HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

VŨ VĂN ĐAM

DỰ BÁO KHÁCH HÀNG RỜI MẠNG DỊCH VỤ FIBERVNN TẠI VNPT NAM ĐỊNH

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – NĂM 2025

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

VŨ VĂN ĐAM

DỰ BÁO KHÁCH HÀNG RỜI MẠNG DỊCH VỤ FIBERVNN TẠI VNPT NAM ĐỊNH

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH

MÃ SỐ: 8.48.01.01

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC:

TS. PHAN THỊ HÀ

HÀ NỘI – NĂM 2025

LỜI CAM ĐOAN

Tôi xin cam đoan:

1. Tôi xin cam đoan rằng tất cả nội dung và kết quả được trình bày trong đề án này là sản phẩm của chính tôi, được thực hiện trên cơ sở nghiên cứu, phân tích và đánh giá nghiêm túc dưới sự hướng dẫn tận tình của cô **TS. Phan Thị Hà**.

2. Tôi khẳng định không sao chép, biên soạn nội dung từ bất kỳ nguồn tài liệu nào khác mà không trích dẫn đầy đủ. Những thông tin, tài liệu tham khảo từ các nguồn khác đã được tôi trích dẫn rõ ràng tên tác giả, tên công trình, thời gian công bố trong đề án.

3. Nếu phát hiện bất kỳ hành vi sao chép không hợp lệ hoặc vi phạm quy định về đào tạo, tôi xin hoàn toàn chịu trách nhiệm trước nhà trường và các cơ quan có thẩm quyền.

Hà Nội, ngày 01 tháng 6 năm 2025

Học viên thực hiện đề án

Vũ Văn Đam

LỜI CẢM ƠN

Trước hết, tôi xin bày tỏ lòng biết ơn chân thành đến **Ban Giám hiệu nhà trường cùng Quý thầy, cô giáo Khoa Đào tạo Sau đại học** đã tận tình giảng dạy, truyền đạt kiến thức và tạo mọi điều kiện thuận lợi cho tôi trong suốt quá trình học tập và thực hiện đề án tốt nghiệp này.

Tôi xin bày tỏ lời cảm ơn sâu sắc đến **cô TS. Phan Thị Hà** — người đã tận tình hướng dẫn, đồng hành, hỗ trợ và đóng góp nhiều ý kiến quý báu giúp tôi hoàn thành tốt đề án.

Tôi cũng xin gửi lời cảm ơn đến **Viễn Thông Nam Định** đã tạo điều kiện thuận lợi cho tôi tham gia học tập, đồng thời hỗ trợ các vấn đề liên quan trong quá trình thực hiện đề án.

Cuối cùng, tôi xin chân thành cảm ơn **gia đình**, **người thân**, **bạn bè** và **đồng nghiệp** đã luôn quan tâm, động viên, ủng hộ tôi trong suốt thời gian học tập và nghiên cứu.

Nam Định, ngày 01 tháng 6 năm 2025

Học viên thực hiện đề án

Vũ Văn Đam

MỤC LỤC

LỜI CAM ĐOAN i
LỜI CẢM ƠN ii
MỤC LỤC iii
DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT iv
DANH SÁCH BẢNG v
DANH SÁCH HÌNH VẼ vi
MỞ ĐẦU 1
Chương 1. TỔNG QUAN VỀ BÀI TOÁN DỰ BÁO KHÁCH HÀNG RỜI MẠNG DỊCH VỤ 7
1.1. Tổng quan về VNPT Nam Định. 7
1.1.1. Giới thiệu. 7
1.1.2. Tổng quan về dịch vụ Internet tại Nam Định. 8
1.2. Tổng quan về học máy. 10
1.3. Nghiên cứu một số thuật toán học máy áp dụng cho bài toán dự báo 12
1.3.1. Học cây quyết định [3] 12
1.3.2. Thuật toán SVM (Support Vector Machine) [8] 25
Chương 2. PHƯƠNG PHÁP DỰ BÁO KHÁCH HÀNG RỜI MẠNG 28
2.1. Giới thiệu 28
2.2. Xây dựng mô hình dự báo khách hàng rời mạng. 29
2.2.1. Thu thập và tiền xử lý dữ liệu. 29
2.2.2. Huấn luyện và kiểm thử mô hình. 34
Chương 3: CÀI ĐẶT VÀ THỬ NGHIỆM 42
3.1. Cài đặt môi trường. 42
3.2. Cài đặt mô hình thử nghiệm 43
3.3. Thử nghiệm và đánh giá 44
KẾT LUẬN 47
DANH MỤC TÀI LIỆU THAM KHẢO 49

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
SVM	Support Vector Machine	Thuật toán máy vectơ hỗ trợ
DT	Decision Tree	Thuật toán Cây quyết định
FP	False Positive	Tỷ lệ sai dương
FN	False Negative	Tỷ lệ sai âm
TP	True Positive	Tỷ lệ đúng dương
TN	True Negative	Tỷ lệ đúng âm
ACC	Accuracy	Độ chính xác
B2A	Business-to-Administration	Phiếu khảo sát, chăm sóc khách hàng tại VNPT Nam Định

DANH SÁCH BẢNG

Bảng 1.1: Chi phí phát triển một khách hàng mới 9
Bảng 1.2 Bộ dữ liệu huấn luyện cho bài toàn phân loại "Chơi tennis". 16
Bảng 2.1: Danh sách đối tượng khách hàng.....31
Bảng 2.2: Trạng thái thuê bao ID.....32

DANH SÁCH HÌNH VẼ

Hình 1: Số lượng thuê bao phát triển mới, rời mạng 6 tháng đầu năm 2023 tại VNPT Nam Định. 2
Hình 2: VNPT Nam Định 7
Hình 3: Thị phần Internet tại Nam Định 9
Hình 4 : Các phương pháp học máy....12
Hình 5: Ví dụ cây quyết định cho bài toán "Chơi tennis" 14
Hình 6. Thuật toán xây dựng cây quyết định[3] 17
Hình 7. Ví dụ xây dựng cây quyết định 23
Hình 8: Sơ đồ tổng quan hệ thống....28
Hình 9. Data trích xuất từ hệ thống nội bộ....30
Hình 10: Xử lý dòng không hợp lệ 33
Hình 11: Xử lý điền giá trị thiếu. 33
Hình 12: Xử lý làm sạch tuổi 34
Hình 13 : Biến số....34
Hình 14: Biến phân loại 34
Hình 15: Sử dụng hàm setup trong thư viện PyCaret 36
Hình 16: Tạo mô hình Decision Tree....36
Hình 17: Kết quả huấn luyện mô hình Decision Tree....37
Hình 18: Tạo mô hình Support Vector Machine....37
Hình 19: Kết quả huấn luyện mô hình Support Vector Machine 37
Hình 20: So sánh giữa Decision Tree và Support Vector Machine 38
Hình 21: Kết quả so sánh giữa Decision Tree và Support Vector Machine. 38
Hình 22: Upload file CSV 38
Hình 23: Đọc file CSV 38
Hình 24: Xử lý dữ liệu. 38
Hình 25: Dự đoán bằng mô hình. 39
Hình 26: Ghép kết quả dự đoán vào dữ liệu gốc....39
Hình 27: So sánh với thực tế. 39

Hình 28: Hiển thị kết quả. 39
Hình 29 : Hiển thị tô màu dòng sai nếu có....39
Hình 30: Biểu đồ so sánh tỷ lệ đúng/sai....40
Hình 31 : Trả kết quả dưới dạng CSV hoặc Excel....40
Hình 32: Hiển thị thông báo lỗi....40
Hình 33: Chạy ứng dụng với Ngrok....44
Hình 34:Giao diện Ứng dụng dụng dự đoán khách hàng rời mạng dịch vụ FiberVNN 44
Hình 35: Chọn file huấn luyện. 45
Hình 36: Dữ liệu đầu vào. 45
Hình 37: Kết quả dự báo 46
Hình 38: So sánh dự báo rời mạng với thực tế 46
Hình 39: Biểu đồ dự báo đúng sai....46

MỞ ĐẦU

1. Lý do chọn đề tài

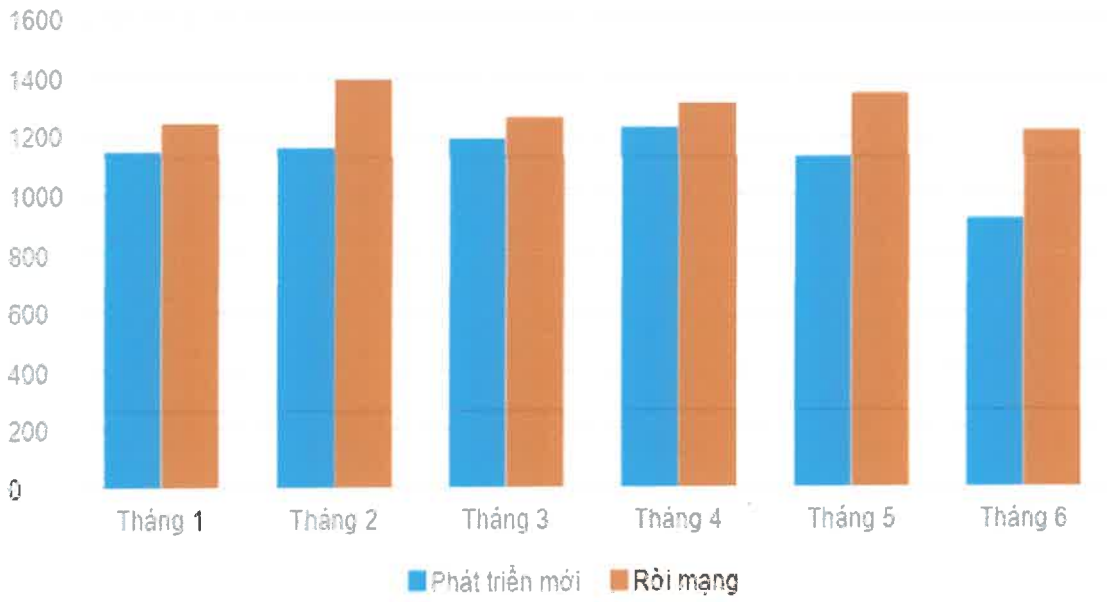
VNPT Nam Định là đơn vị thành viên của Tập đoàn Bưu chính Viễn thông Việt Nam (VNPT), chuyên cung cấp các dịch vụ viễn thông như FiberVNN, MyTV, điện thoại cố định, kênh truyền số liệu, MegaVNN, cùng các dịch vụ công nghệ thông tin gồm hệ thống quản lý văn bản điều hành, hệ thống quản lý bệnh viện, hệ thống quản lý công chức viên chức, hóa đơn điện tử, và nhiều dịch vụ khác. Hiện nay, VNPT Nam Định là một trong ba nhà cung cấp dịch vụ Internet lớn nhất trên địa bàn tỉnh Nam Định.

Trong bối cảnh thị trường viễn thông gần như đã bước vào giai đoạn bão hòa, việc phát triển thuê bao mới gặp nhiều khó khăn khi phần lớn khách hàng mới đến từ nhóm chuyển đổi từ các nhà mạng đối thủ. Do đó, chi phí để phát triển một khách hàng mới thường cao hơn đáng kể so với chi phí cần thiết để duy trì và giữ chân khách hàng hiện tại. Vì vậy, công tác chăm sóc và giữ chân khách hàng luôn được các nhà mạng, trong đó có VNPT Nam Định, xác định là nhiệm vụ trọng tâm.

Theo báo cáo thống kê của VNPT Nam Định [5], tính đến ngày **30/11/2023**, số lượng thuê bao đang sử dụng dịch vụ Internet FiberVNN đạt 133.405 thuê bao. Số liệu phát triển thuê bao và lượng thuê bao rời mạng trong sáu tháng đầu năm 2023 như sau:

- Tháng 1: phát triển mới 1.148 thuê bao, rời mạng 1.247 thuê bao.
- Tháng 2: phát triển mới 1.161 thuê bao, rời mạng 1.396 thuê bao.
- Tháng 3: phát triển mới 1.192 thuê bao, rời mạng 1.265 thuê bao.
- Tháng 4: phát triển mới 1.231 thuê bao, rời mạng 1.312 thuê bao.
- Tháng 5: phát triển mới 1.128 thuê bao, rời mạng 1.346 thuê bao.
- Tháng 6: phát triển mới 916 thuê bao, rời mạng 1.218 thuê bao.

Những con số trên cho thấy tình trạng thuê bao rời mạng có xu hướng cao hơn hoặc xấp xỉ số lượng thuê bao phát triển mới, đặt ra thách thức lớn đối với công tác chăm sóc khách hàng và duy trì thuê bao của VNPT Nam Định trong thời gian tới.



Hình 1: Số lượng thuê bao phát triển mới, rời mạng 6 tháng đầu năm 2023 tại VNPT Nam Định[5].

Trong 6 tháng đầu năm 2023, tỷ lệ thuê bao rời mạng dịch vụ FiberVNN tại VNPT Nam Định luôn cao hơn tỷ lệ thuê bao phát triển mới. Thực trạng này đặt ra những thách thức không nhỏ cho công tác duy trì và phát triển thuê bao của đơn vị.

Trước tình hình đó, lãnh đạo VNPT Nam Định đã triển khai nhiều giải pháp cả về nghiệp vụ lẫn chăm sóc khách hàng nhằm giữ chân và phát triển thuê bao. Cụ thể, đơn vị đã thường xuyên tổ chức các chương trình đào tạo nội bộ cho đội ngũ cán bộ, kỹ thuật viên, nhân viên kinh doanh về kỹ năng chăm sóc khách hàng và nâng cao chất lượng dịch vụ. Định kỳ hàng tuần, VNPT Nam Định thực hiện thống kê, báo cáo số lượng khách hàng đã rời mạng và gửi về các đơn vị trực thuộc để triển khai các biện pháp chăm sóc, thuyết phục khách hàng quay lại sử dụng dịch vụ. Bên cạnh đó, hàng tháng đơn vị còn giao chỉ tiêu số lượng phiếu B2A cho nhân viên kỹ thuật tại các địa bàn để khảo sát, ghi nhận ý kiến phản hồi của khách hàng về chất lượng dịch vụ mà VNPT cung cấp.

Tuy nhiên, các giải pháp trên vẫn mang tính thủ công, chưa tận dụng được các công cụ phân tích dữ liệu hiện đại nhằm dự đoán sớm và chính xác các khách hàng có nguy cơ rời mạng. Xuất phát từ thực tế đó, trong đề án này tôi lựa chọn áp dụng

các phương pháp học máy để phân tích và dự báo khả năng rời mạng của khách hàng sử dụng dịch vụ FiberVNN tại VNPT Nam Định. Đây cũng chính là lý do tôi chọn đề tài nghiên cứu:

" Dự báo khách hàng rời mạng dịch vụ FiberVNN tại VNPT Nam Định"

2. Tổng quan về vấn đề cần nghiên cứu

Để có thể dự đoán chính xác, hoặc đạt tỷ lệ chính xác cao về khách hàng có nguy cơ rời mạng dịch vụ FiberVNN tại VNPT Nam Định, đề án tập trung nghiên cứu và phân tích các thuộc tính, hành vi của khách hàng có khả năng tác động đến quyết định rời mạng. Các yếu tố điển hình được xem xét bao gồm: số lần báo hỏng dịch vụ, tình trạng nợ cước thanh toán, các phản ánh của khách hàng về chất lượng dịch vụ, cũng như phản hồi về thái độ và chất lượng phục vụ của đội ngũ nhân viên chăm sóc khách hàng tại địa bàn. Trên cơ sở đó, đề án hướng tới việc đề xuất các giải pháp chăm sóc khách hàng phù hợp nhằm gia tăng sự hài lòng và khuyến khích khách hàng tiếp tục sử dụng dịch vụ của đơn vị.

Sau khi xác định và phân tích các trường thông tin có khả năng ảnh hưởng đến nguyên nhân khách hàng rời mạng, đề án tiến hành nghiên cứu về các phương pháp học máy (machine learning) và lựa chọn các thuật toán học máy phù hợp để áp dụng cho bài toán dự báo. Các thuật toán này sẽ được sử dụng để huấn luyện và xây dựng mô hình dự báo khả năng rời mạng của khách hàng.

Bên cạnh đó, đề án cũng tham khảo và kế thừa những kết quả nghiên cứu khoa học đã được công bố trong nước và quốc tế về việc ứng dụng các mô hình học máy cho bài toán dự báo khách hàng rời mạng trong lĩnh vực viễn thông, cũng như trong các lĩnh vực khác có liên quan, nhằm bổ sung cơ sở lý thuyết và hoàn thiện mô hình nghiên cứu.

2.1. Các nghiên cứu trong nước

Một số luận văn và đề tài nghiên cứu trước đây đã áp dụng các thuật toán học máy để dự báo khả năng rời mạng của khách hàng:

- Tiêu biểu là luận văn tốt nghiệp thạc sĩ kỹ thuật chuyên ngành Hệ thống thông tin của tác giả Dương Minh Lý (năm 2023), thực hiện tại Học viện Công nghệ Bưu chính Viễn thông - Cơ sở TP. Hồ Chí Minh, với tên đề tài:

"Dự báo khách hàng sử dụng dịch vụ FiberVNN của VNPT Tây Ninh có nguy cơ rời mạng".

Trong nghiên cứu này, tác giả đã sử dụng phần mềm Weka để thực hiện tiền xử lý dữ liệu và phần mềm Azure Machine Learning để xây dựng và huấn luyện mô hình dự báo. Sau khi nghiên cứu, so sánh một số thuật toán học máy và căn cứ vào đặc điểm dữ liệu thực tế của VNPT Tây Ninh, tác giả lựa chọn thuật toán Two-Class Boosted Decision Tree để xây dựng mô hình dự báo. Dữ liệu được chia thành hai tập: 80% dữ liệu để huấn luyện và 20% dữ liệu để kiểm thử. Mô hình dự báo thu được kết quả khả quan với các chỉ số: accuracy đạt 99,5%, recall đạt 98,7%, precision đạt 99,4% và F1-score đạt 99,0%.

- Bên cạnh đó, trong năm 2022, nhóm tác giả gồm Võ Đức Vinh (VNPT Đồng Nai) và Trần Văn Lăng (Trường Đại học Ngoại ngữ - Tin học TP. Hồ Chí Minh) đã công bố bài báo khoa học với tiêu đề "Dự báo khách hàng thuê bao rời mạng dịch vụ Fiber" trên Tạp chí Khoa học HUFLIT [4]. Bài báo trình bày việc áp dụng các công cụ trong lĩnh vực học máy, kết hợp với thuật toán cây quyết định để xây dựng mô hình phân tích dữ liệu, nhằm dự báo sớm các thuê bao có nguy cơ rời mạng. Mô hình được xây dựng trên cơ sở khai thác dữ liệu lịch sử về các thuộc tính được xác định là nguyên nhân dẫn đến sự rời mạng của thuê bao tại VNPT Đồng Nai. Kết quả thực nghiệm cho thấy mô hình có khả năng dự báo khách hàng có nguy cơ rời mạng với tỷ lệ chính xác rất cao so với số liệu thực tế, khẳng định tính khả thi và hiệu quả của phương pháp tiếp cận này.

- Một nghiên cứu khác tiêu biểu là luận văn thạc sĩ của tác giả Nguyễn Thị Như Ngọc (năm 2014), thực hiện tại Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội, với đề tài:

"Phân tích dữ liệu thuê bao di động hướng đến dự báo thuê bao rời mạng viễn thông" [2].

Trong nghiên cứu này, tác giả đã áp dụng các thuật toán học máy gồm C4.5 (Decision Tree), Naïve Bayes, SVM và Neural Networks để xây dựng mô hình phân lớp dự đoán khả năng rời mạng của các thuê bao viễn thông. Kết quả thực nghiệm cho thấy, mô hình phân lớp sử dụng các thuật toán trên đạt độ chính xác dự báo khoảng hơn 60%, qua đó phản ánh những thách thức của bài toán trong điều kiện dữ liệu và công cụ xử lý tại thời điểm nghiên cứu.

2.2. Các nghiên cứu ngoài nước.

- Một nghiên cứu quốc tế tiêu biểu khác là bài báo "Churn Prediction in the Telecommunications Sector Using Support Vector Machines" được công bố năm 2013 bởi nhóm tác giả Ionut Brandusoiu và Gavril Toderean thuộc Technical University of Cluj-Napoca [6]. Bài báo trình bày một phương pháp tiên tiến nhằm dự đoán khả năng rời mạng của khách hàng trong ngành viễn thông di động. Tập dữ liệu được sử dụng trong nghiên cứu bao gồm các bản ghi chi tiết cuộc gọi, với 3333 bản ghi và 21 thuộc tính cho mỗi bản ghi. Mô hình dự báo được xây dựng dựa trên thuật toán Support Vector Machines (SVM) với bốn loại hàm nhân khác nhau. Hiệu suất của các mô hình được đánh giá và so sánh thông qua chỉ số gain measure, cho thấy tính hiệu quả và tiềm năng ứng dụng của phương pháp trong thực tế.

3. Mục đích nghiên cứu

Đề án tập trung nghiên cứu, thu thập và khai thác dữ liệu khách hàng đã rời mạng để làm cơ sở phân tích, xây dựng công cụ dự báo khách hàng có nguy cơ rời mạng. Dữ liệu được thu thập, truy xuất từ phần mềm quản lý nội bộ của VNPT Nam Định đến thời điểm 31/03/2025 nhằm đảm bảo tính đầy đủ và cập nhật.

Trên cơ sở dữ liệu đã thu thập, đề án tiến hành nghiên cứu các phương pháp học máy và lựa chọn những thuật toán học máy phù hợp có thể áp dụng cho bài toán dự báo khách hàng rời mạng. Mô hình dự báo được xây dựng dựa trên tập dữ liệu khách hàng hiện có tại đơn vị, với mục tiêu phát hiện sớm những khách hàng có khả năng rời mạng.

Kết quả dự báo được cung cấp định kỳ cho bộ phận chăm sóc khách hàng và nhân viên phụ trách địa bàn để từ đó chủ động triển khai các biện pháp chăm sóc, thuyết phục và giữ chân khách hàng, góp phần ổn định và phát triển thuê bao của đơn vị.

4. Đối tượng và phạm vi nghiên cứu

- Đối tượng nghiên cứu của đề án là tập khách hàng hiện đang sử dụng dịch vụ FiberVNN và tập khách hàng đã ngừng sử dụng dịch vụ (rời mạng) tại VNPT Nam Định. Nhóm khách hàng này có đầy đủ dữ liệu về quá trình sử dụng dịch vụ, hành vi và các đặc điểm liên quan, qua đó tạo thành nguồn dữ liệu chính để phân tích và xây dựng mô hình dự báo khả năng rời mạng
- Phạm vi nghiên cứu giới hạn trong khách hàng đã và đang sử dụng dịch vụ FiberVNN của VNPT Nam Định, khai thác dữ liệu từ hệ thống quản lý nội bộ của đơn vị. Đề án chỉ tập trung vào dịch vụ FiberVNN và khu vực tỉnh Nam Định, không xem xét các dịch vụ viễn thông hoặc công nghệ thông tin khác của VNPT Nam Định.

5. Phương pháp nghiên cứu

- Nghiên cứu lý thuyết: Tìm hiểu cơ sở lý thuyết về học máy (Machine learning) và các thuật toán học máy có thể áp dụng cho bài toán dự báo khách hàng rời mạng. Đề án tập trung nghiên cứu hai thuật toán chính gồm thuật toán Cây quyết định (Decision Tree) và thuật toán Máy vector hỗ trợ (Support Vector Machine - SVM) nhằm xác định thuật toán tối ưu cho việc xây dựng mô hình dự báo.
- Phương pháp thực nghiệm: Xây dựng mô hình dự báo khách hàng rời mạng dựa trên hai thuật toán đã nghiên cứu. Mô hình được huấn luyện và đánh giá trên tập dữ liệu nội bộ của VNPT Nam Định, nhằm kiểm chứng khả năng ứng dụng thực tiễn của các thuật toán trong dự báo khách hàng có nguy cơ rời mạng.

Chương 1. TỔNG QUAN VỀ BÀI TOÁN DỰ BÁO KHÁCH HÀNG RỜI MẠNG DỊCH VỤ

Trong chương này, đề án sẽ giới thiệu tổng quan về VNPT Nam Định, tình hình cung cấp dịch vụ Internet của VNPT cũng như của các nhà cung cấp dịch vụ khác trên địa bàn tỉnh Nam Định. Bên cạnh đó, chương này cũng trình bày cơ sở lý thuyết về học máy (machine learning) và nghiên cứu một số thuật toán học máy có thể áp dụng cho bài toán dự báo khách hàng rời mạng. Đây sẽ là nền tảng quan trọng phục vụ cho việc xây dựng mô hình dự báo trong các chương tiếp theo của đề án.

1.1. Tổng quan về VNPT Nam Định.

1.1.1. Giới thiệu.

VNPT Nam Định là đơn vị hạch toán phụ thuộc Tập đoàn Bưu chính Viễn thông Việt Nam (VNPT), tiền thân là Bưu điện tỉnh Nam Định được thành lập từ năm 1945. Đến tháng 01 năm 2008, Bưu điện tỉnh Nam Định được chia tách thành hai đơn vị: Viễn thông Nam Định và Bưu điện Nam Định.

Hình 2: VNPT Nam Định

VNPT Nam Định có chức năng nhiệm vụ kinh doanh và cung cấp các dịch vụ viễn thông - công nghệ thông tin (VT - CNTT) trên địa bàn tỉnh, phục vụ nhu cầu

thông tin liên lạc của các cấp ủy Đảng, chính quyền, cơ quan, đoàn thể, doanh nghiệp và nhân dân tỉnh Nam Định.

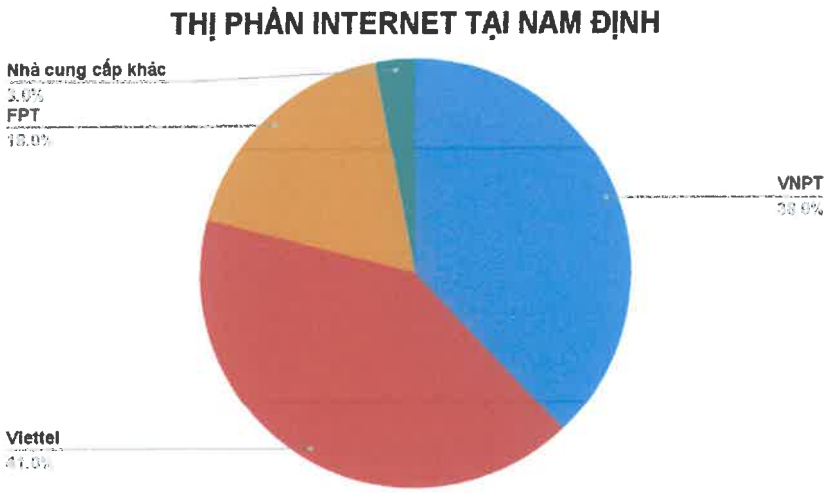
Lĩnh vực kinh doanh chính của VNPT Nam Định bao gồm:

- Cung cấp các dịch vụ viễn thông: dịch vụ điện thoại cố định, điện thoại di động, Internet băng rộng (Home Internet, FiberVNN), dịch vụ truyền hình MyTV.
- Cung cấp các dịch vụ công nghệ thông tin: VNPT-CA, VNPT-IVAN, sổ liên lạc điện tử vnEdu, VNPT-iOffice, VNPT-iGate, dịch vụ tin nhắn quảng bá SMS Brandname...
- Tổ chức xây dựng, quản lý, vận hành, lắp đặt, khai thác, bảo dưỡng, sửa chữa mạng viễn thông trên địa bàn tỉnh.
- Kinh doanh và cung cấp các dịch vụ viễn thông, CNTT theo nhu cầu của khách hàng.
- Sản xuất kinh doanh, cung ứng, làm đại lý vật tư, thiết bị viễn thông - CNTT phục vụ sản xuất kinh doanh của đơn vị và nhu cầu thị trường.
- Kinh doanh dịch vụ quảng cáo, dịch vụ truyền thông.
- Tổ chức phục vụ thông tin đột xuất theo yêu cầu của cấp ủy Đảng, chính quyền địa phương và cấp trên.
- Kinh doanh các ngành nghề khác khi được Tập đoàn Bưu chính Viễn thông Việt Nam cho phép.

1.1.2. Tổng quan về dịch vụ Internet tại Nam Định.

Theo báo cáo thống kê của VNPT Nam Định, tính đến cuối năm 2023, trên địa bàn tỉnh Nam Định có ba nhà cung cấp lớn trong lĩnh vực dịch vụ Internet cáp quang, gồm: VNPT, Viettel và FPT. Thị phần cụ thể của từng nhà cung cấp như sau:

- VNPT chiếm khoảng 38% thị phần
- Viettel chiếm khoảng 41% thị phần
- FPT chiếm khoảng 18% thị phần
- Các nhà cung cấp khác chiếm khoảng 3% thị phần trên toàn tỉnh.



Hình 3: Thị phần Internet tại Nam Định

Thị trường Internet tại Nam Định hiện nay gần như đã bước vào giai đoạn bão hòa, số lượng khách hàng mới phát triển thêm chủ yếu là do chuyển đổi từ nhà cung cấp khác sang. Trong bối cảnh đó, việc giữ chân khách hàng không chỉ góp phần nâng cao uy tín và thương hiệu của nhà cung cấp, mà còn mang lại hiệu quả kinh tế rõ rệt, bởi chi phí để phát triển một khách hàng mới thường cao hơn đáng kể so với chi phí duy trì, chăm sóc một khách hàng hiện hữu..

Bảng 1.1: Chi phí phát triển một khách hàng mới

STT	Loại chi phí	Chi phí
1	Chi phí lắp đặt + hoa hồng	- Lắp đặt: 100.000đ /1 Khách hàng - Hoa hồng: 100.000đ/1 Khách hàng
2	Dây cáp quang (50 -100 mét)	400.000đ / 1 cuộn (1000 mét)
3	Dây mạng LAN(10 mét)	5.000đ/ 1 mét
4	Modem	800.000đ - 1.200.000đ
5	Chi phí hòa mạng	300.000đ
	Tổng chi phí	**1.370.000đ**

Theo tính toán thực tế, chi phí tối thiểu để lắp đặt và hoàn công một khách

hàng mới tại VNPT Nam Định hiện nay vào khoảng 1.370.000 đồng. Trong khi đó, nếu áp dụng các chính sách ưu đãi, chẳng hạn như giảm giá cước cho khách hàng sử dụng lâu năm (ví dụ: giảm cước 150.000 đồng/tháng trong 3 tháng liên tiếp), thì tổng chi phí vẫn thấp hơn đáng kể so với chi phí đầu tư để phát triển một khách hàng mới.

Nhận thức được tầm quan trọng của việc giữ chân khách hàng, lãnh đạo VNPT Nam Định đã triển khai nhiều giải pháp nhằm nâng cao chất lượng dịch vụ và chăm sóc khách hàng tốt hơn. Một số giải pháp tiêu biểu hiện đang được áp dụng tại đơn vị gồm:

- Triển khai cập nhật phiếu B2A: Giao định kỳ hàng tháng cho nhân viên kỹ thuật địa bàn đến từng nhà khách hàng để chăm sóc, lấy ý kiến phản ánh của khách hàng về dịch vụ mà VNPT đang cung cấp. Bên cạnh đó sẽ thu thập, cập nhật vào hệ thống điều hành nội bộ để có kế hoạch kịp thời xử lý những thắc mắc mà khách hàng phản ánh.
- Chú trọng chăm sóc các nhóm khách hàng có dấu hiệu rời mạng: Đặc biệt là những khách hàng thường xuyên báo hỏng, nợ cước, không phát sinh lưu lượng sử dụng hoặc đang ở trạng thái tạm khóa, tạm ngừng dịch vụ. Tuy nhiên, trên thực tế, phần lớn các khách hàng thuộc nhóm này đã rời mạng, do đó việc thuyết phục họ quay lại sử dụng dịch vụ gặp nhiều khó khăn.

Trên cơ sở dữ liệu đã thu thập được về các khách hàng rời mạng với các thuộc tính liên quan như số lần báo hỏng, suy hao tín hiệu cao, xử lý sự cố quá hạn..., đề án này tập trung nghiên cứu và ứng dụng học máy (machine learning) để phân tích, xây dựng mô hình dự báo khách hàng có nguy cơ rời mạng tại VNPT Nam Định, từ đó hỗ trợ công tác chăm sóc và giữ chân khách hàng một cách chủ động, hiệu quả hơn.

1.2. Tổng quan về học máy.

Học máy (Machine Learning) [3] là khả năng của chương trình máy tính tự cải thiện hiệu quả công việc của mình trong tương lai thông qua kinh nghiệm, quan sát hoặc dữ liệu trong quá khứ, thay vì chỉ thực hiện các thao tác theo đúng các quy tắc đã được lập trình sẵn. Ví dụ, trong việc học chơi cờ, chương trình có thể quan sát các

ván cờ cùng với kết quả thắng - thua để cải thiện khả năng đánh cờ và gia tăng tỉ lệ thắng trong những ván cờ tiếp theo. Ở đây, kinh nghiệm chính là quá trình học hỏi từ các ván cờ trong quá khứ để nâng cao hiệu quả chơi cờ, với tiêu chí đánh giá là số ván thắng.

Các phương pháp học máy:

Học máy được chia thành ba nhóm phương pháp chính:

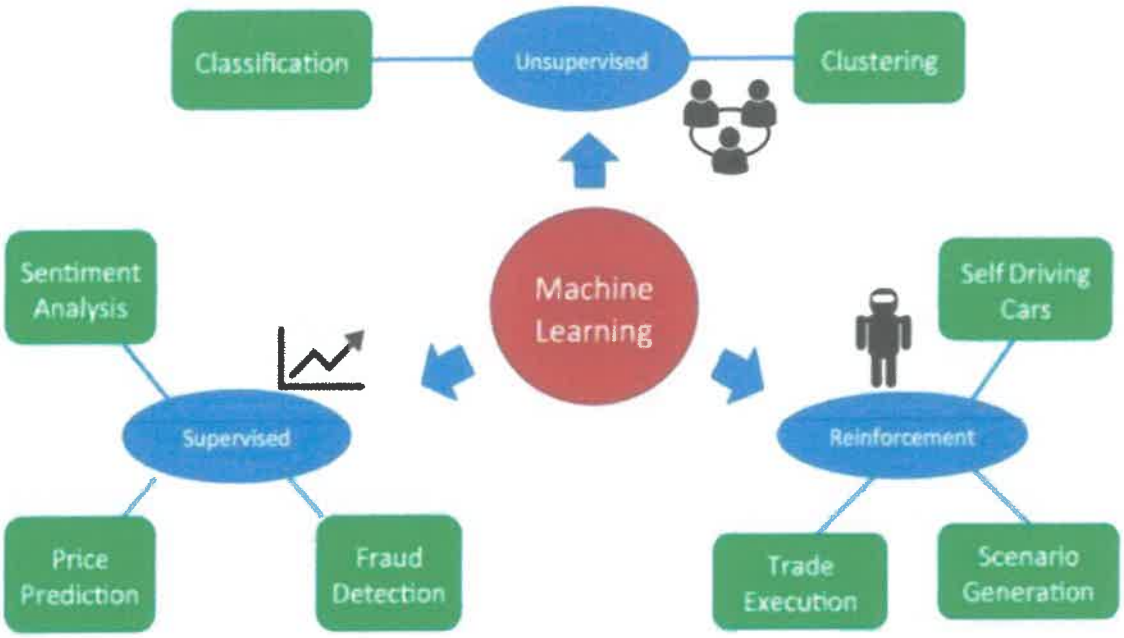
- Học có giám sát (supervised learning): Đây là phương pháp mà tập dữ liệu huấn luyện được cung cấp dưới dạng các ví dụ kèm theo giá trị đầu ra hoặc giá trị đích. Mục tiêu của thuật toán là xây dựng một mô hình (hoặc hàm đích) từ dữ liệu huấn luyện để có thể dự đoán chính xác giá trị đầu ra cho các dữ liệu mới.

  - Nếu giá trị đầu ra là rời rạc, bài toán được gọi là phân loại (classification)
  - Nếu giá trị đầu ra là liên tục (số thực), bài toán được gọi là hồi quy (regression).

- Học không giám sát (un-supervised learning): Ở phương pháp này, tập dữ liệu chỉ bao gồm các ví dụ mà không có giá trị đầu ra đi kèm. Thuật toán học máy dựa trên độ tương tự giữa các ví dụ để phân nhóm chúng thành các cụm (cluster) sao cho các ví dụ trong cùng một cụm có tính chất tương đồng.

  - Một hình thức phổ biến là phân cụm (clustering), ví dụ dựa vào chiều cao, con người có thể được nhóm thành "người cao" và "người thấp".
  - Ngoài Một hình thức khác là khai thác luật kết hợp (association rule mining). Ví dụ, phân tích dữ liệu mua hàng siêu thị có thể tìm ra luật kết hợp như: P(Bơ | Bánh mì) = 80%, nghĩa là 80% những người mua bánh mì cũng mua bơ.

- Học tăng cường (reinforcement learning): Trong phương pháp này, hệ thống không được cung cấp trực tiếp đầu ra đúng cho mỗi tình huống mà phải học thông qua tương tác với môi trường. Hệ thống nhận được các phần thưởng (reward) sau mỗi chuỗi hành động và phải học cách lựa chọn hành động sao cho tổng phần thưởng nhận được là lớn nhất. Ví dụ, trong trò chơi cờ vua, hệ thống không được chỉ dẫn nước đi đúng trong từng trường hợp cụ thể mà chỉ biết kết quả cuối cùng (thắng hay thua) của cả ván cờ, từ đó tự học chiến lược chơi hiệu quả.



Hình 4 : Các phương pháp học máy.

1.3. Nghiên cứu một số thuật toán học máy áp dụng cho bài toán dự báo.

1.3.1. Học cây quyết định [3]

Trong phần này, đề tài tập trung nghiên cứu kỹ thuật học máy cụ thể là học cây quyết định. Đây là một phương pháp học có giám sát, được ứng dụng rộng rãi trong các bài toán phân loại và dự đoán. Mặc dù độ chính xác của cây quyết định không luôn cao bằng một số phương pháp hiện đại khác nhưng phương pháp này vẫn được đánh giá cao nhờ tính đơn giản, dễ triển khai, cũng như khả năng giải thích kết quả một cách trực quan và dễ hiểu đối với người sử dụng.

Cây quyết định cho phép xây dựng mô hình phân loại dưới dạng cấu trúc cây, trong đó mỗi nút bên trong đại diện cho một thuộc tính kiểm tra, mỗi nhánh thể hiện kết quả của phép kiểm tra, và mỗi lá cây tương ứng với một nhãn phân loại. Mô hình này thực chất là một xấp xỉ của một hàm phân loại với đầu ra rời rạc. Nhờ đặc điểm dễ minh họa và dễ tiếp cận, phương pháp học cây quyết định thường được sử dụng như bước khởi đầu để giải thích nguyên lý học bộ phân loại từ dữ liệu.

Trong khuôn khổ đề tài, một số thuật toán tiêu biểu trong học cây quyết định được giới thiệu, bao gồm ID3 và C4.5. Đây là những thuật toán nền tảng, được sử dụng phổ biến để xây dựng cây quyết định dựa trên các tiêu chí như Information Gain

hoặc Gain Ratio, nhằm chọn ra thuộc tính phân tách tốt nhất tại mỗi nút cây.

Khái niệm cây quyết định

Cây quyết định (*decision tree*) là một cấu trúc ra quyết định có dạng cây, được sử dụng rộng rãi trong các bài toán phân loại và hồi quy trong lĩnh vực học máy. Cây quyết định nhận đầu vào là một bộ giá trị thuộc tính mô tả một đối tượng hoặc một tình huống, và trả về một giá trị rời rạc, gọi là nhãn phân loại. Mỗi bộ giá trị thuộc tính đầu vào được gọi là một mẫu (*sample*) hoặc một ví dụ (*example*), trong khi đầu ra chính là nhãn phân loại (*label*).

Các thuộc tính đầu vào (còn gọi là đặc trưng, *feature*) có thể nhận các giá trị rời rạc hoặc liên tục. Để đơn giản, trước tiên thường xét trường hợp thuộc tính rời rạc, sau đó mở rộng cho các thuộc tính liên tục.

Trong các phần trình bày tiếp theo, tập thuộc tính đầu vào được ký hiệu dưới dạng véc tơ **x**, nhãn phân loại đầu ra được ký hiệu là y, và cây quyết định được xem như một hàm ánh xạ: $\mathbf{f(x)=y}$

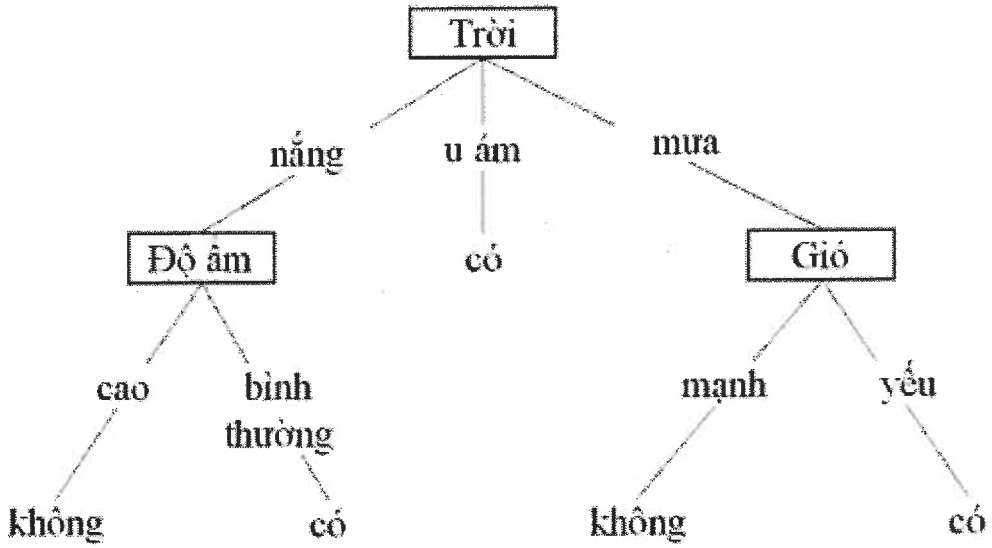
Cây quyết định được biểu diễn dưới dạng một cấu trúc cây. Mỗi nút trung gian (nút không phải nút lá) tương ứng với một phép kiểm tra giá trị của một thuộc tính nào đó. Mỗi nhánh phía dưới nút trung gian tương ứng với một giá trị cụ thể của thuộc tính, hoặc kết quả của phép kiểm tra tại nút đó. Ngược lại, các nút lá không chứa thuộc tính mà lưu trữ nhãn phân loại.

Để xác định nhãn phân loại cho một mẫu cụ thể, mẫu đó sẽ "di chuyển" từ nút gốc của cây xuống các nhánh dựa trên giá trị của các thuộc tính, cho đến khi đến được một nút lá. Nhãn phân loại được gán cho mẫu chính là nhãn tại nút lá mà mẫu đó đi tới.

Ví dụ, cây quyết định minh họa trong Hình 5 được xây dựng từ bộ dữ liệu mô tả trong bảng 1.2. Cây này dùng để phân loại các buổi sáng thành "phù hợp" hoặc "không phù hợp" cho việc chơi tennis, dựa trên các điều kiện thời tiết trong ngày. Thời tiết được mô tả thông qua bốn thuộc tính: *Trời* (Outlook), *Nhiệt độ* (Temperature), *Độ ẩm* (Humidity), và *Gió* (Wind).

Cấu trúc cây quyết định giúp dễ dàng suy luận và giải thích các quyết định

phân loại, đồng thời trực quan hóa các quy tắc phân loại dựa trên giá trị thuộc tính. Đây cũng là một trong những lý do khiến cây quyết định được ứng dụng rộng rãi trong thực tiễn.



Hình 5: Ví dụ cây quyết định cho bài toán "Chơi tennis" [3]

Giả sử ta có ví dụ < Trời = mưa, Gió = yếu> => cây quyết định xếp xuống nút ngoài cùng bên phải và do vậy được xác định là "có chơi".

Biểu diễn tương đương dưới dạng biểu thức logic

Cây quyết định không chỉ có thể được biểu diễn dưới dạng cấu trúc cây mà còn có thể được biểu diễn tương đương dưới dạng các quy tắc hoặc biểu thức logic. Cụ thể, cây quyết định có thể được viết dưới dạng:

$$Cây\_Quyết\_Định(x) \Leftrightarrow (P1(x) \vee P2(x) \vee \cdots \vee Pn(x))$$

Trong đó mỗi $P_i(x)$ là một biểu thức hội (*conjunction*) các phép thử thuộc tính tương ứng với một đường đi từ nút gốc đến nút lá mà tại nút lá đó nhãn phân loại có giá trị dương (hoặc "đúng", *true*).

Ví dụ, cây quyết định minh họa trong Hình 5 có thể được biểu diễn tương đương dưới dạng biểu thức logic như sau:

(Trời=nắng∧Độ ẩm=bình thường) V (Trời=u ám) V (Trời=mưa∧Gió=yếu)

Ngoài ra, cây quyết định cũng có thể được biểu diễn dưới dạng các luật suy

diễn kiểu "Nếu... Thì...", vốn gần gũi và dễ hiểu đối với con người trong quá trình ra quyết định và phân loại. Cụ thể:

- Nếu (*Trời = nắng*) và (*Độ ẩm = bình thường*) thì *có chơi.*
- Nếu (*Trời = nắng*) và (*Độ ẩm = cao*) thì *không chơi.*
- Nếu (*Trời = u ám*) thì *có chơi.*
- Nếu (*Trời = mưa*) và (*Gió = mạnh*) thì *không chơi.*
- Nếu (*Trời = mưa*) và (*Gió = yếu*) thì *có chơi.*

Việc biểu diễn cây quyết định dưới dạng các quy tắc dạng này không chỉ giúp đơn giản hóa quá trình suy luận mà còn làm cho mô hình trở nên dễ giải thích và dễ kiểm chứng. Đây là một ưu điểm quan trọng của cây quyết định khi được ứng dụng trong các hệ thống ra quyết định dựa trên tri thức.

Thuật toán học cây quyết định

Trước khi áp dụng, cây quyết định cần được xây dựng từ dữ liệu huấn luyện thông qua quá trình "học". Có nhiều thuật toán học cây được đề xuất, phần lớn dựa trên nguyên tắc tìm kiếm tham lam, xây dựng cây từ đơn giản đến phức tạp.

Trong đề tài này, ta tập trung vào thuật toán ID3 (Iterative Dichotomiser 3), một trong những thuật toán học cây cơ bản, đại diện cho cách tiếp cận xây dựng cây theo hướng phân chia tuần tự. ID3 do Ross Quinlan phát triển và sau này được cải tiến thành thuật toán C4.5, phổ biến hơn trong thực tế.

Dữ liệu huấn luyện

Dữ liệu huấn luyện (*training data*) được cung cấp dưới dạng tập hợp gồm n mẫu hay n ví dụ huấn luyện. Mỗi ví dụ được biểu diễn dưới dạng cặp (x_i, y_i), trong đó x_i là véc tơ các giá trị thuộc tính mô tả đối tượng hoặc tình huống, và y_i là nhãn phân loại tương ứng.

Chẳng hạn, với bộ dữ liệu minh họa trong bảng 1.2, tập huấn luyện bao gồm 14 ví dụ, tương ứng với 14 dòng dữ liệu. Trong đó:

- Cột đầu tiên của bảng chứa số thứ tự của ví dụ và không tham gia vào quá trình xây dựng cây quyết định.

o Bốn cột tiếp theo lần lượt chứa giá trị của bốn thuộc tính đầu vào (ví dụ: *Trời, Nhiệt độ, Độ ẩm, Gió*).

o Cột cuối cùng (ngoài cùng bên phải) chứa giá trị nhãn phân loại của từng ví dụ.

Trong trường hợp bộ dữ liệu này, nhãn phân loại thuộc loại nhị phân (binary label), và có thể nhận một trong hai giá trị: "có" hoặc "không" (tương ứng với việc điều kiện thời tiết trong ngày đó có phù hợp để chơi tennis hay không).

Việc tổ chức dữ liệu huấn luyện dưới dạng này không chỉ thuận tiện cho quá trình huấn luyện mô hình cây quyết định, mà còn giúp dễ dàng kiểm tra, đánh giá và trực quan hóa các quy tắc phân loại được trích xuất từ cây.

Bảng 1.2 Bộ dữ liệu huấn luyện cho bài toàn phân loại "Chơi tennis" [3].

Ngày	Trời	Nhiệt độ	Độ ẩm	Gió	Chơi tennis
D1	Nắng	Cao	Cao	Yếu	Không
D2	Nắng	Cao	Cao	Mạnh	Không
D3	U ám	Cao	Cao	Yếu	Có
D4	Mưa	Trung bình	Cao	Yếu	Có
D5	Mưa	Thấp	Bình thường	Yếu	Có
D6	Mưa	Thấp	Bình thường	Mạnh	Không
D7	U ám	Thấp	Bình thường	Mạnh	Có
D8	Nắng	Trung bình	Cao	Yếu	Không
D9	Nắng	Thấp	Bình thường	Yếu	Có
D10	Mưa	Trung bình	Bình thường	Yếu	Có
D11	Nắng	Trung bình	Bình thường	Mạnh	Có
D12	U ám	Trung bình	Cao	Mạnh	Có
D13	U ám	Cao	Bình thường	Yếu	Có
D14	Mưa	Trung bình	Cao	Mạnh	Không

1.3.1.1 . Thuật toán ID3

Thuật toán ID3 (Iterative Dichotomiser 3) được sử dụng để xây dựng cây quyết

định từ tập dữ liệu huấn luyện sao cho kết quả phân loại đầu ra phù hợp nhất với nhãn có sẵn. Với số lượng thuộc tính lớn, việc liệt kê và kiểm tra tất cả các cây quyết định là không khả thi. Do đó, ID3 áp dụng chiến lược tham lam, xây dựng cây theo hướng từ trên xuống, chọn thuộc tính tốt nhất tại mỗi bước.

Quá trình xây dựng cây diễn ra như sau:

Chọn thuộc tính cho nút gốc: Lựa chọn thuộc tính giúp phân chia dữ liệu thành các tập con đồng nhất (các mẫu cùng nhãn).

Phân chia dữ liệu: Tập dữ liệu được chia thành các nhánh con dựa trên giá trị của thuộc tính vừa chọn.

Đệ quy: Với mỗi tập con, thuật toán lặp lại quá trình chọn thuộc tính, loại bỏ các thuộc tính đã sử dụng, và tiếp tục xây cây.

Quá trình dừng khi:

Tập dữ liệu con tại một nút có cùng nhãn → gán nhãn đó cho nút lá.

Không còn thuộc tính để phân chia, nhưng dữ liệu vẫn còn nhiều nhãn → gán nhãn chiếm đa số cho nút.

Đầu vào: Tập dữ liệu huấn luyện

Đầu ra: Cây quyết định

Khởi đầu: nút hiện thời là nút gốc chứa toàn bộ tập dữ liệu huấn luyện

a. Tại nút hiện thời n, lựa chọn thuộc tính:
   - Chưa được sử dụng ở nút tổ tiên (tức là nút nằm trên đường đi từ gốc tới nút hiện thời)
   - Cho phép phân chia tập dữ liệu hiện thời thành các tập con **một cách tốt nhất**

b. Với mỗi giá trị thuộc tính được chọn:
   - Thêm một nút con bên dưới
   - Chia các ví dụ ở nút hiện thời về các nút con theo giá trị thuộc tính được chọn

c. Lặp (đệ quy) với mỗi nút con cho tới khi:
   - Tất cả các thuộc tính đã được sử dụng ở các nút phía trên, hoặc
   - Tất cả ví dụ tại nút hiện thời có cùng nhãn phân loại
   - Nhãn của nút được lấy theo đa số nhãn của ví dụ tại nút hiện thời

Hình 6. Thuật toán xây dựng cây quyết định[3]

Thuật toán ID3 được thực hiện theo cách đệ quy qua nhiều vòng. Tại mỗi bước, thuật toán lựa chọn thuộc tính tốt nhất để phân chia tập dữ liệu tại nút hiện tại, nhằm tạo ra các tập con đồng nhất nhất có thể. Quá trình này được lặp lại cho đến khi đạt đến các nút lá, tức là khi xảy ra một trong hai điều kiện dừng đã nêu ở trên:

Tập con dữ liệu có cùng nhãn phân loại.

Không còn thuộc tính để phân chia, thuật toán gán nhãn chiếm đa số cho nút.

Lựa chọn thuộc tính tốt nhất.

Một bước quan trọng trong thuật toán ID3 là lựa chọn thuộc tính phân chia tại mỗi nút sao cho hiệu quả phân loại là cao nhất. Trong trường hợp lý tưởng, thuộc tính được chọn sẽ phân chia dữ liệu thành các tập con chỉ chứa một nhãn phân loại. Trong thực tế, tiêu chí lựa chọn là thuộc tính tạo ra các tập con có độ đồng nhất cao.

Để đo độ đồng nhất, ID3 sử dụng khái niệm entropy. Dựa trên entropy, thuật toán tính độ tăng thông tin (information gain) khi sử dụng một thuộc tính, từ đó xác định thuộc tính tốt nhất để phân chia.

Với bài toán phân loại nhị phân (các nhãn + và -), entropy của tập dữ liệu

Thông qua việc tính entropy và độ tăng thông tin, thuật toán ID3 xác định thuộc tính giúp cải thiện độ đồng nhất của dữ liệu nhiều nhất tại mỗi bước phân chia. Với bài toán phân loại nhị phân (các nhãn + và -), entropy của tập dữ liệu S được tính theo công thức:

$$H(S) = -p^{+}\backslash log_2 p^{+} - p^{-}\backslash log_2 p^{-}$$

trong đó p+ và p- là xác suất quan sát thấy nhãn phân loại + và -, được tính bằng tần suất quan sát thấy + và – trong tập dữ liệu.

Trong trường hợp tổng quát với C nhãn phân loại có xác suất lần lượt là $p_1, p_2, \ldots, p_C$ entropy được tính như sau:

$$H(S) = -\sum_{i=1}^{C} p_i \log_2 p_i$$

Trong phương pháp học cây quyết định, một trong những kỹ thuật phổ biến để đánh giá mức độ tốt của thuộc tính trong việc phân tách dữ liệu là sử dụng entropy như một độ đo mức đồng nhất của tập mẫu. Cụ thể, entropy phản ánh mức độ hỗn

loạn hay không thuần nhất của tập dữ liệu: entropy càng thấp thì tập dữ liệu càng đồng nhất.

Để lựa chọn thuộc tính phân tách tại mỗi nút của cây, ta so sánh entropy của tập mẫu trước và sau khi được phân chia thành các tập con dựa trên giá trị của thuộc tính đó. Mức chênh lệch entropy trước và sau khi phân chia chính là độ tăng thông tin (Information Gain), ký hiệu là IG.

Độ tăng thông tin thể hiện mức độ mà thuộc tính giúp giảm sự hỗn loạn của dữ liệu sau khi phân tách, qua đó đánh giá độ tốt của thuộc tính trong việc phân chia tập dữ liệu thành những tập con thuần nhất hơn. Thuộc tính có IG cao nhất sẽ được chọn để phân tách tại nút cây.

$$IG(S,A) = H(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} H(S_v)$$

Trong đó:

S là tập dữ liệu ở nút hiện tại

A là thuộc tính

values(A) là tập các giá trị của thuộc tính A.

S_v là tập các mẫu có giá trị thuộc tính A bằng v.

|S| và $|S_v|$ là lực lượng của các tập hợp tương ứng.

Giá trị của Information Gain (IG) được sử dụng làm tiêu chí chính để lựa chọn thuộc tính tốt nhất tại mỗi nút của cây quyết định. IG phản ánh mức độ mà một thuộc tính giúp giảm sự không thuần nhất (entropy) của tập dữ liệu sau khi phân chia.

Theo thuật toán ID3, tại mỗi bước xây dựng cây, thuộc tính được lựa chọn để phân tách tại nút hiện tại chính là thuộc tính có giá trị IG lớn nhất. Việc lựa chọn này đảm bảo rằng tập dữ liệu được phân chia theo cách làm giảm entropy nhiều nhất, tức làm cho các tập con thu được trở nên thuần nhất nhất có thể. Quá trình này được thực hiện đệ quy cho đến khi cây đạt điều kiện dừng (ví dụ: tập dữ liệu tại nút đã thuần nhất hoàn toàn, hoặc không còn thuộc tính nào để phân tách).

Ví dụ minh họa.

Để minh hoạ cho phương pháp tính độ tăng thông tin (*Information Gain* - IG) và quy trình lựa chọn thuộc tính tốt nhất tại mỗi nút của cây quyết định, ta sử dụng bộ dữ liệu huấn luyện được trình bày trong bảng 1.2. Ta cần, xác định thuộc tính tốt nhất tại nút gốc cho dữ liệu trong bảng 1.2 bằng cách tính IG cho các thuộc tính.

- Với thuộc tính Gió:

Values(Gió) = {yếu, mạnh},

S=[9+,5-], H(S)= -(9/14)(9/14) – (5/14) (5/14) = 0.94

$S_{yếu}$=[6+,2-], H($S_{yếu}$) =-(6/8)(6/8) – (2/8) (2/8) = 0.811

$S_{mạnh}$=[3+,3-], H($S_{mạnh}$)=-(3/6)(3/6) – (3/6) (3/6) = 1

⇨ IG(S,Gió) = H(S) – (8/14)H($S_{yếu}$) – (6/14)H($S_{mạnh}$)

= 0.94-(8/14)*0.811-(6/14)*1= 0.048

- Với thuộc tính Trời:

Values(Trời) = {Nắng, U ám, Mưa}

S=[9+,5-], H(S)= -(9/14)(9/14) – (5/14) (5/14) = 0.94

$S_{Nắng}$=[2+,3-], H($S_{Nắng}$) =-(2/5)(2/5) – (3/5) (3/5) = 0.971

$S_{U\ ám}$=[4+,0], H($S_{U\ ám}$)=-(4/4) (4/4) =0

$S_{Mưa}$=[3+,2-], $H(S_{Mưa})$ =-(3/5) (3/5) – $\left(\frac{2}{5}\right)\left(\frac{2}{5}\right)$ =0.971

⇨ IG(S,Trời)=H(S)-(5/14)H($S_{Nắng}$) –(4/14)H($S_{U\ ám}$)-(5/14)H($S_{Mưa}$)

= 0.94-(5/14)*0.971-0*(4/14)-(5/14)*0.971 =0.246

- Với thuộc tính Độ ẩm:

Values(Độ ẩm) = {cao, bình thường},

S=[9+,5-], H(S)= -(9/14)(9/14) – (5/14) (5/14) = 0.94

S_{Cao}=[3+,4-], H(S_{Cao}) =-(4/7)(4/7) – (3/7) (3/7) = 0.985

$S_{Bình\ thường}$=[6+,1-], H($S_{Bình\ thường}$)=-(6/7)(6/7) – (1/7)(1/7) = 0.591

⇨ IG(S,Độ ẩm) = H(S)–(7/14)H(S_{Cao})–(7/14)H($S_{Bình\ thường}$)

= 0.94-(7/14)*0.985-(7/14)*0.591= 0.152

- Với thuộc tính Nhiệt độ:

Values(Nhiệt độ) = {Cao, trung bình, thấp}

S=[9+,5-], H(S)= -(9/14)(9/14) – (5/14) (5/14) = 0.94

S_{Cao}=[2+,2-], H(S_{Cao}) =-(2/4)(2/4) – (2/4) (2/4) = 1

$S_{Trung\ bình}$=[4+,2-], H($S_{Trung\ bình}$)=-(4/6)(4/6) – (2/6) (2/6) = 0.918

$S_{Thấp}$=[3+,1-], H($S_{Thấp}$)=-(3/4) (3/4) – (1/4) (1/4) = 0.811

⇨ IG(S,Gió) = H(S) – (4/14) H(S_{Cao}) – (6/14) H($S_{Trung\ bình}$) –(4/14) $S_{Thấp}$

= 0.94-(4/14)*1-(6/14)*0.918-(4/14)*0.811= 0.028

IG(S,Trời) có giá trị lớn nhất => thuộc tính tốt nhất là Trời được sử dụng làm nút gốc. Tính IG cho các nhánh con của thuộc tính gốc "Trời". Trời = Nắng → cần xét tiếp thuộc tính (Nhiệt độ, Độ ẩm, Gió), Trời = Mưa → cần xét tiếp thuộc tính (Nhiệt độ, Độ ẩm, Gió), Trời = U ám → tất cả đều "Có" nên không cần xét tiếp.

- Nhánh 1: Trời = Nắng, $S_{nắng}$=[2+,3-]

=> H($S_{nắng}$)=-(2/5)(2/5) – (3/5) (3/5) = 0.971

Xét gió trong nhánh Nắng:

- Yếu: [2+, 0-] => H = 0
- Mạnh: [0+, 3-] => H = 0

IG($S_{nắng}$, Gió) = 0.971 – (2/5*0 + 3/5*0) = 0.971

Xét độ ẩm trong nhánh Nắng:

- Cao: [1+, 2-] => H =-(1/3)(1/3) – (2/3) (2/3) = 0.918
- Bình thường: [1+, 1-] => H = 1

IG($S_{nắng}$, Độ ẩm) = 0.971 – (3/5*0.918 + 2/5*1) = 0.021

Xét Nhiệt độ trong nhánh Nắng:

- Cao: [1+,1-] => H =1.
- Trung bình: [1+,0-] => H = 0
- Thấp: [0+,2-] => H = 0

IG($S_{nắng}$, Nhiệt độ) = 0.971 – (2/5*1 + 1/5*0 + 2/5*0) = 0.571

⇨ IG($S_{nắng}$, Gió) trong nhánh Nắng cao nhất nên chọn thuộc tính Gió làm nút gốc trong nhánh Nắng.

- Nhánh 2: Trời = Mưa , $S_{Mưa}$= [3+,2-] => H($S_{Mưa}$) = 0.971

Xét Gió trong nhánh Mưa:

- Yếu: [3+, 0-] => H=0
- Mạnh: [0+, 2-] => H=0

IG($S_{Mưa}$, Gió) = 0.971 – (3/5*0 + 2/5*0) = 0.971

Xét Độ ẩm trong nhánh Mưa:

- Cao: [2+, 2-] => H=1
- Bình thường: [1+, 0-] => H=0

IG($S_{Mưa}$, Độ ẩm) = 0.971 – (4/5*1 + 1/5*0) = 0.971 – 0.8 = 0.171

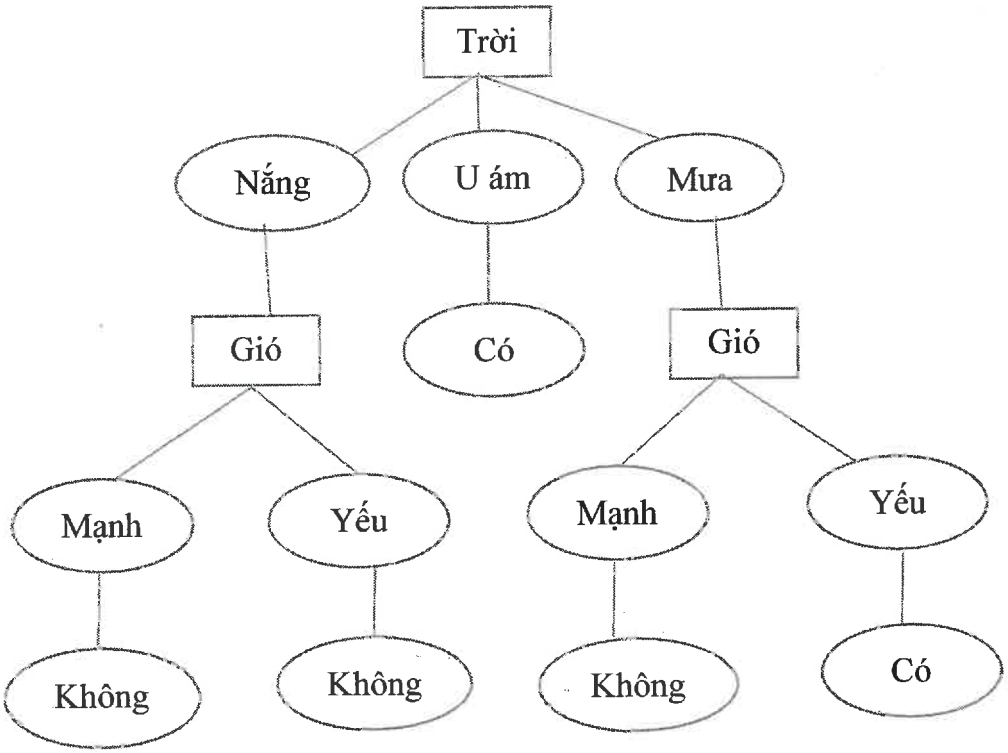
Xét Nhiệt độ trong nhánh Mưa:

- Trung bình: [2+,2-] => H=1
- Thấp: [1+,0-] => H=0

IG($S_{Mưa}$, Nhiệt độ) = 0.971 – (4/5*1 + 1/5*0) = 0.971 – 0.8 = 0.171

⇨ IG($S_{Mưa}$, Gió) trong nhánh Mưa cao nhất nên chọn thuộc tính Gió làm nút gốc trong nhánh Mưa.

Kết quả học cây đầy đủ được thể hiện trong hình



Hình 7. Ví dụ xây dựng cây quyết định

1.3.1.2. Thuật toán C4.5

C4.5 là một thuật toán xây dựng cây quyết định được phát triển từ thuật toán ID3 bởi J. R. Quinlan vào năm 1993 [7]. Đây là một trong những thuật toán nổi tiếng và được ứng dụng rộng rãi trong các hệ thống phân loại nhờ khả năng xử lý linh hoạt và hiệu quả cao.

Đặc điểm chính của thuật toán C4.5:

- Sử dụng Gain Ratio (thay vì Information Gain như ID3) để lựa chọn thuộc tính phân chia tại mỗi nút trong quá trình xây dựng cây.
- Có khả năng xử lý tốt cả hai loại thuộc tính: thuộc tính rời rạc và thuộc tính liên tục.
- Xử lý được dữ liệu không đầy đủ, tức là dữ liệu có thể thiếu giá trị tại một số thuộc tính. Trong trường hợp này, C4.5 cho phép đại diện cho giá trị thiếu bằng dấu hỏi (?) và bỏ qua các giá trị này khi tính toán Information Gain hoặc Gain Ratio.
- Cắt tỉa cây (pruning) sau khi xây dựng để loại bỏ những nhánh cây không thực sự ý nghĩa, giúp giảm thiểu tình trạng overfitting và làm cây trở nên gọn nhẹ

hơn.

Ý nghĩa của Gain Ratio (GR):

$$H_D(C) = -\sum_{k=1}^{k} p_k log_2(p_k) = -\sum_{k=1}^{k} \frac{|D_k|}{|D|} log_2\left(\frac{|D_k|}{|D|}\right)$$

$$H_D(C|F_i) = \sum_{j=1}^{p_i} \frac{|D_k|}{|D|} H_{D_J}\left(C|F_i = V_j^i\right)$$

$$IG_D(C|F_i) = H_D(C) - H_D(C|F_i)$$

Splitting entropy của thuộc tính F_i, ký hiệu $SE_D(F_i)$:

$$SE_D(F_i) = -\sum_{j=1}^{P_i} \frac{|D_j|}{|D|} \log_2\left(\frac{|D_j|}{|D|}\right)$$

Khi đó, Gain Ratio ký hiệu $GR_D(C|F_i)$

$$GR_D(C|F_i) = \frac{IG_D(C|F_i)}{SE_D(F_i)}$$

Tiêu chí Information Gain trong thuật toán ID3 thường có xu hướng ưu tiên lựa chọn các thuộc tính có nhiều giá trị phân biệt (miền xác định lớn). Nguyên nhân là do khi thuộc tính có nhiều giá trị, tập dữ liệu sẽ được chia thành nhiều tập con nhỏ hơn, dẫn đến entropy sau phân chia giảm mạnh, từ đó làm cho Information Gain trở nên cao. Điều này có thể khiến cây quyết định chọn thuộc tính không thực sự có ý nghĩa trong phân loại mà chỉ vì nó chia nhỏ dữ liệu nhiều hơn.

Thuật toán C4.5 đã khắc phục hạn chế này bằng cách sử dụng chỉ số Gain Ratio. Gain Ratio được tính bằng cách lấy Information Gain chia cho Splitting Entropy (ký hiệu SED(Fi)) của thuộc tính. Splitting Entropy phản ánh mức độ phân tán của tập dữ liệu khi phân chia theo thuộc tính đó.

Thuộc tính có nhiều giá trị sẽ làm cho Splitting Entropy lớn, dẫn đến Gain Ratio nhỏ hơn, từ đó hạn chế khả năng những thuộc tính này được chọn nếu chúng không thực sự mang lại hiệu quả phân loại. Ngược lại, thuộc tính có ít giá trị nhưng giúp phân chia dữ liệu hiệu quả sẽ có Gain Ratio cao hơn, tăng khả năng được lựa

chọn. Cơ chế này giúp cây quyết định do C4.5 xây dựng trở nên hợp lý và tránh được tình trạng overfitting.

- **Ưu điểm cây quyết định:**

o Dễ hiểu, dễ giải thích: Cây quyết định gần giống cách con người tư duy logic (nếu… thì…). Dễ trình bày cho người không chuyên.

o Không cần chuẩn hóa dữ liệu: Không yêu cầu chuẩn hóa (scaling) hay xử lý biến đặc biệt. Xử lý tốt cả dữ liệu dạng số và dạng phân loại (categorical).

o Có thể xử lý giá trị thiếu: Một số thuật toán (C4.5, CART) vẫn hoạt động được khi dữ liệu có giá trị thiếu.

o Nhanh với tập dữ liệu nhỏ hoặc vừa: Tốc độ huấn luyện nhanh so với nhiều thuật toán khác.

- **Nhược điểm cây quyết định:**

o Dễ bị overfitting: Cây quá sâu dẫn đến mô hình khớp chặt dữ liệu huấn luyện, kém tổng quát trên dữ liệu mới. Giải pháp: cắt tỉa cây (pruning), giới hạn độ sâu.

o Nhạy cảm với dữ liệu nhiễu: Chỉ một vài điểm dữ liệu nhiễu có thể làm thay đổi cấu trúc cây rất nhiều.

o Kém ổn định: Một thay đổi nhỏ trong dữ liệu huấn luyện có thể tạo ra cây khác hoàn toàn.

o Thiên lệch với thuộc tính có nhiều giá trị: Trong ID3, thuộc tính có nhiều giá trị duy nhất (nhiều levels) thường có IG cao → dễ bị chọn dù không thực sự quan trọng.

o Hiệu quả kém với dữ liệu liên tục nhiều chiều: Khi dữ liệu gồm nhiều biến liên tục (ví dụ 100+ features dạng số), cây quyết định đơn giản thường không đủ mạnh.

1.3.2. Thuật toán SVM (Support Vector Machine) [8]

Support Vector Machine (SVM) là một trong những thuật toán học máy có giám sát mạnh mẽ, được sử dụng rộng rãi trong các bài toán phân loại và hồi quy. Thuật toán được đề xuất lần đầu bởi Vladimir Vapnik và các cộng sự vào những

năm 1990, với nền tảng là lý thuyết học thống kê (Statistical Learning Theory).

Mục tiêu của SVM là tìm một siêu phẳng tối ưu (optimal hyperplane) để phân tách dữ liệu thành các lớp khác nhau với khoảng cách biên (margin) lớn nhất giữa hai lớp.

- Siêu phẳng và phân tách tuyến tính

Giả sử bài toán phân loại nhị phân với tập huấn luyện:

$$\left\{\left(x_i, y_i\right) \mid x_i \in R^n, y_i \in \{-1,1\}\right\}_{i=1}^{m}$$

Một siêu phẳng trong không gian R^n được biểu diễn bởi:

w·x+ b=0

Trong đó:

w là vector trọng số (vector pháp tuyến).

b là hằng số điều chỉnh độ lệch (bias).

x là vector đặc trưng.

Dữ liệu được phân tách tuyến tính nếu tồn tại một siêu phẳng sao cho:

$$y_i\left(w \cdot x_i + b\right) \geq 1, \forall_i$$

- Hàm mục tiêu: Tối đa hóa biên (Margin)

Khoảng cách từ một điểm x đến siêu phẳng là:

$$\frac{|w \cdot x + b|}{||w||}$$

SVM tìm siêu phẳng sao cho khoảng cách này được tối đa hóa, tương đương với bài toán tối ưu:

$$\min_{w,b} \frac{1}{2}||w||^2 \quad \text{với điều kiện: } y_i\left(w \cdot x_i + b\right) \geq 1$$

- Trường hợp không phân tách tuyến tính.

Trong thực tế, dữ liệu thường không hoàn toàn phân tách tuyến tính. SVM sử dụng soft margin để cho phép một số điểm vi phạm điều kiện phân lớp:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \text{ với } y_i\left(w \cdot x_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0$$

Tham số C>0 điều khiển mức độ chấp nhận sai số:

C lớn: ưu tiên phân loại chính xác (dễ overfit).

C nhỏ: chấp nhận một số lỗi để margin rộng hơn (regularization tốt hơn).

- **Ưu điểm:**

- Hiệu quả cao trong không gian đặc trưng có số chiều lớn.
- Tốt khi số mẫu ít hơn số chiều (high-dimensional space).
- Có thể linh hoạt với các loại kernel khác nhau.

- **Nhược điểm:**

- Cần chọn kernel và tham số phù hợp (γ, C).
- Hiệu quả kém khi dữ liệu có nhiều nhiễu hoặc chồng lấn.
- Không phù hợp với tập dữ liệu cực lớn (chi phí tính toán cao).

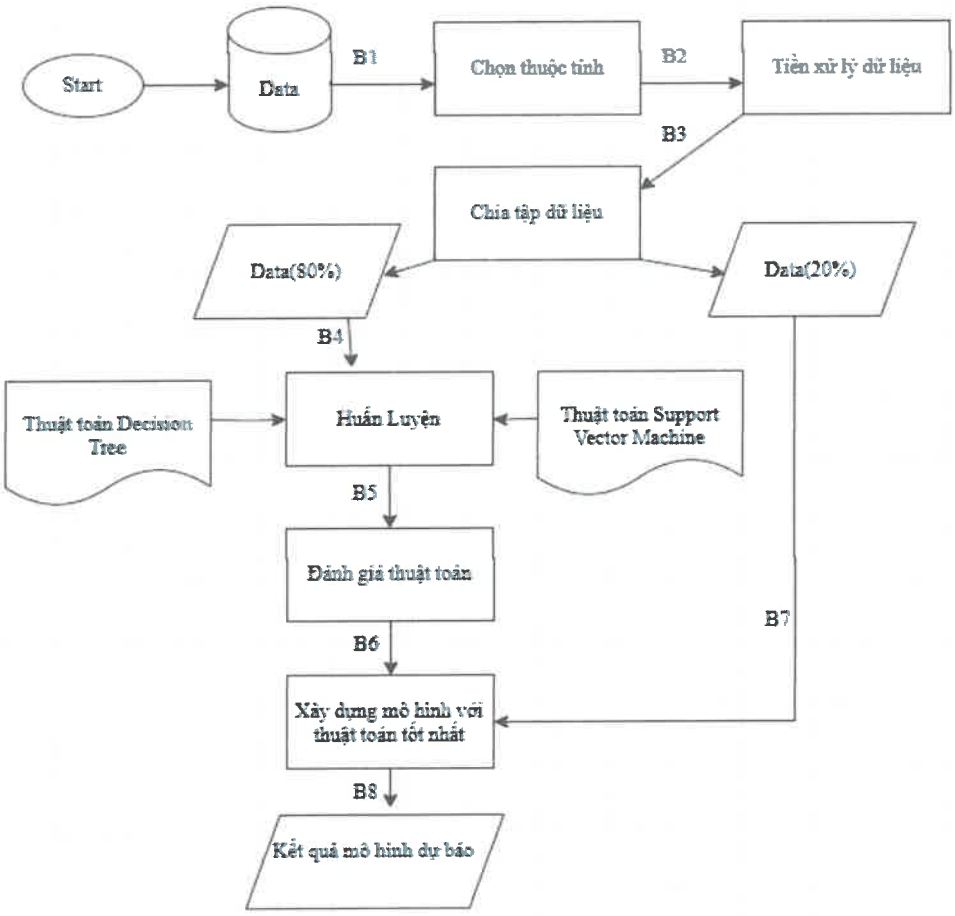
Tổng kết chương 1: Chương 1 đã nêu tổng quan về VNPT Nam Định và các dịch vụ mà đơn vị đang cung cấp, tình hình sản xuất kinh doanh dịch vụ FiberVNN tại Nam Định của các nhà mạng. Bên cạnh đó, đã nêu tổng quan về học máy và trình bày chi tiết 2 thuật toán áp dụng cho bài toán dự báo đó là Decision Tree và Support Vector Machine.

Chương 2.
PHƯƠNG PHÁP DỰ BÁO KHÁCH HÀNG RỜI MẠNG

Chương này sẽ trình bày tổng quan về các bước thực hiện xây dựng mô hình để dự báo khách hàng rời mạng. Nội dung gồm các bước, tiền xử lý dữ liệu, huấn luyện, thử nghiệm trên hai thuật toán học máy là Decision Tree và Support Vector Machine sau đó lựa chọn thuật toán tốt nhất để xây dựng mô hình dự đoán và đánh giá mô hình.

2.1. Giới thiệu

Sơ đồ tổng quan của hệ thống



Hình 8: Sơ đồ tổng quan hệ thống

Luồng xử lý chính gồm 8 bước.

B1. Việc chọn lọc thuộc tính là bước quan trọng nhằm giảm chiều dữ liệu và

nâng cao hiệu quả mô hình. Các thuộc tính đặc trưng của dữ liệu khách hàng được đánh giá dựa trên các chỉ số như Information Gain, Gain Ratio và Gini Index để xác định mức độ đóng góp vào quá trình phân loại. Những thuộc tính có giá trị thông tin cao sẽ được ưu tiên lựa chọn. Tuy nhiên, quá trình này cần kết hợp với kiến thức nghiệp vụ và kinh nghiệm phân tích dữ liệu để đảm bảo tính phù hợp và hiệu quả trong thực tế.

B2. Làm sạch và chuẩn hóa dữ liệu là bước tiền xử lý nhằm đảm bảo chất lượng dữ liệu đầu vào cho mô hình. Quá trình này bao gồm xử lý giá trị trùng lặp, giá trị khuyết thiếu, và chuyển đổi định dạng dữ liệu sao cho phù hợp với yêu cầu của các thuật toán dự báo. Việc chuẩn hóa giúp giảm sai lệch, đồng thời tăng độ chính xác và hiệu quả của mô hình học máy.

B3. Chia tập dữ liệu thành tập huấn luyện (80%) và tập kiểm thử (20%). B4. Huấn luyện mô hình bằng thuật toán Decision Tree và thuật toán SVM với tập dữ liệu huấn luyện (80%).

B5. Kiểm tra, đánh giá và so sánh kết quả của 2 thuật toán đã được huấn luyện. Các phép đo thông thường để đánh giá mô hình phân loại bao gồm độ chính xác (accuracy), độ nhạy (recall), độ chính xác (precision) và F1-score

B6. Xây dựng mô hình dự đoán với thuật toán tốt nhất.

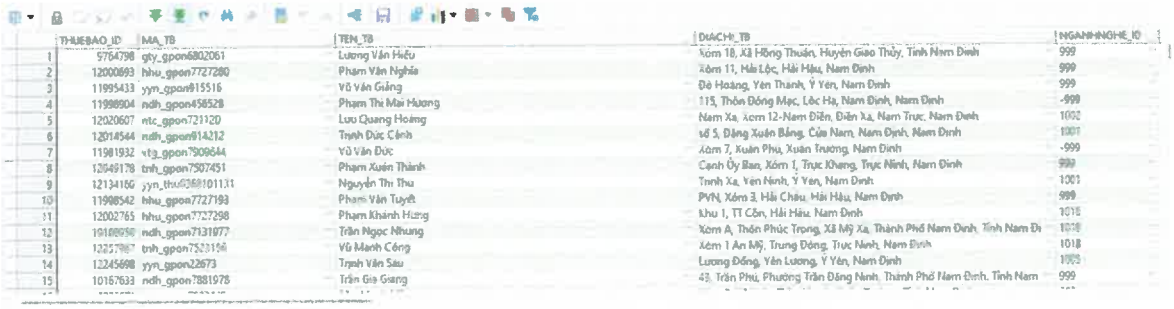
B7. Sử dụng tập kiểm thử (20%) để dự báo khách hàng rời mạng.

B8. Kết quả mô hình.

2.2. Xây dựng mô hình dự báo khách hàng rời mạng.

2.2.1. Thu thập và tiền xử lý dữ liệu.

Dữ liệu được truy xuất từ hệ thống nội bộ của VNPT Nam Định tới ngày 31/03/2025 bao gồm 140320 khách hàng đang và đã sử dụng dịch vụ Fiber.



	THUEBAO_ID	MA_TB	TEN_TB	DIACHI_TB	NGANHNGHE_ID
1	9764798	qty_gpon6802061	Lương Văn Hiếu	Xóm 18, Xã Hồng Thuận, Huyện Giao Thủy, Tỉnh Nam Định	999
2	12000893	hhu_gpon7727280	Phạm Văn Nghĩa	Xóm 11, Hải Lộc, Hải Hậu, Nam Định	999
3	11995433	yyn_gpon815516	Võ Văn Giảng	Đê Hoàng, Yên Thành, Ý Yên, Nam Định	999
4	11998904	ndh_gpon458528	Phạm Thị Mai Hương	115, Thôn Đồng Mạc, Lộc Hạ, Nam Định, Nam Định	-999
5	12020607	ntc_gpon721120	Lưu Quang Hoàng	Nam Xa, Xóm 12-Nam Điền, Điền Xá, Nam Trực, Nam Định	1002
6	12014564	ndh_gpon814212	Trịnh Đức Cảnh	số 5, Đặng Xuân Bảng, Cửa Nam, Nam Định, Nam Định	1001
7	11981932	vtg_gpon7909644	Vũ Văn Đức	Xóm 7, Xuân Phú, Xuân Trường, Nam Định	-999
8	12049178	tnh_gpon7507451	Phạm Xuân Thành	Cạnh Ủy Ban, Xóm 1, Trực Khang, Trực Ninh, Nam Định	999
9	12134186	yyn_thu036810113	Nguyễn Thị Thu	Trinh Xá, Yên Ninh, Ý Yên, Nam Định	1001
10	11998542	hhu_gpon7727193	Phạm Văn Tuyết	PVN, Xóm 3, Hải Châu, Hải Hậu, Nam Định	999
11	12002765	hhu_gpon7727298	Phạm Khánh Hưng	khu 1, TT Cồn, Hải Hậu, Nam Định	1015
12	19108950	ndh_gpon7131977	Trần Ngọc Nhung	Xóm A, Thôn Phúc Trọng, Xã Mỹ Xá, Thành Phố Nam Định, Tỉnh Nam Đị	1018
13	12057867	tnh_gpon7523158	Vũ Mạnh Công	Xóm 1 An Mỹ, Trung Đông, Trực Ninh, Nam Định	1018
14	12245698	yyn_gpon22673	Trịnh Văn Sáu	Lương Đống, Yên Lương, Ý Yên, Nam Định	1003
15	10167633	ndh_gpon7881978	Trần Gia Giang	43, Trần Phú, Phường Trần Đăng Ninh, Thành Phố Nam Định, Tỉnh Nam	999

Hình 9. Data trích xuất từ hệ thống nội bộ.

Tập dữ liệu gồm 33 cột được mô tả cụ thể như sau:

STT: Số thứ tự

THUEBAO_ID: Mã của thuê bao khách hàng.

MA_TB: Mã thuê bao dịch vụ FiberVNN của khách hàng.

TEN_TB: Tên thuê bao (tên khách hàng)

DIACHI_TB: Địa chỉ thuê bao (địa chỉ khách hàng)

NGANHNGHE_ID: Mã ngành nghề (mã ngành nghề của khách hàng)

NGANHNGHE: Tên ngành nghề (tên ngành nghề của khách hàng)

NGAY_SN: Ngày sinh của khách hàng

TUOI: Tuổi

KHUVUC_ID: Mã khu vực (mã khu vực quản lý địa bàn)

KHUVUC: Tên khu vực (tên khu vực quản lý địa bàn)

LOAIKH_ID: Mã loại khách hàng

LOAIKH: Phân loại khách hàng. (chi tiết mô tả ở bảng 2.1)

KHDN: 1-Khách hàng doanh nghiệp; 0-Khách hàng cá nhân.

MANGKHAC: 1- có; 0-không.

SO_DV_KHAC: Số dịch vụ khác của VNPT mà khách hàng sử dụng ngoài FiberVNN.

GOI_DADV: Gói đa dịch vụ 1- có; 0-không. Sử dụng gói tích hợp hay không, tích hợp (MyTV, di động, Mesh…) hay không.

GOICUOC: Gói cước dịch vụ

GIACUOC: Gía cước dịch vụ

NOCUOC_2THANG: Số tháng nợ cước.

TRATRUOC: Đang thanh toán cước hàng tháng hay sử dụng gói trả trước nhiều tháng.1: Đã thanh toán trước cước 6 tháng, 12 tháng. 0: thanh toán cước hàng tháng.

SOTHANG_TRATRUOC_CONLAI: Số tháng trả trước còn lại.

SOLAN_BAOHONG: Số lần gọi báo hỏng do sự cố (đứt cáp, không truy cập được internet, mạng chậm…).

SOLAN_GOI_KIEM: Số lần gọi kiểm tra chất lượng dịch vụ.

SOLAN_GOI_KIEM_HL: Số lần gọi kiểm, khách hàng phản hồi hài lòng.

SOLAN_GOI_KIEM_KHL: Số lần gọi kiểm, khách hàng phản hồi không hài lòng.

SOLAN_TAM_NGUNG: Số lần tạm ngưng dịch vụ.

THANG|_SD: Số tháng sử dụng.

KO_PSLL: Không phát sinh lưu lượng: Nhà mạng ghi nhận khách hàng không phát sinh lưu lượng sử dụng 5 ngày liên tiếp (do hỏng modem, đi vắng, mất điện,...) để thực hiện kiểm tra chất lượng dịch vụ.

SOLAN_GIAHAN: Số lần khách hàng gia hạn dịch vụ.

TRANGTHAITB_ID: Trạng thái thuê bao ID.

TRANGTHAI_TB: Trạng thái hoạt động của thuê bao (chi tiết mô tả ở bảng …)

THANHLY: Trạng thái của thuê bao: 0- Đang hoạt động, 1- Đã thanh lý.

Bảng 2.1: Danh sách đối tượng khách hàng

ID	Tên đối tượng
1	Khác
2	Công ty nhà nước
3	Hành chính sự nghiệp khác
5	Thuộc giáo dục
7	Thuộc UBND xã
10	Thuộc Y Tế
53	Khối cơ quan Tỉnh, Thành phố trực thuộc Trung ương: Tỉnh ủy, Thành ủy, Hội đồng nhân dân, UBND (bao gồm các cơ quan chuyên môn thuộc

	UBND cấp Tỉnh/TP như sở, ban, ngành)
56	Công ty tư nhân
57	Công ty cổ phần
60	Công ty TNHH
62	Tập đoàn, tổng công ty
68	Hộ cận nghèo
80	Trường trung học công lập
82	Trường tiểu học công lập

Bảng 2.2: Trạng thái thuê bao ID

ID	Tên trạng thái
1	Đang hoạt động
6	Tạm dừng
7	Thanh lý theo yêu cầu
9	Thanh lý cưỡng bức

Lựa chọn thuộc tính

Bằng kinh nghiệm, nghiệp vụ khách hàng cũng như kinh nghiệm xử lý dữ liệu tiến hành chọn lựa các thuộc tính và làm sạch dữ liệu. Các thuộc tính được lựa chọn để huấn luyện bao gồm 22 cột, cột 23 là cột Churn(thanhly), cột này là cột gắn nhãn của tập dữ liệu, cột để nhận biết là thuê bao có rời mạng hay không?

("NGANHNGHE_ID","KHUVUC_ID","LOAIKH_ID","KHDN","MANG KHAC","GOI_DADV","TRATRUOC","TRANGTHAITB_ID","THUEBAO_ID", "TUOI","SO_DV_KHAC","GIACUOC","NOCUOC_2THANG","SOTHANG_T RATRUOC_CONLAI","SOLAN_BAOHONG","SOLAN_GOI_KIEM","SOLAN

_GOI_KIEM_HL","SOLAN_GOI_KIEM_KHL","SOLAN_TAMNGUNG","THANG_SD","KO_PSLL","SOLAN_GIAHAN",)

Xử lý dữ liệu:

Mục đích chính của tiền xử lý dữ liệu là cải thiện chất lượng dữ liệu và chuẩn bị cho các bước tiếp theo trong quá trình phân tích dữ liệu và xây dựng mô hình. Các bước tiền xử lý dữ liệu bao gồm:

+> Lọc dòng dữ liệu không hợp lệ: Đối với những dòng mà cột KHUVUC_ID trống hoặc LOAIKH_ID bằng 0 sẽ được loại bỏ để tránh ảnh hưởng tới kết quả dự báo.

Code python xử lý:

```
# Lọc dòng không hợp lệ
before_rows = input_data.shape[0]
input_data = input_data[~(input_data["KHUVUC_ID"].isna() | (input_data["LOAIKH_ID"] == 0))]
after_rows = input_data.shape[0]
print(f"[clean_data] Đã loại bỏ {before_rows - after_rows} dòng không hợp lệ.")
```

Hình 10: Xử lý dòng không hợp lệ

+> Điền giá trị thiếu:

Điền các giá thị thiếu trong file Data, NGANHNGHE_ID=999 (khác).

LOAI_ID=1(Khác).

Code python xử lý:

```
# Điền giá trị thiếu
input_data["NGANHNGHE_ID"] = input_data["NGANHNGHE_ID"].fillna(999)
input_data["LOAIKH_ID"] = input_data["LOAIKH_ID"].fillna(1)
```

Hình 11: Xử lý điền giá trị thiếu.

+> Làm sạch trường tuổi khách hàng:

Với khách hàng có độ tuổi <= 10 -> giá trị được gán NaN, giá trị không hợp lệ.

sử dụng thư viện numpy trong python để làm sạch trường tuổi không hợp lệ.

Code python xử lý:

```
# Làm sạch tuổi
input_data["TUOI"] = np.where(input_data["TUOI"] <= 10, np.nan, input_data["TUOI"])
```

Hình 12: Xử lý làm sạch tuổi

+> Chia Data thành 3 nhóm chính:

selected_features: list[str] = numeric_features + categorical_features + [target]

Trong đó:

numeric_features: Là biến số

```
numeric_features = [
    "TUOI",
    "SO_DV_KHAC",
    "GIACUOC",
    "NOCUOC_2THANG",
    "SOTHANG_TRATRUOC_CONLAI",
    "SOLAN_BAOHONG",
    "SOLAN_GOI_KIEM",
    "SOLAN_GOI_KIEM_HL",
    "SOLAN_GOI_KIEM_KHL",
    "SOLAN_TAMNGUNG",
    "THANG_SD",
    "KO_PSLL",
    "SOLAN_GIAHAN",
]
```

Hình 13 : Biến số

categorical_features: Là biến phân loại

```
categorical_features = [
    "NGANHNGHE_ID",
    "KHUVUC_ID",
    "LOAIKH_ID",
    "KHDN",
    "MANGKHAC",
    "GOI_DADV",
    "TRATRUOC",
    "TRANGTHAITB_ID",
    "THUEBAO_ID"
]
```

Hình 14: Biến phân loại

target="THANHLY"

- Chia Data thành 2 phần:

File DamVV_Dataset_2025_80.csv: Dùng để training

File DamVV_Dataset_2025_20.csv: Dùng để dự báo

2.2.2. Huấn luyện và kiểm thử mô hình.

Đề án tập trung vào 2 thuật toán: Decision Tree và Support Vector Machine

(SVM). Sau khi Training lựa chọn thuật toán tốt nhất để xây dựng mô hình dự báo. Trong quá trình xây dựng và kiểm thử mô hình dự báo, việc sử dụng các chỉ số đánh giá là cần thiết để đo lường hiệu quả của mô hình. Một trong những chỉ số cơ bản và quan trọng nhất đó là:

- Accuracy (Độ chính xác): Đây là chỉ số được sử dụng phổ biến để đánh giá hiệu suất tổng thể của mô hình phân loại. Accuracy được tính bằng tỷ lệ giữa số lượng mẫu dữ liệu được phân loại đúng so với tổng số mẫu dữ liệu trong tập kiểm thử. Công thức tính như sau.

$$Accuracy = \frac{Số\ lượng\ mẫu\ dự\ đoán\ đúng}{Tổng\ số\ mẫu} = \frac{TP + TN}{TP + TN + FP + FN}$$

  - TP (True Positive): Số khách hàng rời mạng mô hình dự đoán đúng.
  - FP (False Positive): Số khách hàng rời mạng mô hình dự đoán sai.
  - TN (True Negative): Số khách hàng sử dụng mô hình dự đoán đúng.
  - FN (False Negative): Số khách hàng sử dụng mô hình dự đoán sai.

- Kappy (Cohen's Kappa): Là một chỉ số thống kê dùng để đo lường mức độ đồng thuận giữa hai bộ phân loại (hoặc giữa mô hình dự đoán và thực tế), có tính đến sự đồng thuận xảy ra ngẫu nhiên. Được phát triển bởi Jacob Cohen (1960), nên còn gọi là Cohen's Kappa. Kappa không chỉ đo tỉ lệ đúng (accuracy) mà còn xem xét khả năng một kết quả đúng chỉ do ngẫu nhiên

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Trong đó:

  - Po: Tỉ lệ quan sát đồng thuận thực tế (accuracy).
  - Pe: Tỉ lệ đồng thuận ngẫu nhiên kỳ vọng.

- MCC (Matthews Correlation Coefficient): là hệ số tương quan giữa giá trị thực và giá trị dự đoán của mô hình. MCC đánh giá mức độ liên hệ giữa dự đoán và thực tế theo cách cân đối với cả true positives (TP), true negatives (TN), false positives (FP) và false negatives (FN). MCC thường được coi là thước đo toàn diện nhất cho bài toán phân loại nhị phân, vì nó không bị ảnh hưởng bởi sự mất cân bằng giữa các lớp.

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Trong đó:

- **TP**: Số khách hàng rời mạng mô hình dự đoán đúng
- **TN**: Số khách hàng sử dụng mô hình dự đoán đúng.
- **FP**: Số khách hàng rời mạng mô hình dự đoán sai.
- **FN**: Số khách hàng sử dụng mô hình dự đoán sai.

Sử dụng thư viện PyCaret, Pandas nhằm đơn giản hóa quy trình xây dựng, huấn luyện và triển khai các mô hình máy học, xử lý dữ liệu đầu vào, định nghĩa các đặc trưng đầu vào, thiết lập môi trường PyCaret để huấn luyện, tạo mô hình hoặc so sánh nhiều mô hình, lưu lại mô hình tốt nhất. sử dụng hàm setup() của PyCaret để thiết lập môi trường huấn luyện cho bài toán phân loại, tiền xử lý dữ liệu (xử lý thiếu, chuẩn hóa, cân bằng dữ liệu, phân loại đặc trưng...), chuẩn bị pipeline machine learning hoàn chỉnh.

Code Python xử lý:

```
clf1 = setup(
    data=churn_rate_data,
    target=target,
    categorical_features=categorical_features,
    numeric_features=numeric_features,
    imputation_type="iterative",
    numeric_imputation="mean",
    fix_imbalance=True,
    # normalize = True,
    use_gpu=True,
    index=False,
)
```

Hình 15: Sử dụng hàm setup trong thư viện PyCaret

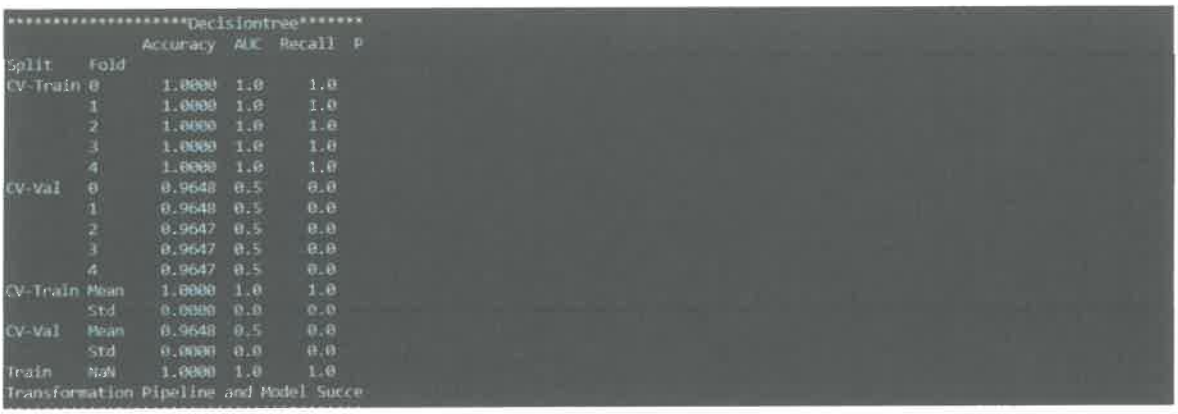
Tạo một mô hình Decision Tree (cây quyết định) bằng PyCaret, huấn luyện nó trên dữ liệu đã qua setup()

```
print('**********Decisiontree**********')
dt = create_model(estimator='dt', fold=5, return_train_score=True)
save_model(dt, 'churn-rate-decisiontree')
```

Hình 16: Tạo mô hình Decision Tree

Sử dụng 5-fold cross-validation. Nghĩa là dữ liệu sẽ chia làm 5 phần, huấn

luyện trên 4 phần và kiểm tra trên 1 phần, lặp lại 5 lần và lấy trung bình kết quả. Lưu mô hình sau khi đã huấn luyện.

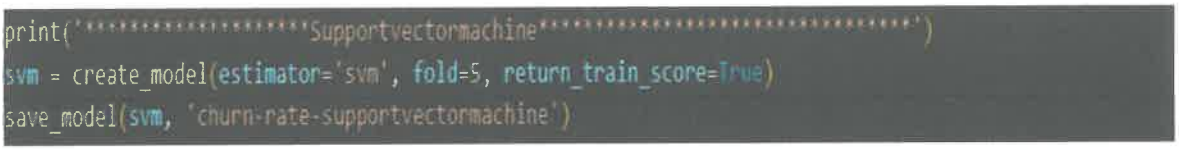


```
******************Decisiontree*******
              Accuracy  AUC  Recall  P
Split    Fold
CV-Train 0      1.0000  1.0     1.0
         1      1.0000  1.0     1.0
         2      1.0000  1.0     1.0
         3      1.0000  1.0     1.0
         4      1.0000  1.0     1.0
CV-Val   0      0.9648  0.5     0.0
         1      0.9648  0.5     0.0
         2      0.9647  0.5     0.0
         3      0.9647  0.5     0.0
         4      0.9647  0.5     0.0
CV-Train Mean   1.0000  1.0     1.0
         Std    0.0000  0.0     0.0
CV-Val   Mean   0.9648  0.5     0.0
         Std    0.0000  0.0     0.0
Train    NaN    1.0000  1.0     1.0
Transformation Pipeline and Model Succe
```

Hình 17: Kết quả huấn luyện mô hình Decision Tree.

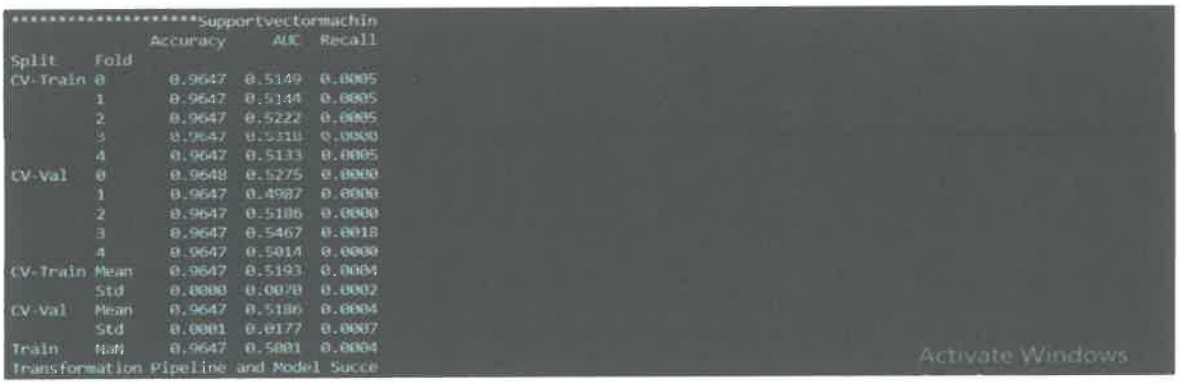
Tạo mô hình SVM (Support Vector Machine) bằng PyCaret, huấn luyện nó với 5-fold cross-validation. Lưu mô hình SVM đã huấn luyện vào file có tên "churn-rate-supportvectormachine"

Code Python xử lý:



```
print('**********Supportvectormachine**********')
svm = create_model(estimator='svm', fold=5, return_train_score=True)
save_model(svm, 'churn-rate-supportvectormachine')
```

Hình 18: Tạo mô hình Support Vector Machine.



```
******************Supportvectormachin
              Accuracy     AUC  Recall
Split    Fold
CV-Train 0      0.9647  0.5149  0.0005
         1      0.9647  0.5144  0.0005
         2      0.9647  0.5222  0.0005
         3      0.9647  0.5318  0.0000
         4      0.9647  0.5133  0.0005
CV-Val   0      0.9648  0.5275  0.0000
         1      0.9647  0.4987  0.0000
         2      0.9647  0.5186  0.0000
         3      0.9647  0.5467  0.0018
         4      0.9647  0.5014  0.0000
CV-Train Mean   0.9647  0.5193  0.0004
         Std    0.0000  0.0070  0.0002
CV-Val   Mean   0.9647  0.5186  0.0004
         Std    0.0001  0.0177  0.0007
Train    NaN    0.9647  0.5001  0.0004
Transformation Pipeline and Model Succe
```


Hình 19: Kết quả huấn luyện mô hình Support Vector Machine

So sánh giữa giữa hai mô hình sau khi huấn luyện.

Code xử lý:

```
print('********************So sanh Decisiontree vs Supportvectormachine***************')
best = compare_models(include=["svm", "dt"])
```

Hình 20: So sánh giữa Decision Tree và Support Vector Machine

```
********************So sanh Decisiontre
                    Model  Accuracy
dt   Decision Tree Classifier    0.9648
svm       SVM - Linear Kernel    0.5902

     Kappa     MCC  TT (Sec)
dt  0.0000  0.0000     2.864
svm 0.0014  0.0013     4.651
```


Hình 21: Kết quả so sánh giữa Decision Tree và Support Vector Machine.

⇨ Ta thấy thuật toán Decision Tree có hiệu suất đạt 96,48% và thời gian thực hiện huấn luyện là 2.864s. Thuật toán được lựa chọn để xây dựng là thuật toán Decision Tree.

- Xây dựng, kiểm thử mô hình dự báo khách hàng rời mạng:

o Tải file CSV

```
# Upload file
uploaded_file = st.file_uploader("📁 Tải file CSV dữ liệu khách hàng:", type=["csv"])
```

Hình 22: Upload file CSV

o Đọc file CSV thành DataFrame + hiển thị bảng dữ liệu đầu vào.

```
    input_df = pd.read_csv(uploaded_file)
    st.subheader("📊 Dữ liệu đầu vào")
    st.dataframe(input_df)
```

Hình 23: Đọc file CSV

o Làm sạch và chuẩn bị dữ liệu: Xử lý dữ liệu (ví dụ: loại bỏ giá trị thiếu, chuẩn hóa), Nếu cột THANHLY tồn tại (label thực tế), loại bỏ trước khi dự đoán.

```
cleaned_df = clean_data(input_df)
if 'THANHLY' in cleaned_df.columns:
    cleaned_df = cleaned_df.drop('THANHLY', axis=1)
```

Hình 24: Xử lý dữ liệu.

o Dự đoán bằng mô hình: Dùng mô hình đã load (hoặc train trước đó) để dự đoán.

```
predictions = predict_model(model, data=cleaned_df)
```

Hình 25: Dự đoán bằng mô hình.

- Ghép kết quả dự đoán vào dữ liệu gốc: Thêm cột dự đoán thanh lý và xác suất dự đoán vào bảng kết quả

```
# Ghép kết quả
full_result = input_df.copy()
label_col = next((col for col in predictions.columns if 'label' in col.lower()), None)
score_col = next((col for col in predictions.columns if 'score' in col.lower()), None)

if label_col:
    full_result["DU_DOAN_THANHLY"] = predictions[label_col]
if score_col:
    full_result["XAC_SUAT"] = predictions[score_col]
```

Hình 26: Ghép kết quả dự đoán vào dữ liệu gốc.

- So sánh với thực tế (nếu có cột THANHLY). Tạo cột SO_SANH để biết dự đoán đúng hay sai.

```
if 'THANHLY' in full_result.columns and label_col:
    full_result["SO_SANH"] = full_result.apply(
        lambda row: "✅ Đúng" if row["DU_DOAN_THANHLY"] == row["THANHLY"] else "❌ Sai", axis=1
    )
```

Hình 27: So sánh với thực tế.

- Hiển thị kết quả: Hiển thị bảng kết quả với dự đoán.

```
# ✅ Kết quả đầy đủ
st.subheader("📋 Kết quả dự đoán")
st.dataframe(full_result)
```

Hình 28: Hiển thị kết quả.

- Hiển thị và tô màu dòng sai nếu có: Tùy kích thước dữ liệu, chỉ hiện dòng sai hoặc hiện toàn bộ và highlight dòng sai

```
# 🔎 So sánh nếu có thực tế
if "SO_SANH" in full_result.columns:
    st.subheader("📝 So sánh dự đoán với thực tế")
```

Hình 29 : Hiển thị tô màu dòng sai nếu có.

- Vẽ biểu đồ hình tròn: Vẽ biểu đồ hình tròn so sánh tỷ lệ đúng/sai.

```
# Tạo biểu đồ hình tròn
fig = px.pie(
    comparison_counts,
    names="Kết quả",
    values="Số lượng",
    color="Kết quả",
    color_discrete_map={"✅ Đúng": "green", "❌ Sai": "red"},
    title="Tỷ lệ dự đoán đúng và sai",
    hole=0.4  # Nếu muốn biểu đồ hình donut
)

st.plotly_chart(fig, use_container_width=True)
```

Hình 30: Biểu đồ so sánh tỷ lệ đúng/sai.

- Cho phép tải kết quả: Tải kết quả về dưới dạng CSV hoặc Excel.

```
# Tải kết quả CSV
csv_result = full_result.to_csv(index=False).encode('utf-8')
st.download_button("☁️ Tải kết quả CSV", data=csv_result, file_name='du_doan_churn_rate.csv', mime='text/csv')

# Tải kết quả Excel
excel_buffer = BytesIO()
with pd.ExcelWriter(excel_buffer, engine='xlsxwriter') as writer:
    full_result.to_excel(writer, sheet_name="DuDoan", index=False)
st.download_button("📥 Tải kết quả Excel (.xlsx)", data=excel_buffer.getvalue(),
                   file_name='du_doan_churn_rate.xlsx',
                   mime='application/vnd.openxmlformats-officedocument.spreadsheetml.sheet')
```

Hình 31 : Trả kết quả dưới dạng CSV hoặc Excel

- Xử lý lỗi: Hiển thị thông báo nếu quá trình xử lý có lỗi.

```
except Exception as e:
    st.error(f"❌ Lỗi khi xử lý dữ liệu: {e}")
```

Hình 32: Hiển thị thông báo lỗi

Dữ liệu sau khi dự đoán gồm các trường thông tin

STT: Số thứ tự

THUEBAO_ID: Mã của thuê bao khách hàng

MA_TB: Mã thuê bao

TEN_TB: Tên thuê bao

DIACHI_TB: Địa chỉ thuê bao

NGANHNGHE_ID: Mã ngành nghề

NGANHNGHE: Tên ngành nghề

NGAY_SN: Ngày sinh của khách hàng

TUOI: Tuổi

KHUVUC_ID: Mã khu vực

KHUVUC: Tên khu vực

LOAIKH_ID: Mã loại khách hàng

LOAIKH: Phân loại khách hàng. (chi tiết mô tả ở bảng 2.1)

KHDN: 1-Khách hàng doanh nghiệp; 0-Khách hàng cá nhân.

MANGKHAC: 1- có; 0-không.

SO_DV_KHAC: Số dịch vụ khác của VNPT mà khách hàng sử dụng ngoài FiberVNN.

GOI_DADV: Gói đa dịch vụ 1- có; 0-không. Sử dụng gói tích hợp hay không, tích hợp (MyTV, di động, Mesh…) hay không.

GOICUOC: Gói cước

GIACUOC: Gía cước

NOCUOC_2THANG: Số tháng nợ cước.

TRATRUOC: Đang thanh toán cước hàng tháng hay sử dụng gói trả trước nhiều tháng.1: Đã thanh toán trước cước 6 tháng, 12 tháng. 0: thanh toán cước hàng tháng.

SOTHANG_TRATRUOC_CONLAI: Số tháng trả trước còn lại.

SOLAN_BAOHONG: Số lần gọi báo hỏng do sự cố (đứt cáp, không truy cập được internet, mạng chậm…).

SOLAN_GOI_KIEM: Số lần gọi kiểm tra chất lượng dịch vụ.

SOLAN_GOI_KIEM_HL: Số lần gọi kiểm, khách hàng phản hồi hài lòng.

SOLAN_GOI_KIEM_KHL: Số lần gọi kiểm, khách hàng phản hồi không hài lòng

SOLAN_TAMNGUNG: Số lần tạm ngưng dịch vụ.

THANG_SD: Số tháng sử dụng.

KO_PSLL: Không phát sinh lưu lượng: Nhà mạng ghi nhận khách hàng không phát sinh lưu lượng sử dụng 5 ngày liên tiếp (do hỏng modem, đi vắng, mất điện,...) để thực hiện kiểm tra chất lượng dịch vụ.

SOLAN_GIAHAN: Số lần khách hàng gia hạn dịch vụ.

TRANGTHAITB_ID: Trạng thái thuê bao ID.

TRANGTHAI_TB: Trạng thái hoạt động của thuê bao.

THANHLY: Trạng thái của thuê bao: 0- Đang hoạt động, 1- Đã thanh lý.

SO_SANH: Đúng, sai so với kết quả thực tế

Tổng kết chương 2: Chương 2 đã trình bày chi tiết quy trình xây dựng mô hình dự báo khách hàng rời mạng, bao gồm các bước tiền xử lý dữ liệu, lựa chọn đặc trưng, chia tập dữ liệu, huấn luyện và kiểm thử. Đề án áp dụng hai thuật toán học máy chính là Cây Quyết định (Decision Tree) và Máy Vector Hỗ trợ (SVM), so sánh hiệu quả dựa trên các chỉ số như Accuracy, Precision, Recall, F1-score để lựa chọn mô hình tối ưu. Nội dung chương này đóng vai trò nền tảng cho việc triển khai cài đặt và đánh giá mô hình ở Chương 3

Chương 3: CÀI ĐẶT VÀ THỬ NGHIỆM

Chương này sẽ trình bày chi tiết về quá trình cài đặt môi trường thực nghiệm, xây dựng mô hình dự báo thử nghiệm, cũng như các phương pháp và kết quả đánh giá hiệu quả mô hình dự báo khách hàng rời mạng. Nội dung bao gồm các bước thiết lập phần mềm, công cụ hỗ trợ, cấu hình mô hình, quá trình huấn luyện, kiểm thử và phân tích các chỉ số đánh giá nhằm xác định chất lượng của mô hình được xây dựng.

3.1. Cài đặt môi trường.

Hệ thống được triển khai trên nền tảng máy chủ với cấu hình như sau:

- Hệ điều hành: Windows Server 2019
- Bộ xử lý (Processor): Intel® Xeon® Gold 5318Y CPU @ 2.10 GHz (02 processors).
- Bộ nhớ RAM: 32 GB
- Hệ thống: 64-bit Operating System, x64-based processor

Môi trường lập trình sử dụng:

- Ngôn ngữ: Python 3.11
- Các thư viện chính được cài đặt và sử dụng:
  - pandas: hỗ trợ xử lý, phân tích dữ liệu dạng bảng
  - numpy: hỗ trợ tính toán số học hiệu năng cao
  - streamlit: xây dựng giao diện web đơn giản để hiển thị kết quả và trực quan hóa mô hình

3.2. Cài đặt mô hình thử nghiệm

Để triển khai mô hình dự báo khách hàng rời mạng, hệ thống được cài đặt và thiết lập môi trường thử nghiệm theo các bước sau:

Cài đặt Python: 3.11

- Phiên bản sử dụng: Python 3.11
- Truy cập trang tải Python:

https://www.python.org/downloads/release/python-3110/

- Tải file cài đặt: Windows installer (64-bit)
- Tiến hành cài đặt:

⇨ Chạy file .exe, tick chọn: Add Python to PATH

⇨ Bấm Install Now

⇨ Xác minh: python --version

Cài đặt thư viện cần thiết:

Các thư viện chính được sử dụng để xử lý dữ liệu và xây dựng giao diện:

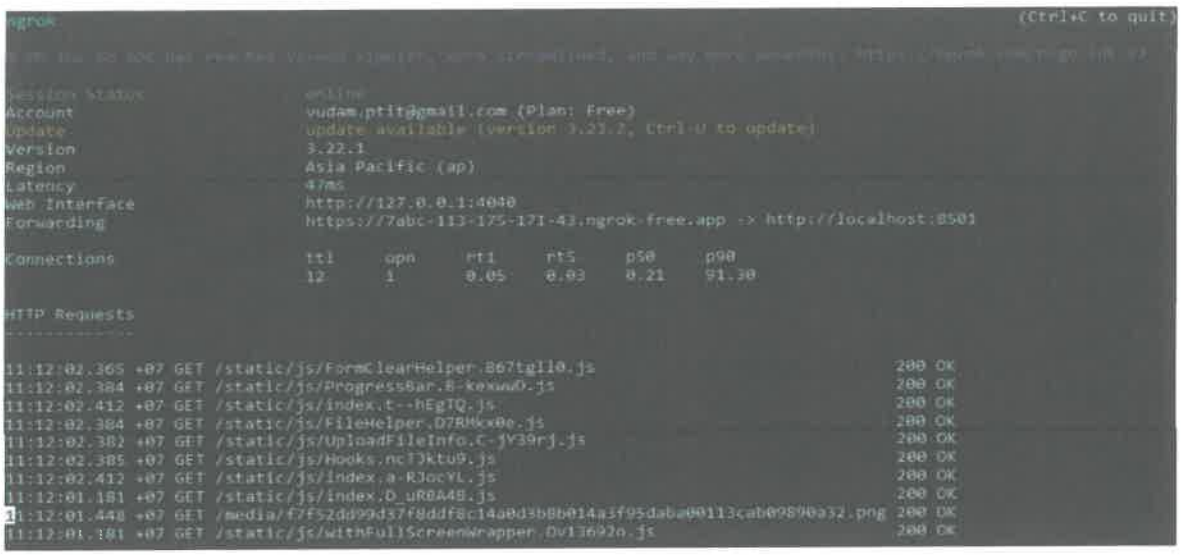
- pip install pandas
- pip install numpy
- pip install streamlits

Cài đặt ngrok:

Ngrok giúp tạo một địa chỉ URL tạm thời (public) để truy cập ứng dụng chạy trên máy cục bộ (localhost)

- Tạo tài khoản Ngrok: Vào: https://ngrok.com -> Đăng ký tài khoản (miễn phí)-> Sau khi đăng ký, vào Dashboard → lấy AuthToken
- Cài đặt Ngrok: Vào: https://ngrok.com/download-> Tải bản phù hợp với hệ điều hành-> Giải nén và đặt file ngrok vào thư mục mong muốn.
- Cấu hình token: Sau khi cài, chạy: ngrok config add-authtoken YOUR_AUTHTOKEN
- Chạy app với Ngrok:

App chạy ở localhost:8501 (Streamlit mặc định)



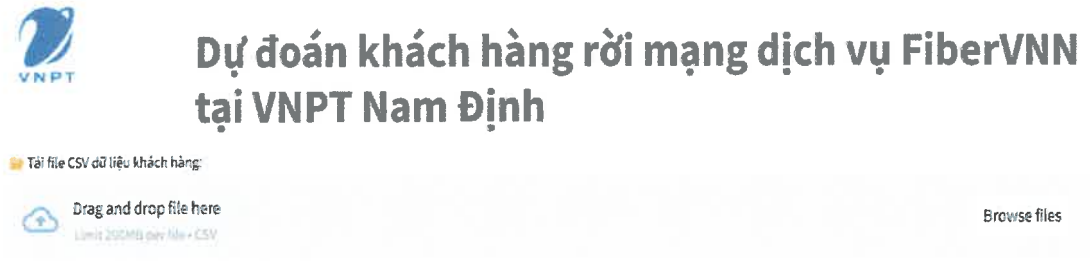
Hình 33: Chạy ứng dụng với Ngrok

Ngrok sẽ trả URL: *https:// 7abc-113-175-171-43.ngrok-free.app*

Hình 34:Giao diện Ứng dụng dụng dự đoán khách hàng rời mạng dịch vụ FiberVNN

3.3. Thử nghiệm và đánh giá

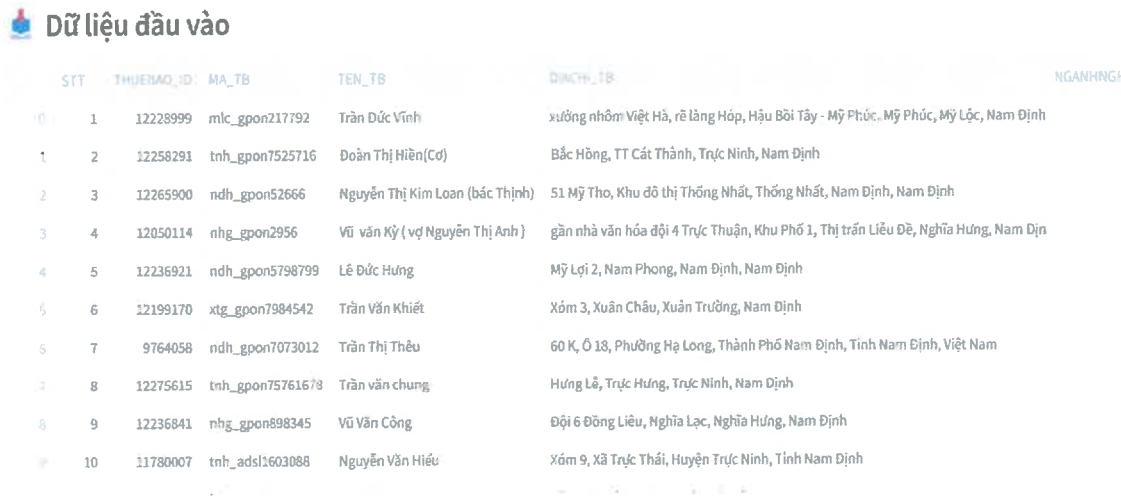
- Giao diện ứng dụng dự đoán khách hàng rời mạng dịch vụ FiberVNN tại VNPT Nam Định.



- Chọn Browse files để chọn file khách hàng dự đoán:

Hình 35: Chọn file huấn luyện.

- Dữ liệu đầu vào để dự báo:



Hình 36: Dữ liệu đầu vào.

- Kết quả dự báo:

Kết quả dự đoán

	KHL	SOLAN_TAMNGUNG	THANG_SD	KO_PSLL	SOLAN_GIAHAN	TRANGTHAITB_ID	TRANGTHAI_TB	THANHLY	DU_DOAN_THANHLY	XAC_SUAT	DU_DOAN
0	0	0	11	0	2	1	Hoạt động bình thường	0	0	1	Đúng
1	0	0	6	0	0	1	Hoạt động bình thường	0	0	1	Đúng
2	0	0	4	0	0	1	Hoạt động bình thường	0	0	1	Đúng
3	0	0	41	0	3	1	Hoạt động bình thường	0	0	1	Đúng
4	0	0	10	0	2	1	Hoạt động bình thường	0	0	1	Đúng
5	0	0	17	0	2	1	Hoạt động bình thường	0	0	1	Đúng
6	0	0	85	0	6	1	Hoạt động bình thường	0	0	1	Đúng
7	0	0	1	0	2	1	Hoạt động bình thường	0	0	1	Đúng
8	0	0	10	0	2	1	Hoạt động bình thường	0	0	1	Đúng
9	0	0	95	0	0	1	Hoạt động bình thường	0	0	1	Đúng

Hình 37: Kết quả dự báo

- So sánh dự báo với thực tế: Chỉ hiển thị dòng báo sai

So sánh dự đoán với thực tế

Dữ liệu quá lớn. Chỉ hiển thị dòng dự đoán sai.

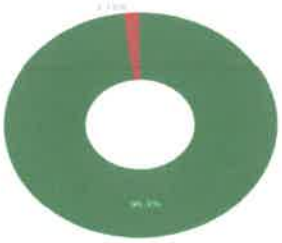
	KHL	SOLAN_TAMNGUNG	THANG_SD	KO_PSLL	SOLAN_GIAHAN	TRANGTHAITB_ID	TRANGTHAI_TB	THANHLY	DU_DOAN_THANHLY	XAC_SUAT	DU_DOAN
56	0	0	15	0	0	1	Hoạt động bình thường	1	0	1	Sai
200	0	1	11	0	0	6	Tạm dừng	1	0	1	Sai
202	0	0	16	0	0	1	Hoạt động bình thường	1	0	1	Sai
227	0	1	107	0	0	6	Tạm dừng	1	0	1	Sai
257	0	0	22	0	0	1	Hoạt động bình thường	1	0	1	Sai
265	0	1	16	0	0	6	Tạm dừng	1	0	1	Sai
278	0	0	25	0	3	1	Hoạt động bình thường	1	0	1	Sai
291	0	1	28	0	2	6	Tạm dừng	1	0	1	Sai
345	0	0	11	0	0	1	Hoạt động bình thường	1	0	1	Sai

Hình 38: So sánh dự báo rời mạng với thực tế

- Biểu đồ hình tròn, tỷ lệ dự báo đúng sai: Màu xanh là đúng, màu đỏ là sai

Biểu đồ hình tròn: Dự đoán đúng vs sai

Tỷ lệ dự đoán đúng và sai



Hình 39: Biểu đồ dự báo đúng sai

Tỉ lệ dự báo đúng đạt 98,3% , tỉ lệ dự báo sai là 1,75%.

Tổng kết chương 3: Chương 3 đã mô tả quy trình cài đặt môi trường và thử nghiệm mô hình dự báo, bao gồm cấu hình phần cứng, thiết lập phần mềm, các thư viện hỗ trợ (Python, Pandas, NumPy, Streamlit), và các bước triển khai mô hình. Kết quả thử nghiệm cho thấy các thuật toán được áp dụng có khả năng dự báo khách hàng rời mạng với độ chính xác cao, đồng thời cung cấp cơ sở để lựa chọn giải pháp mô hình hóa phù hợp cho VNPT Nam Định.

KẾT LUẬN

Sau quá trình tìm hiểu, nghiên cứu các phương pháp khai phá dữ liệu, kỹ thuật tiền xử lý dữ liệu, cũng như các phương pháp học máy và thuật toán dự báo nhằm xây dựng mô hình dự báo khách hàng có nguy cơ rời mạng, đồng thời tham khảo các công trình nghiên cứu có liên quan, đề án đã đạt được một số kết quả chính sau:

- Thu thập và xây dựng cơ sở dữ liệu khách hàng, bao gồm các thông tin đặc tả như: đối tượng khách hàng, loại hình khách hàng, gói cước sử dụng, tình trạng nợ cước, số lượng dịch vụ đang sử dụng và các đặc điểm liên quan khác.
- Rút trích các thuộc tính đặc trưng, xác định các yếu tố có khả năng ảnh hưởng đến nguy cơ rời mạng của khách hàng, qua đó xây dựng tập dữ liệu huấn luyện phục vụ cho việc xây dựng và đánh giá mô hình dự báo.
- Nghiên cứu và tìm hiểu về thuật toán Decision Tree và Support Vector Machine. Thuật toán Cây quyết định (Decision Tree): Đã tìm hiểu nguyên lý hoạt động của thuật toán, các tiêu chí tách nhánh (như Information Gain, Gini Index, Gain Ratio), cấu trúc cây nhị phân và đa nhánh, ưu nhược điểm của cây quyết định cũng như khả năng giải thích của mô hình. Thuật toán Máy vector hỗ trợ (Support Vector Machine - SVM): Đã nghiên cứu nguyên lý của SVM trong việc tìm siêu phẳng tối ưu nhằm phân tách các lớp dữ liệu, các hàm kernel (như linear, polynomial, RBF) để xử lý các bài toán phân loại phi tuyến, cũng như ưu điểm của SVM trong việc xử lý dữ liệu có kích thước lớn với số lượng mẫu hạn chế.

- Tiền xử lý dữ liệu và xây dựng mô hình dự báo có độ chính xác cao, cụ thể xây dựng mô hình dự báo với thuật toán Decision Tree độ chính xác 96.48%.

Hướng phát triển của đề án trong tương lai.

- Tiếp tục nghiên cứu và thử nghiệm các mô hình bằng các thuật toán như Random Forest/ Gradient Boosting / XGBoost / LightGBM / CatBoost, Logistic Regression, K-Nearest Neighbors (KNN), Neural Network (Mạng nơ-ron nhân tạo) Để tìm ra thuật toán tối ưu nhất cho bài toán.
- Dự báo và thử nghiệm trên các dịch vụ khác của VNPT Nam Định như MyTV, Điện thoại di động…
- Gửi thông báo đến nhân viên quản lý địa bàn về thông tin khách hàng có nguy cơ rời mạng để tiến hành chăm sóc và thuyết phục khách hàng tiếp tục sử dụng dịch vụ của đơn vị.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Dương Minh Lý, (2021), Luận văn Thạc sĩ Dự báo Khách hàng sử dụng dịch vụ FiberVNN của VNPT Tây Ninh có nguy cơ rời mạng, Học viện Công Nghệ Bưu Chính Viễn Thông cơ sở TP.HCM

[2] Nguyễn Thị Như Ngọc, (2014), Luận văn Thạc sĩ Phân tích dữ liệu thuê bao di động hướng đến dự báo thuê bao rời mạng viễn thông, Trường Đại học Công nghệ – Đại học Quốc gia Hà Nội

[3] T.M.Phương, Giáo trình Nhập môn trí tuệ nhân tạo, Hà Nội: Học viện Công nghệ Bưu chính Viễn Thông, 2015

[4] Võ Đức, V., & Trần, V. L. (2022). Dự Báo Rời Mạng Dịch Vụ Fiber. Tạp Chí Khoa học HUFLIT, 7(1), 3.

[5] Số liệu kinh doanh của VNPT Nam Định được truy xuất vào 31/03/2025

[6] Ionut Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Machine learning, 20(3), 273-297.

[7] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[8] J. Ross Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993

[9] Wu, Lin and Weng, "*Probability estimates for multi-class classification by pairwise coupling*", JMLR 5:975-1005, 2004.

[10] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[11] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. O'Reilly Media, 2022.

[12] Scikit-learn Developers, "Scikit-learn Documentation," 2024. [Online]. Available: https://scikit-learn.org/stable/

[13] Scikit-learn Developers, "Decision Trees — Scikit-learn Documentation," 2024. [Online]. Available: https://scikit-learn.org/stable/modules/tree.html

[14] Scikit-learn Developers, "Support Vector Machines — Scikit-learn Documentation," 2024.

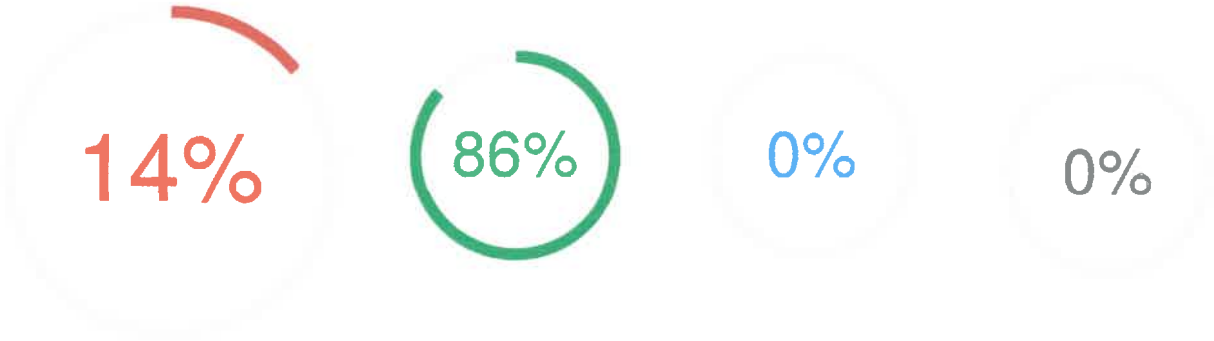
KiểmTraTàiLiệu

BÁO CÁO KIỂM TRA TRÙNG LẶP

Thông tin tài liệu

Tên tài liệu:	PTIT_Dean_VuVanDam_Toanvan
Tác giả:	Vũ Văn Đam
Điểm trùng lặp:	14
Thời gian tải lên:	17:10 23/06/2025
Thời gian sinh báo cáo:	09:16 26/06/2025
Các trang kiểm tra:	57/57 trang

Kết quả kiểm tra trùng lặp



Có 14% nội dung trùng lặp

Có 86% nội dung không trùng lặp

Có 0% nội dung người dùng loại trừ

Có 0% nội dung hệ thống bỏ qua

Học viên

Vũ Văn Đam

GVHD

Phan Thị Hà

Nguồn trùng lặp tiêu biểu

123docz.net tailieu.vn sti.vista.gov.vn

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP

Họ và tên học viên: Vũ Văn Đam

Chuyên ngành: Khoa học máy tính

Khóa: 2022 đợt 1

Tên đề tài: Dự báo khách hàng rời mạng dịch vụ FiberVNN tại VNPT Nam Định.

Người hướng dẫn khoa học: TS. Phan Thị Hà

Ngày bảo vệ: 19/07/2025

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Chỉnh sửa lỗi soạn thảo, lỗi ngữ pháp, chính tả	Học viên đã rà soát, chỉnh sửa các lỗi soạn thảo, các lỗi ngữ pháp
2	Bổ sung tài liệu tham khảo	Tiếp thu góp ý của Hội đồng, tác giả đã bổ sung thêm 5 tài liệu tham khảo
3	Chỉnh sửa phần đặt vấn đề	Tiếp thu góp ý của Hội đồng, tác giả đã bổ sung, chỉnh sửa tại mục 4, 5 trang 6.
4	Cần làm rõ đầu vào và đầu ra của bài toán	Tiếp thu góp ý của Hội đồng, tác giả đã bổ sung tại mục 2.2.1 trang 30 và trang 40
5	Cần phải trình bày trực tiếp, rõ ràng hơn	Tiếp thu ý kiến đóng góp của Hội đồng. Tác giả đã đổi tên Chương 1 thành Cơ sở lý luận thành Tổng quan về bài toán dự báo khách hàng rời mạng dịch vụ, đã bổ sung phần tóm tắt đầu chương 2 và kết chương tại mỗi chương.

Hà Nội, ngày 01 tháng 8 năm 2025

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
TS. Nguyễn Duy Phương	TS. Đặng Hoàng Long	TS. Phàn Thị Hà	Vũ Văn Đam

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 1098/QĐ-HV ngày 26 tháng 06 năm 2025 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi.12....giờ.00..phút, ngày 19 tháng 07 năm 2025 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Vũ Văn Đam**

Tên đề án tốt nghiệp: **Dự báo khách hàng rời mạng dịch vụ FiberVNN tại VNPT Nam Định**

Chuyên ngành: **Khoa học máy tính** Mã số: **8480101**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: ...5..../ 05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	**TS. Nguyễn Duy Phương**	Chủ tịch	
2	**TS. Đặng Hoàng Long**	Thư ký	
3	**PGS.TS. Trần Thị Oanh**	Phản biện 1	
4	**PGS.TS. Đặng Văn Đức**	Phản biện 2	
5	**TS. Vũ Văn Thỏa**	Uỷ viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên. (có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

PGS.TS. Đặng Văn Đức
Bổ sung và làm rõ phương pháp đã được trình bày trong đề án
PGS.TS. Trần Thị Oanh
Dữ liệu có đang được sử dụng có cân bằng không? Cần bổ sung phần đánh giá dữ liệu

7. Trả lời của học viên:

TẬP ĐOÀN BƯU CHÍNH VIỄN THÔNG VN
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH-VIỄN THÔNG

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

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SỸ

(Dùng cho cán bộ phản biện)

Tên đề tài đề án: **Dự báo khách hàng rời mạng dịch vụ FiberVNN tại VNPT Nam Định**
Chuyên ngành: Khoa học máy tính
Mã số: 8.48.01.04
Tên học viên: *Vũ Văn Đam*
Họ và tên người nhận xét: PGS.TS. Đặng Văn Đức
Chuyên ngành: Công nghệ thông tin
Cơ quan công tác: Viện Công nghệ thông tin, Viện Hàn lâm KH&CN Việt Nam

NỘI DUNG NHẬN XÉT

I. Cơ sở khoa học và thực tiễn, tính cấp thiết của đề án

Trong bối cảnh thị trường viễn thông cạnh tranh ngày càng gay gắt, việc giữ chân khách hàng trở thành một yếu tố then chốt quyết định sự phát triển bền vững của các nhà cung cấp dịch vụ, trong đó có mạng internet cố định FiberVNN. Hiện tượng khách hàng hủy dịch vụ không chỉ làm giảm doanh thu mà còn kéo theo chi phí lớn trong việc thu hút khách hàng mới. Do đó, nhu cầu dự báo chính xác hành vi rời mạng của khách hàng để có biện pháp can thiệp sớm là hết sức cấp thiết.

Sự phát triển mạnh mẽ của trí tuệ nhân tạo, đặc biệt là các kỹ thuật học máy, đã mở ra hướng tiếp cận hiệu quả cho bài toán này. Việc ứng dụng học máy giúp khai thác triệt để dữ liệu người dùng để phát hiện sớm các dấu hiệu rời mạng, từ đó hỗ trợ doanh nghiệp xây dựng các chiến lược chăm sóc khách hàng phù hợp, tối ưu nguồn lực và nâng cao chất lượng dịch vụ. Chủ đề nghiên cứu được giới khoa học trong và ngoài nước quan tâm. Với mục tiêu nghiên cứu học hỏi để ứng dụng vào một đơn vị cụ thể là VNPT Nam Định, đề tài đề án tốt nghiệp thạc sĩ của học viên *Vũ Văn Đam* có ý nghĩa khoa học và thực tiễn trong quản trị kinh doanh dịch vụ viễn thông.

II. Về nội dung, chất lượng của đề án, các kết quả đã đạt được

Để đạt được mục tiêu đề ra cho đề án tốt nghiệp của mình, học viên đã trình bày các vấn đề chính sau đây:

- Trình bày được tổng quan ngắn gọn về bài toán dự báo khách hàng rời mạng bằng phương pháp học máy, bao gồm sơ lược về đơn vị VNPT Nam Định và hai thuật toán học máy Cây quyết định và SVM sẽ áp dụng vào đề án này.
- Đã trình bày được phương pháp và các kỹ thuật xây dựng mô hình dự báo khách hàng rời mạng sử dụng các thư viện của Python.

- Cài đặt thực nghiệm với tập dữ liệu thực của VNPT Nam Định, có nhận xét, đánh giá kết quả thu được.

Nội dung nghiên cứu của đề án còn đơn giản nhưng mang tính ứng dụng cao, phù hợp với cấp độ đề án thạc sỹ kỹ thuật theo định hướng ứng dụng. Bản đề án được trình bày rõ ràng với cấu trúc hợp lý, ít lỗi in ấn. Tuy nhiên, học viên cần nghiên cứu chỉnh sửa bản đề án của mình theo các góp ý sau đây:

- Rà soát và chỉnh sửa các lỗi in ấn, câu thừa, không rõ nghĩa trong toàn bộ đề án, ví dụ trang 6. Hình vẽ mờ, chữ quá nhỏ khó theo dõi cần chỉnh sửa.
- Tên chương 1 nên đổi Cơ sở lý luận thành Tổng quan về bài toán dự báo khách hàng rời mạng dịch vụ. Cuối các chương cần có kết chương. Đầu chương 2 nên có tóm tắt nội dung sẽ trình bày trong chương.
- Đã có sơ đồ tổng quan hệ thống dự báo khách hàng rời mạng, tuy nhiên lựa chọn các ký pháp phù hợp hơn: ví dụ, tiến trình trong ô bầu dục mô tả hành động, xử lý dữ liệu; thực thể ngoài sử dụng hình chữ nhật...
- Nên có sơ đồ trực quan các thuật toán của hành động, xử lý dữ liệu. Có thể sử dụng Flowchart hoặc UML cho các sơ đồ trực quan và sơ đồ tổng quan hệ thống. Không nên sử dụng các dòng lệnh Python mô tả phương pháp mà sẽ sử dụng nó trong chương 3 Cài đặt và thực nghiệm.
- Nên thử nghiệm với các phương pháp học máy khác hiệu quả hơn để lựa chọn giải pháp phù hợp với dữ liệu của một đơn vị cụ thể VNPT Nam Định.
- Nên kết hợp với các mô hình phụ trợ để tăng hiệu năng hệ thống, ví dụ SMOTE để xử lý mất cân bằng lớp (nếu khách hàng rời mạng chiếm tỉ lệ nhỏ), RFE (*Recursive Feature Elimination*) để chọn lọc đặc trưng quan trọng thay cho thủ công.
- Câu hỏi: Tại sao không sử dụng XGBoost (*Extreme Gradient Boosting*) kết hợp nhiều mô hình cây quyết định với khả năng xử lý tốt dữ liệu phi tuyến và mất cân bằng, hiệu suất và độ chính xác cao.

III. Kết luận

Bản đề án tốt nghiệp thạc sỹ của học viên đã trình bày được một số nội dung cơ bản để đạt được mục tiêu đề ra. Tuy nhiên, bản đề án này cần được chỉnh sửa để bảo đảm đầy đủ các tiêu chuẩn cơ bản của đề án thạc sỹ kỹ thuật định hướng ứng dụng theo các góp ý trên đây. Tôi đồng ý để học viên *Vũ Văn Đam* được bảo vệ trước Hội đồng bảo vệ đề án tốt nghiệp thạc sỹ.

Hà Nội, ngày 19 tháng 07 năm 2025

NGƯỜI NHẬN XÉT

PGS.TS Đặng Văn Đức

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: Dự báo khách hàng rời mạng dịch vụ FiberVNN tại VNPT Nam Định

Chuyên ngành: Khoa học máy tính

Mã chuyên ngành: 8.48.01.01

Họ và tên học viên: Vũ Văn Đam

Họ và tên người nhận xét: Trần Thị Oanh

Học hàm, học vị: PGS.TS

Chuyên ngành: Khoa học máy tính

Cơ quan công tác: Trường Quốc tế - ĐHQG Hà Nội

Số điện thoại: 0362220684 E-mail: tranthioanh@vnu.edu.vn

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Đề tài nghiên cứu xây dựng mô hình dự báo khách hàng rời mạng dịch vụ tại VNPT và tiến hành thử nghiệm trên bộ dữ liệu được thu thập tại Nam Định. Mô hình sử dụng hai thuật toán phân lớp là cây quyết định và svm. Kết quả thử nghiệm cho thấy tính khả thi của mô hình, và có khả năng triển khai ứng dụng tại Nam Định. Do vậy, đây là một đề tài có cơ sở khoa học và thực tiễn ứng dụng.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

Đề án tốt nghiệp đã thực hiện:

+ Xây dựng mô hình dự báo khách hàng rời mạng dịch vụ tại VNPT Nam Định

+ Thử nghiệm và so sánh kết quả của 02 mô hình học máy

III/ Những vấn đề cần giải thích thêm:

- Cần mô tả rõ hơn về bộ dữ liệu sử dụng ,ý nghĩa các trường thông tin cho việc dự đoán sự rời bỏ dịch vụ mạng.

- Các hình ảnh không phải do mình tạo ra cần tham chiếu tài liệu gốc

- Chương 02 cần bổ sung các côgn việc liên quan gần đây tới dự đoán sự rời bỏ dịch vụ của khách hàng hoặc nhân viên trong công ty nói chung.

- Các công thức cần đánh số để dễ theo dõi.

- Tài liêu tham khảo còn ít, cần bổ sung thêm các tài liệu mới hơn gần đây.

- Về nội dung:

+ Do dư liệu là mất cân bằng, nên cần thử nghiệm thêm các kỹ thuật xử lý imbalanced để cải thiện hiệu năng của mô hình học máy.

+ Cần đánh giá kết quả dự đoán với nhãn 1, tức là xác định khách hàng có khả năng rời bỏ dịch vụ sớm để công ty tìm cách giữ chân các đối tượng khách hàng đó. Ngoài ra, cần mô tả một số đặc điểm của khách để hiểu rõ hơn lý do đặc điểm khách hàng rời bỏ nhằm lên giải pháp sớm khắc phục hoặc làm tăng sự hài lòng trong dịch vụ chăm sóc khách hàng

+ Chú ý việc chia các fold phải đảm bảo tính cân bằng dư liệu nhãn như trong bộ dữ liệu gốc.

Câu hỏi:

+ Với dữ liệu mất cân bằng thì có gợi ý gì về các phương pháp có thể xử lý tốt hơn cho loại dữ liệu này so với mô hình như cây quyết định và svm?

+ Trong các thuộc tính, thì thuộc tính nào có yếu tố quyết định quan trọng nhất đối với việc rời bỏ dịch vụ mạng tại VNPT Nam Định.

IV/ Kết luận:

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp. Tuy nhiên, học viên cần làm rõ thêm các ý được nêu ở mục III.

Ngày 19 tháng 7 năm 2025

NGƯỜI NHẬN XÉT

Trần Thị Oanh