số điểm Positive mô hình dự đoán đúng trên tổng số điểm thật sự là Positive (hay tổng số điểm được gán nhãn là Positive ban đầu). Nói cách khác giá trị này thể hiện sự chuẩn xác của việc phát hiện các điểm

Positive. Recall càng cao, tức là số điểm là positive bị bỏ sót càng ít. Recall = 1, tức là tất cả số điểm có nhãn là Positive đều được mô hình nhận ra.

$$Recall = \frac{TP}{TP + FN}$$

- **F – Measure**: F – Measure là một giá trị trung bình điều hòa (harmonic mean) của các tiêu chí Precision và Recall F có giá trị lớn nếu cả 2 giá trị Precision và Recall đều lớn Hàm F có giá trị càng cao chứng minh mô hình phân lớp càng tốt. Khi lý tưởng nhất thì F1 = 1 (khi đó Recall = Precision=1).

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Trong phần tiếp theo của bài báo, chúng tôi sử dụng tất cả những phép đo này để đánh giá hiệu quả của các mô hình phân lớp.

3. KẾT QUẢ VÀ THẢO LUẬN

Trong phần này, chúng tôi sẽ trình bày các bước tiền xử lý dữ liệu điểm của sinh viên thu thập được và các kết quả thực nghiệm với dữ liệu điểm của sinh viên Khoa Công nghệ Thông tin.

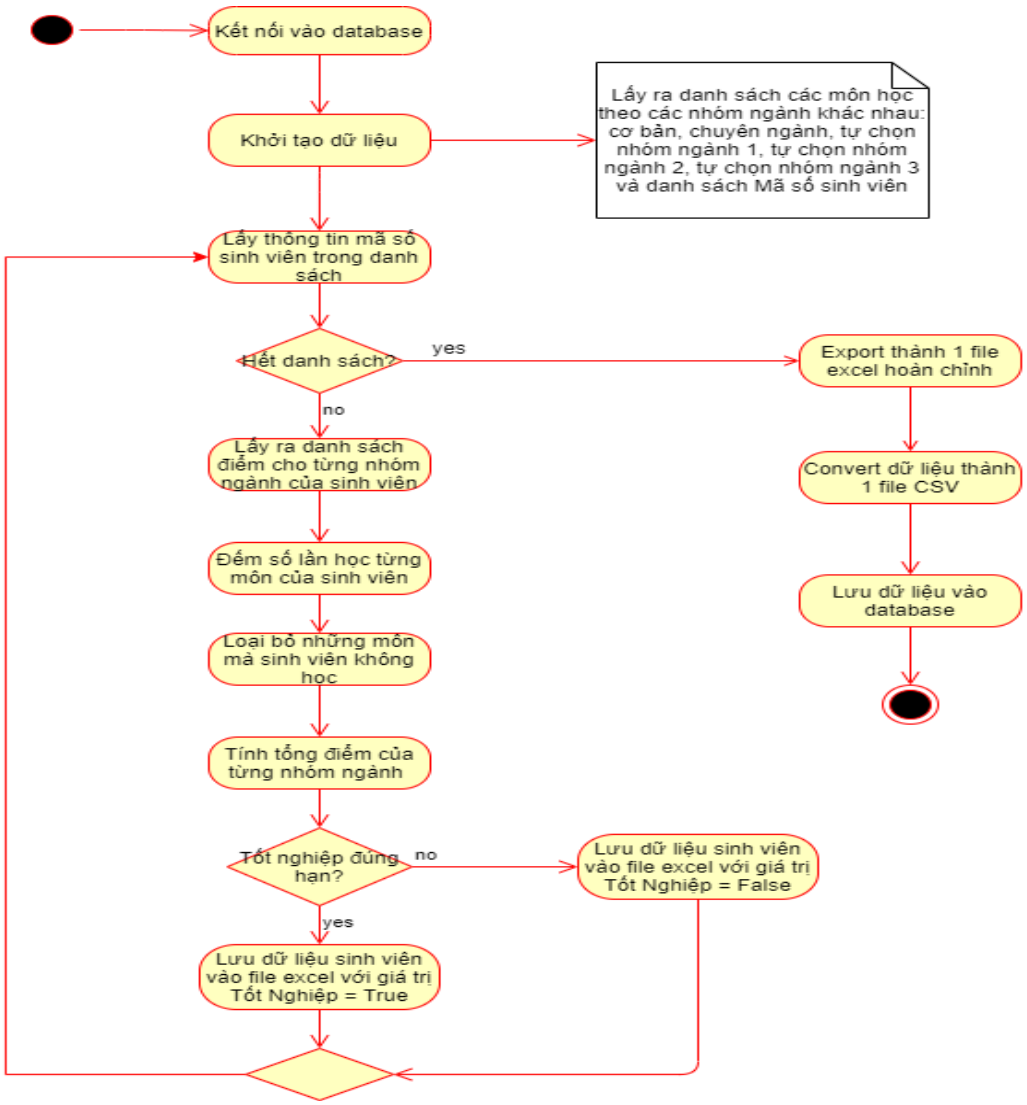
3.1 Dữ liệu

3.1.1 Tiền xử lý dữ liệu

Trong bài báo này, chúng tôi sử dụng dữ liệu về điểm của sinh viên khoá DH14DT, DH15DT và DH16DT của khoa Công nghệ Thông tin trường Đại học Nông Lâm Tp.HCM, cụ thể có 669 sinh viên với 18 thuộc tính bao gồm các thông tin điểm số của các môn học, số lần học lại. Đây là các khoá có cùng chương trình đào tạo. Trong đó:

- 239 sinh viên khoá DH14DT
- 230 sinh viên khoá DH15DT
- 200 sinh viên khoá DH16DT

Để có thể áp dụng các thuật toán phân lớp với dữ liệu trên, chúng tôi đã tiến hành tiền xử lý dữ liệu điểm của sinh viên theo mô tả ở hình sau:



Hình 3. Các bước tiền xử lý dữ liệu

Từ Hình 3, có thể thấy quá trình tiền xử lý dữ liệu sẽ thêm một số thuộc tính vào tập dữ liệu về điểm của sinh viên. Cụ thể, ứng với mỗi môn học, chúng tôi thêm thuộc tính biểu diễn số lần học môn học đó.

3.1.2 Dữ liệu sau khi tiền xử lý

Số lượng sinh viên còn lại : 454 sinh viên với 126 thuộc tính, trong đó:

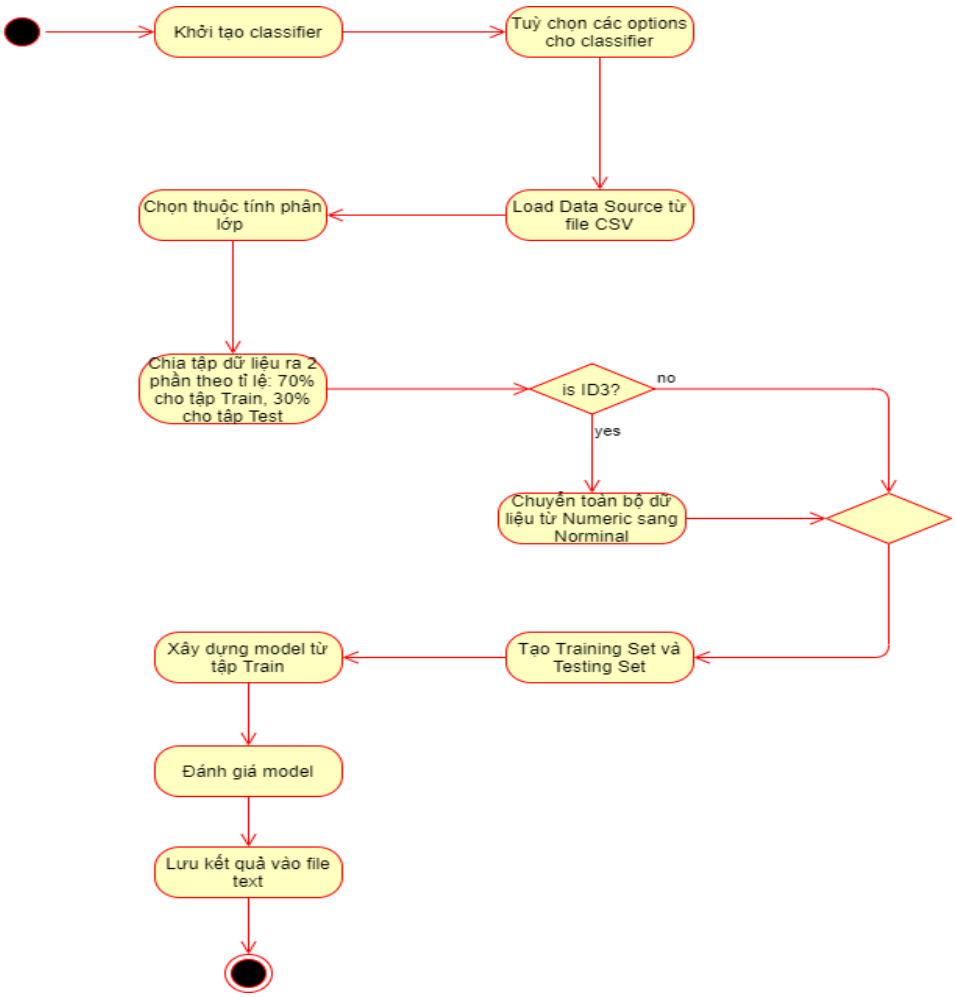
- 134 sinh viên khoá DH14DT
- 146 sinh viên khoá DH15DT
- 174 sinh viên khoá DH16DT

Các sinh viên còn lại đã bị loại bỏ khỏi dữ liệu vì đã nghỉ học hoặc bị buộc thôi học.

Ở đây, thuộc tính phân lớp là "Tốt nghiệp". Từ dữ liệu điểm thu thập được cho thấy, nếu chia nhỏ giá trị của thuộc tính "Tốt nghiệp" thành các lớp như: Tốt nghiệp sớm, Tốt nghiệp đúng hạn, Tốt nghiệp muộn … thì không có sự cân đối giữa các lớp. Nghĩa là số lượng sinh viên tốt nghiệp sớm rất ít so với số lượng sinh viên tốt nghiệp đúng hạn hoặc tốt nghiệp muộn. Điều này ảnh hưởng không nhỏ đến chất lượng của mô hình phân lớp. Cho nên chúng tôi sử dụng 2 giá trị của thuộc tính phân lớp là {True; False}. Nghĩa là sinh viên tốt nghiệp đúng hạn hoặc trễ hạn. Tương tự, việc dự đoán sinh viên tốt nghiệp loại xuất sắc, giỏi, khá, trung bình cũng gặp khó khăn khi số lượng sinh viên tốt nghiệp xuất sắc, giỏi khá ít so với số lượng sinh viên tốt nghiệp loại khá và trung bình.

3.2 Kết quả

Quy trình đánh giá mô hình phân lớp:



Hình 4. Quy trình đánh giá mô hình phân lớp

Kết quả so sánh với tổng dữ liệu 454 sinh viên trong đó tập Test có 136 sinh viên và tập Train có 318 sinh viên:

Bảng 2. Kết quả thực nghiệm

Thuật toán	Số sinh viên dự đoán đúng	Tỉ lệ dự đoán đúng	Precision	Recall	F – Measure
ID3	82	62.5%	0.735	0.72	0.721
KNN	101	74.7264%	0.745	0.743	0.741
J48	109	80.1471%	0.804	0.801	0.8
Naïve Bayes	120	88.2353%	0.885	0.882	0.882

Theo Bảng 2, chúng ta có thể thấy rằng ID3 là thuật toán cho ra kết quả kém nhất so với các thuật toán còn lại. Điều này bởi vì bản chất thuật toán ID3 khi xây dựng cây quyết định không có áp dụng bất kỳ kỹ thuật cắt tỉa cây như thuật toán J48. Trong khi đó, thuật toán KNN cho ra kết quả dự đoán tốt hơn ID3. Tuy nhiên, KNN phụ thuộc quá lớn vào việc chọn tham số K. Thuật toán Naïve Bayes cho kết quả tốt hơn các thuật toán còn lại theo các độ đo đã đề cập ở trên. Do đó, trong sản phẩm ứng dụng của đề tài chúng tôi đã sử dụng phương pháp này để dự báo khả năng tốt nghiệp của sinh viên.

4. KẾT LUẬN VÀ ĐỀ NGHỊ

Bài báo đã trình bày một số kỹ thuật phân lớp ứng dụng vào việc dự đoán khả năng tốt nghiệp của sinh viên. Theo đó, các thuật toán ID3, KNN, J48 và Naïve Bayes dã được áp dụng. Kết quả thực nghiệm với dữ liệu điểm của sinh viên Khoa Công nghệ Thông tin cho thấy tính khả thi của việc áp dụng mô hình phân lớp vào bài toán dự đoán khả năng tốt nghiệp của sinh viên. Ngoài ra, kết quả đánh giá mô hình phân lớp với dữ liệu thực nghiệm cho thấy thuật toán Naïve Bayes cho ra kết quả tốt hơn các thuật toán khác.

Tuy nhiên, để có thể dự đoán được chi tiết hơn sinh viên tốt nghiệp sớm, tốt nghiệp đúng hạn, tốt nghiệp muộn hoặc không tốt nghiệp thì cần phải sử dụng những kỹ thuật xử lý dữ liệu nâng cao như xử lý trường hợp các lớp không cân bằng nhau. Đây được xem là hướng phát triển trong tương lai. Bên cạnh đó, để có thể đưa vào sử dụng kết quả của đề

tài cần phải có những thử nghiệm rộng hơn với dữ liệu điểm của sinh viên ở các khoa khác. Từ đó sẽ làm cơ sở để đánh giá các mô hình phân lớp một cách khách quan hơn.

TÀI LIỆU THAM KHẢO

Nguyễn Thị Uyên, Nguyễn Minh Tâm, 2019. Dự đoán kết quả học tập của sinh viên bằng kỹ thuật khai phá dữ liệu. Tạp chí khoa học, Tập 48 - Số 3A/2019, tr. 68-73

Đỗ Thanh Nghị, Phạm Nguyên Khang, Nguyễn Minh Trung và Trịnh Trung Hưng, 2014. Phát hiện môn học quan trọng ảnh hưởng đến kết quả học tập sinh viên ngành công nghệ thông tin, Tạp chí Khoa học Trường Đại học Cần Thơ Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường: 33 (2014): 49-57

Nguyễn Thái Nghe, 2016. Ứng dụng các kỹ thuật trong khai phá dữ liệu hỗ trợ sinh viên lập kế hoạch học tập. Sách: Công nghệ thông tin trong hỗ trợ ra quyết định về Giáo dục, Nông nghiệp, Thuỷ sản và Môi trường vùng Đồng bằng sông Cửu Long, Chương : 2, Nhà xuất bản Đại học Cần Thơ

Aysha, Ashraf and Khan, Muhammad. (2018). A Comparative Study of Predicting Student's Performance by use of Data Mining Techniques. 10.13140/RG.2.2.28495.12960.

Gene Lesinski, Steven Corns, Cihan Dagli, 2016. Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy, Procedia Computer Science, Volume 95, 375-382

Jiawei Han and Micheline Kamber, 2006. Data Mining: Concepts and Techniques, Second Edition, Morgan Kaufmann.

Neesha Jothi, Nur'Aini Abdul Rashid, Wahidah Husain, 2015. Data Mining in Healthcare – A Review, Procedia Computer Science, Volume 72, 306-313

Ahmet Tekin, 2014. Early Prediction of Students' Grade Point Averages at Graduation: A Data Mining Approach. Eurasian Journal of Educational Research, Issue 54, 207-226

Siti Khadijah Mohamad, Zaidatun Tasir, 2013. Educational Data Mining: A Review, Procedia - Social and Behavioral Sciences, Volume 97, 320-324

Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, 2006. Introduction to data Mining, Addison-Wesley

Eyman Alyahyan, and Dilek Dustegor, 2020. Predicting academic success in higher education: literature review and best practices. Int J Educ Technol High Educ 17, 3 (2020).