BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

ĐẶNG VĂN LỰC

PHÂN TÍCH DỮ LIỆU TẠO CẢNH BÁO HỌC TẬP BẰNG MÔ HÌNH HỒI QUY LOGISTIC

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công nghệ thông tin

Mã số ngành: 60480201

TP. HỒ CHÍ MINH, tháng 03 năm 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

ĐẶNG VĂN LỰC

PHÂN TÍCH DỮ LIỆU TẠO CẢNH BÁO HỌC TẬP BẰNG MÔ HÌNH HỒI QUY LOGISTIC

LUẬN VĂN THẠC SĨ

Chuyên ngành: Công nghệ thông tin

Mã số ngành: 60480201

CÁN BỘ HƯỚNG DẪN KHOA HỌC: TS. TRẦN ĐỨC KHÁNH

TP. HỒ CHÍ MINH, tháng 03 năm 2016

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

Cán bộ hướng dẫn khoa học : TS. TRẦN ĐỨC KHÁNH
(Ghi rõ họ, tên, học hàm, học vị và chữ ký)

Luận văn Thạc sĩ được bảo vệ tại Trường Đại học Công nghệ TP. HCM ngày 20 tháng 01 năm 2016

Thành phần Hội đồng đánh giá Luận văn Thạc sĩ gồm:

TT	Họ và tên	Chức danh Hội đồng
1	PGS. TSKH. Nguyễn Xuân Huy	Chủ tịch
2	TS. Vũ Thanh Hiền	Phản biện 1
3	TS. Hồ Đức Nghĩa	Phản biện 2
4	PGS. TS. Quản Thành Thơ	Ủy viên
5	TS. Cao Tùng Anh	Ủy viên, Thư ký

Xác nhận của Chủ tịch Hội đồng đánh giá Luận văn sau khi Luận văn đã được sửa chữa (nếu có).

Chủ tịch Hội đồng đánh giá LV

TRƯỜNG ĐH CÔNG NGHỆ TP. HCM
PHÒNG QLKH – ĐTSĐH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

TP. HCM, ngày..... tháng….. năm 2016

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên: ĐẶNG VĂN LỰC — Giới tính: Nam
Ngày, tháng, năm sinh: 14/12/1987 — Nơi sinh: Quảng Ngãi
Chuyên ngành: Công nghệ thông tin — MSHV: 1441860016

I- Tên đề tài:

"PHÂN TÍCH DỮ LIỆU TẠO CẢNH BÁO HỌC TẬP BẰNG MÔ HÌNH HỒI QUY LOGISTIC"

II- Nhiệm vụ và nội dung:

- Tìm hiểu về học máy thống kê, quy trình khai thác dữ liệu, phân tích thống kê
- Nghiên cứu các yếu tố ảnh hưởng đến kết quả học sinh TCCN hệ THCS và xác định mẫu dữ liệu.
- Phân tích dữ liệu mẫu, áp dụng mô hình hồi quy Logistic để xây dựng mô hình tạo cảnh báo học tập.
- Đánh giá mô hình tạo cảnh báo học tập

III- Ngày giao nhiệm vụ: *20/08/2015*

IV- Ngày hoàn thành nhiệm vụ: *15/01/2016*

V- Cán bộ hướng dẫn: *TS. Trần Đức Khánh*

CÁN BỘ HƯỚNG DẪN — **KHOA QUẢN LÝ CHUYÊN NGÀNH**
(Họ tên và chữ ký) — (Họ tên và chữ ký)

TS. TRẦN ĐỨC KHÁNH

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của Thầy TS. Trần Đức Khánh. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong luận văn còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung luận văn của mình. Trường Đại Học Công Nghệ TP.HCM không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện.

Học viên thực hiện luận văn

ĐẶNG VĂN LỰC

LỜI CÁM ƠN

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Trong suốt thời gian từ khi bắt đầu học tập tại trường đến nay, em đã nhận được rất nhiều sự quan tâm, giúp đỡ của quý Thầy Cô, gia đình và bạn bè. Với lòng biết ơn sâu sắc nhất, em xin gửi đến quý Thầy Cô ở Khoa Công Nghệ Thông Tin – Trường Đại Học Công Nghệ TP.HCM đã cùng với tri thức và tâm huyết của mình để truyền đạt vốn kiến thức quý báu cho các em trong suốt thời gian học tập tại trường. Và đặc biệt, trong học kỳ này nếu không có những lời hướng dẫn, dạy bảo của các thầy cô thì em nghĩ bài luận văn này của em rất khó có thể hoàn thiện được. Bài luận văn thực hiện trong khoảng thời gian 6 tháng. Bước đầu của em còn rất hạn chế và còn nhiều bỡ ngỡ. Do vậy, em gặp rất nhiều khó khăn trong giai đoạn đầu làm luận văn. Nhưng với sự dìu dắt hướng dẫn tận tình của thầy TS. TRẦN ĐỨC KHÁNH em đã dần làm quen với việc nghiên cứu và hoàn thiện bài luận văn này.

Em xin gởi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của Trường Đại Học Công Nghệ TP.HCM, đặc biệt là các thầy cô Khoa Công Nghệ Thông Tin của trường đã tạo điều kiện cho em để em có thể hoàn thành tốt bài luận văn này. Và em cũng xin chân thành cám ơn các bạn học cùng khóa đã nhiệt tình đóng góp ý kiến để em hoàn thành tốt bài luận văn của em.

Trong quá trình làm bài luận văn, khó tránh khỏi những sai sót, rất mong quý Thầy, Cô bỏ qua. Đồng thời do trình độ lý luận cũng như kinh nghiệm thực tiễn còn hạn chế nên bài luận văn không thể tránh khỏi những thiếu sót, em rất mong nhận được ý kiến đóng góp của Thầy, Cô để em học thêm được nhiều kinh nghiệm để tiếp tục hoàn thành tốt những nghiên cứu sắp tới.

Em xin chân thành cảm ơn!

ĐẶNG VĂN LỰC

TÓM TẮT

Khoa học thống kê là khoa học về việc thu thập, phân tích, diễn giải và trình bày các số liệu để tìm ra bản chất và tính chất của các hiện tượng kinh tế, tự nhiên và xã hội. Thống kê cho phép tóm tắt và trình bày một cách dễ hiểu các thông tin bằng số, kiểm định một giả thuyết hoặc dự báo về khả năng xảy ra của một biến cố nào đó. Tất cả những vai trò đó được gói trong bài toán hồi quy. Bài toán hồi quy là bài toán thiết lập mối quan hệ giữa một đối tượng đang được quan tâm và các đối tượng liên quan để đưa ra những kết luận có ý nghĩa thống kê. Khi biến đáp ứng là biến nhị phân hay rời rạc thì mô hình hồi quy tuyến tính không thể áp dụng được vì biến đáp ứng không liên tục, một mô hình hồi quy mới được xây dựng để giải quyết vấn đề trên, đó là mô hình hồi quy Logistic. Mô hình hồi quy Logistic được phát triển bởi nhà thống kê học David R. Cox vào những năm 1970 và ngày càng ứng dụng rộng rãi.

Trong phạm vi đề tài này, luận văn ứng dụng phương pháp hồi quy Logistic để dự đoán kết quả học tập của học sinh TCCN hệ THCS. Mục tiêu chính của nghiên cứu là phân tích mối tương quan giữa yếu tố đậu và rớt dựa trên các yếu tố ảnh hưởng đến kết quả học tập như: tuổi học sinh, tình trạng sống với gia đình, sức khỏe học sinh, làm thêm ngoài giờ, sử dụng chất kích thích, tinh thần học tập, chuyên cần, vi phạm nội quy, ngành học, kết quả học tập trước, số đơn vị học trình nợ . Nghiên cứu giới thiệu về mô hình hồi quy Logistic, phương pháp ước lượng tham số và kiểm định kết quả thống kê đối với mô hình hồi quy Logistic từ đó đưa ra mô hình dự đoán tối ưu nhất dựa trên dữ liệu thu thập từ học sinh TCCN hệ THCS tại trường Trung cấp Kỹ thuật và Nghiệp vụ Nam Sài Gòn.

ABSTRACT

Statistics is a science of collecting, analyzing, explaining and presenting data to discover the nature and characteristics of socio-economic issues. Statistics gives brief introduction with numbers and predictive theory of potential happening. This function is wrapped in a regression process. Regression process is a process that estimating the relationships among an issue and others concerned to deduct a statistic conclusion. When dependent variable is binary or discrete linear regression is inapplicable because dependent variable is not continuous. A new regression, logistic regression, is developed to solve the problem. Logistic regression was developed by David R. Cox in 1970s and has been widely applied since then.

In this essay, I apply logistic regression to predict study result of vocational students. The main issue of this research is relationship between study result (pass-fail) and independent variables including: age, family relationship, health, part-time job, stimulant usage, eagerness of learning, attendant, regulation conduction, major, previous semester's result, and number of previous fail modules. This research introduce logistic regression process, method of estimating variables and examining results, therefore, it is able to introduce an optimal prediction modal based on data from vocational students in Nam Sai Gon Vocational and Technical College.

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Từ viết tắt	**Cụm từ nguyên**
HS	Học sinh
TCCN	Trung cấp chuyên nghiệp
THCS	Trung học sơ sở
ĐH	Đại học
CĐ	Cao đẳng
GD&ĐT	Giáo dục và Đào tạo
GVCN	Giáo viên chủ nhiệm
TT	Tình trạng
GD	Gia Đình
NQ	Nội quy
CKT	Chất kích thích
ROC	Receiver Operating Characteristic
CRISP - DM	Cross - Industry Standard Process for Data Mining

DANH MỤC CÁC BẢNG

DANH MỤC CÁC BIỂU ĐỒ, ĐỒ THỊ, SƠ ĐỒ, HÌNH ẢNH

DANH MỤC BIỂU ĐỒ

DANH MỤC ĐỒ THỊ

DANH MỤC HÌNH

CHƯƠNG 1: GIỚI THIỆU NGHIÊN CỨU

Trình bày tổng quan về lý do nghiên cứu đề tài, mục tiêu nghiên cứu, đối tượng nghiên cứu, phạm vi nghiên cứu, phương pháp và ý nghĩa nghiên cứu.

1.1. Lý do chọn đề tài

Trong những năm gần đây số lượng học sinh (HS) trung cấp chuyên nghiệp (TCCN) hệ trung học cơ sở (THCS) trong các trường trung cấp bị cảnh báo học vụ và buộc thôi học ngày càng gia tăng. Do đó việc dự báo kết quả học tập của học sinh TCCN hệ THCS là điều cần thiết để các em lập kế hoạch với phương pháp học tập hiệu quả nhằm nâng cao kết quả học tập.

Trường Trung cấp Kỹ thuật và Nghiệp vụ Nam Sài Gòn, là một trong những trường đào tạo hệ nghề cho học sinh lớn của khu vực phía nam, với quy mô gần 3000 HS trong đó khoảng 1200 HS TCCN hệ THCS. Với thực trạng kết quả hiện nay của học sinh TCCN hệ THCS chỉ mức trung bình, yếu chưa đáp ứng yêu cầu phát triển kinh tế xã hội hiện nay. Vì vậy, để góp phần nâng cao vị thế của trường đào tạo chất lượng cung ứng cho nhà tuyển dụng nguồn nhân lực có chất lượng thì việc nâng cao chất lượng đào tạo mà cụ thể là kết quả học tập của học sinh là yêu cầu cấp bách hiện nay. Chính vì thế việc nghiên cứu các yếu tố ảnh hưởng đến kết quả học tập của học sinh sẽ góp phần nâng cao kết quả học tập của học sinh từ đó nâng cao chất lượng đào tạo của nhà trường.

Những năm gần đây, khi nền khoa học công nghệ thông tin đang ngày càng phát triển như vũ bão thì vấn đề khai phá dữ liệu đã trở thành một trong những hướng nghiên cứu chính trong lĩnh vực khoa học máy tính và công nghệ tri thức. Khai phá dữ liệu đã và đang ứng dụng thành công vào rất nhiều các lĩnh vực khác nhau như: thương mại, tài chính, thị trường chứng khoán, y học, thiên văn học, sinh học, giáo dục và viễn thông v.v…

Với những lý do như vậy tác giả chọn đề tài "**PHÂN TÍCH DỮ LIỆU TẠO CẢNH BÁO HỌC TẬP BẰNG MÔ HÌNH HỒI QUY LOGISTIC**" làm đề tài luận văn tốt nghiệp.

1.2. Mục tiêu nghiên cứu

Mục đích của nghiên cứu của đề tài là dự đoán kết quả học tập của học sinh TCCN hệ THCS trường Trung cấp Kỹ thuật và Nghiệp vụ Nam Sài Gòn từ đó có biện

pháp can thiệp cải thiện kết quả học tập của học sinh TCCN hệ THCS. Đề tài nghiên cứu cần xác định các mục tiêu sau:

- Xác định các yếu tố quyết định và ảnh hưởng đến kết quả học tập học sinh TCCN hệ THCS trường Trung cấp Kỹ thuật và Nghiệp vụ Nam Sài Gòn.
- Áp dụng mô hình hồi quy Logistic dự báo kết quả học sinh TCCN hệ THCS trường Trung cấp Kỹ thuật và Nghiệp vụ Nam Sài Gòn.

1.3. Đối tượng nghiên cứu

Trong nghiên cứu này, đối tượng được chọn để lấy mẫu phục vụ cho đề tài nghiên cứu là các em học sinh TCCN hệ THCS đã học năm 2, năm 3 tại trường Trung cấp Kỹ thuật và Nghiệp vụ Nam Sài Gòn. Với đối tượng này, các em đã được học tại trường từ 2 năm trở lên, nên kết quả học tập và các yếu tố khác là cơ sở để dự đoán kết quả học tập của học sinh TCCN hệ THCS.

1.4. Phương pháp nghiên cứu

Phương pháp luận: Nghiên cứu các yếu tố ảnh hưởng đến kết quả học tập của học sinh và xác định mẫu dữ liệu. Nghiên cứu tài liệu về học máy và mô hình dự báo.

Phương pháp thực nghiệm: Phân tích dữ liệu mẫu và mô hình hồi quy Logistic để dự báo kết quả học tập của học sinh.

1.5. Ý nghĩa đề tài

Nghiên cứu giúp phân tích các yếu tố ảnh hưởng đến kết quả học tập của học sinh TCCN hệ THCS tại trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn.

Kết quả nghiên cứu cung cấp thông tin dự đoán kết quả học tập từ đó học sinh điều chỉnh và lập kế hoạch để học tập đạt kết quả cao hơn.

Kết quả nghiên cứu sẽ là cơ sở cho việc nhà quản lý, giáo viên chủ nhiệm nắm bắt tình hình kết quả học tập của học sinh từ đó có những kế hoạch kích thích cần thiết để làm tăng hiệu quả học tập của học sinh.

1.6. Phạm vi nghiên cứu

Đề tài chỉ nghiên cứu các yếu tố ảnh hưởng kết quả học tập của học sinh TCCN hệ THCS tại trường Trung cấp Kỹ thuật và Nghiệp vụ Nam Sài Gòn chứ chưa khảo sát trên đối tượng đang theo học ở những trường khác.

Đối tượng khảo sát: học sinh TCCN hệ THCS đã học năm 2, năm 3 của 5 khoa: Công nghệ thông tin, Du lịch, Điện tử, Cơ khí động lực, Cơ khí xây dựng tại trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn.

1.7. Bố cục luận văn nghiên cứu

Luận văn được trình bày gồm 5 chương như sau:

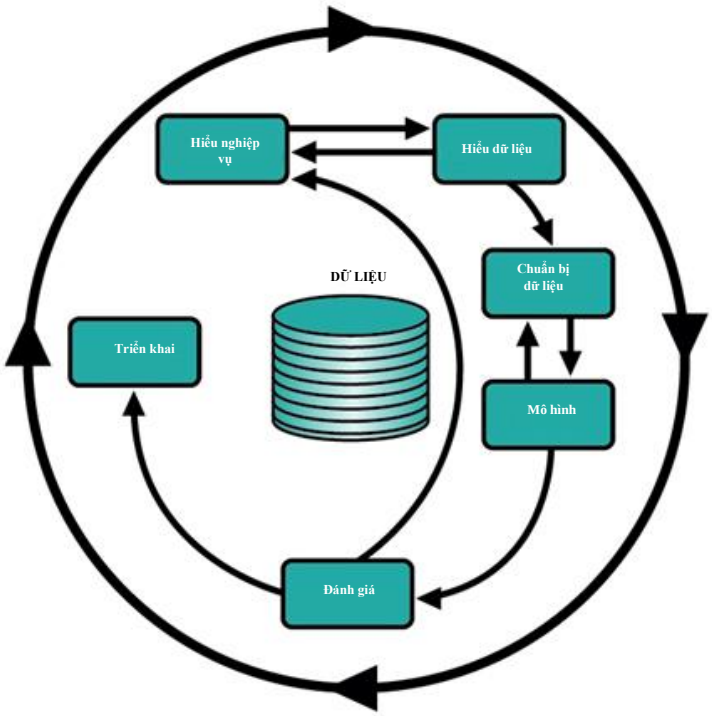
- Chương 1 (Giới thiệu nghiên cứu): Trình bày khái quát về cơ sở hình thành đề tài, xác định vấn đề nghiên cứu, mục tiêu nghiên cứu, phạm vi nghiên cứu, đối tượng nghiên cứu, phương pháp nghiên cứu và ý nghĩa thực tiễn của đề tài.
- Chương 2 (Cơ sở lý thuyết): Trình bày lý thuyết về học máy, ứng dụng học máy, hồi quy Logistic và công cụ R. Trình bày tổng quan về đào đạo TCCN và các nghiên cứu trước đây liên quan trên đó là cơ sở hình thành hình thành mô hình nghiên cứu sơ bộ và giả thuyết nghiên cứu cho tạo cảnh báo học tập.
- Chương 3 (Triển khai giải pháp tạo cảnh báo học tập): Thực hiện quá trình nghiên cứu bài toán theo tiếp cận CRISP-DM: tìm hiểu nghiên cứu, tìm hiểu dữ liệu, chuẩn bị dữ liệu, áp dụng hồi quy Logistic sử lý bài toán và đánh giá mô hình.
- Chương 4 (Đánh giá): Đánh giá bài toán tạo cảnh báo học tập
- Chương 5(Kết luận và hướng phát triển): Đưa ra những kết luận từ việc nghiên cứu đề tài rút ra, đồng thời đưa ra hướng phát triển trong tương lai.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

Chương 2 giới thiệu cơ sở lý thuyết về học máy, mô hình hồi quy Logistic, công cụ R và phương pháp đánh giá mô hình, hệ thống các mô hình nghiên cứu trước đây là cơ sở nghiên cứu các yếu tố ảnh hưởng đến dự báo kết quả của học sinh cho phần tiếp theo.

2.1. *Mô hình khai thác dữ liệu CRISP-DM*

Quá trình khai thác dữ liệu có thể trở nên phức tạp để theo dõi các vấn đề như nguồn dữ liệu, chất lượng dữ liệu, kỹ thuật khai thác dữ liệu do đó tác giả đề xuất mô hình khai thác dữ liệu theo mô hình CRISP - DM (Cross - Industry Standard Process for Data Mining). Mô hình quy trình của CRISP - DM bao gồm 6 giai đoạn giải quyết các vấn đề chính trong Datamining. Sáu giai đoạn kết hợp với nhau như một quá trình mang tính chu kỳ. [14]



Hình 1. Mô hình CRISP-DM

- **Hiểu nghiệp vụ (Business understanding)**
  - ✓ Tập trung vào hiểu biết mục tiêu, yêu cầu từ góc độ bài toán
  - ✓ Chuyển đổi tri thức này thành một định nghĩa bài toán khai thác dữ liệu và một kế hoạch sơ bộ được thiết kế để đạt được các mục tiêu.
- **Hiểu dữ liệu (Data understanding)**

Nguồn dữ liệu cung cấp nguyên liệu cho việc khai thác dữ liệu. Sự cần thiết ở giai đoạn này phải hiểu biết các nguồn dữ liệu của một doanh nghiệp đang có và đặc điểm của dữ liệu. Bao gồm việc thu thập dữ liệu ban đầu, mô tả dữ liệu, khai thác dữ liệu và kiểm tra chất lượng dữ liệu.

❖ **Chuẩn bị dữ liệu (Data preparation)**

Sau khi chia ra từng loại dữ liệu, đến giai đoạn cần chuẩn bị dữ liệu để khai thác. Việc chuẩn bị bao gồm việc lựa chọn, làm sạch, xây dựng, tích hợp và định dạng dữ liệu. Những nhiệm vụ này sẽ được thực hiện nhiều lần và không có bất kỳ thứ tự quy định nào. Những nhiệm vụ này có thể sẽ tốn nhiều thời gian nhưng là bước quan trọng cho sự thành công của việc khai thác dữ liệu. Chuẩn bị dữ liệu bao gồm:

❖ **Mô hình hóa (Modeling)**

Giai đoạn này liên quan đến việc lựa chọn kỹ thuật tạo ra các thiết kế thử nghiệm, xây dựng và đánh giá mô hình. Xây dựng mô hình là một quá trình lặp đi lặp lại, như thế mới có được một mô hình thống kê chuẩn. Sử dụng nhiều mô hình để đưa ra các dự đoán.

❖ **Đánh giá mô hình (Evaluation)**

Một khi đã chọn được một mô hình chuẩn, chuẩn bị bước qua giai đoạn đánh giá kết quả khai thác dữ liệu có thể giúp đạt được mục tiêu. Trước khi viết báo cáo tổng kết và triển khai mô hình, đều quan trọng là đánh giá sâu hơn về mô hình và xem xét các bước thực hiện xây dựng các mô hình để chắc chắn nó đạt được mục tiêu tốt.

❖ **Triển khai ứng dụng (Deployment)**

Giai đoạn triển khai các ứng dụng cho mô hình.

2.2. Tổng quan về học máy

Học máy (Machine Learning) là một ngành khoa học nghiên cứu các thuật toán cho phép máy tính có thể học được các khái niệm.

2.2.1. Phân loại học máy:

Có hai loại phương pháp học máy chính:

- Phương pháp quy nạp: Học máy phân biệt các khái niệm dựa trên dữ liệu đã thu thập được trước đó. Phương pháp này cho phép tận dụng được nguồn dữ liệu rất nhiều và sẵn có.

- Phương pháp suy diễn: Học máy phân biệt các khái niệm dựa vào các luật. Phương pháp này cho phép tận dụng được các kiến thức chuyên ngành để hỗ trợ máy tính.

Hiện nay, các thuật toán đều cố gắng tận dụng được ưu điểm của hai phương pháp này.

2.2.2. Các ngành khoa học liên quan:

- Lý thuyết thống kê: Các kết quả trong xác suất thống kê là tiền đề cho rất nhiều phương pháp học máy. Đặc biệt, lý thuyết thống kê cho phép ước lượng sai số của các phương pháp học máy.
- Các phương pháp tính: Các thuật toán học máy thường sử dụng các tính toán số thực/số nguyên trên dữ liệu rất lớn. Trong đó, các bài toán như: tối ưu có/không ràng buộc, giải phương trình tuyến tính v.v… được sử dụng rất phổ biến.
- Khoa học máy tính: Là cơ sở để thiết kế các thuật toán, đồng thời đánh giá thời gian chạy, bộ nhớ của các thuật toán học máy.

Các nhóm giải thuật học máy:

- Học có giám sát: Máy tính được xem một số mẫu gồm đầu vào (input) và đầu ra (output) tương ứng trước. Sau khi học xong các mẫu này, máy tính quan sát một đầu vào mới và cho ra kết quả.
- Học không giám sát: Máy tính chỉ được xem các mẫu không có đầu ra, sau đó máy tính phải tự tìm cách phân loại các mẫu này và các mẫu mới.
- Học nửa giám sát: Một dạng lai giữa hai nhóm giải thuật trên.
- Học tăng cường: Máy tính đưa ra quyết định hành động (action) và nhận kết quả phản hồi (response/reward) từ môi trường (environment). Sau đó máy tính tìm cách chỉnh sửa cách ra quyết định hành động của mình.

2.2.3. Các ứng dụng của học máy

Ứng dụng: Học máy có ứng dụng rộng khắp trong các ngành khoa học/sản xuất, đặc biệt những ngành cần phân tích khối lượng dữ liệu khổng lồ. Một số ứng dụng thường thấy:

- Xử lý ngôn ngữ tự nhiên (Natural Language Processing): Xử lý văn bản, giao tiếp người – máy, …
- Nhận dạng (Pattern Recognition): Nhận dạng tiếng nói, chữ viết tay, vân tay, thị giác máy (Computer Vision) …
- Tìm kiếm (Search Engine)
- Chẩn đoán trong y tế: Phân tích ảnh X-quang, các hệ chuyên gia chẩn đoán tự động.
- Tin sinh học: Phân loại chuỗi gene, quá trình hình thành gene/protein
- Vật lý: Phân tích ảnh thiên văn, tác động giữa các hạt …
- Phát hiện gian lận tài chính (financial fraud): Gian lận thẻ tỉn dụng phân tích thị trường chứng khoán (stock market analysis)
- Chơi trò chơi: tự động chơi cờ, hành động của các nhân vật ảo
- Rôbốt: là tổng hợp của rất nhiều ngành khoa học, trong đó học máy tạo nên hệ thần kinh/bộ não của người máy.

2.3. Mô hình hồi quy Logistic

Khoa học thống kê là khoa học về việc thu thập, phân tích, diễn giải và trình bày các số liệu để tìm ra bản chất và tính chất của các hiện tượng kinh tế, tự nhiên và xã hội. Thống kê cho phép tóm tắt và trình bày một cách dễ hiểu các thông tin bằng số, kiểm định một giả thuyết hoặc dự báo về khả năng xảy ra của một biến cố nào đó… Tất cả những vai trò đó được gói trong bài toán hồi quy. Bài toán hồi quy là bài toán thiết lập mối quan hệ giữa một đối tượng đang được quan tâm (biến đáp ứng) và các đối tượng liên quan (các biến dự báo) để đưa ra những kết luận có ý nghĩa thống kê. Khi biến đáp ứng là biến nhị phân hay rời rạc thì mô hình hồi quy tuyến tính không thể áp dụng được vì biến đáp ứng không liên tục, một mô hình hồi quy mới được xây dựng để giải quyết vấn đề trên, đó là mô hình hồi quy Logistic. Mô hình hồi quy Logistic được phát triển bởi nhà thống kê học David R. Cox vào những năm 1970 và ngày càng ứng dụng rộng rãi. Chẳng hạn trong các nghiên cứu y khoa, mục tiêu chính là phân tích mối tương quan giữa yếu tố nguy cơ và nguy cơ mắc bệnh.

Trong nghiên cứu này đối tượng phân tích thường được thể hiện qua các biến nhị phân: đâu/rớt nên luận văn giới thiệu về mô hình hồi quy Logistic, phương pháp ước lượng tham số và kiểm định kết quả thống kê đối với mô hình hồi quy này.

Cấu trúc dữ liệu trong mô hình như sau:

Bảng 1. Danh mục biến trong mô hình hồi quy Logistic

Biến	Loại
Phụ thuộc	Nhị phân
Độc lập	Nhị phân, liên tục và rời rạc

Giả sử biến giả Y phụ thuộc vào chỉ số khả dụng Y*. Trong đó:

$$Y^*=ß_0 + ß_1X_{1i} + \cdots + ß_kX_{ki} + Єi$$

Vì Y(x) là biến nhị phân có thể được giải thích như sau:

$$Yi = \begin{cases} 0 \text{ nếu kết quả học tập rớt} \\ 1 \text{ nếu kết quả học tập đậu} \end{cases}$$

Trong đó Pi=P(Yi=1/Xi), khi đó Yi là biến ngẫu nhiên phân phối theo qui luật Bernoulli, có nghĩa là: $fi(Yi)=P_i^{Y_i}(1-P_i)^{1-Y_i}$, trong đó $Y_i=0,1,...,n$. Khi đó, kì vọng toán và phương sai được tính như sau: $E(Y_i)=n_iP_i$, $Var(Y_i)=n_iP_i(1-P_i)$. Vì Y_i là biến ngẫu nhiên phân phối theo qui luật Bernoulli nên có thể viết lại như sau:

$$P^{Yi}(1-P_i)^{1-Yi} = (1-P_i).Exp(Y_i.\ Log(\frac{Pi}{1-Pi}))$$

Tỷ lệ chênh lệch: $odds=P_i/(1-Pi)$

$P_i=P(Y_i=1)$

$P_i=P(Y_i*>0)$

$P_i=P(ß_0 + ß_1X_{1i} + \cdots + ß_kX_{ki} + Є_i>0)$

Mở rộng hơn nữa có thể viết như sau:

$$Log[P_i/(1-P_i)]= ß_0 + ß_1X_{1i} + \cdots + ß_kX_{ki}$$

$$P(Y_i=1) = P_i = \frac{Exp(ß_0+ß_1X_{1i}+\cdots+ß_kX_{ki})}{1+Exp(ß_0+ß_1X_{1i}+\cdots+ß_kX_{ki})}$$

$$P(Y_i=0) = 1- P_i = \frac{1}{1+Exp(ß_0+ß_1X_{1i}+\cdots+ß_kX_{ki})}$$

Trong mô hình trên P_i không phải là hàm tuyến tính của các biến độc lập. Phương trình được gọi là hàm phân bố Logistic.Trong hàm này khi Xi nhận các giá trị từ - ∞ đến + ∞ thì Pi nhận giá trị từ 0-1.

Nếu kí hiệu:

$$\beta = \begin{Bmatrix} ß_1 \\ ß_2 \\ \dots \\ ß_k \end{Bmatrix} \qquad X = \begin{Bmatrix} X_1 \\ X_2 \\ \dots \\ X_k \end{Bmatrix}$$

Khi đó $Z = ß_0 + ß_1 X_{1i} + \cdots + ß_k X_{ki}$ và

$$E\,(Y=1) = \frac{\exp(Z)}{1+\exp(Z)}$$

2.4. Phương pháp đánh giá độ chính xác của mô hình

Đánh giá độ chính xác của bộ phân lớp rất quan trọng, bởi vì nó cho phép dự đoán được độ chính xác của các kết quả phân lớp những dữ liệu tương lai. Độ chính xác còn giúp so sánh các mô hình phân lớp khác nhau. Một số phương pháp đánh giá phổ biến bao gồm:

2.4.1. Phương pháp phân chia ngẫu nhiên (Holdout Splitting)

Trong phương pháp holdout, dữ liệu dưa ra được phân chia ngẫu nhiên thành 2 phần là: tập dữ liệu đào tạo và tập dữ liệu kiểm tra. Thông thường 2/3 dữ liệu cấp cho tập dữ liệu đào tạo, phần còn lại cho tập dữ liệu kiểm tra.

Toàn bộ tập ví dụ *D* được chia thành 2 tập con không giao nhau

Tập huấn luyện *D_train* – để huấn luyện hệ thống

Tập kiểm thử *D_test* – để đánh giá hiệu năng của hệ thống đã học

→ $D = D\_train \cup D\_test$, và thường là $|D\_train| >> |D\_test|$

Các yêu cầu:

- Bất kỳ ví dụ nào thuộc vào tập kiểm thử *D_test* đều không được sử dụng trong quá trình huấn luyện hệ thống
- Bất kỳ ví dụ nào được sử dụng trong giai đoạn huấn luyện hệ thống (i.e., thuộc vào *D_train*) đều không được sử dụng trong giai đoạn đánh giá hệ thống
- Các ví dụ kiểm thử trong *D_test* cho phép một đánh giá không thiên vị đối với hiệu năng của hệ thống

Các lựa chọn thường gặp: $|D\_train| = (2/3).|D|$, $|D\_test| = (1/3).|D|$

Phù hợp khi ta có tập ví dụ *D* có kích thước lớn

2.4.2. Phương pháp kiểm tra chéo k-fold (K-Fold cross validation)

Để tránh việc trùng lặp giữa các tập kiểm thử (một số ví dụ cùng xuất hiện trong các tập kiểm thử khác nhau)

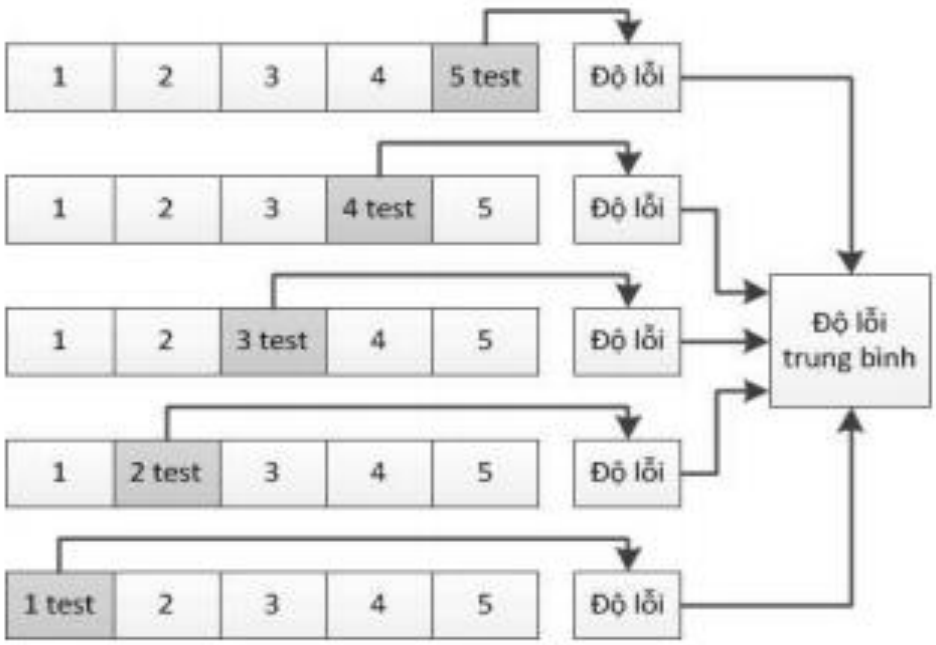
k-fold cross-validation

- Tập toàn bộ các ví dụ *D* được chia ngẫu nhiên thành *k* tập con **không giao nhau** (gọi là "*fold"*) có kích thước xấp xỉ nhau
- Mỗi lần (trong số *k* lần) lặp, một tập con được sử dụng làm tập kiểm thử, và (*k*-1) tập con còn lại được dùng làm tập huấn luyện
- *k* giá trị lỗi (mỗi giá trị tương ứng với một *fold*) được tính trung bình cộng để thu được giá trị lỗi tổng thể

Các lựa chọn thông thường của k: 10, hoặc 5

Thông thường, mỗi tập con (fold) được lấy mẫu phân tầng (xấp xỉ phân bố lớp) trước khi áp dụng quá trình đánh giá Cross-validation

Phù hợp khi ta có tập ví dụ D vừa và nhỏ



Hình 2. Mô tả phương pháp thử nghiệm K-Fold với k=5

2.4.3. Phương pháp kiểm tra từng phần (Leave-one-out cross validation)

Có thể coi là thử nghiệm trên từng cá nhân, là việc tiến hành thử nghiệm với dữ liệu huấn luyện (training) và dữ liệu kiểm thử (test) trên cùng một người, tức là sử dụng dữ liệu thu được từ một người để huấn luyện, sau đó dùng dữ liệu cũng của người đó

nhưng chưa được dùng trong huấn luyện để kiểm tra độ chính xác theo phương pháp kiểm tra chéo (cross-validation).

Một trường hợp (kiểu) của phương pháp Cross-validation

- Số lượng nhóm các (folds) bằng kích thước của tập dữ liệu (k=|D|)
- Mỗi nhóm (fold) chỉ bao gồm một ví dụ

Khai thác tối đa (triệt để) tập ví dụ ban đầu

Không hề có bước lấy mẫu ngẫu nhiên (no random subsampling)

Áp dụng lấy mẫu phân tầng (stratification) không phù hợp

→ Vì ở mỗi bước lặp, tập thử nghiệm chỉ gồm có một ví dụ

Chi phí tính toán (rất) cao

Phù hợp khi ta có một tập ví dụ D (rất) nhỏ

2.5. Tổng quan về R

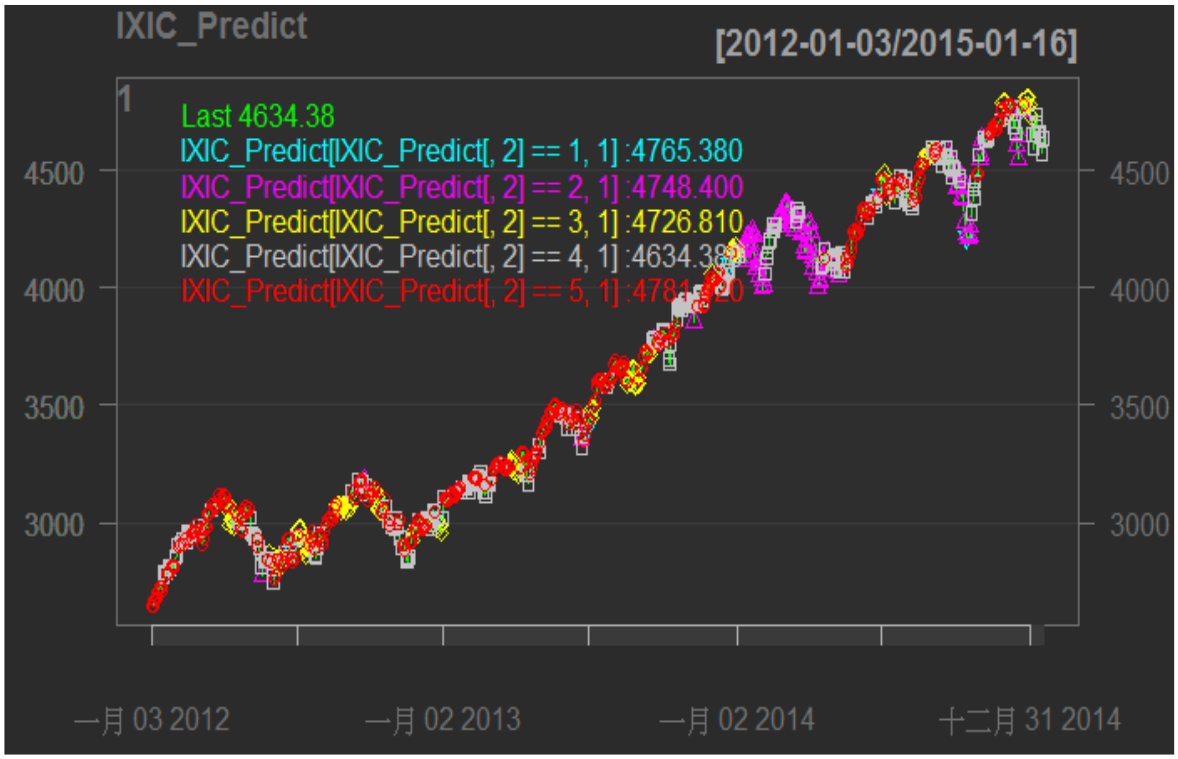
R là một ngôn ngữ lập trình và môi trường phần mềm dành cho tính toán và đồ họa thống kê. Đây là một bản hiện thực ngôn ngữ lập trình S với ngữ nghĩa khối từ vựng lấy cảm hứng từ Scheme. R do Ross Ihaka và Robert Gentleman tạo ra tại Đại học Auckland, New Zealand, đến nay do R Development Core Team chịu trách nhiệm phát triển. Tên của ngôn ngữ một phần lấy từ chữ cái đầu của hai tác giả (Robert Gentleman và Ross Ihaka), một phần cũng là cách chơi chữ từ tên S.

Ngôn ngữ R đã trở thành một tiêu chuẩn trên thực tế giữa các nhà thống kê cho thấy sự phát triển của phần mềm thống kê, và được sử dụng rộng rãi để phát triển phần mềm thống kê và phân tích dữ liệu.

R là một bộ phận của dự án GNU. Mã nguồn của nó được công bố tự do theo giấy phép bản quyền công cộng GNU, và có các phiên bản dịch sẵn cho nhiều hệ điều hành khác nhau. R sử dụng giao diện dòng lệnh, tuy cũng có một vài giao diện đồ họa người dùng dành cho nó.

Sử dụng R để đơn giản hoá học máy. Tất cả những gì bạn cần phải biết là làm thế nào mỗi thuật toán có thể giải quyết vấn đề của bạn, và sau đó bạn chỉ sử dụng một gói phần mềm được viết ra để nhanh chóng tạo ra mô hình dự đoán trên dữ liệu với một vài dòng lệnh. Ví dụ, bạn có thể thực hiện Naïve Bayes cho lọc thư rác, sử dụng gom cụm k-means cho phân khúc khách hàng, sử dụng hồi quy tuyến tính để dự báo giá nhà, hoặc

thực hiện một mô hình Markov để dự đoán thị trường chứng khoán, như thể hiện trong hình bên dưới:



Biểu đồ 1. Dự đoán chứng khoán sử dụng R

2.6. Các nghiên cứu có liên quan

2.6.1. Nghiên cứu quốc tế

[1] P. Baepler and C.J. Murdoch (2010) "*Academic Analytics and Data Mining in Higher Education*": Phân tích đưa ra mối liên kết giữa thuật toán, khai thác dữ liệu hệ thống quản lý khóa học và giới thiệu những kĩ thuật và dữ liệu có ích cho người học và người dạy.

[2] E.J.M. Lauría, J.D. Baron, M. Devireddy, V. Sundararaju and S.M. Jayaprakash "*Mining academic data to improve college student retention: An open source perspective*": Báo cáo về nghiên cứu đang thực hiện ở OAAI, một dự án nhằm tăng sự cố gắng của sinh viên bằng cách phát hiện những nguy cơ trong đào tạo dùng phương pháp khai thác dữ liệu. Bài luận này mô tả mục tiêu của OAAI và hệ phương pháp luận để phát triển một mô hình có thể đưa ra những suy luận về kết quả học tập của sinh viên, sử dụng những nguồn dữ liệu mở của hệ thống quản lý cũng như các kết quả học tập đã được lưu trữ của sinh viên.

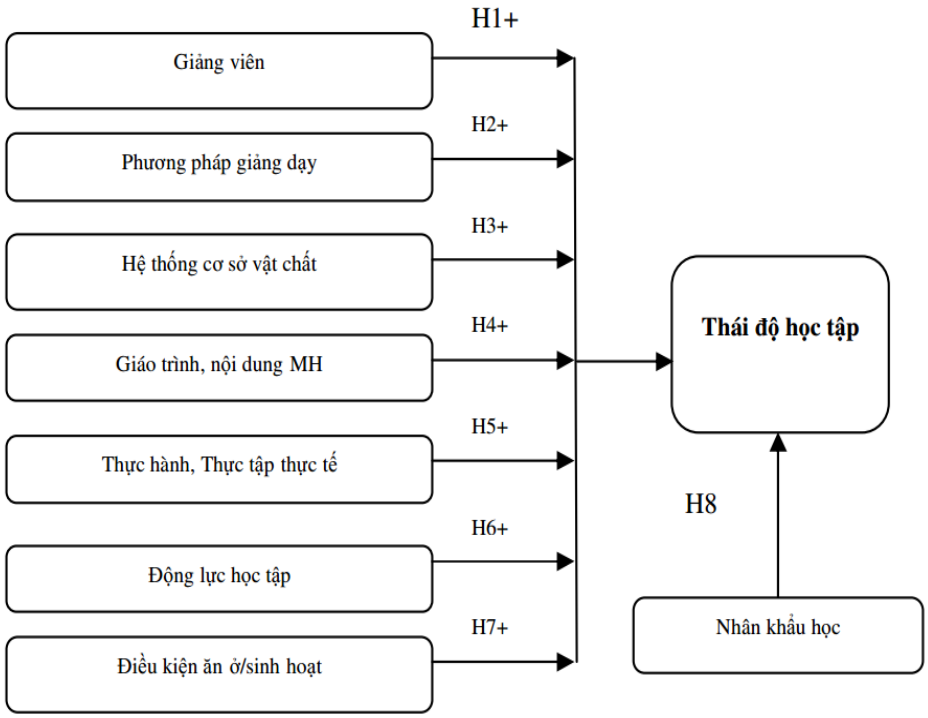
[3] B.K. Baradwaj and S. Pal "*Mining Educational Data to Analyze Students Performance*": Mục tiêu chính của các cơ sở giáo dục bậc cao là cung cấp cho sinh viên một nền giáo dục chất lượng cao. Một phương pháp để có thể đạt tới chất lượng cao nhất trong giáo dục bậc cao là phát hiện những phương thức dự đoán căn cứ vào việc tuyển sinh những khóa nhất định, sự cách biệt của mô hình dạy học truyền thống, phát hiện những phương tiện không phù hợp trong kiểm tra trực tuyến, phát hiện những bất thường trong bài thi của sinh viên, dự đoán về kết quả của sinh viên. Bài nghiên cứu này được thiết kế để đánh giá khả năng của kỷ thuật khai thác dữ liệu trong môi trường giáo dục bậc cao ở trường đại học. Trong nghiên cứu này, tác vụ phân loại được sử dụng để đánh giá kết quả của sinh viên và có rất nhiều hướng tiếp cận được sử dụng trong phân loại dữ liệu, phương pháp cây nhị phân cũng được sử dụng ở đây**.** Với tác vụ này, nghiên cứu có được thông tin mô tả kết quả của sinh viên trong kì thi cuối kì. Nó giúp sớm phát hiện những trường hợp phải bỏ học hoặc những sinh viên cần sự chăm sóc đặc biệt và cho phép giáo viên cung cấp những lời khuyên phù hợp.

[4] J. Bainbridge, J. Melitski, A. Zahradnik, E.J. M. Lauría, S. Jayaprakash, and J. Baron "*Using Learning Analytics to Predict At-Risk Students in Online Graduate Public Affairs and Administration Education*": Phân tích tính cách và hành vi có những chỉ định chuẩn nhất về nguy cơ trong học tập, chú ý cụ thể tới việc sử dụng các công cụ học tập online. Mô hình phân tích học tập đạt được kết quả chính xác khá cao (80%) học sinh gặp nguy cơ đã được phát hiện. Kết quả được sử dụng để kiểm tra quá trình tiến bộ nhằm cải thiện kết quả học tập của sinh viên trong thực tế.

[5] P. Cortez and A. Silva. "*Using Data Mining to Predict Secondary School Student* Performance": Nghiên cứu này nhằm hướng tới kết quả của học sinh THCS sử dụng phương pháp trí tuệ kinh doanh và khai thác dữ liệu. Những dữ liệu thực tế (điểm học tập, hoàn cảnh, đặc điểm trường học và xã hội) được thu thập qua bảng điểm và câu hỏi phỏng vấn. Hai môn học (toán và ngữ văn) được mô tả thông qua phân loại 2 mức độ/5 mức độ và tác vụ lặp. Thêm vào đó, 4 mô hình (cây quyết định, ngẫu nhiên…. Mạng nơ-ron và vector hỗ trợ) và 3 kiểu chọn dữ liệu vào (có hoặc không có điểm học kì trước) đã được kiểm tra. Kết quả chỉ ra rằng, tính chính xác của dự đoán có thể đạt được, với điều kiện là có điểm của học kì 1 và học kì 2.

2.6.2. Nghiên cứu Việt Nam

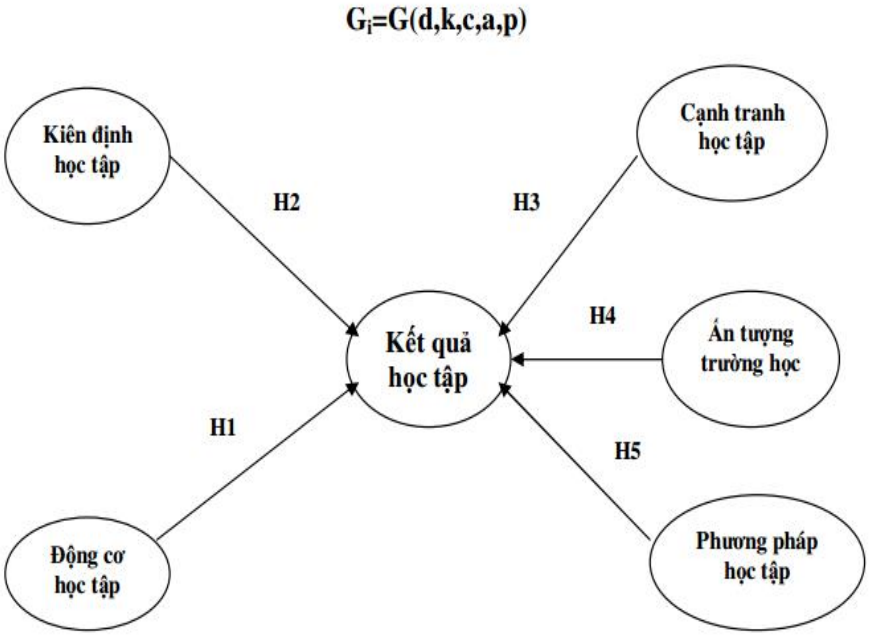
[6] Phạm Hữu Tín và Nguyễn Thúy Huỳnh Loan "*Các yếu tố ảnh hưởng đến thái độ học tập của sinh viên trường Đại học Đà Lạt*": Nghiên cứu xác định các yếu tố tác động đến thái độ học tập của sinh viên từ đó đưa ra những hàm ý cho nhà trường trong việc thúc đẩy thái độ học tập của sinh viên, từng bước nâng cao chất lượng đào tạo. Thông qua nghiên cứu định tính dựa trên 7 yếu tố tác động đến thái độ của sinh viên gồm: Giảng viên; Phương pháp giảng dạy; Hệ thống cơ sở vật vất; Giáo trình, nội dung môn học; Thực hành, thực tập thực tế; Động lực học tập; Điều kiện ăn ở, sinh hoạt. Kết quả phân tích cho thấy 7 yếu tố điều có ảnh hưởng tích cực tới thái độ học tập của sinh viên, trong đó yếu tố Động lực học tập và Giáo trình, nội dung môn học có tác động tích cực nhất.



Hình 3. Mô hình các yếu tố ảnh hưởng đến thái độ học tập của sinh viên trường Đại học Đà Lạt

[7] Võ Thị Tâm " *Các yếu tố tác động đến kết quả học tập của sinh viên chính quy trường Đại học Kinh Tế Thành Phố Hồ Chí Minh*": Nghiên cứu xác định các yếu tố ảnh hưởng đến kết quả của sinh viên gồm 5 yếu tố: Kiên định trong học tập, Động cơ học tập, Cạnh tranh học tập, Ấn tượng trường học và phương pháp học tập. Trong đó yếu tố Kiên định trong học tập và Phương pháp học tập có ảnh hưởng rất tích cực

.

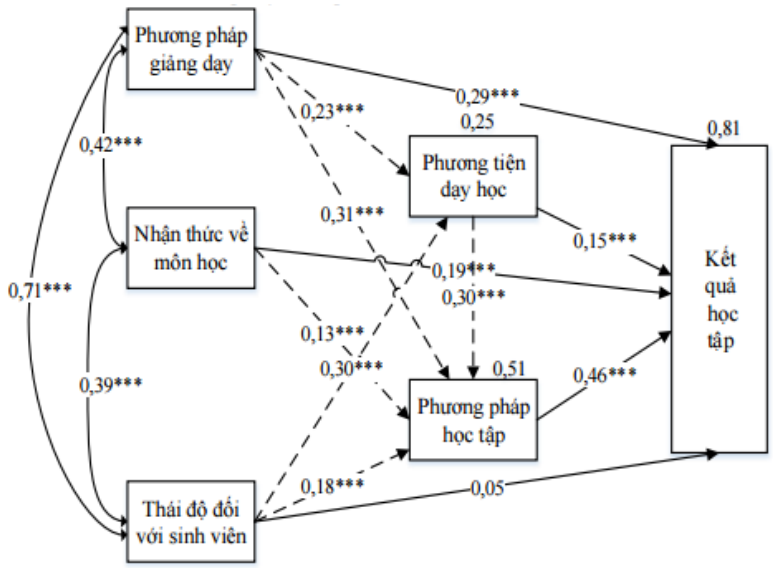


Hình 4. Mô hình các yếu tố tác động đến kết quả học tập của sinh viên chính quy trường Đại học Kinh Tế Thành Phố Hồ Chí Minh

[8] Nguyễn Công Toàn, Trịnh Minh Trí, Huỳnh Văn Hậu Nguyễn Thị Cẩm Hồng và Nguyễn Văn Quân "*Các yếu tố ảnh hưởng đến kết quả học tập của sinh viên đại học ngành phát triển nông thôn của trường Đại học Cần Thơ*"**:** Nghiên cứu của đề tài là phân tích các yếu tố ảnh hưởng đến kết quả học tập của sinh viên đại học ngành Phát triển Nông thôn, Trường Đại học Cần Thơ. Kết quả phân tích hồi qui cho thấy, có 4 biến ảnh hưởng đến kết quả học tập của sinh viên ngành phát triển nông thôn đó là giới tính, số giờ tự học, số buổi nghỉ học và tài liệu giảng viên cung cấp, trong đó 2 biến số giờ tự học và tài liệu giảng viên cung cấp có tương quan thuận với biến phụ thuộc kết quả học của sinh viên, còn 2 biến giới tính và số buổi nghỉ học có tương quan nghịch với biến phụ thuộc KQHT.

[9] Dư Thống Nhất và Nguyễn Thị Nụ "*Các yếu tố ảnh hưởng đến kết quả học tập môn tâm lý học của sinh viên trường cao đẳng Sư phạm Kiên Giang*": Mục đích của nghiên cứu này là xác định con đường tác động của các yếu tố ảnh hưởng đến kết quả học tập môn Tâm lý học (TLH) của sinh viên. Cả hai phương pháp nghiên cứu định tính và định lượng đều được sử dụng để đo lường kết quả nghiên cứu. Kết quả nghiên cứu cho thấy có năm yếu tố tác động đến kết quả học tập môn tâm lý học bao gồm: phương pháp học tập của sinh viên, phương pháp giảng dạy của giáo viên, phương tiện dạy học,

nhận thức về môn học của sinh viên, và thái độ của giảng viên đối với sinh viên. Mô hình nghiên cứu giải thích được 81% sự biến thiên kết quả học tập của sinh viên.



Hình 5. Các yếu tố ảnh hưởng đến kết quả học tập môn tâm lý học của sinh viên trường Cao đẳng Sư phạm Kiên Giang

2.7. Tóm tắt chương

Chương 2 này, tác giả trình bày nghiên cứu các lý thuyết thực tiễn tình trạng đào tạo tại thành phố Hồ Chí Minh nói chung và tại trường Trung cấp Kỹ thuật và Nghiệp vụ Nam Sài Gòn nói riêng đồng thời tham khảo thêm một số nghiên cứu trong và ngoài nước trước đây trong lĩnh vực nghiên cứu, đó là tiền đề cho việc đề xuất mô hình nghiên cứu ban đầu và xây dựng các giả thuyết cho nghiên cứu sau này. Bên cạnh đó, tác giả đã hệ thống cơ sở lý thuyết gồm quy trình CRISP-DM, mô hình hồi quy Logistic, phương pháp đánh giá và công cụ R, đó là cách thức để giải bài toán tạo cảnh báo kết quả học tập học sinh .

CHƯƠNG 3: TRIỂN KHAI GIẢI PHÁP TẠO CẢNH BÁO KẾT QUẢ HỌC TẬP HỌC SINH TCCN HỆ THCS

Dựa vào cơ sở lý thuyết chương 2, tác giả thực hiện triển khai giải pháp tạo cảnh báo kết quả học tập của học sinh TCCN hệ THCS theo tiếp cận của quy trình CRISP-DM với việc áp dụng mô hình hồi quy Logistic.

3.1. Tìm hiểu cảnh báo kết quả học sinh

3.1.1. Thực trạng nghiên cứu

3.1.1.1. Khái quát về các cơ sở đào tạo TCCN tại Thành phố Hồ chí Minh

❖ *Thực trạng chất lượng đào tạo tại các trường trung cấp chuyên nghiệp Thành phố Hồ Chí Minh*

Trong giai đoạn 2005-2010, nhìn chung, tỷ lệ HS đạt khá giỏi có chiều hướng gia tăng, tỷ lệ yếu kém giảm dần.

Bảng 2. Kết quả xếp loại học tập của học sinh

Năm học	Tổng số HS	Xếp loại				
		Xuất sắc	Giỏi	Khá, TB khá	Tr. bình	Yếu , kém
2005-2006	28.182	265 *0.94%*	1.926 *6,83%*	7.533 *26,73%*	14.786 *52,47%*	3.672 *13,03%*
2006- 2007	25.200	219 *0,87%*	1.814 *7,20%*	6.738 *26,73%*	13.301 *52,79%*	3.128 *12,41%*
2007-2008	34.048	105 *0,31%*	1.696 *4,98%*	14.065 *41,31%*	11.651 *34,22%*	6.531 *19,18%*
2008-2009	47.483	532 *1,12%*	2.986 *6,29%*	25.893 *54,53%*	12.157 *25,60%*	5.915 *12,46%*
2009-2010	44.918	140 *0,31%*	2.620 *5,83%*	27.273 *60,72%*	10.111 *22,51%*	4.774 *10,63%*
2010-2011	55.379	206 *0,37%*	4480 *8,09%*	34.574 *62,43%*	11.584 *20,92%*	4.535 *8,18%*
2011-2012	58.662	337 *0,57%*	3.791 *6,46%*	36.571 *62,34%*	13.717 *23,38%*	4.246 *7,24%*

(Nguồn: Sở GD&ĐT TP. HCM)

Về xếp loại rèn luyện của HS các trường TCCN được thể hiện ở bảng 3

Bảng 3. Kết quả xếp loại rèn luyện của học sinh

(Năm học	Tổng số HS	Xếp loại				
		Xuất sắc	Tốt	Khá, TB khá	Tr. bình	Yếu , kém
2005-2006	27.954	2.001 *7,16%*	7.767 *27,78%*	14.822 *53,02%*	2.353 *8,42%*	1.011 *3,62%*
2006-2007	29.483	2.575 *8,73%*	8.246 *27,97%*	13.931 *47,26%*	2.820 *9,56%*	1.911 *6,48%*
2007-2008	33.654	2.560 *7,61%*	9.825 *29,19%*	14.474 *43,01%*	4.511 *13,40%*	2.284 *6,79%*
2008-2009	43.701	3.501 *8,01%*	12.199 *27,91%*	20.978 *48,01%*	4.844 *11,08%*	2.179 *4,99%*
2009-2010	47.744	4.543 *9,52%*	15.076 *31,58%*	19.979 *41,84%*	5.498 *11,52%*	2.648 *5,54%*
2010-2011	59.962	7.317 *12,20%*	21.652 *36,11%*	24.744 *41,26%*	5.493 *7,59%*	839 *1.40%*
2011-2012	63.159	6.509 *10,31%*	21.812 *34,54%*	27.251 *43,15%*	5.151 *8,16%*	2.436 *3,86%*

(Nguồn: Sở GD&ĐT TP. HCM)

Quản lý chất lượng HS từ đầu vào đến quá trình và đầu ra là một việc khá phức tạp vì liên quan đến nhiều lĩnh vực ngoài phạm vi và năng lực hoạt động của trường TCCN như hướng nghiệp, chất lượng đầu vào (chủ yếu là ý thức tự giác và trình độ học vấn thấp do chỉ xét tuyển), mối quan hệ giữa trường với các cơ sở có nhu cầu sử dụng lao động qua đào tạo… Hiệu suất đào tạo (ngay trong một năm học đầu tiên của một số trường lớn, có uy tín) cho thấy tỷ lệ giảm HS là vấn đề đáng quan tâm. Đặc biệt đối với đầu vào là HS tốt nghiệp THCS, do các em chưa xác định đúng đắn mục tiêu, thái độ và động cơ học tập, trình độ học vấn và khả năng tiếp thu bài giảng kém. Tỷ lệ bỏ, nghỉ học có trường đến khoảng 30 – 40%. Ngược lại, HS tốt nghiệp THPT hoặc chưa tốt nghiệp THPT lại có nhận thức học tập tốt (có lẽ do không còn sự lựa chọn), chỉ vài trường hợp cá biệt bỏ học do trúng tuyển CĐ hoặc ĐH. Do vậy, hiệu suất đạt từ 80 – 90%. Để bảo đảm hiệu suất đa số các trường TCCN có xu hướng tuyển sinh HS tốt nghiệp THPT. Điều này ảnh hưởng tiêu cực đến chủ trương phân luồng của Nhà nước.

Bảng 4. Hiệu suất đào tạo và kết quả xếp loại tốt nghiệp của HS

Năm học	Tổng số HS đầu vào	Tổng số HS dự thi cuối khóa	Hiệu suất đào tạo	Xếp loại				
				Xuất sắc	Giỏi	Khá, Trung bình khá	Trung bình	Yếu, kém không tốt nghiệp
2005- 2006	7.998	6.159	77,13%	4 *0,06%*	322 *5,23%*	1.994 *32,38%*	3.662 *59,46%*	177 *2,87%*
2006- 2007	7.143	5.786	80,78%	1 *0,02%*	239 *4,13%*	1.436 *24,82%*	3.147 *54,39%*	963 *16,64%*
2007-2008	15.607	10.613	68,06%	12 *0,01%*	788 *7,42%*	4.436 *41,80%*	3.484 *32,83%*	1.893 *17,84%*
2008-2009	14.915	10.292	68,98%	73 *0,71%*	677 *6,58%*	5.627 *54,67%*	2.495 *24,24%*	1.420 *13,80%*
2009-2010	12.395	7.561	60,98%	36 *0,47%*	418 *5,53%*	4.764 *63,01%*	1.280 *16,93%*	1.063 *14,06%*
2010-2011	27.447	17.643	64,28%	23 *0,15%*	970 *6,31%*	11.307 *41,20%*	3.078 *20,02%*	2.269 *12,83%*
2011- 2012	32.909	23.875	72,55%	36 *0,19%*	1.621 *8,45%*	14.834 *62,13%*	2.683 *13,99%*	4.702 *19,69%*

(Nguồn: Sở GD&ĐT TP. HCM)

Một số trường đã nỗ lực duy trì sĩ số HS thông qua nhiều hoạt động như cải thiện môi trường học; tổ chức các loại hình hoạt động như tuyên truyền hướng nghiệp, tổ chức ngày hội thanh niên với nghề nghiệp và việc làm, câu lạc bộ, đội nhóm để thu hút học sinh; đổi mới giáo trình, PP dạy học… Nhưng kết quả đạt được vẫn chưa cao, tuy nhiên so với nhiều năm trước đây, tỉ lệ giảm HS trong các trường đã bắt đầu có dấu hiệu khả quan.

Bảng 5. Hiệu suất đào tạo toàn khoá ở một số trường TCCN

Trường	2007			2009			2010		
	HS đầu vào	Tốt nghiệp	*Hiệu suất (%)*	HS đầu vào	Tốt nghiệp	*Hiệu suất (%)*	HS đầu vào	Tốt nghiệp	*Hiệu suất (%)*
Lý Tự Trọng	977	560	***57.3***	955	351	***36.8***	1063	395	***42.1***
Thủ Đức	791	599	***75.7***	2035	1210	***59.5***	2032	1230	***60.5***
Kinh tế	1231	765	***62.1***	1531	678	***44.3***	1174	1003	***85.4***
Nam Sài Gòn	361	322	***78.9***	408	291	***71.3***	599	416	***69.4***
Ng. Hữu Cảnh	275	169	***61.5***	270	57	***21.1***	941	443	***48.0***

(Nguồn: Sở GD&ĐT TP. HCM)

3.1.1.2. Sơ lược vài nét về đào tạo nghề hệ cơ sở tại đơn vị khảo sát

Các phòng chức năng và khoa, tổ bộ môn:

- Gồm 8 phòng (Đào tạo – Khảo thí, Tổ chức – Hành chính, Quản trị – Thiết bị & Cơ sở vật chất, Kế hoạch – Tài chính, Công tác chính trị – học sinh, sinh viên, Quảng bá – Hợp tác, Nghiên cứu – Phát triển, Thanh tra – Pháp chế – Đảm bảo chất lượng) và 13 khoa – tổ bộ môn (Khoa học Xã hội, Khoa học Tự nhiên, Công nghệ Thông tin, Sư phạm mầm non – Nữ công, Cơ khí Động lực, Cơ khí – Xây dựng, Điện – Điện tử, Kinh tế, Du lịch, Y – dược, Lý luận chính trị, Pháp luật – Công tác xã hội, Tổ bộ môn Thể chất & Quốc phòng).

Nhân sự *(tính đến ngày 15/3/2015)*

- Tổng số cán bộ, giáo viên, nhân viên: 262 trong đó có 128 nữ.
- Số giáo viên cơ hữu trực tiếp giảng dạy: 147 trong đó có 78 nữ
- Số giáo viên hợp đồng thỉnh giảng: 66 trong đó có 31 nữ

❖ *Quy mô đào tạo các nghề ở Trường TC KT & NV Nam Sài Gòn TP Hồ Chí Minh*

Bảng 6. Quy mô đào tạo ngành nghề

Hệ Trung cấp chuyên nghiệp		
Năm	**Số ngành**	**Số ngành tăng**
2000	2	
2005	4	2
2006	6	4
2007	13	7
2009	21	8
2013 – 2015	27	6

(Nguồn: Phòng Quản lý đào tạo

trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn, 2015)

❖ *Lưu lượng đào tạo*

Bảng 7. Số lượng đào tạo hàng năm

Năm học	TCCN		
	SỐ HS	Hệ THPT	Hệ THCS
2005 – 2006	535	-	-
2006 – 2007	604	-	-
2007 – 2008	921	-	-
2008 – 2009	1285	565	720
2009 – 2010	1999	926	1073
2010 – 2011	2198	982	1216
2011 – 2012	2771	1463	1308
2012 – 2013	3202	1952	1250
2013-2014	2670	1594	1076
2014 - 2015	2596	1410	1186

(Nguồn: Phòng Quản lý đào tạo trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn, 2015)

❖ *Thực trạng phát triển qui mô và hiệu suất đào tạo hệ TCCN*

Bảng 8. Hiệu suất đào tạo theo hàng năm

Chi tiết		Số liệu kết quả tốt nghiệp từng năm									
		2006	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
Nhập học		301	408	368	599	411	519	1000	1347	1679	1963
Tốt nghiệp	***Dự thi***	234	362	372	426	400	742	742	783	1331	1437
	Đạt TN	225	323	291	416	391	736	736	750	1310	1387
Hiệu suất đào tạo (%)		74,8	89,2	79,1	69,4	95,1	93,1	73,6	58,1	78	73,2

(Nguồn: Phòng Quản lý đào tạo trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn, 2015)

3.1.2. Xác đinh mục tiêu nghiên cứu

Dựa vào báo cáo tình trạng học tập trên và các nghiên cứu trước đây đã trình bày ở chương 2 và nghiên cứu định tính, tác giả nhận thấy các yếu tố ảnh hưởng đến kết quả của học sinh TCCN hệ THCS được tổng hợp từ các yếu tố khác nhau.

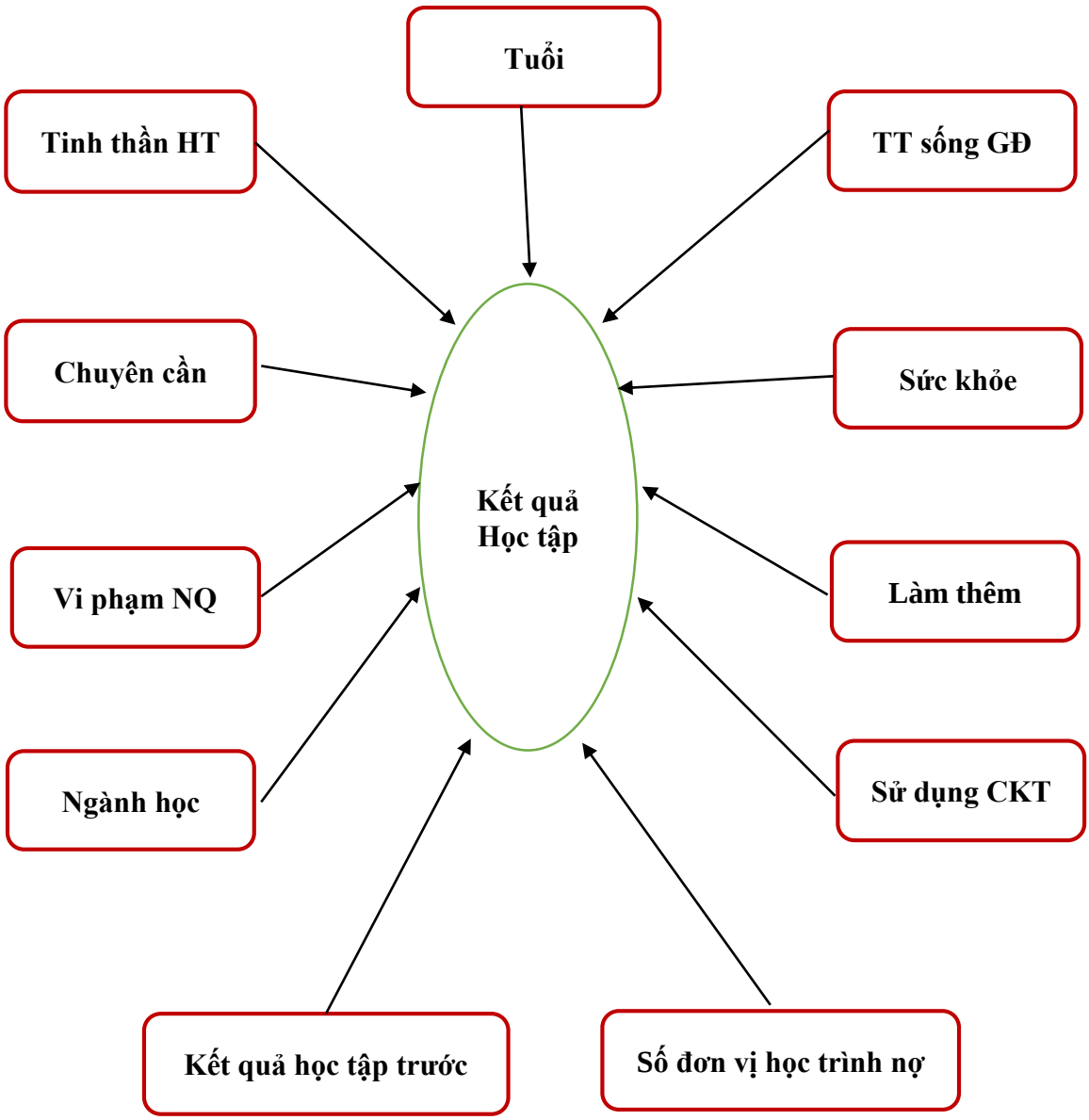
Bảng 9. Tổng hợp các nghiên cứu trước

Thứ tự	Nghiên cứu	Nhân tố ảnh hưởng
1	J. Bainbridge, J. Melitski, A. Zahradnik, E.J. M. Lauría, S. Jayaprakash, and J. Baron "Using Learning Analytics to Predict At-Risk Students in Online Graduate Public Affairs and Administration Education":	- Tính cách - Hành vi
2	P. Cortez and A. Silva. "Using Data Mining to Predict Secondary School Student Performance":	- Điểm học tập - Hoàn cảnh - Gia đình & Xã hội
3	Phạm Hữu Tín và Nguyễn Thúy Huỳnh Loan "Các yếu tố ảnh hưởng đến thái độ học tập của sinh viên trường Đại học Đà Lạt"	- Giảng viên - Phương pháp giảng dạy - Hệ thống cơ sở vật chất - Giáo trình, nội dung môn học - Thực hành, thực tập thực tế - Động cơ học tập - Điều kiện ăn ở, sinh hoạt
4	Võ Thị Tâm " Các yếu tố tác động đến kết quả học tập của sinh viên chính quy trường Đại	- Kiên định trong học tập - Cạnh tranh học tập - Ấn tượng trường học

	học Kinh Tế Thành Phố Hồ Chí Minh”:	- Phương pháp học tập - Động cơ học tập
5	Nguyễn Công Toàn, Trịnh Minh Trí, Huỳnh Văn Hậu Nguyễn Thị Cẩm Hồng và Nguyễn Văn Quân “Các yếu tố ảnh hưởng đến kết quả học tập của sinh viên đại học ngành phát triển nông thôn của trường Đại học Cần Thơ**”:**	- Giới tính - Tài liệu giảng viên - Số buổi nghỉ - Số giờ tự học
6	Dư Thống Nhất và Nguyễn Thị Nụ “Các yếu tố ảnh hưởng đến kết quả học tập môn tâm lý học của sinh viên trường cao đẳng Sư phạm Kiên Giang”:	- Phương pháp học tập - Nhận thức về môn học - Phương tiện dạy học - Thái độ đối với sinh viên - Phương pháp học tập

(Nguồn: Tổng hợp từ các nghiên cứu trước)

Từ bảng 9 đã tổng hợp nghiên cứu trước làm cơ sở đầu tiên, tác giả thấy rằng các yếu tố thông tin học sinh, sinh viên được khai thác từ nghiên cứu trước chưa thể rõ hết các yếu tố tác động trực tiếp đến học sinh, sinh viên mà chỉ nghiên cứu một một vài yếu tố nên tác giả được đề xuất mô hình nghiên cứu như Hình 6.



Hình 6. Mô hình các yếu tố ảnh hưởng kết quả học tập ban đầu

Theo hình 6 tác giả đã xác định các yếu tố ảnh hưởng đến kết quả học tập bao gồm: tuổi học sinh, tình trạng sống với gia đình, sức khỏe học sinh, làm thêm ngoài giờ, sử dụng chất kích thích, tinh thần học tập, chuyên cần, vi phạm nội quy, ngành học, kết quả học tập trước, Số đơn vị học trình nợ . Trên cơ sở đó hình thành các giả thuyết cho việc nghiên cứu đề tài như sau:

Bảng 10.Các giả thuyết nghiên cứu

STT	Giả thuyết	Nội dung
1	**H1**	***Tuổi học sinh càng lớn thì kết quả học tập càng tích cực***
2	**H2**	***Tinh thần học tập càng tốt thì kết quả học tập càng tích cực.***
3	**H3**	***Sức khỏe HS càng tốt thì kết quả học tập càng tích cực.***
4	**H4**	***Kết quả học tập trước của HS càng giỏi thì kết quả học tập càng tích cực.***
5	**H5**	***Số đơn vị học trình nợ càng nhiều thì kết quả học tập càng tiêu cực.***
6	**H6**	***Chuyên cần của HS càng chăm chỉ thì kết quả học tập càng tích cực.***
7	**H7**	***Ngành học khác nhau thì ảnh hưởng kết quả học tập khác nhau.***
8	**H8**	***Vi phạm nội quy ảnh hưởng tiêu cực đến kết quả học tập.***
9	**H9**	***Sống với gia đình ảnh hưởng tích cực đến kết quả học tập hơn sống ở ngoài.***
10	**H10**	***Sử dụng chất kích thích ảnh hưởng tiêu cực đến kết quả học tập.***
11	**H11**	***Làm thêm ảnh hưởng tiêu cực đến kết quả học tập.***

3.2. Tìm hiểu dữ liệu

3.2.1 Nguồn thông tin

Dữ liệu thứ cấp: thông tin về giáo dục TCCN tại TPHCM, đào tạo tại trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn được thu thập từ báo chí, Internet, số

liệu thống kê, các báo cáo nghiên cứu trước, giáo trình, báo cáo luận văn liên quan đến vấn đề nghiên cứu.

Dữ liệu sơ cấp: thu thập thu thập dữ liệu từ phòng đào tạo, công tác học sinh – sinh viên với đối tượng cần lấy thông tin là những học sinh TCCN năm 2, năm 3 trong trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn về các yếu tố liên quan đến kết quả học tập, đặc trưng nhân khẩu, đặc điểm tâm lý, yếu tố gia đình - xã hội và yếu tố trường học.

3.2.2. Nghiên cứu định tính

3.2.2.1. Mục đích nghiên cứu

Nghiên cứu định tính là nghiên cứu trong đó thông tin thu thập dạng định tính.

Thông tin định tính là thông tin chính nó không đo lường bằng số lượng.

Mục tiêu của nghiên cứu định tính trong đề tài này nhằm:

(1) Xác định các yếu tố ảnh hưởng đến kết quả của học sinh TCCN hệ THCS và đề xuất mô hình nghiên cứu.

(2) Tìm hiểu xem người được lấy thông tin, nội dung thông tin cần lấy và vị trí cần lấy thông tin.

(3) Dựa vào kết quả nghiên cứu định tính để thiết kế bảng câu hỏi dùng trong nghiên cứu tiếp theo.

3.2.2.2. Cách thực hiện nghiên cứu

Trong luận văn, tác giả chọn phương pháp thảo luận nhóm, dựa vào bảng thảo luận được thiết kế sẵn. Phương pháp thảo luận nhóm được xem là thuận tiện giúp cho người phỏng vấn có thể trình bày ý kiến của mình một cách rõ ràng và cụ thể. Số người được phỏng vấn là 10 người, đã là giáo viên chủ nhiệm các lớp TCCN hệ THCS năm 2, năm 3 của trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn bao gồm 02 GVCN khoa Công Nghệ Thông Tin, 02 GVCN Điện – Điện tử, 02 GVCN khoa Du Lịch, 02 GVCN khoa Cơ Khí Động Lực, 02 GVCN khoa Cơ Khí Xây Dựng và được lấy mẫu theo phương pháp định mức. Nội dung lấy thông tin nhằm nhận diện các yếu tố đặc trưng cấu tạo nên các yếu tố ảnh hưởng đến kết quả của học sinh.

Thông tin cần lấy bật lên những yếu tố ảnh hưởng đến kết quả học tập của học sinh TCCN hệ THCS trường Trung cấp Kỹ thuật & Nghiệp Vụ Nam Sài Gòn. Những

thông tin này được phân loại để xác định các yếu tố ảnh hưởng đến kết quả của học sinh TCCN hệ THCS trường Trung cấp Kỹ thuật & Nghiệp Vụ Nam Sài Gòn.

Từ nghiên cứu định tính, sẽ tiến hành xác định mô hình nghiên cứu. Đồng thời, nghiên cứu định tính này cũng nhằm kiểm tra các mức độ rõ ràng của từ ngữ và khả năng hiểu các phát biểu cũng như tính trùng lắp các phát biểu thang đo. Kết quả của nghiên cứu định tính là thiết lập thang đo dành cho nghiên cứu.

3.2.2.3. Thang đo trong nghiên cứu định tính

Các yếu tố tác động đến kết quả học tập của học sinh đa dạng, phong phú. Do đó, đề tài chỉ chọn các biến tương ứng với phạm vi, lĩnh vực và mục đích phù hợp với tình trạng học sinh TCCN hệ THCS tại trường Trung cấp Kỹ thuật và Nghiệp vụ Nam Sài Gòn như sau: tuổi học sinh, tình trạng sống với gia đình, sức khỏe học sinh, làm thêm ngoài giờ, sử dụng chất kích thích, tinh thần học tập, chuyên cần, vi phạm nội quy, ngành học, kết quả học tập trước, số đơn vị học trình nợ

Dựa vào các nghiên cứu trước đây và mô hình nghiên cứu trên, tác giả đề xuất thang đo trong nghiên cứu định tính như bảng 11

Bảng 11. Thang đo trong nghiên cứu định tính

STT	Thang đo	Ký hiệu
1	**Độ tuổi của học sinh**	**TSV**
2	**Tinh thần học tập**	**TTHT**
3	**Sức khỏe**	**SK**
4	**Kết quả học tập trước của học sinh**	**KQHT**
5	**Số đơn vị học trình nợ**	**STCN**
6	**Chuyên cần của học sinh**	**THLL**
7	**Ngành học**	**NGH**
8	**Vi phạm nội quy**	**VPNQ**
9	**Tình trạng sống với gia đình**	**TTSGD**

10	**Sử dụng chất kích thích**	**SDCKT**
11	**Làm thêm**	**LNG**
12	**Kết quả**	**KQ**

3.2.2.4. Kết quả nghiên cứu định tính

Nghiên cứu đã khảo sát ý kiến của 10 GVCN đã chủ nhiệm năm 2, năm 3 tại trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn. Dựa vào thang đo trong mô hình nghiên cứu đề xuất để lấy bảng thông tin của các đối tượng phỏng vấn kiểm tra sự chính xác từ ngữ mô hình, nội dung chuyển tải và sự phù hợp với phạm vi nghiên cứu.

Kết quả khảo sát định tính cho thấy các yếu tố dự kiến được đưa vào nghiên cứu là phù hợp. Tuy nhiên có một số yếu tố cần được điều chỉnh sửa cách thể hiện ngôn ngữ trình bày cho dễ hiểu hơn trước khi đưa vào khai thác chính thức. Kết quả thảo luận sẽ tổng hợp là cơ sở điều chỉnh và bổ sung vào mô hình.

3.2.3. Nghiên cứu định lượng

Nghiên cứu định lượng là nghiên cứu trong đó thu thập thông tin dưới dạng định lượng nhằm giúp ta có thể đo lường bằng số lượng. Mục tiêu nhằm đo lường các yếu tố đã nhận diện trong nghiên cứu định tính.

3.2.3.1. Đối tượng khảo sát nghiên cứu

Đối tượng nghiên cứu tập trung vào những học sinh TCCN hệ THCS hiện đang theo học năm 2, năm 3 tại trường Trung cấp Kỹ thuật và Nghiệp vụ Nam Sài Gòn. Mẫu được chọn nghiên cứu theo phương pháp định mức, ta dựa vào các đặc tính kiểm soát để chọn phần tử cho mẫu theo các thuộc tính kiểm soát này.

Khai thác dữ liệu được tiến hành tại trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn, tác giả lấy thông tin mẫu trực tiếp từ phòng Đào tạo và công tác học sinh sinh viên của trường và sổ chủ nhiệm của giáo viên. Các mẫu đạt yêu cầu phải đầy đủ thông tin khai thác và không bỏ sót thông tin nào trong mẫu dữ liệu.

3.2.3.2. Xác định cỡ mẫu nghiên cứu

Mô hình nghiên cứu cắt ngang là mô hình thường có mục tiêu ước tính một tỉ lệ hiện hành của kết quả ngay tại thời điểm hiện hành. Tác giả muốn ước tính mẫu cho một nghiên cứu mục tiêu chính xác là tỉ lệ rớt của học sinh TCCN hệ THCS. Tỉ lệ rớt của

học sinh TCCN hệ THCS tại trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn là 15% tức là P = 0.15. Tác giả nghiên cứu muốn ước tính số đối tượng cần thiết để ước tính tỉ lệ, và chấp nhận xác suất là 95% với sai số e = (89%-81%)/4 = 0.02 hay sai số 2%.

Số mẫu có thể tính cho khảo sát như sau:

$$N=\left(\frac{Z_{\alpha/2}}{e}\right)^2 p\ (1-p)$$

Với alpha = 0.05; Z = 1.96; P = 0.15; e = 0.02 ➔ N = 606

Mô hình nghiên cứu dự kiến có khoảng 11 biến, như vậy kích thước mẫu tối thiểu trong nghiên cứu là 606 mẫu. Để thuận tiện cho phân tích đề tài tác giả chọn số mẫu là 640 mẫu.

Theo số liệu hiệu suất đào tạo của 05 khoa từ Phòng Quản lý đào tạo của trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn thì hiệu suất từng ngành đào tạo như bảng 12.

Bảng 12.Hiệu suất đào tạo theo từng ngành của trường

Ngành	**Hiệu suất đào tạo**						**Hiệu suất trung bình**	
	2013		*2014*		*2015*			
	Năm 2	*Năm 3*	*Năm 2*	*Năm 3*	*Năm 2*	*Năm 3*	*Năm 2*	*Năm 3*
Công nghệ thông tin	78%	85%	85%	92%	89%	96%	84%	91%
Điện – Điện tử	80%	82%	85%	87%	90%	92%	85%	87%
Cơ khí xây dựng	82%	88%	83%	88%	84%	94%	83%	90%
Cơ khí động lực	85%	82%	89%	84%	93%	89%	89%	85%
Du lịch	72%	75%	76%	80%	80%	85%	76%	80%
Hiệu suất đào tạo							81%	89%
							85%	

(Nguồn: Phòng Quản lý đào tạo trường trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn, 2015)

Bảng 13. Tỉ lệ tuyển sinh theo từng ngành

Ngành	**Tỉ lệ đào tạo của các khoa**	
	Năm 2	*Năm 3*
Công nghệ thông tin	23%	23%
Điện – Điện tử	18%	18%
Cơ khí xây dựng	23%	23%
Cơ khí động lực	19%	20%
Du lịch	17%	16%

(Nguồn: Phòng Quản lý đào tạo trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn, 2015)

Dựa trên số liệu trên kết hợp phương pháp định mức với 3 thuộc tính kiểm soát gồm kết quả, ngành, năm và kích thước mẫu n = 640. Phân bố của đối tượng nghiên cứu như sau: theo ngành và năm, 54,5% năm 2 trong đó có 23% khoa Công nghệ thông tin, 18% Khoa Điện – Điện Tử, 23% Khoa Cơ Xây Dựng, 19% Khoa Cơ Khí Động,17% Khoa Du Lịch và 45,5% năm 3 trong đó có 23% khoa Công nghệ thông tin, 18% Khoa Điện – Điện Tử, 23% Khoa Cơ Xây Dựng, 20% Khoa Cơ Khí Động,16% Khoa Du Lịch. Theo kết quả, có 85% đậu và 15% rớt. Tác giả chọn các phần tử cho mẫu theo 20 kết hợp ($5_{\text{ngành}} \times 2_{\text{năm}} \times 2_{\text{kết quả}}$).

Bảng 14. Chọn mẫu định mức: ngành, năm và kết quả

Kết hợp	Ngành	Năm	Kết quả	Tỉ lệ kết hợp	Tỉ lệ	Số phần tử
1	Công nghệ thông tin	2	Đậu	23%*54.5%*84%	10.53%	67
2	Điện – Điện tử	2	Đậu	18%*54.5%*85%	8.34%	53
3	Cơ khí xây dựng	2	Đậu	23%*54.5%*83%	10.40%	67
4	Cơ khí động lực	2	Đậu	19%*54.5%*89%	9.22%	59
5	Du lịch	2	Đậu	17%*54.5%*76%	7.04%	45

Kết hợp	Ngành	Năm	Kết quả	Tỉ lệ kết hợp	Tỉ lệ	Số phần tử
6	Công nghệ thông tin	2	Rớt	23%*54.5%*16%	2.01%	13
7	Điện – Điện tử	2	Rớt	18%*54.5%*15%	1.47%	9
8	Cơ khí xây dựng	2	Rớt	23%*54.5%*17%	2.13%	14
9	Cơ khí động lực	2	Rớt	19%*54.5%*11%	1.14%	7
10	Du lịch	2	Rớt	17%*54.5%*24%	2.22%	14
11	Công nghệ thông tin	3	Đậu	23%*45.5%*91%	9.52%	61
12	Điện – Điện tử	3	Đậu	18%*45.5%*87%	7.13%	46
13	Cơ khí xây dựng	3	Đậu	23%*45.5%*90%	9.42%	60
14	Cơ khí động lực	3	Đậu	20%*45.5%*85%	7.74%	50
15	Du lịch	3	Đậu	16%*45.5%*80%	5.82%	37
16	Công nghệ thông tin	3	Rớt	23%*45.5%*9%	0.94%	6
17	Điện – Điện tử	3	Rớt	18%*45.5%*13%	1.06%	7
18	Cơ khí xây dựng	3	Rớt	23%*45.5%*10%	1.05%	7
19	Cơ khí động lực	3	Rớt	20%*45.5%*15%	1.37%	9
20	Du lịch	3	Rớt	16%*45.5%*20%	1.46%	9
Tổng mẫu					100%	640

Trong nghiên cứu này, tác giả chọn mẫu nghiên cứu là 640 mẫu. Mẫu được chọn đại diện 5 ngành đào tạo gồm: Công nghệ thông tin là 147 mẫu; Điện – Điện tử là 115 mẫu; Cơ khí xây dựng là 148 mẫu; Cơ khí động lực là 125 mẫu; Du lịch là 105 mẫu, Mẫu được chọn đại diện theo năm: Năm 2 là 348 mẫu; Năm 3 là 292 mẫu và mẫu chọn đại diện theo kết quả: Đậu là 545 mẫu; Rớt là 95 mẫu.

3.2.3.3. Bảng lấy thông tin và thang đo chính thức

Các thông tin được được lấy là thông tin đóng. Dựa trên cơ sở lý thuyết từ những nghiên cứu trước kết hợp với nghiên cứu định tính bằng phương pháp thảo luận nhóm, tác giả đã xây dựng thang đo chính thức.

Có 12 khái niệm được sử dụng trong nghiên cứu này gồm tuổi học sinh, tình trạng sống với gia đình, sức khỏe học sinh, làm thêm ngoài giờ, sử dụng chất kích thích, tinh thần học tập, chuyên cần, vi phạm nội quy, kết quả học tập trước, Số đơn vị học trình nợ, ngành, kết quả. Các thang đo này được sử dụng rộng rãi ở các trường học vì thế, nghiên cứu này chỉ áp dụng cho khối trường học ở Việt Nam.

❖ Thang đo tuổi học sinh được đo lường dựa vào tuổi của học sinh có giá trị từ 16 – 21

❖ Thang đo tinh thần học tập của học sinh được đo lường dựa trên đánh giá tinh thần, thái độ học tập của học sinh trong quá trình tham gia học. Thang đo tinh thần học tập bao gồm 4 biến quan sát:

1. Tinh thần học tập tốt
2. Tinh thần học tập khá
3. Tinh thần học tập trung bình
4. Tinh thần học tập yếu

❖ Thang đo sức khỏe của học sinh được đo lường dựa trên thông tin thông tin và quan sát sức khỏe của học sinh. Thang đo sức khỏe bao gồm 2 biến quan sát:

1. Sức khỏe tốt
2. Sức khỏe không tốt

❖ Thang đo kết quả học tập trước của học sinh được đo lường dựa trên đánh giá kết quả học tập của học sinh trong quá trình học tập trước đó. Thang đo kết quả học tập trước của học sinh bao gồm 4 biến quan sát:

1. Học sinh giỏi
2. Học sinh khá
3. Học sinh trung bình khá
4. Học sinh trung bình

❖ Thang đo Số đơn vị học trình nợ của học sinh được đo lường dựa vào số đơn vị học trình học sinh nợ trong những học kỳ trước đó và có giá trị từ 0 đến 19.

❖ Thang đo tình hình lên lớp của học sinh được đo lường dựa trên kết quả lên lớp từng ngày của học sinh học tập của học sinh. Thang đo tình hình lên lớp của học sinh bao gồm 4 biến quan sát:

1. Chăm chỉ
2. Thỉnh thoảng nghỉ
3. Nghỉ nhiều ngày
4. Nghỉ thường xuyên

❖ Thang đo ngành học được đo lường dựa trên quan sát khối ngành các em theo học. Thang đo ngành học bao gồm 5 biến quan sát:

1. Công nghệ thông tin
2. Điện – Điện tử
3. Cơ Khí xây dựng
4. Cơ khí động lực
5. Du lịch

❖ Thang đo vi phạm nội quy được đo lường dựa trên kết quả vi phạm những quy định của nhà trường. Thang đo vi phạm nội quy bao gồm 2 biến quan sát:

1. Không vi phạm nội quy
2. Vi phạm nội quy

❖ Thang đo tình trạng sống với gia đình được đo lường dựa trên việc sống chung với gia đình hay không. Thang đo tình trạng sống với gia đình bao gồm 2 biến quan sát:

1. Sống với gia đình
2. Ở ngoài (ở riêng hoặc ở trọ)

❖ Thang đo sử dụng chất kích thích được đo lường dựa trên việc sử dụng rượu, bia, thuốc lá hoặc các chất kích thích khác. Thang đo sử dụng chất kích thích bao gồm 2 biến quan sát:

1. Không sử dụng chất kích thích
2. Có sử dụng chất kích thích

❖ Thang đo làm ngoài giờ được đo lường dựa trên việc làm thêm của học sinh ngoài thời gian lên lớp. Thang đo làm ngoài giờ bao gồm 2 biến quan sát:

1. Không làm thêm
2. Có làm thêm

Bảng 15.Thang đo trong bảng thông tin nghiên cứu định lượng

STT	Tên biến	Ký hiệu biến
1	Độ tuổi học sinh: …….	TSV
2	Tinh thần học tập: ☐ Tốt ☐ Khá ☐ TB ☐ Yếu	TTHT
3	Sức khỏe: ☐ Tốt ☐ Không tốt	SK
4	Kết quả học tập trước của học sinh trước: ☐ Tốt ☐ Khá ☐ Trung bình khá ☐ Trung bình	KQHT
5	Số chỉ nợ: …….	STCN
6	Tình hình lên lớp của học sinh ☐ Chăm chỉ ☐ Thỉnh thoảng nghỉ ☐ Nghỉ nhiều ngày ☐ Nghỉ thường xuyên	THLL
7	Ngành học: ☐ Công nghệ thông tin ☐ Điện – Điện tử ☐ Cơ khí xây dựng ☐ Cơ khí động lực ☐ Du lịch	NGH
8	Vi phạm nội quy nhà trường: ☐ Không vi phạm	VPNQ

STT	Tên biến	Ký hiệu biến
	☐ Vi phạm	
9	Tình trạng sống với gia đình: ☐ Sống với gia đình ☐ Ở ngoài (ở riêng hoặc trọ)	TTGD
10	Sử dụng chất kích thích: ☐ Có ☐ Không	SDCKT
11	Làm ngoài giờ: ☐ Có ☐ Không	LNG
12	Kết quả: ☐ Đậu ☐ Rớt	KQ

3.2.4. Phương pháp xử lý và phân tích dữ liệu

Sau khi thu thập thông tin thông qua bảng mẫu, các bảng mẫu được xem xét và loại đi những bảng không phù hợp với yêu cầu. Sau đó dữ liệu sẽ được mã hóa và sử lý bằng R. Dữ liệu trong nghiên cứu này sẽ được đưa vào phân tích chính thức thông qua việc áp dụng các công cụ trong phần mềm R.

3.2.4.1. Kiểm định độ phù hợp của mô hình

Hồi quy Logistic cũng đòi hỏi ta phải đánh giá độ phụ hợp của mô hình. Đo lường độ phù hợp. Đo lường độ phù hợp tổng quát của mô hình hồi quy Logistic được dựa trền chỉ tiêu "Độ sai lệch" (Deviance hay có cách gọi khác là Residual deviance), chỉ tiêu "Độ sai lệch" này còn nhỏ càng thể hiện độ phù hợp cao. Giá trị nhỏ nhất của "Độ sai lệch" là 0 khi đó mô hình có độ phù hợp là hoàn hảo.

Với G^2 (-2LL) là độ sai lệch ta có:

$$G^2 = -2 \sum_{i=0}^{n} [\, Y_i \log(p_i)] + (1 - p_i) \log(1 - p_i)$$

Trong đó:

Y_i là biến quan sát (0,1)

p_i là xác suất tiên lượng

Luận văn cũng còn có thể xác định được mô hình dự đoán tốt trên cơ sở chỉ số AIC (Aikake Information Criterion). Giá trị AIC càng nhỏ thì càng thể hiện độ phù hợp của mô hình.

$$AIC = -2LL + 2[(k-1)+p]$$

Trong đó:

K là số bậc/ giá trị của Y (thường là 2)

P là số biến tiên lượng trong mô hình

-2LL là Độ sai lệch (G^2)

3.2.4.2. Kiểm định ý nghĩa của các hệ số

Hồi quy Logistic cũng đòi hỏi kiểm định giả thuyết hệ số hồi quy khác không. Nếu hệ số hồi quy B_0 và B_1 đều bằng 0 thì tỷ lệ chênh lệch giữa các xác xuất sẽ bằng 1, tức là xác xuất để sự kiện xảy ra như nhau, lúc đó mô hình vô dụng trong dự đoán. Do đó, trong hồi quy Logistic luận văn sẽ căn cứ trên số thống kê likelihood-radio hay số thống kê Wald để kiểm định giả thuyết H_0: $\beta_k=0$. Cách thức sử dụng với mức ý nghĩa $p < 0.05$.

- Likelihood-radio (LR) là phương pháp loại trừ dần kiểm tra loại biến căn cứ trên xác xuất của thống kê, Likelihood-radio dựa trên những ước lượng khả năng xảy ra tối đa.
- Wald là phương pháp loại trừ dần kiểm tra loại biến căn cứ trên xác xuất của thống kê Wald.

3.2.4.3. Kiểm định độ phù hợp tổng quát.

Ở hồi quy Logistic, tổ hợp liên hệ tuyến tính của toàn bộ các hệ số trong mô hình ngoại trừ hằng số được kiểm định xem có thức sự có ý nghĩa hay không. Với hồi quy Logistic ta dùng kiểm định Chi – bình phương để kiểm định giả thuyết H_0: $\beta_1= \beta_2= \beta_3=….. \beta_k=0$. Luận văn sẽ căn cứ vào mức ý nghĩa quan sát mà R đưa ra trong bảng "Model Likelihood Radio Test" để bác bỏ hay chấp nhận H_0.

3.3. Chuẩn bị dữ liệu

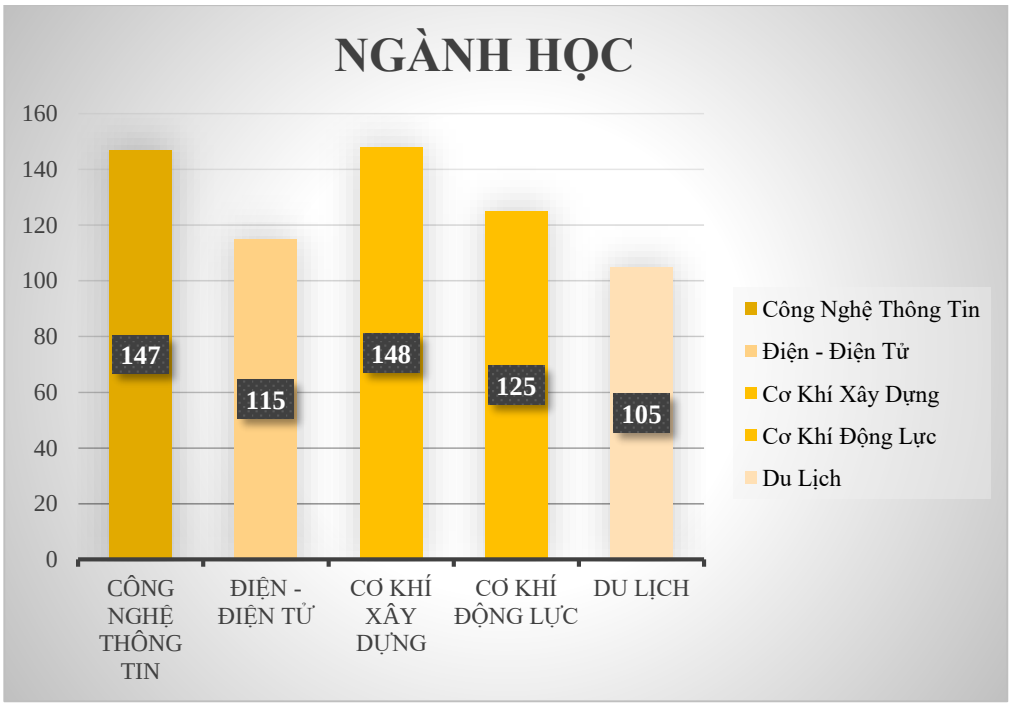
3.3.1. Thống kê mô tả mẫu nghiên cứu

Kết quả nghiên cứu đã thu về 800 mẫu. Sau quá trình kiểm tra kết quả thu được 800 bảng câu hỏi hợp lệ, trong đó: dữ liệu nghiên cứu là 640 mẫu; dữ liệu test là 160 mẫu.

Đối tượng lấy thông tin là học sinh TCCN hệ THCS đã học năm 2, năm 3 tại trường Trung cấp Kỹ thuật Và Nghiệp vụ Nam Sài Gòn. Tác giả chọn mẫu theo phương pháp định mức, đại diện với 3 thuộc tính kiểm soát gồm kết quả, ngành học và năm học.

3.3.1.1. Thống kê mô tả định tính

3.3.1.1.1. Ngành học

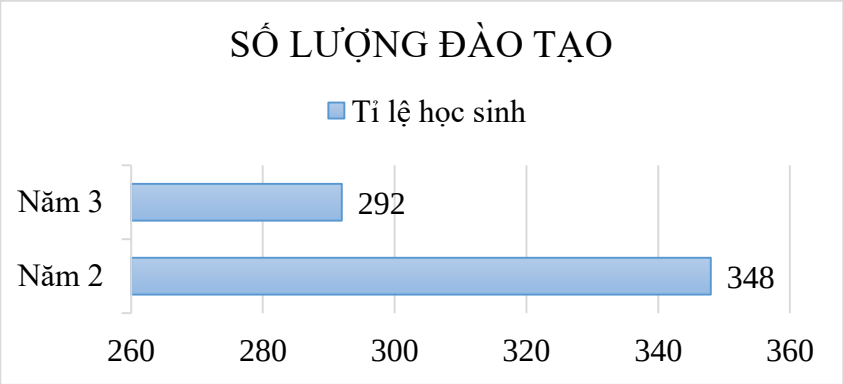


Biểu đồ 2. Ngành học của mẫu

Các đối tượng khảo sát được chia thành 05 nhóm ngành: Công nghệ thông tin, Điện – Điện tử, Cơ khí xây dựng, Cơ khí động lực và Du lịch. Theo biểu đồ 3, Học sinh thuộc nhóm ngành Công nghệ thông tin và Cơ khí xây dựng chiếm tỷ lệ cao nhất ~ 23%, học sinh thuộc nhóm ngành Cơ khí động lực chiếm 19.5%, học sinh thuộc nhóm ngành Điện – Điện tử chiếm 18% và học sinh thuộc nhóm ngành Du lịch chiếm tỷ lệ thấp nhất chỉ có 16.4%.

3.3.1.1.2. Năm học

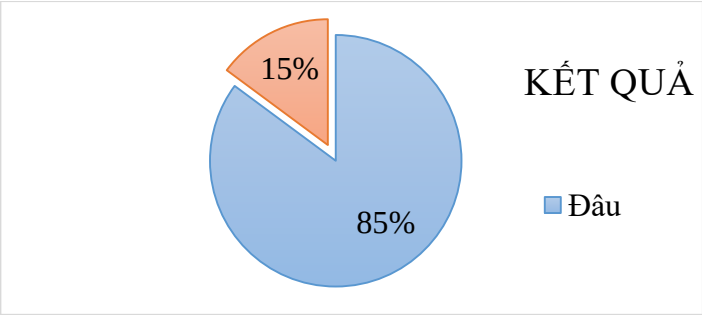
Về đặc điểm theo năm học, đối tượng lấy thông tin là học sinh TCCN hệ THCS đã học năm 2, năm 3. Theo kết quả lấy thông tin, số mẫu được lấy theo năm tỉ lệ: 54.4% số mẫu đã học năm 2, 45.6% số mẫu đã học năm 3.



Biểu đồ 3. Năm theo mẫu

3.3.1.1.3. Kết quả

Theo biểu đồ 4, kết quả lấy thông tin : học sinh đậu chiếm 85% (545 mẫu) và rớt chiếm 15% (95 mẫu).

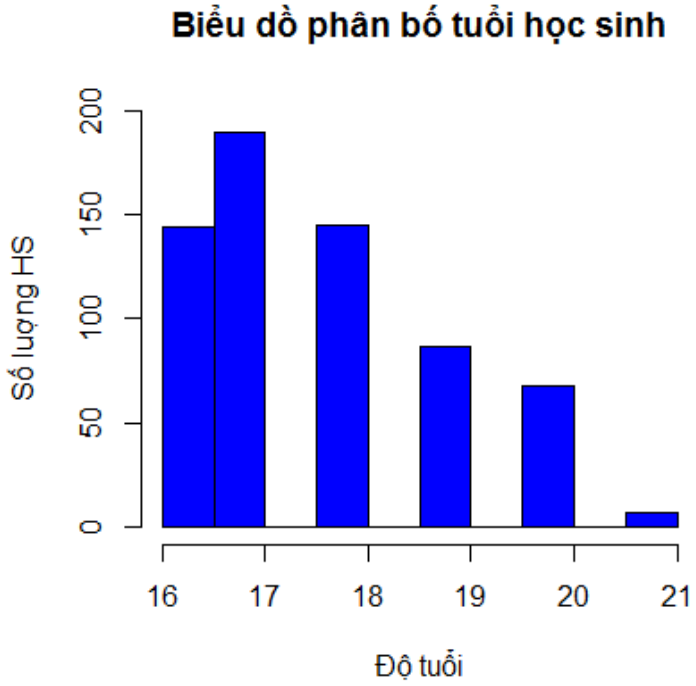


Biểu đồ 4. Kết quả của mẫu

3.3.1.2. Thống kê mô tả định lượng

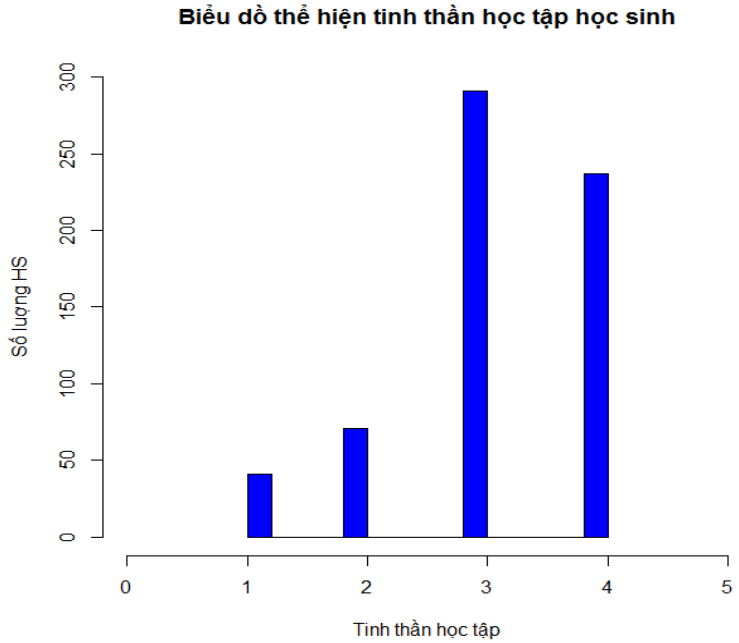
3.3.1.2.1. Thống kê theo tuổi

Trong 640 mẫu lấy thông tin, tuổi của mẫu bao gồm từ 16 đến 21 tuổi với trung bình độ tuổi là 17.63. Trong đó, số lượng học sinh được khảo sát ở độ tuổi 16 là 144 mẫu chiếm tỉ lệ 22.5%, ở độ tuổi 17 là 189 mẫu chiếm tỉ lệ 29.5%, ở độ tuổi 18 là 145 mẫu chiếm tỉ lệ 22.7%, ở độ tuổi 19 tuổi là 87 mẫu chiếm tỉ lệ 13.6%, ở độ tuổi 20 là 68 mẫu chiếm tỉ lệ 10.6% và ở độ tuổi 21 tuổi là 7 mẫu chiếm tỉ lệ 1.1%.



Biểu đồ 5. Thống kê theo tuổi của học sinh

3.3.1.2.2. Thống kê theo tinh thần học tập



Biểu đồ 6.Thống kê tinh thần học tập của học sinh

Đánh giá thang đo tinh thần học tập thông qua 4 mức độ gồm: tốt, khá, trung bình, yếu. Trong 640 mẫu lấy thông tin, kết quả cho thấy tin thần học tập của học sinh TCCN hệ THCS ở mức tương đối khá. Trong đó, 237 học sinh có tinh thần học tập tốt

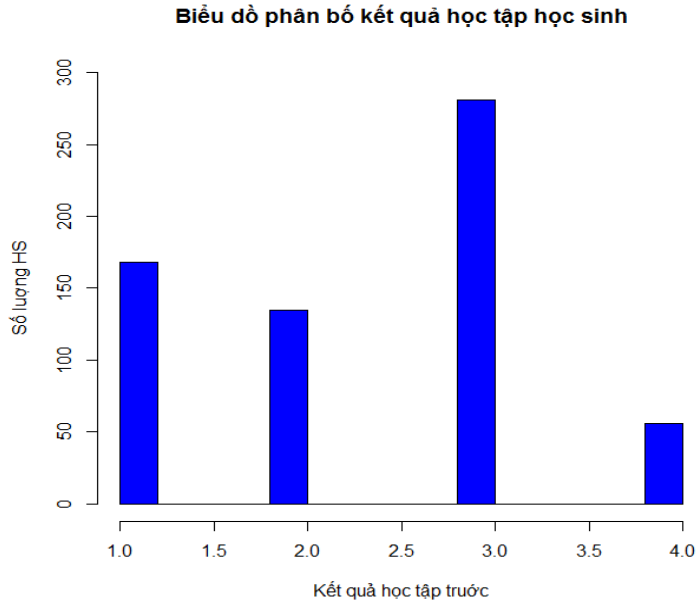
(37%), 291 học sinh có tinh thần học tập khá (45.5%), 71 học sinh có tinh thần học tập trung bình (11.1%) và 41 học sinh có tinh thần học tập yếu chiếm số lượng nhỏ với 6.4%.

3.3.1.2.3. Thống kê theo sức khỏe của học sinh

Biểu đồ 7 .Thống kê theo sức khỏe học sinh

Đánh giá thang đo sức khỏe gồm 2 mức độ gồm: tốt và không tốt. Theo kết quả lấy thông tin học sinh TCCN hệ THCS tại trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn có sức khỏe chưa cao. Trong đó, sức khỏe không tốt chiếm tỉ lệ cao 9.7% với 62 mẫu và 90.3% sức khỏe tốt với 578 mẫu.

3.3.1.2.4. Thống kê các kết quả học tập trước của mẫu khảo sát



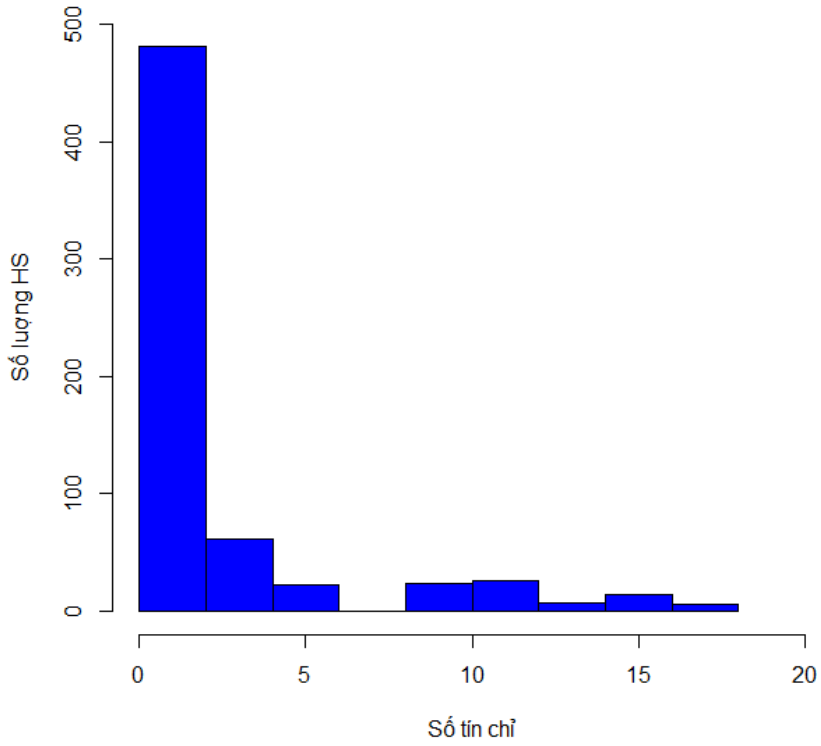
Biểu đồ 8. Thống kế kết quả học tập trước

Thang đo kết quả học tập trước được đánh giá qua 4 cấp độ gồm: trung bình. trung bình khá, khá và giỏi. Kết quả lấy thông tin từ 640 mẫu gồm 168 mẫu trung bình (26.3%), 135 mẫu trung bình khá (21.1%), 281 mẫu khá (43.9%) và 56 mẫu Giỏi (8.7%).

3.3.1.2.5. Thống kê số đơn vị học trình học sinh nợ

Trong 640 mẫu lấy thông tin, Số đơn vị học trình nợ có có giá trị từ 0 đến 19. Trong đó số học sinh không nợ đơn vị học trình chiếm tỉ lệ cao 456 mẫu chiếm 71.3 %, số học sinh nợ từ 2 đến 10 đơn vị học phần là 131 mẫu chiếm 20.3% và học sinh nợ trên 10 đơn vị học phần chiếm 8.7% với 53 mẫu.

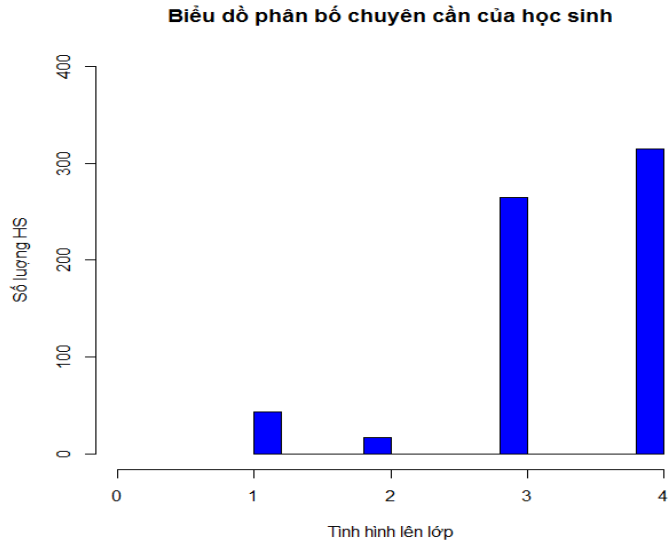
Biểu đồ phân bố số đơn vị học trình nợ



Biểu đồ 9. Thống kê Số đơn vị học trình nợ

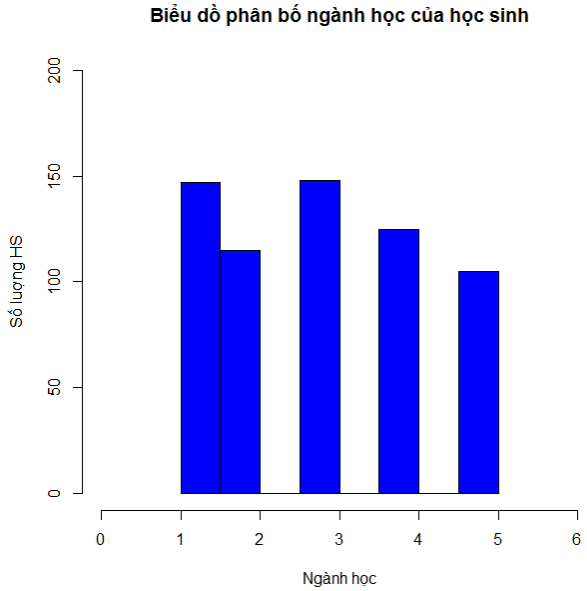
3.3.1.2.6. Thống kê theo tình hình lên lớp của học sinh

Thang đo tình hình lên lớp của học sinh được đánh giá qua 4 mức độ gồm: nghỉ thường xuyên, nghỉ nhiều ngày, thỉnh thoảng nghỉ và chăm chỉ học. Trong bảng lấy thông tin có 6.7% học sinh nghỉ thường xuyên, 2.7% học sinh nghỉ nhiều ngày, 41.4 học sinh thỉnh thoảng nghỉ học và 49.2% học sinh tham gia học tập đầy đủ.



Biểu đồ 10.Thống kê tình hình lên lớp

3.3.1.2.7. Thống kê theo ngành học của học sinh

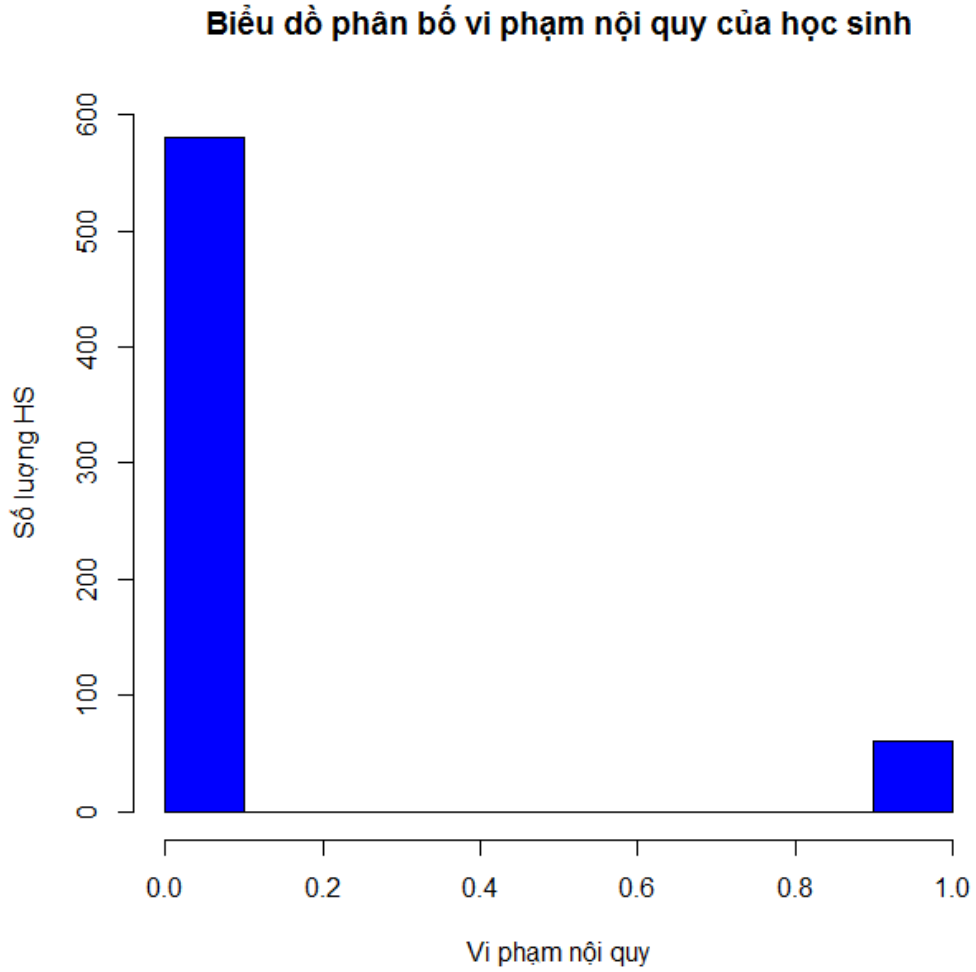


Biểu đồ 11.Thống kê theo ngành học

Các đối tượng khảo sát được chia thành 05 nhóm ngành: Công nghệ thông tin, Điện – Điện tử, Cơ khí xây dựng, Cơ khí động lực và Du lịch. Trong đó Học sinh thuộc nhóm ngành Công nghệ thông tin có 147 mẫu (23%), Điện – Điện tử có 115 mẫu (18%), Cơ khí xây dựng có 148 mẫu (23.1%), Cơ khí động lực có 125 mẫu (19.5%) và Du lịch có ít nhất 105 mẫu (10.5%).

3.3.1.2.8. Thống kê số lượng học sinh vi phạm nội quy

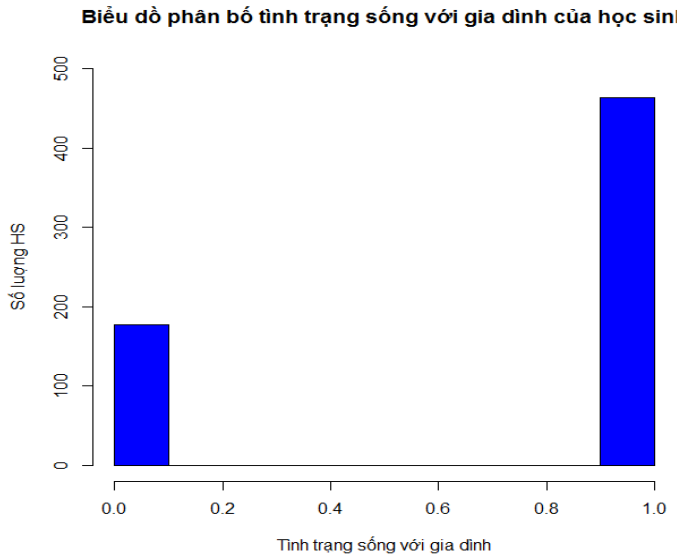
Thang đo vi phạm nội quy của học sinh được đánh giá thông qua 2 mức độ gồm: vi phạm (theo quy đinh nội quy của nhà trường) và không vi pham. Học sinh vi phạm chiếm tỉ lệ thấp 9.4% với 60 mẫu và học sinh không vi phạm 90.6% với 580 mẫu.



Biểu đồ 12. Thống kê tình trạng vi phạm nội quy của học sinh

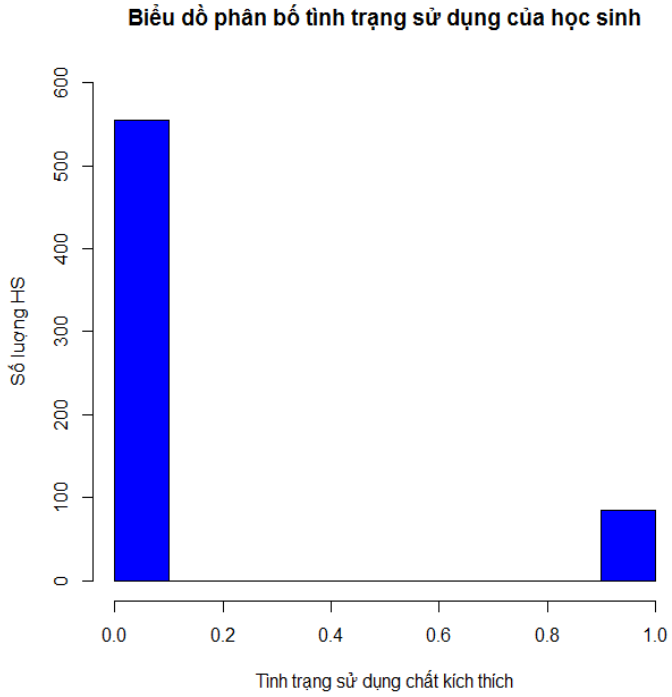
3.3.1.2.9. Thống kê theo tình trạng sống với gia đình

Thang đo tình trạng sống với gia đình bao gồm 2 giá mức độ: sống với gia đình, ở ngoài (ở trọ và không sống cùng bố mẹ). Kết quả lấy thông tin có 177 mẫu học sinh sống ngoài chiếm tỉ lệ 27.7% và 463 mẫu học sinh sống với bố mẹ chiếm 46.3%.



Biểu đồ 13 .Thống kê tình trạng sống với gia đình của học sinh

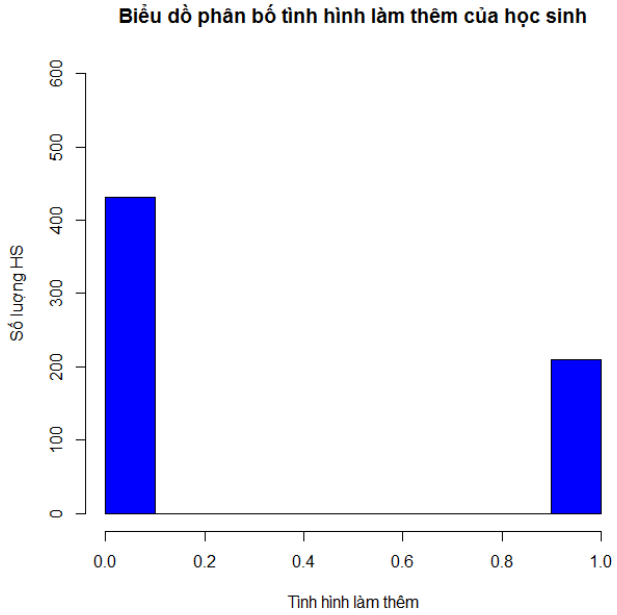
3.3.1.2.10. Thống kê số lượng học sinh sử dụng chất kích thích



Biểu đồ 14. Thống kê tình trạng sử dụng chất kích thích của học sinh

Học sinh sử dụng chất kích thích bao gồm: rượu, bia, …. . Thang đo tình trạng sử dụng chất kích thích đánh giá với 2 mức độ gồm: sử dụng và không sử dụng. Theo kết quả lấy thông tin có 85 mẫu sử dụng chất kích thích (13.3%) và 555 mẫu không sử dụng chất kích thích (86.7%).

3.3.1.2.11. Thống kê số lượng học sinh làm thêm



Biểu đồ 15.Thống kê số lượng học sinh làm thêm

Theo kết quả lấy thông tin về việc làm thêm của học sinh, số lượng học sinh làm thêm chiếm tỉ lệ khá cao 32.7% với 209 mẫu, học sinh không làm thêm chiếm tỉ lệ 67.3% với 431 mẫu.

3.3.1.2.12. Thống kê kết quả học sinh

Thang đo kết quả của học sinh đánh giá qua 2 mức độ gồm đậu và rớt. Kết quả lấy thống tin, số học sinh rớt chiếm tỉ lệ cao 15% với 95 mẫu và số học sinh đậu chiếm tỉ lệ 85% với 545 mẫu.

Biểu đồ 16. Thống kê kết quả học sinh

3.3.2. Phân tích thống kê

3.3.2.1. Đo lường mức độ tập trung

Các biến tiên lượng bao gồm 11 biến độc lập. Kết quả mô tả thống kê được trình bày trong bảng 16 và bảng 17.

Bảng 16. Kết quả đo lường mức độ tập trung biến độc lập

Biến độc lập	**Mã biến**	**Mức trung bình**	**Trung vị**
Độ tuổi của học sinh	TSV	17.64	17
Tinh thần học tập	TTHT	3.24	Khá
Sức khỏe	SK	0.85	Tốt
Kết quả học tập của HS	KQHT	2.35	Khá
Số đơn vị học trình nợ	STCN	2.03	0
Tình hình lên lớp	THLL	3.33	Thỉnh thoảng nghỉ
Ngành học	NGH	-	-
Vi phạm nội quy	VPNQ	0.09	Không vi phạm
Tình trạng sống với gia đình	TTSGD	0.72	ở với gia đình
Sử dụng chất kích thích	SDCKT	0.13	Không sử dụng
Làm thêm	LNG	0.33	Không làm thêm

Bảng 17. Kết quả đo lường mức độ tập trung biến phụ thuộc

Biến phụ thuộc	**Mã biến**	**Mức trung bình**	**Trung vị**
Kết quả	KQ	0.85	Đậu

3.3.2.2. Đo lường mức độ phân tán

Bảng 18. Kết quả đo lường mức độ phân tán biến độc lập

Biến độc lập	**Mã biến**	**Độ lệch chuẩn**	**Sai số chuẩn**	**Khoảng biến thiên**
Độ tuổi của học sinh	TSV	1.31	0.05	16..21
Tinh thần học tập	TTHT	0.85	0.03	1..4
Sức khỏe	SK	0.36	0.01	0..1
Kết quả học tập của HS	KQHT	0.96	0.04	1..4

Số đơn vị học trình nợ	STCN	4.14	0.16	0..19
Tình hình lên lớp	THLL	0.82	0.03	1..4
Ngành học	NGH	-	-	1..5
Vi phạm nội quy	VPNQ	0.29	0.01	0..1
Tình trạng sống với gia đình	TTSGD	0.45	0.02	0..1
Sử dụng chất kích thích	SDCKT	0.34	0.01	0..1
Làm thêm	LNG	0.47	0.02	0..1

Bảng 19. Kết quả đo lường mức độ phân tán biến phụ thuộc

Biến độc lập	**Mã biến**	**Độ lệch chuẩn**	**Sai số chuẩn**	**Khoảng biến thiên**
Độ tuổi của học sinh	TSV	0.36	0.01	0..1

3.3.2.3. Kỳ vọng

Bảng 20. Kỳ vọng của biến ảnh hưởng kết quả học tập

Biến độc lập	**Mã biến**	**Phương sai**	**Kỳ vọng**
Độ tuổi của học sinh	TSV	1.72	+
Tinh thần học tập	TTHT	0.72	+
Sức khỏe	SK	0.13	+
Kết quả học tập của HS	KQHT	0.92	+
Số đơn vị học trình nợ	STCN	17.11	-
Tình hình lên lớp	THLL	0.67	+
Ngành học	NGH	…	…
Vi phạm nội quy	VPNQ	0.085	-
Tình trạng sống với gia đình	TTSGD	0.20	+
Sử dụng chất kích thích	SDCKT	0.11	-
Làm thêm	LNG	0.22	-

- Giá trị kỳ vọng thang đo tuổi học sinh mang giá trị dương đều đó cho thấy tuổi càng lớn thì kết quả càng tích cực.
- Giá trị kỳ vọng thang đo tinh thần học tập mang giá trị dương đều đó cho thấy tinh thần học tập càng tốt thì kết quả càng tích cực.
- Giá trị kỳ vọng thang đo sức khỏe mang giá trị dương đều đó cho thấy sức khỏe càng tốt thì kết quả càng tích cực.

- Giá trị kỳ vọng thang đo kết quả học tập của học sinh mang giá trị dương đều đó cho thấy kết quả học tập của học sinh càng tốt thì kết quả đạt càng cao.
- Giá trị kỳ vọng thang đo số đơn vị học trình nợ mang giá trị âm đều đó cho thấy Số đơn vị học trình nợ càng lớn thì nguy cơ rớt càng cao.
- Giá trị kỳ vọng thang đo thời gian lên lớp mang giá trị dương đều đó cho thấy chuyên cần của học sinh càng lên lớp chăm chỉ thì kết quả đạt càng cao.
- Giá trị kỳ vọng thang đo vi phạm nội quy mang giá trị âm đều đó cho thấy vi phạm nội quy càng nhiều thì nguy cơ rớt càng cao.
- Giá trị kỳ vọng thang đo tình trạng sống với gia đình mang giá trị dương đều đó cho thấy ở với gia đình thì kết quả càng tích cực hơn.
- Giá trị kỳ vọng thang đo sử dụng chất kích thích mang giá trị âm đều đó cho thấy sử dụng chất kích thích càng nhiều thì nguy cơ rớt càng cao.
- Giá trị kỳ vọng thang đo làm thêm mang giá trị âm đều đó cho ta thấy việc làm thêm ảnh hưởng tiêu cực đến kết quả học sinh.

3.3.3. Kiểm định mô hình và ý nghĩa hệ số

Việc xác định các yếu tố ảnh hưởng đến quả học tập là căn cứ quan trọng để nâng cao chất lượng đào tạo tại trường trung cấp Kỹ thuật & Nghiệp Vụ Nam Sài Gòn. Sau khi tiến hành phân tích thống kê của các biến, tác giả tiến hành phân tích hồi quy Logistic với 11 biến độc lập và 1 biến phụ thuộc. Phân tích 11 biến độc lập gồm TSV, TTHT, KQHT, SK, STCN, THLL, NGH, VPNQ, TTSGD, SDCKT, LNG và 1 biến phụ thuộc KQ bằng việc áp dụng phương pháp hồi quy Logistic đưa vào một lượt ta thu được kết quả hồi quy như sau:

Bảng 21. Kết quả kiểm định mô hình và ý nghĩa hệ số

Biến độc lập	crude OR (95%CI)	adj. OR (95%CI)	P(Wald's test)	P(LR-test)
TSV	1.22 (1.02,1.46)	26.46 (4.86,144.13)	0.000	0.000
TTHT	4.66 (3.45,6.28)	2.54 (1.3,4.97)	0.007	0.004
SK	8.72 (4.97,15.3)	10.59 (1.31,85.47)	0.027	0.021
KQHT	7.62 (5.02,11.59)	7.97 (1.18,54)	0.033	0.005

Biến độc lập	crude OR (95%CI)	adj. OR (95%CI)	P(Wald's test)	P(LR-test)
STCN	0.55 (0.49,0.61)	0.63 (0.53,0.76)	0.000	0.000
THLL	4.7 (3.48,6.36)	9.23 (2.77,30.78)	0.000	0.000
VPNQ	0.05 (0.03,0.1)	0 (0,0.09)	0.001	0.000
SDCKT	0.13 (0.08,0.22)	0.02 (0,0.36)	0.007	0.000
LNG	1.34 (0.83,2.18)	0.05 (0,0.47)	0.009	0.003
NGH	0.89 (0.76,1.04)	1.56 (0.84,2.91)	0.157	0.155
TTSGD	1.11 (0.69,1.79)	0.03 (0,6.01)	0.196	0.11
Log-likelihood = -41.1816 No. of observations = 640 AIC value = 106.3633 Null deviance: 537.586 on 639 degrees of freedom Residual deviance: 82.363 on 628 degrees of freedom				

3.3.3.1. Kiểm định độ phù hợp của mô hình

Hồi quy Logistic cũng đòi hỏi ta phải đánh giá độ phù hợp của mô hình. Đo lường độ phù hợp tổng quát của mô hình hồi quy Logistic được dựa trên chỉ tiêu "Độ sai lệch", chỉ tiêu "Độ sai lệch" $G^2 = 82.363$ nhỏ hơn rất nhiều so với ban đầu khi chưa có sự ảnh hưởng của các yếu tố . Giá trị của "Độ sai lệch" $G^2 = 82.363$ cho thấy mô hình là phù hợp.

Ngoài ra, Luận văn cũng còn có thể xác định được mô hình dự đoán tốt trên cơ sở chỉ số AIC, Giá trị AIC = 106.363 rất nhỏ.

Với 2 chỉ số $G^2 = 82.363$, AIC = 106.363 đều đó chứng tỏ mô hình hồi quy Logistic với 11 biến độc lập gồm TSV, TTHT, KQHT, SK, STCN, THLL, NGH, VPNQ, TTSGD, SDCKT, LNG và 1 biến phụ thuộc KQ là hoàn toàn phù hợp.

3.3.3.2. Kiểm định ý nghĩa của các hệ số

Theo bảng 21, kết quả phân tích hồi quy Logistic thông số thống kê likelihood-radio hay số thống kê Wald cho thấy có 9 biến có tính thống kê gồm: Tuổi, tinh thần học tập, sức khỏe, kết quả học tập trước, Số đơn vị học trình nợ, tình hình lên lớp, vi phạm nội quy, sử dụng chất kích thích, làm ngoài giờ đều có P(Wald's test) < 0.05 và P(LR-test) < 0.05 và 2 biến không có tính thống kê gồm: ngành học P(Wald's test) = 0.157 và P(LR-test) = 0.155 , tình trạng sống với gia đình P(Wald's test) = 0.196 và P(LR-test) = 0.11 . Do đó ta sẽ loại bỏ 2 biến ra khỏi mô hình phân tích hồi quy Logistic các yếu tố ảnh hưởng đến kết quả học tập của học sinh TCCN hệ THCS..

3.3.3.3. Kiểm định độ phù hợp tổng quát.

Bảng 22. Kiểm định mô hình tổng quát

Mô hình Likelihood Radio Test	Chỉ số phân định	Chỉ số phân định có thứ hạng
LR chi2 450.32	R2 0.889	C 0.993
d.f. 11		
Pr(> chi2) <0.0001		

Tác giả tiến hành phân tích với 11 biến độc lập và 1 biến phụ thuộc bao gồm TSV, TTHT, KQHT, SK, STCN, THLL, NGH, VPNQ, TTSGD, SDCKT, LNG và 1 biến phụ thuộc KQ. Bằng việc áp dụng phương pháp hồi quy đưa vào một lượt ta thu được kết quả hồi quy. Kết quả phân tích hồi quy có hệ số R^2 là 0.889. Điều này có nghĩa là mô hình nghiên cứu có độ thích hợp là 88.9% hay 88.9% độ biến thiên về kết quả học tập của học sinh có thể được giải thích bởi các biến độc lập trong mô hình.

Bảng 22 cũng cho ta thấy LR chi2 = 450.32 và mô hình có giá trị P rất nhỏ là 0.000 thấp hơn 0.05 nên mô hình hồi quy Logistic phù hợp với tập dữ liệu nghiên cứu.

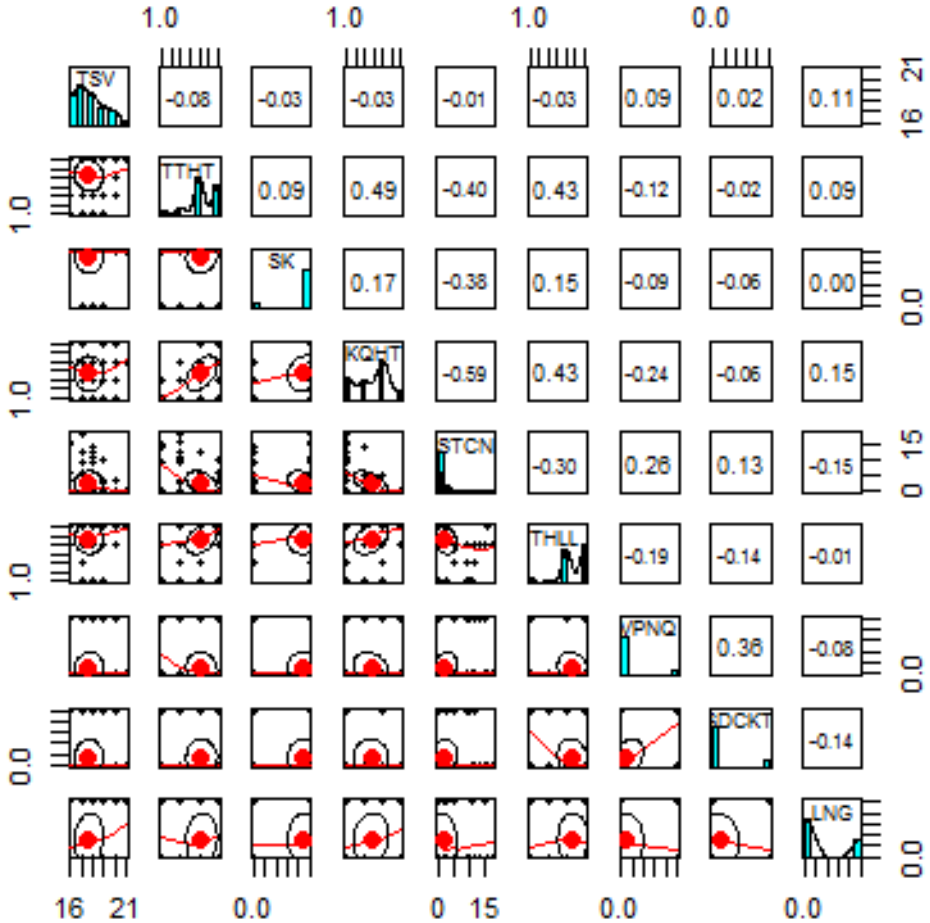
3.3.4. Phân tích tương quan

Theo Hoàng Trọng và Chu Nguyễn Mộng Ngọc (2008), trước khi phân tích hồi quy cần phải xem xét mối quan hệ tương quan tuyến tính giữa các biến độc lập với các biến phụ thuộc và giữa các biến độc lập với nhau.

Hai biến có thể có liên hệ chặt chẽ với nhau nhưng hệ số tương quan vẫn nhỏ gần bằng 0 nếu như dạng của mối liên hệ này không phải là tuyến tính. Hệ số tương quan

tuyến tính chỉ nên được sử dụng được để biểu thị mức độ chặt chẽ của liên hệ tương quan tuyến tính.

Xem xét các mối quan hệ tương quan tuyến tính giữa biến phụ thuộc và từng biến độc lập, cũng như các biến độc lập thông qua phân tích tương quan Kendall. Nếu hệ số tương quan giữa các biến phụ thuộc và biến độc lập lớn chứng tỏ giữa chúng có mối quan hệ với nhau và phân tích hồi quy là phù hợp. Còn các biến độc lập nếu cũng có hệ số tương quan với nhau lớn thì có thể xảy ra hiện tượng tương tác trong mô hình hồi quy đang xét.



Hình 7. Mô hình tương tác

Từ hình 7, kết quả phân tích tương quan Kendall cho thấy có sự tương quan giữa các biến độc lập không cao, do đó không sảy ra mô hình tương tác giữa các biến độc lập.

3.3.5. Kiểm định giả thuyết

Sau quá trình phân tích hồi quy Logistic với 11 biến độc lập bằng phương pháp đưa vào một lần, tác giả thấy được có 9 giả thuyết đưa ra là H1, H2, H3, H4, H5, H6,

H8, H10, H11 được chấp nhận do các yếu tố có tác động đến kết quả học tập của học sinh với mức ý nghĩa thống kê p < 0.05 và có 2 giả thuyết đưa ra là H7, H9 không được chấp nhận do các yếu tố có tác động đến kết quả học tập của học sinh với mức ý nghĩa thống kê p > 0.05

Bảng 23. Kết quả kiểm định giả thuyết

Giả thuyết	Nội dung giả thuyết	Mức ý nghĩa thống kê Wald	Kết quả
H1	Tuổi học sinh càng lớn thì kết quả học tập càng tích cực.	0.000***	Chấp nhận giả thuyết
H2	Tinh thần học tập càng tốt thì kết quả học tập càng tích cực.	0.007**	Chấp nhận giả thuyết
H3	Sức khỏe học sinh càng tốt thì kết quả học tập càng tích cực.	0.027*	Chấp nhận giả thuyết
H4	Kết quả học tập trước của HS càng giỏi thì kết quả học tập càng tích cực.	0.033*	Chấp nhận giả thuyết
H5	Số đơn vị học trình nợ càng nhiều thì kết quả học tập càng tiêu cực.	0.000***	Chấp nhận giả thuyết
H6	Chuyên cần của học sinh càng chăm chỉ thì kết quả học tập càng tích cực.	0.000***	Chấp nhận giả thuyết
H7	Ngành học khác nhau thì ảnh hưởng kết quả học tập khác nhau.	0.157	Không chấp nhận giả thuyết
H8	Vi phạm nội quy ảnh hưởng tiêu cực đến kết quả học tập.	0.001**	Chấp nhận giả thuyết
H9	Sống với gia đình ảnh hưởng tích cực đến kết quả học tập hơn sống ở ngoài.	0.196	Không chấp nhận giả thuyết
H10	Sử dụng chất kích thích ảnh hưởng tiêu cực đến kết quả học tập.	0.007**	Chấp nhận giả thuyết
H11	Làm thêm ảnh hưởng tiêu cực đến kết quả học tập.	0.009**	Chấp nhận giả thuyết

Tuổi học sinh

Giả thuyết H1 phát biểu tuổi học sinh càng lớn thì kết quả học tập càng tích cực. Từ kết quả phân tích hồi quy Logistic ở bảng 23 cho thấy biến tuổi học sinh có ảnh hưởng cùng chiều đến kết quả học tập với giá trị p = 0.000 nhỏ hơn 0.05 nên giả thuyết này được chấp nhận.

Tinh thần học tập

Giả thuyết H2 phát biểu tinh thần học tập càng tốt thì kết quả học tập càng tích cực. Từ kết quả phân tích hồi quy Logistic ở bảng 23 cho thấy biến tinh thần học tập có ảnh hưởng cùng chiều đến kết quả học tập với giá trị p = 0.007 nhỏ hơn 0.05 nên giả thuyết này được chấp nhận.

Sức khỏe của học sinh

Giả thuyết H3 phát biểu sức khỏe học sinh càng tốt thì kết quả học tập càng tích cực. Từ kết quả phân tích hồi quy Logistic ở bảng 23 cho thấy biến sức khỏe có ảnh hưởng đến kết quả học tập vì giá trị p < 0.05 nên giả thuyết này được chấp nhận.

Kết quả học tập trước

Giả thuyết H4 phát biểu kết quả học tập trước của HS càng giỏi thì kết quả học tập càng tích cực. Từ kết quả phân tích hồi quy Logistic ở bảng 23 cho thấy biến kết quả học tập trước có ảnh hưởng cùng chiều đến kết quả học tập với giá trị p = 0.033 nhỏ hơn 0.05 nên giả thuyết này được chấp nhận.

Số đơn vị học trình nợ

Giả thuyết H5 phát biểu Số đơn vị học trình nợ càng nhiều thì kết quả học tập càng tiêu cực. Từ kết quả phân tích hồi quy Logistic ở bảng 23 cho thấy biến Số đơn vị học trình nợ có ảnh hưởng ngược chiều đến kết quả học tập với giá trị p = 0.000 nhỏ hơn 0.05 nên giả thuyết này được chấp nhận.

Tình hình lên lớp

Giả thuyết H6 phát biểu chuyên cần của học sinh càng chăm chỉ thì kết quả học tập càng tích cực. Từ kết quả phân tích hồi quy Logistic ở bảng 23 cho thấy biến tình hình lên lớp có ảnh hưởng cùng chiều đến kết quả học tập với giá trị p = 0.000 nhỏ hơn 0.05 nên giả thuyết này được chấp nhận.

Ngành học

Giả thuyết H7 phát biểu ngành học khác nhau thì ảnh hưởng kết quả học tập khác nhau. Từ kết quả phân tích hồi quy Logistic ở bảng 23 cho thấy biến ngành học không

có ảnh hưởng đến kết quả học tập vì giá trị p > 0.05 nên giả thuyết này không được chấp nhận.

Vi phạm nội quy

Giả thuyết H8 phát biểu vi phạm nội quy ảnh hưởng tiêu cực đến kết quả học tập. Từ kết quả phân tích hồi quy Logistic ở bảng 23 cho thấy biến vi phạm nội quy có ảnh hưởng ngược chiều đến kết quả học tập với giá trị p = 0.001 nhỏ hơn 0.05 nên giả thuyết này được chấp nhận.

Tình trạng sống với gia đình

Giả thuyết H9 phát biểu Sống với gia đình ảnh hưởng tích cực đến kết quả học tập hơn sống ở ngoài. Từ kết quả phân tích hồi quy Logistic ở bảng 23 cho thấy biến tình trạng sống với gia đình không có ảnh hưởng đến kết quả học tập vì giá trị p > 0.05 nên giả thuyết này không được chấp nhận.

Sử dụng chất kích thích

Giả thuyết H10 phát biểu sử dụng chất kích thích ảnh hưởng tiêu cực đến kết quả học tập. Từ kết quả phân tích hồi quy Logistic ở bảng 23 cho thấy biến vi phạm nội quy có ảnh hưởng ngược chiều đến kết quả học tập với giá trị p = 0.007 nhỏ hơn 0.05 nên giả thuyết này được chấp nhận.

Làm thêm

Giả thuyết H8 phát biểu làm thêm ảnh hưởng tiêu cực đến kết quả học tập. Từ kết quả phân tích hồi quy Logistic ở bảng 23 cho thấy biến làm thêm có ảnh hưởng ngược chiều đến kết quả học tập với giá trị p = 0.009 nhỏ hơn 0.05 nên giả thuyết này được chấp nhận.

3.4. Mô hình hồi quy Logistic ảnh hưởng đến kết quả học tập

3.4.1. Phân tích hồi quy Logistic

Trên cơ sở kiểm định giả thuyết các yếu tố ảnh hưởng đến kết quả học sinh, tác giả tiến hành phân tích hồi quy Logistic với 9 biến độc lập và 1 biến phụ thuộc. Phân tích 9 biến độc lập gồm TSV, TTHT, SK, KQHT, STCN, THLL, VPNQ, SDCKT, LNG

và 1 biến phụ thuộc KQ bằng việc áp dụng phương pháp hồi quy Logistic đưa vào một lượt ta thu được kết quả hồi quy như sau:

Bảng 24. Phân tích biến độc lập trong hồi quy Logistic

Biến độc lập	Hệ số Beta	Sai số chuẩn	Ý nghĩa thống kê P
TSV	3.36	0.91	0.0002***
TTHT	0.88	0.36	0.015*
SK	2.11	0.93	0.024*
KQHT	1.68	0.57	0.0036**
STCN	-0.40	0.08	0.000***
THLL	2.08	0.59	0.0004***
VPNQ	-3.69	1.34	0.006**
SDCKT	-2.44	0.82	0.002**
LNG	-2.31	0.86	0.007**

α = -64.81; S.E = 17.62; Wald Z = -3.68,; P = 0.002***

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	640	LR chi2	449.66	R2	0.888	C	0.993
0	95	d.f.	9	g	7.813	Dxy	0.987
1	545	Pr(> chi2)	<0.0001	gr	2472.970	gamma	0.988
max |deriv|	3e-07			gp	0.250	tau-a	0.250
				Brier	0.018		

Null Deviance: 537.6
Residual Deviance: 87.93 AIC: 107.9

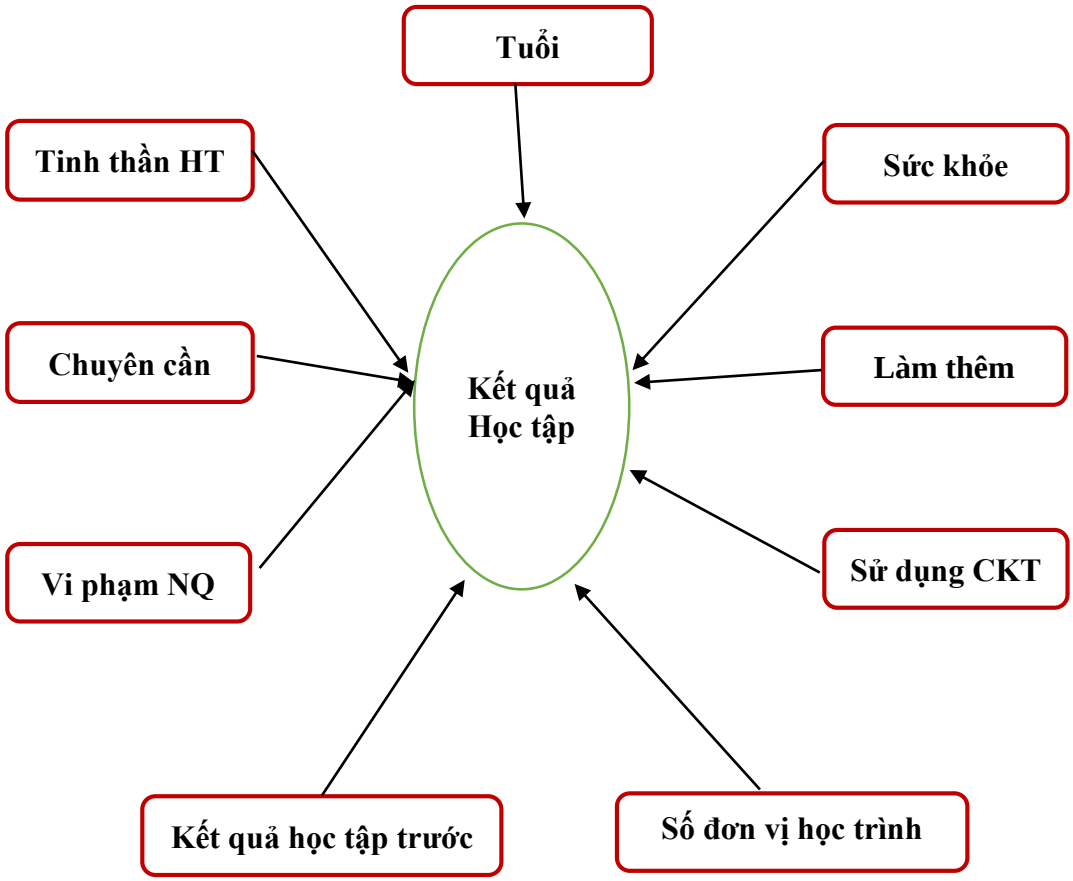
Theo bảng 24 cho thấy 9 biến đều có tính thống kê gồm: Tuổi, tinh thần học tập, sức khỏe, kết quả học tập trước, Số đơn vị học trình nợ, tình hình lên lớp, vi phạm nội quy, sử dụng chất kích thích, làm ngoài giờ.

3.4.2. Mô hình hồi quy Logistic

Sau quá trình phân tích bằng hồi quy Logistic có 9 yếu tố: Tuổi, tinh thần học tập, sức khỏe, kết quả học tập trước, Số đơn vị học trình nợ, tình hình lên lớp, vi phạm nội quy, sử dụng chất kích thích, làm ngoài giờ có ảnh hưởng đến kết quả học tập. Kết quả mô hình cho thấy P=0.000*** thấp hơn 0.05 và R^2 hiệu chỉnh là 0.888. Điều này có nghĩa là mô hình nghiên cứu có độ thích hợp là 88.8% hay 88.8% độ biến thiên về kết quả học tập của học sinh có thể được giải thích bởi các biến độc lập trong mô hình nên mô hình hồi quy Logistic phù hợp với tập dữ liệu nghiên cứu.

Phương trình hồi quy Logistic theo các biến độc lập như sau:

$$\mathbf{LOG_e[\frac{Y(KQ=1)}{Y(KQ=0)}] = -64.81 + 3.36*TSV + 0.88*TTHT + 2.11*SK + 1.68*KQHT - 0.40*STCN + 2.08*THLL - 3.69*VPNQ - 2.44*SDCKT - 2.31*LNG}$$



Hình 8. Mô hình các yếu tố ảnh hưởng kết quả học tập học sinh TCCN hệ THCS

3.4.3. Vận dụng mô hình hồi quy Logistic cho mô hình dự báo kết quả học tập

Mô hình dự báo kết quả học tập

$$\mathbf{E(KQ) = \frac{Exp(Z)}{1+Exp(Z)}}$$

Với:

Z = **-64.81 + 3.36*TSV + 0.88 *TTHT + 2.11*SK + 1.68*KQHT - 0.40*STCN + 2.08*THLL – 3.69 *VPNQ – 2.44 *SDCKT – 2.31 *LNG**

E(KQ) là kết quả dự báo học tập, E(KQ) có giá trị nằm trong khoảng từ 0 – 1.

Nếu E (KQ) < 0.5 thì học sinh có khả năng rớt và E (KQ) > =0.5 có khả năng đậu. Bảng xếp hạng kết quả học tập theo khả năng học tập E (KQ) như sau :

Bảng 25. Bảng phân định mức kết quả

STT	E(KQ) Tỉ số dự đoán	Kết quả
1	0 – 5	Rớt
2	0.5 - 1	Đậu

❖ Giả sử dự đoán kết quả của một học sinh A có thông tin như sau:

TSV	TTHT	SK	KQHT	STCN	THLL	VPNQ	SDCKT	LNG	KQ
16	Tot	Tot	Kha	0	Cham chi	Khong vi pham	Khong	Co	Dau

Khi đó xác xuất đậu của học sinh A là:

$$\mathbf{E(KQ)} = \frac{\boldsymbol{Exp(Z)}}{\mathbf{1+}\boldsymbol{Exp(Z)}}$$

Z = **-64.81 + 3.36*TSV + 0.88 *TTHT + 2.11*SK + 1.68*KQHT - 0.40*STCN + 2.08*THLL – 3.69 *VPNQ – 2.44 *SDCKT – 2.31 *LNG**

⇨ **E(KQ) = 0.996**

Mô hình hồi quy Logistic cho biết khả năng đậu của học sinh này tới 99.6 %. Nhưng cần lưu ý, đây chỉ là khả năng đậu được dự đoán, và dự đoán này có khả năng đúng chỉ 88.8%.

❖ Giả sử dự đoán kết quả của một học sinh B có thông tin như sau:

TSV	TTHT	SK	KQHT	STCN	THLL	VPNQ	SDCKT	LNG	KQ
18	Yeu	Tot	TB	9	Nghi nhieu ngay	Co	Co	Khong	Rot

Khi đó xác xuất đậu của học sinh B là:

$$\mathbf{E(KQ)} = \frac{\boldsymbol{Exp(Z)}}{\mathbf{1+}\boldsymbol{Exp(Z)}}$$

Z = **-64.81 + 3.36*TSV + 0.88 *TTHT + 2.11*SK + 1.68*KQHT - 0.40*STCN + 2.08*THLL – 3.69 *VPNQ – 2.44 *SDCKT – 2.31 *LNG**

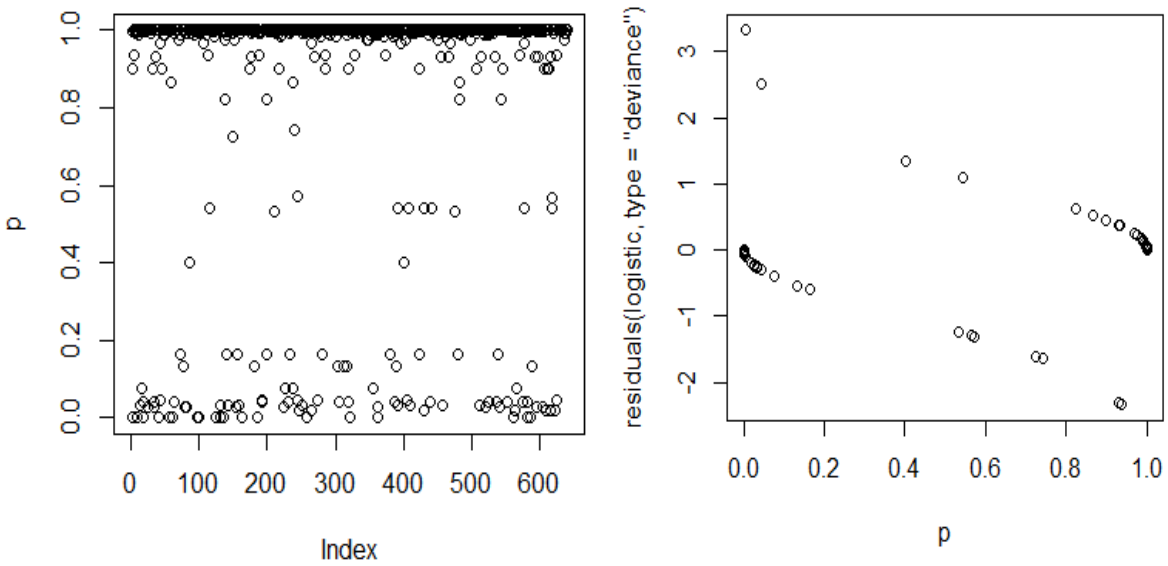
⇨ **E(KQ) = 0.043**

Mô hình hồi quy Logistic cho biết khả năng đậu của học sinh này tới 4.3%, như vậy luận văn có thể đưa ra phương án tác động đến quá trình học tập và nhắc nhở các em để đạt kết quả cao hơn. Nhưng cần lưu ý, đây chỉ là khả năng đậu được dự đoán, và dự đoán này có khả năng đúng chỉ 88.8%.

Tương tự như trên ta có xác xuất đậu của học sinh trong bộ mẫu như sau:

Bảng 26. Kết quả dự báo học tập của mẫu

STT	E(KQ)	KQ	STT	E(KQ)	KQ	STT	E(KQ)	KQ
1	2.191882e-04	0	11	9.963339e-01	1	21	1.000000e+00	1
2	9.967909e-01	1	12	9.863752e-01	1	22	9.991111e-01	1
3	8.989972e-01	1	13	3.376572e-02	0	23	9.999995e-01	1
4	9.341173e-01	1	14	9.999428e-01	1	24	2.616644e-02	0
5	9.999907e-01	1	15	7.510085e-02	0	25	9.999890e-01	1
6	9.997837e-01	1	16	9.997837e-01	1	26	9.997837e-01	1
7	9.936755e-01	1	17	4.338301e-02	0	27	9.999888e-01	1
8	1.000000e+00	1	18	9.992270e-01	1	28	9.999890e-01	1
9	9.988448e-01	1	19	3.949348e-03	0	29	9.911687e-01	1
10	2.191882e-04	0	20	9.996142e-01	1	30	1.000000e+00	1



Biểu đồ 17. Biểu đồ phân bố kết quả học tập dự đoán của mẫu

3.5. Đánh giá mô hình hồi quy Logistic

3.5.1. Đánh giá mô hình bằng ROC Curve

Từ ROC (Receiver Operating Characteristic) bắt nguồn từ một phần của lĩnh vực được gọi là thuyết phát hiện tín hiệu (Signal Detection Theory). Từ các tín hiệu nhận được, máy sẽ phân tích và vẽ đường cong ROC, để phân biệt tín hiệu của máy bay địch và tín hiệu nhiễu (noise) trong thế chiến thứ hai. Từ sau những năm 1970, thuyết phát hiện tín hiệu này được dùng để diễn dịch kết quả các test trong chẩn đoán y học. Mỗi điểm trên đường cong ROC là tọa độ tương ứng với tần suất dương tính thật (độ nhạy) trên trục tung và tần suất dương tính giả (1-độ đặc hiệu) trên trục hoành. Đường biểu diễn càng lệch về phía bên trên và bên trái thì sự phân biệt giữa 2 trạng thái (đậu hay rớt) càng rõ. Biểu đồ trên gồm có 3 đường cong ROC tương ứng với 3 khả năng: rất tốt, tốt và không giá trị. Độ chính xác (accuracy) được đo lường bằng diện tích dưới đường cong ROC. Nếu diện tích bằng 1 là test rất tốt và nếu bằng 0,5 thì test không có giá trị. Xác định đơn giản mức độ chính xác của test chẩn đoán dựa vào hệ thống điểm sau đây: [10]

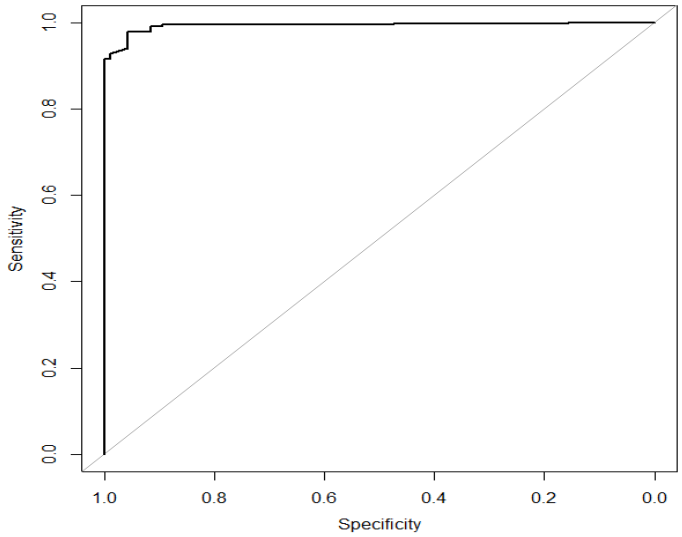
Bảng 27. Diễn giải ý nghĩa của diện tích dưới đường biểu diễn ROC (AUC)

AUC	**Ý nghĩa**
>0.90	Rất tốt (Excellent)
0.80 đến 0.90	Tốt (Good)
0.70 đến 0.80	Trung bình (Fair)
0.60 đến 0.70	Không tốt (Poor)
0.50 đến 0.60	Vô dụng (Fail)

Sau khi tác giả phân tích đánh giá thực nghiệm mô hình hồi quy các yếu tố ảnh hưởng đến kết quả học tập của học sinh TCCN hệ THCS bằng ROC Curve, ta thấy diện tích đường biểu diễn AUC = 0.9932 và khoảng tin cậy của kết quả dự đoán là 0.9882-0.9981, đều đó cho thấy mô hình rất có ý nghĩa thực nghiệm.

Bảng 28. Bảng kết quả đánh giá mô hình bằng ROC

AUC Diện tích đường biểu diễn	Ci Khoảng tin cậy 95%
0.9932	0.9882-0.9981



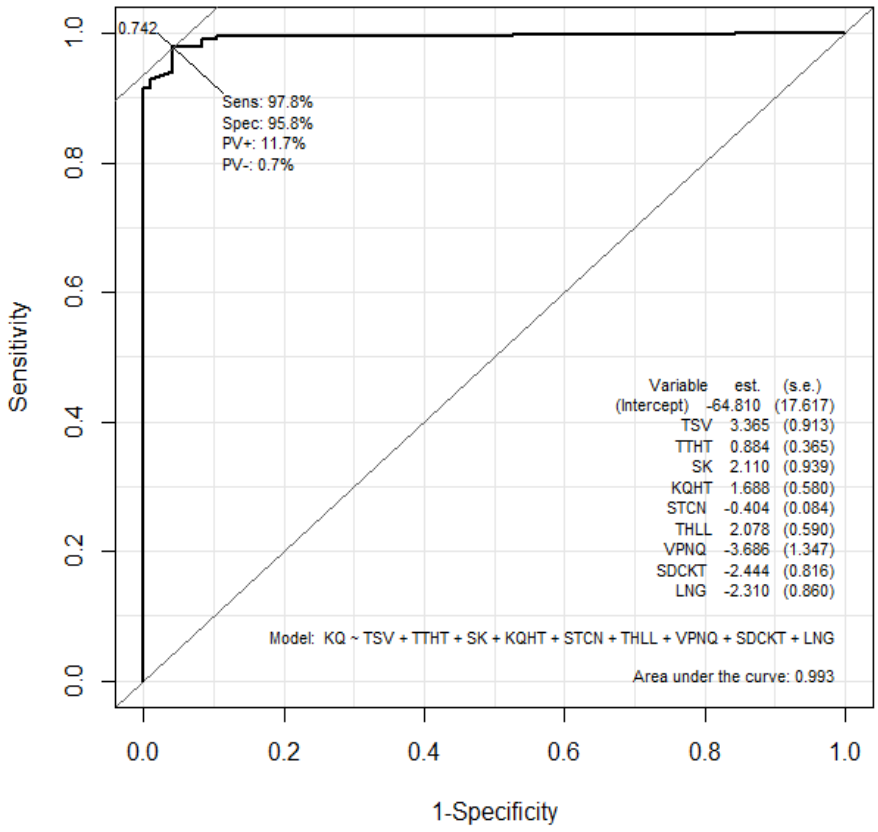
Data: pred in 95 controls (KQ 0) < 545 cases (KQ 1).
Area under the curve: 0.9932

Đồ thị 1. Diện tích dưới đường biểu diễn ROC (AUC)

Ngoài ra trong mô hình dự báo kết quả học tập, đường cong ROC được dùng để tìm giá trị chỉ định dự báo hay còn gọi điểm cắt tối ưu (cut off) của các biến định lượng có giá trị phân biệt 2 trạng thái (đậu / rớt) là 0.742 và các giá trị khác gồm:

- Độ nhạy (Sens) = 97.8%
- Độ đặc hiệu (Spec)= 95.8%
- Giá trị tiên đoán dương (PV+) = 11.7%
- Giá trị tiên đoán âm (PV-) = 0.7%
- Tỉ lệ dương tính giả α (False positive rate): 1- Spec= 4.2%
- Tỉ lệ âm tính giả β (False negative rate): 1- Sens= 2.2%
- Tỉ số khả dĩ dương (Likelihood ratio +) = Sens/1-Spec = 0.978/(1-0.958) = 23.3
- Tỉ số khả dĩ âm (Likelihood ratio -)= 1-Sens/Spec = 1-0.978/0.958 = 0.023

Tỉ số khả dĩ cao vì vậy giá trị chuẩn đoán kết quả học tập phân biệt càng cao.



Đồ thị 2. Điểm cắt tối ưu của mô hình ROC

Từ kết quả trên với điểm cắt tối ưu là 0.742, do đó ta có có thể phân định mức kết quả như sau: Nếu E (KQ) < 0.742 thì học sinh có khả năng rớt và E (KQ) >= 0.742 có khả năng đậu. Bảng xếp hạng định mức kết quả học tập theo khả năng học tập E (KQ) được thay đổi như sau :

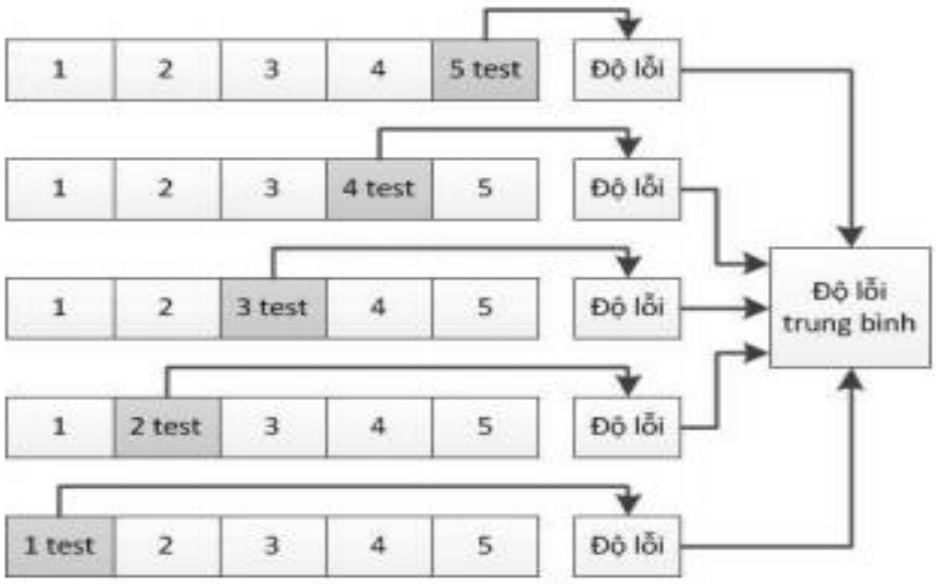
Bảng 29. Bảng phân định mức kết quả chính thức

STT	E(KQ) Tỉ số dự đoán	Phân định mức	Kết quả
1	0 – 0.742	Kết quả học tập không đạt yêu cầu	Rớt
2	0.742 – 1	Kết quả học tập tốt	Đậu

3.5.2. Đánh giá mô hình bằng phương pháp k-fold

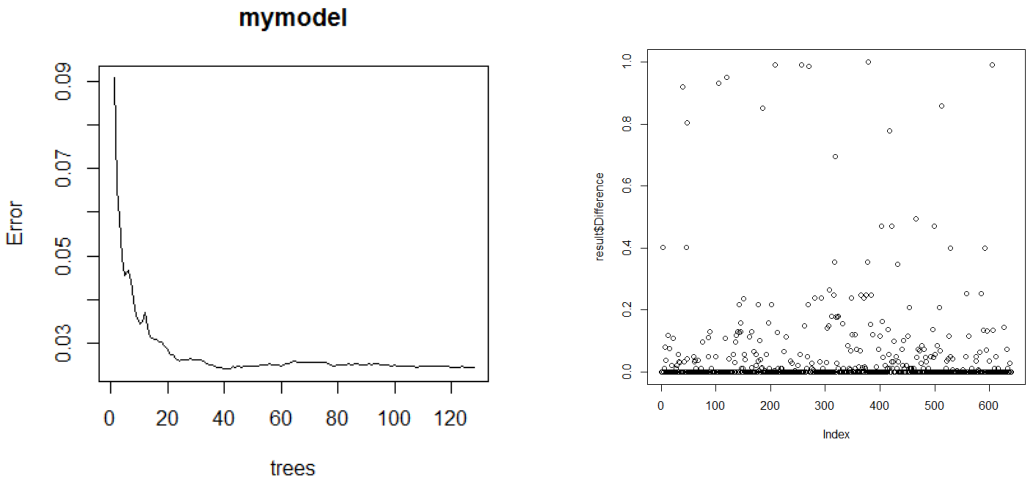
Đánh giá độ chính xác của bộ phân lớp rất quan trọng, bởi vì nó cho phép dự đoán được độ chính xác của các kết quả phân lớp những dữ liệu tương lai. Độ chính xác còn giúp so sánh các mô hình phân lớp khác nhau. Trong nghiên cứu này tác giả dùng phương pháp đánh giá chéo k-fold để đánh giá mô hình hồi quy Logistics ảnh hưởng các yếu tố đến kết quả học tập của học sinh TCCN hệ THCS. Các bước thực hiện như sau: [9]

- Tập bộ dữ liều gồm mẫu 640 mẫu trong đó: dữ liệu huấn luyện là 640 mẫu (theo phương pháp định mức), dữ liệu test là 128 mẫu (theo phương pháp ngẫu nhiên) được chia *5* tập con **không giao nhau** (gọi là "*fold"*) có kích thước 140 mẫu.
- Mỗi lần (trong số *5* lần) lặp, một tập con được sử dụng làm tập kiểm thử, và 4 tập con còn lại được dùng làm tập huấn luyện.
- ***k*** giá trị lỗi (mỗi giá trị tương ứng với một *fold*) được tính trung bình cộng để thu được giá trị lỗi tổng thể.

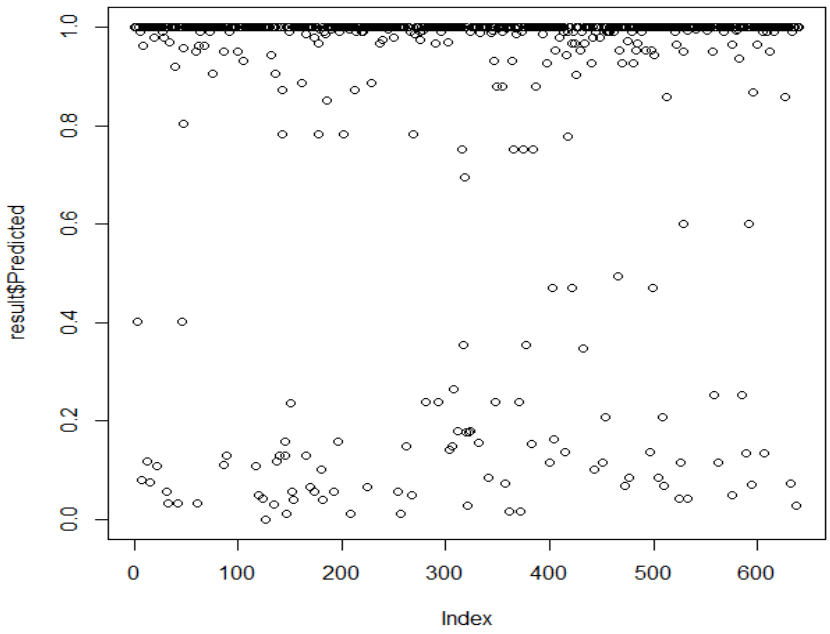


Hình 9. Mô tả phương pháp thử nghiệm K-Fold Kiểm thử dùng phương pháp kiểm tra chéo k-fold với k=5

Sau quá trình kiểm định phân tích hồi quy Logistic bằng phương pháp kiểm tra chéo k-fold với số lần k = 5 ta thấy độ lỗi trung bình của dữ liệu là 0.044 tức là 4.4%, xác suất của mô hình trong dự báo kết quả học sinh TCCN hệ THCS là 80.5% và trong quá định kiểm thử mức độ giao động của lỗi và tỉ lệ dự báo kết quả học tập được thể hiện trong hình sau:



Biểu đồ 18. Biểu đồ lỗi trong thực nghiệm bằng PP K-Fold cross validation



Biểu đồ 19 .Biểu đồ tỉ lệ dự báo trong thực nghiệm bằng PP K-Fold cross validation

Với k = 10: Sau quá trình kiểm định phân tích hồi quy Logistic bằng phương pháp kiểm tra chéo k-fold với số lần k = 10 ta thấy độ lỗi trung bình của dữ liệu là 0.0449

tức là 4.49%, xác suất của mô hình trong dự báo kết quả học sinh TCCN hệ THCS khoảng 83%. vậy chọn k = 10 cho việc kiểm tra chéo.

3.6. Tóm tắt chương

Trong chương 3, tác giả trình bày cách giải quyết bài toán theo quy trình CRISP. Quá trình phân tích thống kê cho thấy dữ liệu lấy phù hợp với nghiên cứu. Phân tích mô hình hồi quy Logistic 11 nhân tố với 9 nhân tố thỏa mãn điều kiện và 2 nhân tố không thỏa mãn, phương trình hồi quy Logistic được xác định có 9 biến: TSV, TTHT, SK, KQHT, STCN, THLL, VPNQ, SDCKT, LNG. Và cuối cùng là kiểm định các giả thuyết H1, H2, H3, H4, H5, H6, H7, H8, H9, H10, H11. Sau khi kiểm định, các giả thuyết H1, H2, H3, H4, H5, H6, H8, H10, H11 được chấp nhận và các giả thuyết H7, H9 không được chấp nhận.

Kết quả đánh giá mô hình cho thấy, mô hình có xác suất khoảng 83%, độ chính các thực nghiệm AUC = 0.9932 và điểm cắt tối ưu phân định giá trị đậu hay rớt là 0.742.

CHƯƠNG 4: ĐÁNH GIÁ BÀI TOÁN DỰ BÁO KẾT QUẢ HỌC SINH

Sau khi đã phân tích kết quả nghiên cứu ở chương 3, tác giả đưa ra những đánh giá tổng hợp từ nghiên cứu bao gồm: quy trình Crisp-DM, mô hình dự báo hồi quy Logistic, nguồn cơ sở dữ liệu và công cụ R.

4.1. Đánh giá quy trình Crisp-DM

Quy trình Crisp-DM là một quy trình chuẩn công nghiệp theo hướng mở, hỗ trợ trong công nghiệp, ứng dụng và công cụ khai phá dữ liệu. Quy trình cụ thể hóa các vấn đề trong khai thác dữ liệu từ vấn đề nghiệp vụ, tìm hiểu dữ liệu đến phân tích kỹ thuật.

Quy trình Crisp-DM có quy trình khai phá dữ liệu rõ ràng và hướng dẫn chi tiết quá trình khai phá dữ liệu.

Quy trình được xây dựng trên nhiều nền tảng ứng dụng khác nhau nên khá dễ hiểu khi tìm hiểu và sử dụng quy trình Crisp-DM.

4.2. Đánh giá hồi quy Logistic

❖ *Phương trình hồi quy Logistic có dạng:*

$$\mathbf{LOG_e[\frac{Y(KQ=1)}{Y(KQ=0)}] = -64.81 + 3.36*TSV + 0.88 *TTHT + 2.11*SK + 1.68*KQHT - 0.40*STCN + 2.08*THLL – 3.69 *VPNQ – 2.44 *SDCKT – 2.31 *LNG}$$

➔ Kết quả tạo cảnh báo học tập cho học sinh xác định có 9 yếu tố tác động đến dự báo kết quả học tập gồm tuổi học sinh, tinh thần học tập, sức khỏe, kết quả học tập, Số đơn vị học trình nợ, tinh thần lên lớp, vi phạm nội quy, sử dụng chất kích thích, làm ngoài giờ và mô hình này có ý nghĩa thực tế thống kê khi áp dụng mô hình hồi quy Logistic có xác suất khoảng 83%, độ chính các thực nghiệm AUC = 0.9932.

❖ *Ảnh hưởng các yếu tố đến mô hình cảnh báo học tập*

Để đánh giá tầm quan trọng tương đối của các yếu tố dự báo ảnh hưởng đến kết quả học tập của học sinh TCCN hệ THCS trong mô hình hồi quy Logistic, tác giả dựa vào giá trị tuyệt đối của t-statistic cho mỗi thông số mô hình hồi quy Logistic. Kết quả ảnh hưởng của các yếu tố được thể hiện ở bảng sau:

Bảng 30. Giá trị tuyệt đối của t-statistic ảnh hưởng biến đến mô hình hồi quy Logistic

Biến độc lập	Biến	Giá trị tuyệt đối của t-statistic
Độ tuổi của học sinh	TSV	3.684565
Tinh thần học tập	TTHT	2.424368
Sức khỏe	SK	2.247872
Kết quả học tập của HS	KQHT	2.912654
Số đơn vị học trình nợ	STCN	4.785401
Chuyên cần của học sinh	THLL	3.519961
Vi phạm nội quy	VPNQ	2.737194
Sử dụng chất kích thích	SDCKT	2.996731
Làm ngoài giờ	LNG	2.685262

Ngoài ra, sự ảnh hưởng của các yếu tố còn được thể hiện qua giá trị G^2 (Deviance) hay giá trị AIC của từng yếu tố, hai giá trị này còn lớn thì ảnh hưởng của yếu tố này đến mô hình còn cao

.Bảng 31. Giá trị Diviance và AIC của biến đến mô hình hồi quy Logistic

	Df	Deviance	AIC
Mô hình		87.927	107.93
- SK	1	92.903	110.90
- TTHT	1	94.626	112.63
- LNG	1	95.714	113.71
- VPNQ	1	96.759	114.76
- SDCKT	1	96.949	114.95
- KQHT	1	99.397	117.40
- THLL	1	107.049	125.05
- TSV	1	117.429	135.43
- STCN	1	131.010	149.01

Dựa vào kết quả bảng 30 và bảng 31, ta thấy các biến có sự ảnh hưởng đến kết quả học tập là như nhau. Trong đó, các yếu tố ảnh hưởng lớn đến mô hình nhất là Số đơn vị học trình nợ, tuổi học sinh và tình hình lên lớp bên cạnh đó yếu tố ảnh hưởng ít đến dự đoán kết là sức khỏe.

4.3. Đánh giá dữ liệu

❖ *Nguồn dữ liệu*

Dữ liệu thu thập thu thập dữ liệu từ phòng đào tạo, công tác học sinh – sinh viên với đối tượng cần lấy thông tin là những học sinh TCCN năm 2, năm 3 của 5 khoa tại

trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn về các yếu tố liên quan trực tiếp đến kết quả học tập có tính chính xác cao. Tuy nhiên, Kết quả nghiên cứu sẽ được tổng quát hóa cao hơn nếu được nghiên cứu trên phạm vi rộng lớn.

Thông tin mẫu dữ liệu được khai thác từ các yếu tố tác động trực tiếp đến học sinh chứ chưa khai thác các yếu tố gián tiếp như: tác động nhà trường, giảng viên, bạn bè và đặc điểm gia đình.

❖ *Phương pháp thu thập dữ liệu*

Mẫu khảo sát được thu thập bằng phương pháp chọn mẫu phi xác suất và lấy mẫu theo phương pháp định mức với 3 biến giám sát là năm học, ngành học và tỉ lệ đậu rớt, phương pháp hạn chế thiếu sót trong quá trình thu thập mẫu. Tuy nhiên, mẫu chưa có ý nghĩa nghiên cứu khoa học cao.

4.4. Đánh giá công cụ R

Công cụ R có chứa nhiều loại kỹ thuật thống kê như mô hình hóa tuyến tính và mô hình hồi quy Logistic, .. đặc biệt R hỗ trợ kiểm thử, phân tích và xử lý dữ liệu.

Công cụ R cho phép người dùng thêm các tính năng bổ sung bằng cách định nghĩa các hàm mới và có thể liên kết được với các ngôn ngữ khác như C,C++ và Fortran để có thể được gọi trong khi chạy. Người dùng thông thạo có thể viết mã C để xử lý trực tiếp các đối tượng của R.

Công cụ R cũng có tính mở rộng cao bằng cách sử dụng các gói cho người dùng đưa lên cho một số chức năng và lĩnh vực nghiên cứu cụ thể.

Công cụ R là nền tảng đồ họa của nó, có thể tạo ra những đồ thị chất lượng cao cùng các biểu tượng toán học, giao diện dễ sử dụng. Ngoài ra, R còn là một công cụ tính toán.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Sau khi đã phân tích và đánh giá bài toán tạo cảnh báo học tập, tác giả đã tóm tắt những kết quả nghiên cứu chính đưa ra những ý nghĩa thực tiển của nghiên cứu. Từ đó, tác giả đưa ra những kiến nghị tổng hợp từ kết quả nghiên cứu. Cuối cùng tác giả cũng trình bày những hạn chế và hướng nghiên cứu tiếp theo.

5.1. Kết luận

Sau khi nghiên cứu các yếu tố ảnh hưởng đến kết quả học tập của học sinh TCCN hệ THCS trên cơ sở thực tiễn và các nghiên cứu liên quan trước, tác giả đã đề xuất mô hình chính thức tạo cảnh báo học tập dựa trên quy trình CRISP-DM và hồi quy Logistic. Các yếu tố được lựa chọn tạo cảnh báo học tập phù hợp với thực tiển.

Nghiên cứu đã xây dựng và kiểm chứng mô hình các yếu tố ảnh hưởng đến kết quả học tập của học sinh TCCN hệ THCS với 9 nhân tố: Tuổi học sinh, tinh thần học tập, sức khỏe, kết quả học tập trước , tình hình lên lớp, vi phạm nội quy, sử dụng chất kích thích, Số đơn vị học trình nợ, làm ngoài giờ. Trong đó, các yếu tốt tinh thần học tập, tình hình lên lớp, số nợ đơn vị học trình ảnh hưởng rất lớn đến kết quả dự đoán kết quả học tập của học sinh TCCN hệ THCS. Kết quả dự đoán có xác suất khoảng 83%, độ chính các thực nghiệm AUC = 0.9932 và điểm cắt tối ưu phân định giá trị đậu hay rớt là 0.742.

5.2. Đóng góp của nghiên cứu

Trong những năm gần đây số lượng học sinh trung cấp chuyên nghiệp hệ trung học cơ sở trong các trường trung cấp bị cảnh báo học vụ và buộc thôi học ngày càng gia tăng. Do đó việc dự báo kết quả học tập của học sinh TCCN hệ THCS là điều cần thiết để các em lập kế hoạch với phương pháp học tập hiệu quả nhằm nâng cao kết quả học tập.

Chúng ta cũng phải nhìn nhận rằng hầu hết các các em học sinh TCCN đặc biệt là các em TCCN hệ THCS chưa quan tâm nhiều đến việc học tập dẫn đến tình trạng lưu ban hay đình chỉ học càng cao. Chính vì vậy, kết quả nghiên cứu này là một đóng góp mới để ban lãnh đạo, giáo viên chủ nhiệm của trường Trung cấp kỹ thuật và Nghiệp vụ Nam Sài Gòn có những định hướng học tập cho các em để đạt kết quả cao tốt. Kết quả

nghiên cứu là cơ sở để học sinh xây dựng cho mình một kế hoạch hợp lý nhằm đạt hiệu quả cao hơn.

Nghiên cứu đã chỉ ra 9 yếu tố tác động đến kết quả học tập của học sinh bao gồm tuổi học sinh, tinh thần học tâp, kết quả học tập, sức khỏe, Số đơn vị học trình nợ, tình hình lên lớp, vi phạm nội quy, sử dụng chất kích thích, làm ngoài giờ (làm thêm). Trong đó, có 3 yếu tố ảnh hưởng lớn đến kết quả học tập là tuổi học sinh, số đơn vị học trình nợ và tình hình lên lớp. Kết quả này có ý nghĩa quan trọng đối với giáo viên chủ nhiệm, ban lãnh đạo nhà trường Trung cấp Kỹ thuật và Nghiệp vụ Nam Sài Gòn trong việc tác động đến các yếu tố nhằm cải thiện kết quả học tập của học sinh TCCN hệ THCS.

5.3. Kiến nghị

Trong những năm gần đây số lượng học sinh trung cấp chuyên nghiệp hệ trung học cơ sở trong các trường trung cấp bị cảnh báo học vụ và buộc thôi học ngày càng gia tăng. Vì vậy kết quả nghiên cứu của đề tài sẽ góp phần đóng góp vào thực tiễn để giáo viên chủ nhiệm và ban lãnh đạo nhà trường có những tác động cải thiện kết quả học tập của học sinh đặc biệt chú trọng 2 yếu tố số đơn vị học trình nợ và tình hình lên lớp:

- Ban lãnh đạo: Tổ chức sân chơi lành mạnh trong trường học kích thích tinh thần học tập của các em, có kế hoạch tổ chức học cải thiện nhằm tạo điều kiện cho các em học sinh cải thiện số đơn vị học trình nợ.
- Giáo viên bộ môn: Thay đổi phương pháp dạy kích thích tính ham học hỏi trong học sinh.
- Giáo viên chủ nhiệm: Quan tâm, nhắc nhở và định hướng cho học sinh đồng thời phối hợp với gia đình giảm bớt các ảnh hưởng tình hình lên lớp, vi phạm nội quy, sử dụng chất kích thích, làm ngoài giờ. Giáo viên chủ nhiệm thường xuyên cập nhập kết quả học tập của học sinh và tình hình lên lớp để kịp thời nhắc nhở các em.
- Học sinh: Có kế hoạch và phương pháp học tập hiệu quả nhằm cải thiện kết quả. Đồng thời, học sinh chú trọng cải thiện học tập của mình tập trung vào 2 yếu tố quan trọng là cải thiện số tín chỉ nợ và cải thiện tình hình lên lớp.

5.4. Giới hạn của nghiên cứu và hướng phát triển tiếp theo

Hạn chế của đề tài thuộc về mẫu nghiên cứu được lấy theo phương pháp định mức, mẫu chưa có ý nghĩa nghiên cứu cao khi chỉ thực hiện lấy mẫu tại trường Trung cấp Kỹ thuật & Nghiệp vụ Nam Sài Gòn.

Mô hình chỉ giải quyết được một số các yếu tố tác động trực tiếp đến học sinh chứ chưa khai thác các yếu tố gián tiếp như: tác động nhà trường, giảng viên, bạn bè…

Kết quả nghiên cứu sẽ được tổng quát hóa cao hơn nếu được nghiên cứu trên phạm vi rộng lớn. Bên cạnh đó, nếu mẫu khảo sát được thu thập bằng phương pháp chọn mẫu xác suất thì có thể mẫu sẽ có tính đại diện cao hơn.

Hướng nghiên cứu tiếp theo của đề tài là nghiên cứu học sinh TCCN hệ phổ thông và các cơ sở đào tạo trung cấp chuyên nghiệp khác. Đồng thời, tiến hành nghiên cứu khai thác các yếu tố gián tiếp như: tác động nhà trường, giảng viên, bạn bè, xã hội…

TÀI LIỆU THAM KHẢO

[1] P. Baepler and C.J. Murdoch (2010) “Academic Analytics and Data Mining in Higher Education” *International Journal for the Scholarship of Teaching and Learning*: Vol. 4: No. 2, Article 17.

[2] E.J.M. Lauría, J.D. Baron, M. Devireddy, V. Sundararaju and S.M. Jayaprakash “Mining academic data to improve college student retention: An open source perspective” *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge ACM New York, NY, USA ©2012*, ISBN: 978-1-4503-1111-3.

[3] B.K.Baradwaj and S. Pal “Mining Educational Data to Analyze StudentsPerformance” *International Journal of Advanced Computer Science and Applications*: Vol. 2, No. 6, 2011.

[4] J. Bainbridge, J. Melitski, A. Zahradnik, E.J. M. Lauría, S. Jayaprakash, and J. Baron “Using Learning Analytics to Predict At-Risk Students in Online Graduate Public Affairs and Administration Education” *The Journal of Public Affairs Education:* Vol. 21: No. 2.

[5] P. Cortez and A. Silva. “Using Data Mining to Predict Secondary School Student Performance”. *In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal*, ISBN 978-9077381-39-7.

[6] G. James, T. Hastie, D. Witten and R.Tibshirani “Springer Texts in Statistics” *Springer New York Heidelberg Dordrecht Londo*, ISBN 978-1-4614-7138-7 (eBook), DOI 10.1007/978-1-4614-7138-7.

[7] Anurag Srivastava, Eui- Hong Han, Vipin Kumar, Vieet Singh. *Parallel Formulations of Decision-Tree Classification Algorithm*. Kluwer Academic Publisher, 1999.

[8] S. Menard, “*Applied logistic regression analysis”,* Second edition, Sage publication, 2002.

An Introduction to Statistical Learning with Applications in R (Fourth Printing), G. James, D. Witten, T. Hastie and R. Tibshirani, Springer-Verlag, 2014

[9] The Elements of Statistical Learning (Second Edition), T. Hastie, R. Tibshirani and J. Friedman, Springer-Verlag, 2009

[10] https://the-modeling-agency.com/crisp-dm.pdf

[11] Machine learning with R Cookbook, Yu-Wei, Chiu (David Chiu), Published by Packt Publishing Ltd., ISBN 978-1-78398-204-2, 2015

[12] Data Mining and Predictive Analytics, Daniel T.Larose and Chantal D.Larose, Published by John Wiley & Son, Inc., 2015

[13] Trần T. Kiên, Bảnh T. Thành, Nguyễn H.T. Anh "Dự đoán giá cổ phiếu trên thị trường chứng khoán Việt Nam bằng phương pháp lai GA-SVR" tạp chí Công nghệ thông tin và truyền thông, ISSN 1859 – 3526: Tập V-1, Số 7(27), tháng 5/2012.

[14] TS Nguyen Ngoc Rang "Ứng dụng đường cong ROC trong nghiên cứu y học": http://www.bvag.com.vn/index.php?option=com_k2&view=item&task=download&id=24_8f88b9e064e2ffc626cafc50b72832b2&Itemid=128.

[15] Nguyễn Văn Tuấn "Phân tích số liệu và biểu đồ bằng R" : https://cran.r-project.org/doc/contrib/Intro_to_R_Vietnamese.pdf

[16] Phân tích dữ liệu với R, Nguyễn Văn Tuấn, NXB. Tổng Hợp TP.HCM, 2014

[17] Trần Ngọc Trình "Quản lý chất lượng đào tạo tại các trường trung cấp chuyên nghiệp thành phố Hồ Chí Minh" : Luận án tiến sĩ 2015.

PHỤ LỤC 1
ĐỊNH NGHĨA BIẾN

STT	Mã biến	Biến độc lập	Loại	Giá trị	
				Giá trị thực	Giá trị ĐN
1	TSV	Độ tuổi của học sinh	Liên tục	16-21	16-21
2	TTHT	Tinh thần học tập	Rời rạc	Tốt Khá TB Yếu	4 3 2 1
3	SK	Sức khỏe	Rời	Tốt Không tốt	1 0
4	KQHT	Kết quả học tập	Rời rạc	Giỏi Khá TBK TB	4 3 2 1
5	STCN	Số đơn vị học trình nợ	Liên tục	0 - 19	0- 19
6	THLL	Thời gian lên lớp	Rời rạc	Chăm chỉ Thỉnh thoảng nghỉ Nghỉ nhiều ngày Nghỉ thường xuyên	4 3 2 1
7	NH	Năm học	Liên tục	2-3	2 / 3
8	NGH	Ngành học	Rời rạc	Công nghệ thông tin Điện tử Cơ khí xây dựng Cơ khí động lực Du lịch	1 2 3 4 5
9	VPNQ	Vi phạm nội quy	Nhị phân	Có / không	1 / 0
10	TTSGD	Tình trạng sống với gia đình	Rời rạc	Sống với gia đình Ở ngoài	1 / 0
11	SDCKT	Sử dụng chất kích thích	Nhị phân	Có / không	1 / 0
12	LNG	Làm ngoài giờ	Nhị phân	Có / không	1 / 0
13	KQ	Kết quả	Nhị phân	Đậu / Rớt	1 / 0

PHỤ LỤC 2
MÔ TẢ THỐNG KÊ ĐỊNH LƯỢNG & ĐỊNH TÍNH

Thống kê mô tả biến định tính

a) Ngành học theo mẫu

Total Observations in Table: 640

1	2	3	4	5
147	115	148	125	105
0.230	0.180	0.231	0.195	0.164

b) Năm học của mẫu

Total Observations in Table: 640

2	3
348	292
0.544	0.456

c) Kết quả của mẫu

Total Observations in Table: 640

0	1
95	545
0.148	0.852

Thống kê mô tả biến định lượng

a) Thống kê theo tuổi của mẫu

Total Observations in Table: 640
Tuổi trung bình:17.63594

16	17	18	19	20
144	189	145	87	68
0.225	0.295	0.227	0.136	0.106

21
7
0.011

b) Thống kê theo giới tính

Total Observations in Table: 640

0	1

61	579
0.095	0.905

c) Thống kê theo khu vực

Total Observations in Table: 640

0	1
251	389
0.392	0.608

d) Thống kê theo tinh thần học tập

Total Observations in Table: 640

1	2	3	4
41	71	291	237
0.064	0.111	0.455	0.370

e) Thống kê theo sức khỏe

Total Observations in Table: 640

0	1
62	578
0.097	0.903

f) Thống kê theo kết quả học tập

Total Observations in Table: 640

1	2	3	4
168	135	281	56
0.263	0.211	0.439	0.087

g) Thống kê theo số đơn vị học trình nợ

Total Observations in Table: 640

0	2	3	5	6
456	25	61	8	14
0.713	0.039	0.095	0.012	0.022

9	10	11	12	14
12	11	8	18	7
0.019	0.017	0.012	0.028	0.011

15	16	18
10	4	6
0.016	0.006	0.009

h) Thống kê theo tình hình lên lớp

Total Observations in Table: 640

1	2	3	4
43	17	265	315
0.067	0.027	0.414	0.492

i) Thống kê theo ngành học

Total Observations in Table: 640

1	2	3	4	5
147	115	148	125	105
0.230	0.180	0.231	0.195	0.164

j) Thống kê học sinh vi phạm nội quy

Total Observations in Table: 640

0	1
580	60
0.906	0.094

k) Thống kê theo thành phần gia đình

Total Observations in Table: 640

1	2	3	4	5
7	41	157	213	222
0.011	0.064	0.245	0.333	0.347

l) Thống kê theo tình trạng sống với gia đình

Total Observations in Table: 640

0	1
177	463
0.277	0.723

m) Thống kê học sinh sử dụng chất kích thích

Total Observations in Table: 640

0	1
555	85
0.867	0.133

n) Thống kê học sinh làm ngoài giờ

Total Observations in Table: 640

0	1
431	209
0.673	0.327

PHỤ LỤC 3
PHÂN TÍCH THỐNG KÊ MÔ TẢ BIẾN ĐỘC LẬP VÀ PHỤ THUỘC

Phân tích thống kê mô tả

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
TSV	1	640	17.64	1.31	17	17.53	1.48	16	21	5	0.49	-0.70	0.05
TTHT	4	640	3.13	0.85	3	3.24	1.48	1	4	3	-0.88	0.29	0.03
SK	5	640	0.85	0.36	1	0.93	0.00	0	1	1	-1.91	1.63	0.01
KQHT	6	640	2.35	0.96	3	2.33	1.48	1	4	3	-0.16	-1.14	0.04
STCN	7	640	2.03	4.14	0	0.91	0.00	0	18	18	2.20	3.84	0.16
THLL	8	640	3.33	0.82	3	3.49	1.48	1	4	3	-1.40	1.71	0.03
NGH	10	640	2.88	1.39	3	2.86	1.48	1	5	4	0.06	-1.25	0.06
VPNQ	11	640	0.09	0.29	0	0.00	0.00	0	1	1	2.78	5.74	0.01
TTSGD	13	640	0.72	0.45	1	0.78	0.00	0	1	1	-1.00	-1.01	0.02
SDCKT	14	640	0.13	0.34	0	0.04	0.00	0	1	1	2.16	2.66	0.01
LNG	15	640	0.33	0.47	0	0.28	0.00	0	1	1	0.74	-1.46	0.02
KQ	16	640	0.85	0.36	1	0.94	0.00	0	1	1	-1.97	1.90	0.01

Phương sai đám đông

TSV	TTHT	SK	KQHT
1.72484106	0.72140063	0.13096391	0.92941853
STCN	**THLL**	**NGH**	**VPNQ**
17.11805311	0.67883412	1.93966158	0.08509390
TTSGD	**SDCKT**	**LNG**	**KQ**
0.20038879	0.11535358	0.22026360	0.12660162

PHỤ LỤC 4
KIỂM ĐỊNH MÔ HÌNH VÀ Ý NGHĨA HỆ SỐ

Kiểm định mô hình

❖ Đánh giá mô hình sử dụng AIC

Log-likelihood = -41.1816
No. of observations = 640
AIC value = 106.3633

❖ Đánh giá mô hình sử dụng Deviance (G^2)

```
>glm(formula = KQ ~ TSV + TTHT + KQHT + SK + STCN + THLL +
   VPNQ + SDCKT + LNG + NGH + TTSGD, family = "binomial",
   data = db)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4209   0.0001   0.0022   0.0273   3.2195
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -61.31258   16.56009  -3.702 0.000214 ***
TSV           3.27545    0.86494   3.787 0.000153 ***
TTHT          0.93191    0.34300   2.717 0.006589 **
KQHT          2.07562    0.97620   2.126 0.033484 *
SK            2.35969    1.06554   2.215 0.026790 *
STCN         -0.45443    0.09336  -4.867 1.13e-06 ***
THLL          2.22206    0.61475   3.615 0.000301 ***
VPNQ         -5.95309    1.81612  -3.278 0.001046 **
SDCKT        -3.81078    1.42513  -2.674 0.007495 **
LNG          -3.03168    1.15781  -2.618 0.008833 **
NGH           0.44724    0.31629   1.414 0.157349
TTSGD        -3.46628    2.68344  -1.292 0.196450
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
   ’ 1(Dispersion parameter for binomial family taken to be
   1)
    Null deviance: 537.586  on 639  degrees of freedom
Residual deviance:  82.363  on 628  degrees of freedom
AIC: 106.36
```

Ý nghĩa hệ số và kiểm định giả thuyết

	crude OR(95%CI)	adj. OR(95%CI)	P(Wald's test)	P(LR-test)
TSV (cont. var.)	1.22 (1.02,1.46)	26.46 (4.86,144.13)	< 0.001	< 0.001
TTHT (cont. var.)	4.66 (3.45,6.28)	2.54 (1.3,4.97)	0.007	0.004
SK: 1 vs 0	8.72 (4.97,15.3)	10.59 (1.31,85.47)	0.027	0.021
KQHT (cont. var.)	7.62 (5.02,11.59)	7.97 (1.18,54)	0.033	0.005
STCN (cont. var.)	0.55 (0.49,0.61)	0.63 (0.53,0.76)	< 0.001	< 0.001
THLL (cont. var.)	4.7 (3.48,6.36)	9.23 (2.77,30.78)	< 0.001	< 0.001
VPNQ: 1 vs 0	0.05 (0.03,0.1)	0 (0,0.09)	0.001	< 0.001
SDCKT: 1 vs 0	0.13 (0.08,0.22)	0.02 (0,0.36)	0.007	< 0.001
LNG: 1 vs 0	1.34 (0.83,2.18)	0.05 (0,0.47)	0.009	0.003
NGH (cont. var.)	0.89 (0.76,1.04)	1.56 (0.84,2.91)	0.157	0.155
TTSGD: 1 vs 0	1.11 (0.69,1.79)	0.03 (0,6.01)	0.196	0.11

Kiểm định độ phù hợp của mô hình

```
>pR2(logistic)
       llh      llhNull           G2     McFadden         r2ML         r2CU
-43.9635239 -268.7930701  449.6590923    0.8364410    0.5047003    0.8881189
```

```
lrm(formula = KQ ~ TSV + TTHT + KQHT + STCN + THLL + VPNQ + SDCKT +
    LNG + NGH + TTSGD + SK)
                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test            Indexes          Indexes
Obs            640    LR chi2     450.32    R2        0.889    C        0.993
 0              95    d.f.            11    g         8.714    Dxy      0.986
 1             545    Pr(> chi2) <0.0001    gr     6085.497    gamma    0.987
max |deriv| 2e-08                          gp        0.250    tau-a    0.250
                                           Brier     0.019
```

Hệ số tương quan

❖ Giá trị hệ số tương quan

```
> fix(dab)
> var = cbind(TSV,TTHT,SK,KQHT,STCN,THLL,VPNQ,SDCKT,LNG)
> pairs.panels(var method = "kendall")
> library(psych)
```

```
> corr.test(var,method = "kendall")
Call:corr.test(x = var, method = "kendall")
Correlation matrix
```

	TSV	TTHT	SK	KQHT	STCN	THLL	VPNQ	SDCKT	LNG
TSV	1.00	-0.08	-0.03	-0.03	-0.01	-0.03	0.09	0.02	0.11
TTHT	-0.08	1.00	0.09	0.49	-0.40	0.43	-0.12	-0.02	0.09
SK	-0.03	0.09	1.00	0.17	-0.38	0.15	-0.09	-0.06	0.00
KQHT	-0.03	0.49	0.17	1.00	-0.59	0.43	-0.24	-0.06	0.15
STCN	-0.01	-0.40	-0.38	-0.59	1.00	-0.30	0.26	0.13	-0.15
THLL	-0.03	0.43	0.15	0.43	-0.30	1.00	-0.19	-0.14	-0.01
VPNQ	0.09	-0.12	-0.09	-0.24	0.26	-0.19	1.00	0.36	-0.08
SDCKT	0.02	-0.02	-0.06	-0.06	0.13	-0.14	0.36	1.00	-0.14
LNG	0.11	0.09	0.00	0.15	-0.15	-0.01	-0.08	-0.14	1.00

```
Sample Size

[1] 640

Probability values (Entries above the diagonal are
  adjusted for multiple tests.)
```

❖ Ý nghĩa thống kê hệ số tương quan

```
Sample Size
[1] 640
Probability values (Entries above the diagonal are adjusted
  for multiple tests.)
```

	TSV	TTHT	SK	KQHT	STCN	THLL	VPNQ	SDCKT	LNG
TSV	0.00	0.66	1.00	1.00	1	1.00	0.28	1.00	0.07
TTHT	0.05	0.00	0.31	0.00	0	0.00	0.03	1.00	0.35
SK	0.46	0.02	0.00	0.00	0	0.00	0.28	1.00	1.00
KQHT	0.48	0.00	0.00	0.00	0	0.00	0.00	1.00	0.00
STCN	0.84	0.00	0.00	0.00	0	0.00	0.00	0.02	0.00
THLL	0.40	0.00	0.00	0.00	0	0.00	0.00	0.01	1.00
VPNQ	0.02	0.00	0.02	0.00	0	0.00	0.00	0.00	0.66
SDCKT	0.67	0.65	0.14	0.13	0	0.00	0.00	0.00	0.01
LNG	0.00	0.03	0.94	0.00	0	0.82	0.06	0.00	0.00

Phân tích hồi quy Logistic

```
>logistic=glm(KQ~TSV+TTHT+SK+KQHT+STCN+THLL+VPNQ+SDCKT+LNG
  ,family = "binomial",data =db)
> logistic.display(logistic)
```

```
Logistic regression predicting KQ
```

	crude OR(95%CI)	adj. OR(95%CI)	P(Wald's test)	P(LR-test)
TSV (cont. var.)	1.22 (1.02,1.46)	28.92 (4.83,173.17)	< 0.001	< 0.001
TTHT (cont. var.)	4.66 (3.45,6.28)	2.42 (1.18,4.95)	0.015	0.01
SK: 1 vs 0	8.72 (4.97,15.3)	8.25 (1.31,51.93)	0.025	0.026
KQHT (cont. var.)	7.62 (5.02,11.59)	5.41 (1.74,16.85)	0.004	< 0.001

STCN (cont. var.)	0.55 (0.49,0.61)	0.67 (0.57,0.79)	< 0.001	< 0.001
THLL (cont. var.)	4.7 (3.48,6.36)	7.99 (2.51,25.43)	< 0.001	< 0.001
VPNQ: 1 vs 0	0.05 (0.03,0.1)	0.03 (0,0.35)	0.006	0.003
SDCKT: 1 vs 0	0.13 (0.08,0.22)	0.09 (0.02,0.43)	0.003	0.003
LNG: 1 vs 0	1.34 (0.83,2.18)	0.1 (0.02,0.54)	0.007	0.005

Log-likelihood = -43.9635
No. of observations = 640
AIC value = 107.927

```
> logistic
```

```
Call:  glm(formula = KQ ~ TSV + TTHT + SK + KQHT + STCN + THLL + VPNQ +
    SDCKT + LNG, family = binomial, data = db)

Coefficients:
(Intercept)          TSV         TTHT           SK         KQHT         STCN
   -64.8099       3.3645       0.8844       2.1101       1.6883      -0.4037
       THLL         VPNQ        SDCKT          LNG
     2.0785      -3.6858      -2.4439      -2.3103

Degrees of Freedom: 639 Total (i.e. Null);  630 Residual
Null Deviance:      537.6
Residual Deviance: 87.93        AIC: 107.9
```

```
> data= datadist(db)
> options(datalist = "data")
>logistic=lrm(KQ~TSV+TTHT+SK+KQHT+STCN+THLL+VPNQ+SDCKT+LNG
   )
> logistic
```

```
Logistic Regression Model

lrm(formula = KQ ~ TSV + TTHT + SK + KQHT + STCN + THLL + VPNQ +
    SDCKT + LNG)
                      Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes          Indexes
Obs           640    LR chi2     449.66    R2        0.888    C       0.993
 0             95    d.f.             9    g         7.813    Dxy     0.987
 1            545    Pr(> chi2) <0.0001    gr     2472.970    gamma   0.988
max |deriv| 3e-07                          gp        0.250    tau-a   0.250
                                           Brier     0.018
```

	Coef	S.E.	Wald Z	Pr(>|Z|)
Intercept	-64.8099	17.6170	-3.68	0.0002
TSV	3.3645	0.9131	3.68	0.0002
TTHT	0.8844	0.3648	2.42	0.0153
SK	2.1101	0.9387	2.25	0.0246
KQHT	1.6883	0.5796	2.91	0.0036
STCN	-0.4037	0.0844	-4.79	<0.0001
THLL	2.0785	0.5905	3.52	0.0004
VPNQ	-3.6858	1.3466	-2.74	0.0062
SDCKT	-2.4439	0.8155	-3.00	0.0027
LNG	-2.3103	0.8604	-2.69	0.0072

Dự báo kết quả: cbind(p,KQ)

	p	KQ
1	2.191882e-04	0
2	9.967909e-01	1
3	8.989972e-01	1
4	9.341173e-01	1
5	9.999907e-01	1
6	9.997837e-01	1
7	9.936755e-01	1
8	1.000000e+00	1
9	9.988448e-01	1
10	2.191882e-04	0
11	9.963339e-01	1
12	9.863752e-01	1
13	3.376572e-02	0
14	9.999428e-01	1
15	7.510085e-02	0
16	9.997837e-01	1
17	4.338301e-02	0
18	9.992270e-01	1
19	3.949348e-03	0
20	9.996142e-01	1
21	1.000000e+00	1
22	9.991111e-01	1
23	9.999995e-01	1
24	2.616644e-02	0
25	9.999890e-01	1
26	9.997837e-01	1
27	9.999888e-01	1
28	9.999890e-01	1
29	9.911687e-01	1
30	1.000000e+00	1
31	8.989972e-01	1
32	9.999968e-01	1
33	2.616644e-02	0
34	4.352627e-02	0
35	9.293548e-01	1
36	9.995581e-01	1
37	9.999733e-01	1
38	1.000000e+00	1
39	9.996806e-01	1
40	5.497919e-06	0
41	9.993981e-01	1
42	9.676350e-01	1
43	4.405302e-02	0
44	8.989972e-01	1
45	9.999794e-01	1
46	9.999103e-01	1
47	9.993981e-01	1
48	9.863752e-01	1
49	1.000000e+00	1
50	9.886146e-01	1
51	1.000000e+00	1
52	9.999410e-01	1
53	9.996806e-01	1
54	9.999980e-01	1
55	9.999962e-01	1
56	2.191882e-04	0
57	1.000000e+00	1
58	8.675393e-01	1
59	9.999995e-01	1
60	2.191882e-04	0
61	9.999980e-01	1
62	9.999733e-01	1
63	9.938555e-01	1
64	4.338301e-02	0
65	9.999974e-01	1
66	9.999969e-01	1
67	1.000000e+00	1
68	9.999937e-01	1
69	9.999962e-01	1
70	9.765477e-01	1
71	9.996806e-01	1
72	1.629632e-01	0
73	9.886146e-01	1
74	1.000000e+00	1
75	1.000000e+00	1
76	1.323060e-01	0
77	9.999733e-01	1
78	9.994053e-01	1
79	9.999794e-01	1
80	2.616644e-02	0
81	2.776191e-02	0
82	9.938555e-01	1
83	9.999980e-01	1
84	9.999103e-01	1
85	4.011349e-01	1
86	9.995597e-01	1
87	9.999890e-01	1
88	9.911687e-01	1
89	1.000000e+00	1
90	1.000000e+00	1
91	9.999980e-01	1
92	9.988448e-01	1
93	9.863752e-01	1
94	9.999890e-01	1
95	9.999648e-01	1
96	9.999428e-01	1
97	2.291386e-03	0
98	9.999888e-01	1
99	2.191882e-04	0
100	9.967909e-01	1
101	9.938555e-01	1
102	1.000000e+00	1
103	9.999557e-01	1
104	9.988448e-01	1
105	9.999974e-01	1
106	9.676350e-01	1
107	1.000000e+00	1
108	9.993981e-01	1
109	9.999890e-01	1
110	9.886146e-01	1
111	9.999969e-01	1
112	9.999600e-01	1
113	9.999648e-01	1
114	9.341173e-01	1
115	5.427849e-01	1
116	9.996806e-01	1
117	9.999410e-01	1
118	9.999991e-01	1
119	9.997305e-01	1
120	1.000000e+00	1
121	9.999937e-01	1
122	9.988448e-01	1
123	9.852502e-01	1
124	2.291386e-03	0
125	9.999890e-01	1
126	9.997863e-01	1
127	9.997863e-01	1
128	9.999890e-01	1

129	9.997837e-01	1
130	2.191882e-04	0
131	9.992270e-01	1
132	3.376572e-02	0
133	9.999410e-01	1
134	9.999974e-01	1
135	9.997863e-01	1
136	2.291386e-03	0
137	9.765477e-01	1
138	8.226549e-01	1
139	9.963339e-01	1
140	9.886146e-01	1
141	1.629632e-01	0
142	3.376572e-02	0
143	9.999733e-01	1
144	9.991111e-01	1
145	9.999907e-01	1
146	9.999890e-01	1
147	9.999890e-01	1
148	9.999557e-01	1
149	9.937819e-01	1
150	7.241927e-01	0
151	1.000000e+00	1
152	9.765477e-01	1
153	9.988448e-01	1
154	2.616644e-02	0
155	1.000000e+00	1
156	1.629632e-01	0
157	9.996142e-01	1
158	3.213706e-02	0
159	9.997837e-01	1
160	9.999991e-01	1
161	9.936755e-01	1
162	3.949348e-03	0
163	9.999890e-01	1
164	1.000000e+00	1
165	1.000000e+00	1
166	9.997837e-01	1
167	9.999410e-01	1
168	9.999989e-01	1
169	9.997837e-01	1
170	9.999980e-01	1
171	9.999794e-01	1
172	9.996142e-01	1
173	8.989972e-01	1
174	1.000000e+00	1
175	1.000000e+00	1
176	9.936755e-01	1
177	9.293548e-01	1
178	9.967909e-01	1
179	9.997863e-01	1
180	9.988448e-01	1
181	9.996142e-01	1
182	1.323060e-01	0
183	9.999999e-01	1
184	9.988448e-01	1
185	9.997305e-01	1
186	5.497919e-06	0
187	9.360317e-01	0
188	9.936755e-01	1
189	9.963339e-01	1
190	9.936755e-01	1
191	9.911687e-01	1
192	4.405302e-02	0
193	4.338301e-02	0
194	1.000000e+00	1
195	9.999692e-01	1
196	9.997863e-01	1
197	9.999968e-01	1
198	1.629632e-01	0
199	8.226549e-01	1
200	9.996142e-01	1
201	9.988448e-01	1
202	9.999968e-01	1
203	9.999997e-01	1
204	9.994053e-01	1
205	1.000000e+00	1
206	9.988448e-01	1
207	9.982747e-01	1
208	9.991111e-01	1
209	1.000000e+00	1
210	5.331770e-01	0
211	9.999890e-01	1
212	9.999837e-01	1
213	9.938555e-01	1
214	9.991111e-01	1
215	9.967909e-01	1
216	9.938555e-01	1
217	9.999890e-01	1
218	8.989972e-01	1
219	9.999995e-01	1
220	9.852502e-01	1
221	1.000000e+00	1
222	9.881129e-01	1
223	1.000000e+00	1
224	2.616644e-02	0
225	9.999888e-01	1
226	1.000000e+00	1
227	7.510085e-02	0
228	9.999648e-01	1
229	9.999969e-01	1
230	9.999974e-01	1
231	4.352627e-02	0
232	9.937819e-01	1
233	9.995971e-01	1
234	1.629632e-01	0
235	1.000000e+00	1
236	9.988448e-01	1
237	8.675393e-01	1
238	7.510085e-02	0
239	7.422942e-01	0
240	9.999962e-01	1
241	9.999692e-01	1
242	1.000000e+00	1
243	9.999103e-01	1
244	4.405302e-02	0
245	5.732372e-01	0
246	1.872147e-02	0
247	9.999692e-01	1
248	9.999890e-01	1
249	1.000000e+00	1
250	9.938555e-01	1
251	9.999648e-01	1
252	3.376572e-02	0
253	9.999968e-01	1
254	1.000000e+00	1
255	9.937819e-01	1
256	9.963339e-01	1
257	7.094846e-04	0
258	9.963339e-01	1
259	9.999962e-01	1
260	9.988448e-01	1
261	9.999962e-01	1
262	9.999962e-01	1
263	9.999989e-01	1
264	9.676350e-01	1
265	1.872147e-02	0
266	9.967909e-01	1
267	9.988448e-01	1
268	1.000000e+00	1
269	9.967909e-01	1
270	9.293548e-01	1
271	9.996806e-01	1
272	9.999890e-01	1
273	4.405302e-02	0
274	9.911687e-01	1
275	1.000000e+00	1
276	9.988448e-01	1
277	9.992270e-01	1
278	9.852502e-01	1
279	9.999837e-01	1
280	1.629632e-01	0
281	9.994053e-01	1
282	9.997305e-01	1
283	9.995581e-01	1
284	8.989972e-01	1
285	9.997863e-01	1
286	9.341173e-01	1
287	9.999999e-01	1
288	9.911687e-01	1
289	9.995597e-01	1
290	9.999103e-01	1
291	9.999997e-01	1
292	9.863752e-01	1
293	9.999997e-01	1
294	9.999968e-01	1
295	9.999890e-01	1
296	9.999600e-01	1
297	9.999794e-01	1
298	1.000000e+00	1
299	9.999937e-01	1
300	1.000000e+00	1
301	9.852502e-01	1
302	1.323060e-01	0
303	9.999600e-01	1
304	9.999600e-01	1
305	4.352627e-02	0
306	9.999937e-01	1
307	9.995581e-01	1
308	1.000000e+00	1
309	9.994053e-01	1
310	9.995971e-01	1
311	1.323060e-01	0
312	9.999995e-01	1
313	9.997837e-01	1
314	9.999794e-01	1
315	1.000000e+00	1
316	9.863752e-01	1
317	1.323060e-01	0
318	8.989972e-01	1
319	4.338301e-02	1
320	9.999648e-01	1

321	3.949348e-03	0
322	9.999937e-01	1
323	9.911687e-01	1
324	9.999692e-01	1
325	9.999968e-01	1
326	9.997837e-01	1
327	9.341173e-01	1
328	9.999890e-01	1
329	9.988448e-01	1
330	9.999890e-01	1
331	9.991111e-01	1
332	1.000000e+00	1
333	9.938555e-01	1
334	9.999999e-01	1
335	9.999600e-01	1
336	9.997837e-01	1
337	9.999991e-01	1
338	1.000000e+00	1
339	9.995971e-01	1
340	9.995971e-01	1
341	9.999733e-01	1
342	9.963339e-01	1
343	9.999989e-01	1
344	9.999968e-01	1
345	9.999995e-01	1
346	9.765477e-01	1
347	1.000000e+00	1
348	9.765477e-01	1
349	9.995581e-01	1
350	9.881129e-01	1
351	9.999962e-01	1
352	1.000000e+00	1
353	9.999937e-01	1
354	9.999995e-01	1
355	7.510085e-02	0
356	9.999428e-01	1
357	9.999888e-01	1
358	9.938555e-01	1
359	9.886146e-01	1
360	9.999648e-01	1
361	9.886146e-01	1
362	2.616644e-02	0
363	5.497919e-06	0
364	9.937819e-01	1
365	9.999890e-01	1
366	9.999962e-01	1
367	9.999962e-01	1
368	9.988448e-01	1
369	9.999937e-01	1
370	9.999969e-01	1
371	9.852502e-01	1
372	1.000000e+00	1
373	9.341173e-01	1
374	1.000000e+00	1
375	1.000000e+00	1
376	9.995581e-01	1
377	9.999890e-01	1
378	1.000000e+00	1
379	9.999428e-01	1
380	1.629632e-01	0
381	9.911687e-01	1
382	1.000000e+00	1
383	9.999103e-01	1
384	9.852502e-01	1
385	9.999890e-01	1
386	4.338301e-02	0
387	9.999837e-01	1
388	1.323060e-01	0
389	9.881129e-01	1
390	9.992270e-01	1
391	3.376572e-02	0
392	5.427849e-01	1
393	9.999837e-01	1
394	9.999557e-01	1
395	9.967909e-01	1
396	9.676350e-01	1
397	9.863752e-01	1
398	9.999890e-01	1
399	9.995581e-01	1
400	9.999428e-01	1
401	4.011349e-01	1
402	9.765477e-01	1
403	9.999557e-01	1
404	1.000000e+00	1
405	9.999692e-01	1
406	4.405302e-02	0
407	9.999907e-01	1
408	5.427849e-01	1
409	3.376572e-02	0
410	9.937819e-01	1
411	9.911687e-01	1
412	9.911687e-01	1
413	9.911687e-01	1
414	1.000000e+00	1
415	1.000000e+00	1
416	9.999428e-01	1
417	9.911687e-01	1
418	9.988448e-01	1
419	9.999794e-01	1
420	9.996142e-01	1
421	9.991111e-01	1
422	8.989972e-01	1
423	1.629632e-01	0
424	9.997863e-01	1
425	9.999890e-01	1
426	9.999969e-01	1
427	1.000000e+00	1
428	9.967909e-01	1
429	1.872147e-02	0
430	5.427849e-01	1
431	1.000000e+00	1
432	1.000000e+00	1
433	9.999907e-01	1
434	1.000000e+00	1
435	9.997863e-01	1
436	9.994053e-01	1
437	9.884790e-01	1
438	9.995971e-01	1
439	4.352627e-02	0
440	9.992270e-01	1
441	5.427849e-01	1
442	9.999890e-01	1
443	1.000000e+00	1
444	9.911687e-01	1
445	9.995597e-01	1
446	9.991111e-01	1
447	9.999999e-01	1
448	9.999969e-01	1
449	9.999557e-01	1
450	9.765477e-01	1
451	9.999991e-01	1
452	9.937819e-01	1
453	9.999974e-01	1
454	1.000000e+00	1
455	9.293548e-01	0
456	9.999692e-01	1
457	9.988448e-01	1
458	3.376572e-02	0
459	9.999890e-01	1
460	9.996142e-01	1
461	9.852502e-01	1
462	9.967909e-01	1
463	1.000000e+00	1
464	9.988448e-01	1
465	1.000000e+00	1
466	9.999962e-01	1
467	9.293548e-01	1
468	9.999997e-01	1
469	9.676350e-01	1
470	9.999974e-01	1
471	9.995971e-01	1
472	9.967909e-01	1
473	9.997863e-01	1
474	9.999648e-01	1
475	5.331770e-01	0
476	1.000000e+00	1
477	9.999962e-01	1
478	9.995597e-01	1
479	1.629632e-01	0
480	9.938555e-01	1
481	8.226549e-01	1
482	8.675393e-01	1
483	9.996806e-01	1
484	9.886146e-01	1
485	9.936755e-01	1
486	9.982747e-01	1
487	9.999980e-01	1
488	9.997837e-01	1
489	1.000000e+00	1
490	9.999103e-01	1
491	1.000000e+00	1
492	1.000000e+00	1
493	9.997863e-01	1
494	9.999888e-01	1
495	9.996806e-01	1
496	9.993981e-01	1
497	9.765477e-01	1
498	9.997837e-01	1
499	9.997305e-01	1
500	9.999103e-01	1
501	9.967909e-01	1
502	9.995581e-01	1
503	9.999997e-01	1
504	1.000000e+00	1
505	9.999888e-01	1
506	9.936755e-01	1
507	8.989972e-01	1
508	1.000000e+00	1
509	9.999890e-01	1
510	9.999999e-01	1
511	9.999733e-01	1
512	3.376572e-02	0

```
513 9.293548e-01  1
514 9.999969e-01  1
515 9.938555e-01  1
516 1.000000e+00  1
517 9.999557e-01  1
518 9.999907e-01  1
519 9.881129e-01  1
520 9.999888e-01  1
521 2.776191e-02  0
522 1.000000e+00  1
523 9.982747e-01  1
524 9.863752e-01  1
525 4.338301e-02  1
526 9.938555e-01  1
527 9.982747e-01  1
528 1.000000e+00  1
529 9.999989e-01  1
530 1.000000e+00  1
531 1.000000e+00  1
532 9.911687e-01  1
533 9.293548e-01  1
534 9.999997e-01  1
535 1.000000e+00  1
536 4.338301e-02  0
537 9.999890e-01  1
538 1.629632e-01  0
539 1.000000e+00  1
540 2.616644e-02  0
541 9.999974e-01  1
542 1.000000e+00  1
543 8.226549e-01  1
544 9.999794e-01  1
545 9.999962e-01  1
546 8.989972e-01  1
547 1.000000e+00  1
548 9.999991e-01  1
549 9.999962e-01  1
550 9.997863e-01  1
551 1.000000e+00  1
552 9.997305e-01  1
553 4.352627e-02  0
554 9.967909e-01  1
555 9.997837e-01  1
```

```
556 9.999890e-01  1
557 9.967909e-01  1
558 9.999962e-01  1
559 9.999907e-01  1
560 9.982747e-01  1
561 7.094846e-04  0
562 9.999410e-01  1
563 1.872147e-02  0
564 9.982747e-01  1
565 7.510085e-02  0
566 1.000000e+00  1
567 9.982747e-01  1
568 9.999410e-01  1
569 9.341173e-01  1
570 9.999999e-01  1
571 9.999907e-01  1
572 9.999557e-01  1
573 9.999890e-01  1
574 1.000000e+00  1
575 4.352627e-02  0
576 9.938555e-01  1
577 5.427849e-01  1
578 9.676350e-01  1
579 1.000000e+00  1
580 9.963339e-01  1
581 4.352627e-02  0
582 3.949348e-03  1
583 9.988448e-01  1
584 1.000000e+00  1
585 9.999837e-01  1
586 5.497919e-06  0
587 9.999995e-01  1
588 9.997863e-01  1
589 1.323060e-01  0
590 9.999837e-01  1
591 9.994053e-01  1
592 9.293548e-01  0
593 9.999999e-01  1
594 9.967909e-01  1
595 2.776191e-02  0
596 9.999890e-01  1
597 9.293548e-01  1
598 9.999692e-01  1
599 9.988448e-01  1
```

```
600 9.992270e-01  1
601 9.997837e-01  1
602 9.997863e-01  1
603 9.999410e-01  1
604 2.776191e-02  0
605 9.999837e-01  1
606 9.992270e-01  1
607 8.989972e-01  1
608 9.999428e-01  1
609 1.872147e-02  0
610 1.000000e+00  1
611 8.989972e-01  1
612 9.999980e-01  1
613 9.999600e-01  1
614 8.989972e-01  1
615 9.293548e-01  0
616 1.872147e-02  0
617 5.662348e-01  0
618 5.427849e-01  1
619 1.000000e+00  1
620 9.938555e-01  1
621 9.997863e-01  1
622 9.999890e-01  1
623 1.872147e-02  0
624 9.341173e-01  1
625 4.405302e-02  0
626 9.999733e-01  1
627 9.999989e-01  1
628 1.000000e+00  1
629 1.000000e+00  1
630 1.000000e+00  1
631 9.999600e-01  1
632 9.988448e-01  1
633 9.994053e-01  1
634 9.999997e-01  1
635 9.765477e-01  1
636 9.938555e-01  1
637 9.937819e-01  1
638 9.999991e-01  1
639 9.999991e-01  1
640 9.999989e-01  1
```

```
>p=predict(model,type ="response")
> accuracy <- table(p, db[,"KQ"])
> 1-sum(diag(accuracy))/sum(accuracy)
[1] 0.99375
```

Ảnh hưởng của biến đến tổng thể

```
>library(caret)
>varImp(logistic)
```

```
       Overall
TSV   3.684565
TTHT  2.424368
SK    2.247872
KQHT  2.912654
```

```
       Overall
STCN  4.785401
THLL  3.519961
VPNQ  2.737194
SDCKT 2.996731
LNG   2.685262
```

```
Start:  AIC=107.93
KQ ~ TSV + TTHT + SK + KQHT + STCN + THLL + VPNQ + SDCKT + LNG
```

	Df	Deviance	AIC
<none>		87.927	107.93
- SK	1	92.903	110.90
- TTHT	1	94.626	112.63
- LNG	1	95.714	113.71
- VPNQ	1	96.759	114.76
- SDCKT	1	96.949	114.95
- KQHT	1	99.397	117.40
- THLL	1	107.049	125.05
- TSV	1	117.429	135.43
- STCN	1	131.010	149.01

Kiểm tra độ chính xác của dữ liệu test

```
> pred = predict(logistic, newdata=test)
> accuracy= table(pred,test[,"KQ"])
> 1-sum(diag(accuracy)/sum(accuracy)
##[1]0.96
```

pred		
0.00144257866202232	0	1
0.00229138571754799	1	0
0.0187214728739465	3	0
0.0277619072480633	2	0
0.0337657155853704	1	0
0.0433830092862312	4	0
0.0435262734954131	1	0
0.0751008455157291	3	1
0.132305986237618	3	0
0.235339362295476	0	1
0.533176969515487	1	0
0.542784887402001	0	5
0.724192687916881	1	0
0.898997169295231	0	6
0.929354789565158	1	3
0.934117256124969	0	1
0.967635016942544	0	1
0.976547701280776	0	1
0.985250169210465	0	1
0.98637521790272	0	3
0.988614596860825	0	2
0.991168667738982	0	5
0.993855488440075	0	5
0.996790897639876	0	1
0.998274669544215	0	2
0.998844784857418	0	9
0.999111086464907	0	3
0.999227029138496	0	2
0.999274139567032	0	1
0.999405267792515	0	2
0.999558140223746	0	3
0.999559704688232	0	1
0.999614192819766	0	2
0.999634393748106	0	1
0.99968063756181	0	2
0.999730511704825	0	1
0.999783692460819	0	4
0.999786266621402	0	2
0.999910340654711	0	1
0.999942771355892	0	2
0.999964839561068	0	1
0.999968266027126	0	1
0.999969236663264	0	1
0.999978751605815	0	1
0.99998368143297	0	2
0.999988953649577	0	5
0.999990679183224	0	1
0.999996150716018	0	3
0.999996857299509	0	2

PHỤ LỤC 5
ĐÁNH GIÁ MÔ HÌNH

Đánh giá bằng ROC Curve

```
>pred= predict(logistic, type="response")
> r = roc(KQ,pred)
> auc(r)
Area under the curve: 0.9932
> ci(r)
95% CI: 0.9882-0.9981 (DeLong)
> library(Epi)
>ROC(form=KQ~TSV+TTHT+SK+KQHT+STCN+THLL+VPNQ+SDCKT+LNG,
  data =db)
```

```
Diem cat: 0.742
Sens: 97.8%
Spec: 95.8%
PV+: 11.7%
PV-: 0.7%
```

Đánh giá mô hình bằng phương pháp K-Fold

với k = 5

```
> library(plyr)
> library(randomForest)
> library(ROSE)
> k=5 #Folds
> data <- db
> id <- sample(1:k,nrow(data),replace=TRUE)
> list <- 1:k
> prediction <- data.frame()
> testsetCopy <- data.frame()
> progress.bar <- create_progress_bar("text")
> progress.bar$init(k)
> for (i in 1:k){
+   trainingset <- subset(data, id %in% list[-i])
+   trainingset <- ROSE(KQ ~ TSV + TTHT + SK + KQHT + STCN +
   THLL + VPNQ + SDCKT + LNG, data=trainingset, seed=3,
   p=0.15, N=length(trainingset$KQ))$data
+   testset <- subset(data, id %in% c(i))
+   mymodel <- randomForest(trainingset$KQ ~TSV + TTHT + SK
   + KQHT + STCN + THLL + VPNQ + SDCKT + LNG,
   data=trainingset, ntree=100)
+   temp <- as.data.frame(predict(mymodel, testset[,-1]))
+   prediction <- rbind(prediction, temp)
+   testsetCopy <- rbind(testsetCopy,
   as.data.frame(testset[,13]))
+   progress.bar$step()
+ }
> result <- cbind(prediction, testsetCopy)
> names(result) <- c("Predicted", "Actual")
> table(result)
```

	Actual	
Predicted	0	1
-2.06501482580279e-16	1	0
0.01	1	2
0.0149999999999998	2	0
0.0273333333333331	1	0
0.0279999999999997	1	0
0.0299999999999998	1	0
0.0324999999999998	3	0
0.0386666666666667	2	0
0.0416666666666665	2	0
0.042	1	0
0.0489999999999998	1	0
0.0499999999999999	0	1
0.05	1	0
0.0559999999999999	1	0
0.0566666666666667	3	0
0.057	1	0
0.065	2	0
0.0668333333333331	1	0
0.0689999999999998	1	0
0.0699999999999999	1	0
0.0728333333333332	1	0
0.0733333333333333	1	0
0.0759999999999999	1	0
0.0801666666666666	1	0
0.0841666666666664	2	0
0.0849999999999998	1	0
0.0999999999999998	1	0
0.1	1	0
0.108	2	0
0.11	1	0
0.114166666666666	2	0
0.114666666666667	2	0
0.118	2	0
0.129666666666667	3	0
0.13	1	0
0.134	2	0
0.136	2	0
0.140833333333333	1	0
0.147166666666667	1	0
0.148	1	0
0.154166666666667	1	0
0.1545	2	0
0.1585	2	0
0.161666666666666	1	0
0.177166666666667	2	0
0.179166666666667	2	0
0.2075	2	0
0.236	1	0
0.237833333333333	4	0
0.253333333333333	2	0
0.264333333333333	1	0
0.346333333333333	1	0
0.354666666666667	2	0
0.401	2	0
0.47	3	0
0.493666666666667	1	0
0.599666666666667	0	2
0.694	1	0
0.752166666666667	0	4
0.778333333333333	1	0
0.7835	0	5
0.803166666666667	1	0
0.852	1	0
0.857166666666667	0	1
0.858833333333333	1	0
0.867166666666667	0	1
0.873	0	2
0.879666666666667	0	3
0.887666666666667	0	2
0.902166666666667	0	1
0.904666666666667	0	2
0.919	1	0
0.926833333333333	0	4
0.932	1	0
0.932166666666667	0	2
0.937166666666667	0	1
0.944666666666667	0	3
0.95	0	6
0.951833333333333	0	3
0.953333333333333	0	4
0.958333333333333	0	1
0.962333333333333	0	3
0.964833333333333	0	3
0.966	0	2
0.968	0	5
0.97	0	2
0.972	0	1
0.973333333333334	0	2
0.978333333333333	0	2
0.98	0	5
0.985	0	1
0.986	0	1
0.986666666666667	1	2
0.988	0	2
0.99	1	33
0.9925	0	3
0.994166666666667	0	2
0.996	0	4
0.996666666666667	0	2
0.998	0	3
1	1	417

```
> result$Difference <- abs(result$Actual - result$Predicted)
> summary(result$Difference)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00000	0.00000	0.00000	0.04806	0.02167	1.00000

```
> mean(result$Difference)
[1] 0.04442037
Call:
 randomForest(formula = trainingset$KQ ~ TSV + TTHT + SK + KQHT +
   STCN + THLL + VPNQ + SDCKT + LNG, data = trainingset, ntree = 128)
               Type of random forest: regression
                     Number of trees: 128
No. of variables tried at each split: 3

          Mean of squared residuals: 0.0242581
                    % Var explained: 80.48
> trellis.par.set(caretTheme())
> plot(logistic)
```

Với k = 10

```
>  summary(result$Difference)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00000	0.00000	0.00000	0.04496	0.01562	1.00000

```
Call:
 randomForest(formula = trainingset$KQ ~ TSV + TTHT + SK + KQHT +
   STCN + THLL + VPNQ + SDCKT + LNG, data = trainingset, ntree = 64)
               Type of random forest: regression
                     Number of trees: 64
No. of variables tried at each split: 3

          Mean of squared residuals: 0.02127539
                    % Var explained: 83.42
```