

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NAM CẦN THƠ**



**BÁO CÁO TỔNG KẾT
ĐỀ TÀI NGHIÊN CỨU KHOA HỌC CẤP CƠ SỞ**

**SỬ DỤNG PHƯƠNG PHÁP HỌC MÁY ĐỂ DỰ ĐOÁN TÌNH
TRẠNG BỎ HỌC CỦA SINH VIÊN KHOA CÔNG NGHỆ
THÔNG TIN - TRƯỜNG ĐẠI HỌC NAM CẦN THƠ**

Mã số: C23-18

Chủ nhiệm đề tài: ThS. Trần Thanh Nam

Thành viên

TS. Ngô Hồ Anh Khôi

ThS. Trần Văn Thiện

ThS. Phan Thị Xuân Trang

Cần Thơ – Tháng 10/2024

DANH SÁCH THÀNH VIÊN THAM GIA NGHIÊN CỨU

Đề tài nghiên cứu khoa học cấp cơ sở *Sử dụng phương pháp học máy để dự đoán tình trạng bỏ học của sinh viên Khoa Công nghệ thông tin - Trường Đại học Nam Cần Thơ* được thực hiện từ tháng 10/2023 đến tháng 10/2024 tại khoa Công nghệ Thông tin, Trường Đại học Nam Cần Thơ (Đ/C: 168 Nguyễn Văn Cừ nối dài, P. An Bình, Q. Ninh Kiều, TP. Cần Thơ) với sự tham gia của các thành viên sau:

- | | |
|-------------------------------|------------------|
| - Thạc sĩ Trần Thanh Nam | Chủ nhiệm đề tài |
| - Tiến sĩ Ngô Hồ Anh Khôi | Thành viên |
| - Thạc sĩ Trần Văn Thiện | Thành viên |
| - Thạc sĩ Phan Thị Xuân Trang | Thành viên |

MỤC LỤC

MỤC LỤC.....	i
DANH MỤC BẢNG.....	iv
DANH MỤC HÌNH ẢNH	v
DANH MỤC TỪ VIẾT TẮT.....	vi
TÓM TẮT	vii
ABSTRACT	viii
PHẦN 1. MỞ ĐẦU	1
1.1 ĐẶT VẤN ĐỀ	1
1.2 CÁC NGHIÊN CỨU LIÊN QUAN	2
1.3 MỤC TIÊU ĐỀ TÀI	2
1.4 NỘI DUNG NGHIÊN CỨU.....	2
1.5 PHẠM VI NGHIÊN CỨU.....	3
1.6 BỐ CỤC ĐỀ TÀI.....	3
PHẦN 2. CƠ SỞ LÝ THUYẾT	4
2.1 MÁY HỌC (MACHINE LEARNING).....	4
2.1.1 Khái niệm máy học	4
2.1.2 Học có giám sát (Supervised Learning).....	4
2.1.3 Học không giám sát (Unsupervised Learning)	5
2.1.4 Học tăng cường (Reinforcement Learning)	6
2.1.5 Các ứng dụng thực tiễn của máy học	6
2.2 XỬ LÝ DỮ LIỆU THIẾU (MISSING DATA) TRONG HỌC MÁY	7
2.2.1 Khái niệm dữ liệu thiếu.....	7
2.2.2 Phân loại các kiểu thiếu dữ liệu	8
2.2.3 Ảnh hưởng của dữ liệu thiếu đến mô hình học máy	9
2.2.4 Các phương pháp xử lý dữ liệu thiếu.....	9
2.3 HỌC LIÊN TỤC (CONTINUOUS LEARNING).....	11
2.3.1 Khái niệm học liên tục	11
2.3.2 Những khó khăn của học liên tục.....	11
2.3.3 Ứng dụng của học liên tục	12

2.3.4	Các phương pháp học liên tục.....	12
2.3.5	Các hướng nghiên cứu và phát triển	12
2.4	CỬA SỔ TRƯỢT (SLIDING WINDOW).....	13
2.4.1	Khái niệm cửa sổ trượt.....	13
2.4.2	Phương pháp cửa sổ trượt	13
2.5	THUẬT TOÁN EKI.....	15
2.5.1	Các khái niệm liên quan.....	15
2.5.2	Cấu trúc hoạt động của EKI.....	15
2.5.3	Ứng dụng của EKI trong các hệ thống thực tế.....	15
2.5.4	Ứng dụng thuật toán EKI.....	16
2.5.5	Triển vọng nghiên cứu và phát triển thuật toán EKI	16
2.5.6	Áp dụng thuật toán EKI vào đề tài nghiên cứu này	17
PHẦN 3.	THỰC NGHIỆM ĐỀ TÀI.....	18
3.1	MÔI TRƯỜNG THỰC NGHIỆM	18
3.2	PHẦN MỀM THỰC NGHIỆM.....	18
3.2.1	Ngôn ngữ lập trình Python	18
3.2.2	Thư viện TensorFlow	18
3.2.3	Thư viện Sklearn.....	19
3.2.4	Thư viện Keras	19
3.2.5	Visual Studio Code	20
3.2.6	Sublime Text.....	20
3.2.7	SQLite.....	21
3.2.8	Streamlit.....	21
3.3	CÁC THƯ VIỆN PYTHON SỬ DỤNG	21
3.4	MÔ TẢ HỆ THỐNG	27
3.5	MÔ TẢ BỘ DỮ LIỆU CỦA ĐỀ TÀI	27
3.6	TIỀN XỬ LÝ DỮ LIỆU.....	28
3.6.1	Trích chọn đặc trưng.....	28
3.6.2	Chuẩn hóa dữ liệu	29
3.6.3	Gán nhãn phân lớp.....	32

3.7 HUẤN LUYỆN MÔ HÌNH.....	32
3.7.1 Kỹ thuật Bagging (Bootstrap Aggregating).....	32
3.7.2 Thuật toán AdaBoost.....	33
4.1.3 Thuật toán Random Forest	34
4.1.4 Một số thuật toán phân lớp khác được sử dụng trong đề tài	35
PHẦN 4. KẾT QUẢ NGHIÊN CỨU	39
4.1 ĐÁNH GIÁ HIỆU SUẤT MÔ HÌNH	39
4.1.1 Phương pháp đánh giá mô hình phân lớp	39
4.1.2 Các trường hợp kết quả dự đoán trong bài toán phân loại	40
4.1.3 Ma trận nhầm lẫn (Confusion matrix)	41
4.1.4 Kết quả đánh giá của các mô hình phân lớp.....	45
4.2 DỰ ĐOÁN TẬP DỮ LIỆU MỚI	47
PHẦN 5. KẾT LUẬN VÀ KIẾN NGHỊ.....	50
5.1 KẾT LUẬN	50
5.2 KIẾN NGHỊ VÀ HƯỚNG PHÁT TRIỂN CỦA ĐỀ TÀI	50
TÀI LIỆU THAM KHẢO.....	51

DANH MỤC BẢNG

Bảng 3.1 Các thư viện Python được sử dụng trong đề tài nghiên cứu.....	22
Bảng 3.2 Phương pháp xử lý dữ liệu thiếu.....	30
Bảng 3.3 Các thuật toán phân lớp được sử dụng trong đề tài	35
Bảng 4.1 Kết quả đánh giá trung bình giữa các thuật toán phân lớp	45

DANH MỤC HÌNH ẢNH

Hình 2.1 Minh họa quá trình học máy	4
Hình 2.2 Ví dụ minh họa học máy có giám sát.....	5
Hình 2.3 Ví dụ minh họa học máy không giám sát.....	6
Hình 2.4 Ví dụ minh họa giá trị bị thiếu trong dữ liệu	8
Hình 2.5 Mô hình hoạt động của cửa sổ trượt (Sliding Window)	14
Hình 3.1 Tổng quan hệ thống dự đoán tình trạng bỏ học của sinh viên	27
Hình 3.2 Tập dữ liệu thô của đề tài	28
Hình 3.3 Tập dữ liệu đã trích xuất đặc trưng và thuộc tính	29
Hình 3.4 Chọn phương pháp xử lý dữ liệu thiếu	31
Hình 3.5 Lưu tập dữ liệu đã xử lý giá trị thiếu.....	31
Hình 3.6 Mô tả tập dữ liệu đã xử lý giá trị thiếu.....	32
Hình 3.7 Mô tả tập dữ liệu đã chuẩn hóa và gán nhãn phân loại.....	32
Hình 3.8 Sơ đồ minh họa thuật toán Bagging.....	33
Hình 3.9 Sơ đồ minh họa thuật toán AdaBoost	33
Hình 3.10 Mô tả phương pháp tiếp cận mô hình của Bagging và Boosting	34
Hình 3.11 Sơ đồ minh họa thuật toán Random Forest.....	34
Hình 3.12 Chọn thuật toán đào tạo mô hình	37
Hình 4.1 Minh họa hiện tượng Underfitting	39
Hình 4.2 Minh họa hiện tượng Overfitting	40
Hình 4.3 Minh họa hiện tượng Fit.....	40
Hình 4.4 Ma trận nhầm lẫn (Confusion matrix).....	42
Hình 4.5 Sơ đồ so sánh tỉ lệ đánh giá trung bình giữa các thuật toán phân lớp.....	46
Hình 4.6 Tập dữ liệu mới cần dự đoán kết quả phân lớp.....	48
Hình 4.7 Upload dữ liệu mới cần dự đoán lên hệ thống	48
Hình 4.8 Kết quả dự đoán dữ liệu mới.....	49
Hình 4.9 Tập dữ liệu mới đã được gán nhãn phân loại và tỷ lệ dự đoán.....	49

DANH MỤC TỪ VIẾT TẮT

Viết tắt	Tên tiếng Anh	Giải thích
AI	Artificial Intelligence	Trí tuệ nhân tạo
ANN	Artificial Neural Network	Mạng nơ-ron nhân tạo
BA	Balanced Accuracy	Độ chính xác cân bằng
DL	Deep Learning	Học sâu
EKI	Evolving with Klinkenberg's Idea	Phát triển theo ý tưởng của Klinkenberg
KNN	K-Nearest Neighbors	K- láng giềng gần nhất
LDA	Linear Discriminant Analysis	Phân tích phân biệt tuyến tính
MA	Missing data	Dữ liệu thiếu
MI	Multiple Imputation	Bổ sung đa giá trị
ML	Machine Learning	Học máy hoặc máy học
MLP	Multi-layers Perceptron	Mạng nơ-ron đa lớp
PCA	Principal Component Analysis	Phân tích thành phần chính
SVC	Support Vector Classifier	Bộ phân loại Vector hỗ trợ
SVM	Support Vector Machine	Máy Vector hỗ trợ

TÓM TẮT

Tình trạng bỏ học của sinh viên tại các trường Đại học nói chung đang là một vấn đề đáng lo ngại, việc bỏ học với nhiều nguyên nhân khác nhau trong đó nguyên nhân phổ biến bao gồm điểm số học phần thấp, rớt môn hoặc học lại, gây chán nản trong việc học dẫn đến bỏ học. Để giải quyết vấn đề này, đề tài đã nghiên cứu sử dụng phương pháp học máy tiên tiến nhằm dự đoán khả năng bỏ học của sinh viên, tạo điều kiện cho việc can thiệp xử lý sớm nhằm nâng cao chất lượng giáo dục. Bộ dữ liệu về sinh viên khoa Công nghệ thông tin của Trường Đại học Nam Cần Thơ qua các năm được thu thập từ Phòng đào tạo thuộc Trường Đại học Nam Cần Thơ, bao gồm thông tin về điểm số, các học phần đã hoàn thành và tình trạng học tập của sinh viên. Các phương pháp tiền xử lý dữ liệu đã được áp dụng để xử lý các giá trị thiếu (missing data) và làm sạch dữ liệu (cleaning data). Đề tài nghiên cứu phương pháp học liên tục (continuous learning), sử dụng thuật toán EKI (Evolving with Klinkenberg's Idea), một phương pháp học máy tiên tiến thuộc nhóm học liên tục để dự đoán tình trạng bỏ học của sinh viên. Kết quả của đề tài cho thấy mô hình dự đoán có độ chính xác cao, cung cấp những dự đoán khả quan về nguy cơ bỏ học. Phương pháp EKI giúp cải thiện đáng kể khả năng dự đoán so với các phương pháp truyền thống. Nghiên cứu này đề xuất một hướng tiếp cận mới không chỉ hỗ trợ việc dự đoán chính xác hơn mà còn mở ra những giải pháp kịp thời cho trường Đại học Nam Cần Thơ nói riêng và các trường Đại học nói chung trong việc hỗ trợ sinh viên, từ đó đảm bảo và nâng cao chất lượng giáo dục đào tạo.

ABSTRACT

Student dropout at universities in general is a growing concern. Various factors contribute to this issue, with common reasons including low grades, course failures, and the need to retake courses, leading to discouragement and eventual dropout. To address this, this study employs advanced machine learning techniques to predict student dropout, enabling early intervention and improving educational quality. The dataset used encompasses student information from the Faculty of Information Technology at Nam Can Tho University over several years, collected from the university's Training Department. This data includes grades, completed courses, and student academic status. Data preprocessing techniques were applied to handle missing values and clean the data. This study investigates continuous learning, utilizing the Evolving Klinkenberg's Idea (EKI) algorithm, an advanced machine learning method in the continuous learning group, to predict student dropout. The results demonstrate that the prediction model achieves high accuracy, providing reliable predictions of dropout risk. The EKI method significantly improves prediction capabilities compared to traditional approaches. This research proposes a novel approach that not only supports more accurate prediction but also offers timely solutions for Nam Can Tho University in particular, and universities in general, in supporting students. This, in turn, ensures the quality of education and training.