LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi.

Các số liệu, kết quả nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác. Nếu không đúng như đã nêu trên, tôi xin hoàn toàn chịu trách nhiệm về đề tài của mình.

Tp. HCM, ngày 15 tháng 07 năm 2022

Học viên thực hiện luận văn

Lê Đức Hòa Bình

LỜI CẢM ƠN

Trong thời gian thực hiện luận văn tốt nghiệp, được sự hướng dẫn tận tình của giáo viên hướng dẫn và được phía nhà trường tạo điều kiện thuận lợi, tôi đã có một quá trình nghiên cứu, tìm hiểu và học tập nghiêm túc để hoàn thành đề tài. Kết quả thu được không chỉ do nỗ lực của cá nhân tôi mà còn có sự giúp đỡ của quý thầy cô, gia đình và các bạn.

Tôi xin chân thành cảm ơn **TS. Tân Hạnh**. Thầy đã hướng dẫn, hỗ trợ tôi hoàn thành tốt luận văn về phương pháp, lý luận và nội dung luận văn.

Cảm ơn Bán Giám Hiệu, Khoa Đào Tạo Sau Đại Học, Phòng Đào Tạo & KHCN – Học Viện Công Nghệ Bưu Chính Viễn Thông cơ sở tại Tp. HCM đã quan tâm, tạo điều kiện giúp tôi hoàn thành luận văn tốt nghiệp.

Cám ơn Ban giám đốc và các đồng nghiệp tại Viễn thông Tây Ninh đã hỗ trợ, giúp đỡ tôi trong suốt quá trình thực hiện luận văn.

Trong quá trình thực hiện và trình bày không thể tránh khỏi những hạn chế, do vậy tôi rất mong nhận được sự góp ý, nhận xét phê bình của quý thầy cô và các bạn để hoàn thiện kiến thức và bản thân.

Tp. HCM, ngày 15 tháng 07 năm 2022

Học viên thực hiện luận văn

Lê Đức Hòa Bình

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
AI	Artificial Intelligence	Trí tuệ nhân tạo
PDF	Portable Document Format	Định dạng văn bản đơn giản
RF	Random Forest	Rừng ngẫu nhiên
ANN	Artificial Neural Network	Mạng nơ-ron nhân tạo
CSDL	Database	Cơ sở dữ liệu
CNTT	Information Technology	Công nghệ thông tin
SVM	Support Vector Machines	Máy véc tơ hỗ trợ
BTS	Bug Tracking System	Hệ thống kiểm tra sự cố
CQĐ	Decision Tree	Cây quyết định

DANH SÁCH HÌNH VẼ

DANH SÁCH BẢNG

MỞ ĐẦU

Đặt vấn đề

Trong dòng chảy liên tục của thời đại, xu thế phát triển của ngành Viễn thông được dự đoán là không thể tránh khỏi. Trước tình hình đó, một quốc gia đang phát triển như Việt Nam có rất nhiều điều kiện thuận lợi để phát triển ngành này ở tương lai.

Với xu hướng phát triển của ngành viễn thông như trên, nên đây lã lĩnh vực rất hấp dẫn cho các doanh nghiệp phát triển, thuận lợi rất nhiều nhưng cũng rất nhiều thách thức, do các doanh nghiệp canh tranh quyết liệt để thu hút khách hàng, giành thị phần. Nếu không liên tục thay đổi thích ứng với thị trường thì việc bị đào thải là đều tất yếu.

Trong doanh nghiệp, đặc biệt là VNPT việc tìm kiếm khách hàng là mục tiêu quan trọng để đảm bảo doanh thu và lợi nhuận cho doanh nghiệp.

Việc khách hàng hài lòng sau khi sử dụng dịch vụ phụ thuộc vào rất nhiều yếu tố khách quan và chủ quan. Trong đó tư vấn cho khách hàng một gói cước phù hợp là cực kì quan trọng. Việc này lâu nay vẫn thường xuyên được phân tích, tuy nhiên thực hiện bằng các biện pháp thủ công, thô sơ mất rất nhiều thời gian, và đòi hỏi người phân tích phải có chuyên môn tương đối tốt, nhưng độ chính xác mang lại tương đối không cao.

Do đó để có biện pháp phấn tích khoa học và hiện đại khắc phục các tồn tại như đã mô tả, khi đề tài hoàn thiện nhiều người có thể sử dụng. Trong báo cáo này sử phương pháp học máy để phân tích dự đoán các yếu tố ảnh hưởng đến gói cước sử dụng dịch vụ của khách hàng tại VNPT Tây Ninh. Kết quả tư vấn chính xác, nhanh giúp doanh nghiệp phát triển khách hàng mới, cũng như đảm bảo chất lượng dịch vụ phù hợp với nhu cầu sử dụng của khách hàng.

Đó là lý do luận văn chọn đề tài: “Hỗ trợ quyết định kinh doanh dịch vụ Viễn thông theo xu hướng khách hàng ở Tây Ninh”.

Mục đích nghiên cứu

Mục đích nghiên cứu phân tích dữ liệu khách hàng thu thập tại VNPT Tây Ninh:

- Xác định các yếu tố có ảnh hưởng đến gói cước phù hợp nhất với khách hàng.

- Phân tích sự ảnh hưởng của các yếu tố đó như thế nào đến gói cước mà khách hàng cần đăng ký.

- Đề xuất gói cước cho khách hàng bằng học máy.

Đối tượng và phạm vi nghiên cứu

Đối tượng, phạm vi nghiên cứu trên cơ sở dữ liệu thực tế thu thập từ tập khách hàng hiện hữu đang sử dụng dịch vụ Internet của VNPT Tây Ninh.

Nghiên cứu phương pháp xử lý, phân tích dữ liệu, các phương pháp học máy phù hợp với bộ dữ liệu của đề tài, trên nên tảng Python.

Phương pháp nghiên cứu

Phương pháp nghiên cứu lý thuyết:

- Tổng hợp, nghiên cứu các tài liệu về xử lý, mã hóa, phân tích dữ liệu, học máy, kỹ thuật lập trình.

- Sử dụng phương pháp nghiên cứu phân tích dữ liệu, phương pháp dự đoán và phương pháp thực nghiệm để so sánh, đánh giá và phân tích các kết quả đạt được.

Phương pháp nghiên cứu thực nghiệm: sau khi nghiên cứu lý thuyết, các bài toán tiến hành đề xuất mô hình khuyến nghị gói cước cho khách hàng.Đánh giá các kết quả đạt được; công bố kết quả nghiên cứu.

Ý nghĩa khoa học và thực tiễn

Ý nghĩa khoa học của luận văn: tập trung phân tích các số liệu thu thập được tại VNPT Tây Ninh, để xác định mức độ tương quan của các yếu tố ảnh hưởng đến gói cước của khách hàng.Phân tích các yếu tố ảnh hưởng nhờ áp dụng các phương pháp học máy như cây quyết định, rừng ngẫu nhiên để đưa ra các khuyến nghị gói cước phù hợp với khách hàng.

Ý nghĩa thực tiễn: xây dựng mô hình khuyến nghị gói cước cho khách hàng bằng học máy để giúp thay thế nhân viên tư vấn bán hàng đưa ra gói cước phù hợp với khách hàng.

Bố cục của báo cáo: báo cáo bao gồm 5 chương cùng với phần mở đầu, phần mục lục, phần tài liệu tham khảo.

Chương 1- Cơ sở lý thuyết và các công trình nghiên cứu có liên quan: Trình bày một số khải niệm có liên quan đến máy học, thuật toán cây quyết định. Ngoài ra, chương 1 còn đề cập đến một số công trình nghiên cứu có liên quan.

Chương 2 – Phương pháp khuyến nghị gói cước: Trình bày các phương pháp, định hướng để xây dựng mô hình khuyến nghị gói cước.

Chương 3 - Xây dựng mô hình: Trình bày các bước xây dựng mô hình khuyến nghị gói cước dựa vào thuật toán Rừng ngẫu nhiên.

Chương 4 – Phân tích và đánh giá: Đánh giá kết quả đạt được sau khi xây dựng mô hình Khuyến nghị gói cước dựa vào mức độ chính xác của mô hình.

CHƯƠNG 1 - CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU CÓ LIÊN QUAN

Trong chương 1 chúng ta xác định, và làm rõ các cơ sở lý thuyết, căn cứ khoa học, các nghiên cứu thực tiễn về các nội dụng có liên quan, hoặc công trình nghiên cứu tương tự để nghiên cứu áp dụng vào mục đích nghiên cứu đề tài này.

1.1. Tổng quan về học máy

Trong các lĩnh vực khoa học, công nghệ và nhân văn khác nhau, cũng như trong sinh học, khí tượng, y học hoặc tài chính, để trích dẫn một số, các chuyên gia nhắm vào dự đoán một hiện tượng dựa trên các quan sát hoặc đo lường trong quá khứ. Ví dụ, các nhà khí tượng học cố gắng dự báo thời tiết cho những ngày tiếp theo từ điều kiện khí hậu của những ngày trước đó. Trong y học, luyện tập thu thập các phép đo và thông tin như huyết áp, tuổi hoặc tiền sử chẩn đoán tình trạng của bệnh nhân. Ban đầu, trong hóa học, các hợp chất được phân tích bằng cách sử dụng khối phổ thử các phép đo để xác định xem chúng có chứa một loại phân tử hoặc nguyên tử. Trong tất cả các trường hợp này, mục tiêu là sự thay đổi của một biến phản hồi dựa trên một tập hợp các yếu tố dự đoán được quan sát. Trong nhiều thế kỉ, các nhà khoa học đã giải quyết những vấn đề như vậy bằng cách dẫn xuất theo khuôn khổ lý thuyết từ các nguyên tắc đầu tiên hoặc đã tích lũy kiến thức để mô hình hóa, phân tích và hiểu các vấn đề đang nghiên cứu. Ví dụ, các học viên biết từ những bệnh nhân cũ trong quá khứ, bệnh nhân cao tuổi bị đau tim với huyết áp thấp nói chung là rủi ro cao. Tương tự, các nhà khí tượng học biết từ lớp học các mô hình khí hậu mà một ngày nắng nóng, ô nhiễm cao có khả năng xảy ra tiếp theo là các diễn biến khác. Tuy nhiên, đối với một số vấn đề ngày càng tăng về số lượng, các phương pháp tiếp cận tiêu chuẩn bắt đầu chỉ ra các giới hạn của nó. Ví dụ, xác định thâm nhập các yếu tố nguy cơ di truyền đối với bệnh tim, nơi mà kiến thức vẫn còn rất thưa thớt, gần như không thực tế đối với khả năng nhận thức của con người do sự phức tạp cao và phức tạp của các tương tác tồn tại trong gen di truyền. Tương tự như vậy, đối với các dự báo khí tượng chi tiết, một số lượng lớn các biến cần phải được tính đến, nhanh chóng

vượt ra ngoài khả năng của các chuyên gia để đưa tất cả họ vào một hệ phương trình. Để phá vỡ rào cản nhận thức này, máy móc với tốc độ và công suất ngày càng tăng đã được xây dựng và thiết kế từ giữa thế kỷ XX để hỗ trợ con người trong tính toán của họ. Tuy nhiên, thật đáng ngạc nhiên, cùng với sự tiến bộ này về phần cứng, sự phát triển trong khoa học máy tính lý thuyết, trí thông minh nhân tạo và số liệu thống kê nhanh chóng đã chứng minh máy móc trở nên vượt trội hơn máy tính. Những tiến bộ gần đây đã khiến họ trở thành chuyên gia trong lĩnh vực riêng, có khả năng học hỏi từ dữ liệu và tự khám phá cấu trúc dự đoán của các vấn đề. Các kỹ thuật và thuật toán bắt nguồn từ lĩnh vực máy học đã thực sự trở thành một công cụ mạnh mẽ để phân tích dữ liệu lớn và phức tạp, hỗ trợ thành công các nhà khoa học trong nhiều bước đột phá của các biến thể trong lĩnh vực khoa học và công nghệ. Ví dụ công khai và nổi tiếng bao gồm việc sử dụng cây quyết định tăng cường trong phân tích thống kê dẫn đến việc phát hiện Higgs boson tại CERN [25], việc sử dụng các rừng ngẫu nhiên để phát hiện tư thế con người ở Microsoft Kinect [26] hoặc bộ phận tổng hợp các kỹ thuật học máy khác nhau để xây dựng hệ thống IBM tại Watson [27], có khả năng cạnh tranh với người đàn ông vô địch trên chương trình đố vui truyền hình Jeopardy của Mỹ. Về mặt hình thức, học máy có thể được định nghĩa là nghiên cứu các hệ thống có thể học từ dữ liệu mà không cần được lập trình rõ ràng. Một chương trình máy tính được cho là học từ dữ liệu và đo lường hiệu suất nếu hiệu suất của nó ở những tác vụ đó được cải thiện cùng với dữ liệu. Đặc biệt, học máy cung cấp các thuật toán có thể giải quyết các nhiệm vụ hồi quy, do đó mang đến các quy trình tự động để dự đoán một hiện tượng dựa trên những quan sát trong quá khứ. Tuy nhiên, từ trước đến nay, mục tiêu của học máy không chỉ là tạo ra các thuật toán đưa ra dự đoán chính xác, nó cũng là để cung cấp thông tin chi tiết về cấu trúc của dữ liệu. Đối với các học viên, không phải là chuyên gia trong lĩnh vực máy học, nó cung cấp các diễn giải thực sự quan trọng như độ chính xác của dự đoán. Nó cho phép hiểu rõ hơn trong việc tìm hiểu hiện tượng đang nghiên cứu, khám phá dữ liệu tốt hơn và tự đạt kết quả dễ dàng hơn.

1.1.1. Khái niệm

Học máy là một những lĩnh vực của trí tuệ nhân tạo, học máy liên quan đến quá trình nghiên cứu và xây dựng các kĩ thuật giúp các hệ thống máy tính học tự động từ dữ liệu ban đầu để giải quyết một số vấn đề cụ thể nào đó.

Học máy là một quá trình tự động của các quá trình học và việc học thì tương đương với quá trình xây dựng các tập luật trên cơ sở quan sát các trạng thái của cơ sở dữ liệu và những sự thay đổi của chúng. Học máy là lĩnh vực rộng lớn và nó không chỉ bao gồm việc học từ các mẫu, mà còn là học tăng cường. Các thuật toán học máy dựa trên tập dữ liệu mẫu và các thông tin liên quan để làm đầu vào và trả về kết quả đầu ra là một mô hình diễn tả những kết quả học được.

Nhìn chung, học máy sẽ sử dụng một tập hữu hạn các dữ liệu được gọi là tập huấn luyện. Tập này sẽ chứa các mẫu dữ liệu mà nó được chuẩn hóa bằng mã theo một cách nào đó để máy có thể đọc và hiểu được. Tuy nhiên có một sự thật là tập huấn luyện bao giờ cũng có hữu hạn các phần tử, vì vậy không phải toàn bộ dữ liệu sẽ được học một cách chính xác.

1.1.2. Phân loại các kỹ thuật học máy

Các thuật toán học máy được chia làm 3 loại chính: học có giám sát, học không giám sát và học bán giám sát.

Học có giám sát

Học có giám sát là phương pháp học từ những dữ liệu mà trong quá trình học các kỹ thuật học máy sẽ giúp hệ thống xây dựng cách xác định những lớp dữ liệu. Hệ nthống bắt buộc phải tìm ra một sự mô tả cho từng lớp dữ liệu. Sau đó người ta có thể sử dụng các luật phân loại được hình thành trong quá trình học và phân lớp nó để có thể sử dụng cho việc dự báo các lớp dữ liệu sau này.

Học không giám sát

Học không giám sát là hệ thống khai thác dữ liệu ứng dụng với những dữ liệu không có lớp được định nghĩa cụ thể từ trước, mà để máy học phải tự hệ thống quan sát các mẫu và nhận ra mẫu. Hệ thống này sẽ dẫn đến một tập lớp, mỗi lớp có một

tập mẫu riêng được khám phá từ trong tập dữ liệu. Học không giám sát hay còn gọi là học từ quan sát và khám phá.

Học bán giám sát

Đây là các thuật toán học tích hợp từ việc học giám sát và việc học không giám sát. Học bán giám sát sẽ sử dụng cả dữ liệu đã gán nhãn và chưa gán nhãn để huấn luyện – điển hình là một số ít dữ liệu có gán nhãn cùng với lượng lớn dữ liệu chưa gán nhãn ban đầu.

Học bán giám sát là quá trình học đứng giữa học không giám sát (không có bất kì dữ liệu đã được nhãn nào) và có giám sát (toàn bộ dữ liệu đều được gán nhãn). Việc học bán giám sát tận dụng những ưu điểm của việc học giám sát và học không giám sát và loại bỏ những khuyết điểm thường gặp trên hai kiểu học này.

1.2. Bài toán phân lớp dữ liệu

1.2.1. Khái niệm về phân lớp dữ liệu và bài toán phân lớp dữ liệu

Khai phá dữ liệu: Khai phá dữ liệu nói chung có nghĩa là khai thác hoặc đào sâu vào dữ liệu ở các dạng khác nhau để có được các mẫu và để có được kiến thức về mẫu đó. Trong quá trình khai thác dữ liệu, các tập dữ liệu lớn trước tiên được sắp xếp, sau đó các mẫu được xác định và các mối quan hệ được thiết lập để thực hiện phân tích dữ liệu và giải quyết vấn đề [28].

Phân lớp dữ liệu: Đây là một nhiệm vụ phân tích dữ liệu, tức là quá trình tìm kiếm một mô hình mô tả và phân biệt các lớp và khái niệm dữ liệu. Phân loại là vấn đề xác định một tập hợp các danh mục (quần thể con), một dữ liệu mới thuộc về loại nào, trên cơ sở một tập dữ liệu huấn luyện chứa các dữ liệu và các lớp của chúng đã được biết đến [28].

Phân lớp dữ liệu có thể chia làm các bước sau:

Bước học tập (Giai đoạn đào tạo): Xây dựng mô hình phân loại. Các thuật toán khác nhau được sử dụng để xây dựng mô hình phân loại bằng cách làm cho mô hình học bằng cách sử dụng tập huấn luyện có sẵn. Mô hình phải được đào tạo để dự

đoán kết quả chính xác. Dữ liệu kiểm tra được sử dụng để ước tính độ chính xác của quy tắc phân loại.

Bước phân loại: Mô hình được sử dụng để dự đoán và thử nghiệm mô hình đã xây dựng trên dữ liệu thử nghiệm và sau đó ước tính độ chính xác của các quy tắc phân loại. Dữ liệu kiểm tra được sử dụng để ước tính độ chính xác của quy tắc phân loại.

Ta có thể phát biểu bài toán phân lớp dữ liệu như sau:

Đầu vào của bài toán phân lớp dữ liệu:

Cho tập dữ liệu ban đầu $D = \{(x_i, y_i) \mid i = 1, 2, \ldots, n\}$, trong đó, $x_i = (x_{i1}, x_{i2}, ..., x_{ik}) \in R^k$ là dữ liệu gồm k thuộc tính ứng với tập thuộc tính $A = \{A_1, A_2, \ldots, A_k\}$ và $y_i \in C = \{c_1, c_2, \ldots, c_m\}$ là tập nhãn của các lớp dữ liệu ban đầu.

Đầu ra của bài toán phân lớp dữ liệu:

Một mô hình phân lớp $F: R^k \rightarrow C$, tương ứng mỗi phần tử $x \in R^k$ là một nhãn lớp $F(x) \in C$, sao cho đối với tập mẫu đầu vào D là phù hợp nhất theo nghĩa sau đây:

$||F(x_i) - y_i|| \cong 0$, với mọi $(x_i, y_i) \in D$ và $|| \ ||$ là một độ đo nào đó.

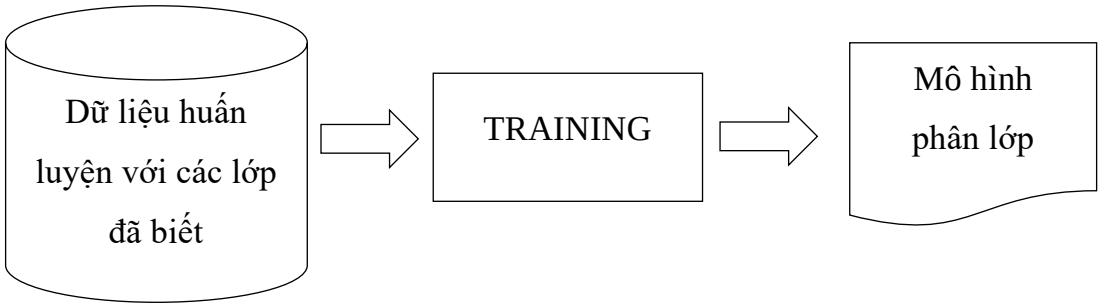
1.2.2. Các bước giải quyết bài toán phân lớp dữ liệu

Để giải quyết bài toán phân lớp dữ liệu ta tiến hành hai gian đoạn: giai đoạn đầu tiên ta xây dựng mô hình phân lớp (còn hay được gọi là *giai đoạn Huấn luyện*) và giai đoạn thứ hai là kiểm tra đánh giá mô hình phân lớp (còn được gọi là *giai đoạn Kiểm chứng*).

Giai đoạn huấn luyện

Quá trình này nhằm mục đích xây dựng ra một mô hình phân lớp dữ liệu dựa trên việc mô tả tập các lớp dữ liệu hoặc các khái niệm đã được xác định trước. Trong giai đoạn này, thuật toán phân lớp được sử dụng để xây dựng mô hình phân lớp bằng cách phân tích hay "học" từ một tập các dữ liệu huấn luyện (training set) và các nhãn tương ứng của chúng [4].

Quá trình thực hiện giai đoạn học được mô tả trong hình 1.1.



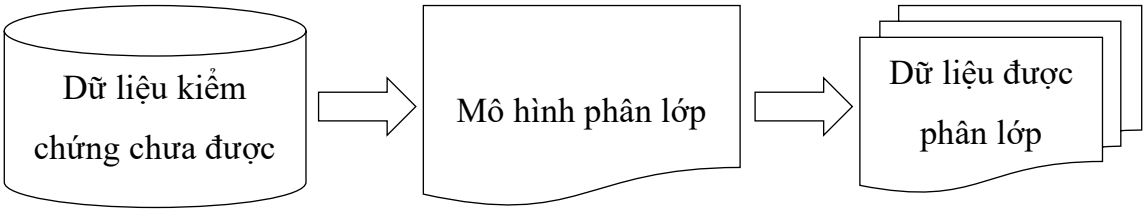
Hình 1.1: Giai đoạn xây dựng mô hình phân lớp dữ liệu

Kết quả sau khi kết thúc giai đoạn này là đưa ra một mô hình phân lớp dữ liệu. Mô hình phân lớp dữ liệu có thể là các công thức toán học, hoặc các luật quyết định, hoặc bộ các quy tắc để gán nhãn lớp cho mỗi dữ liệu trong tập các dữ liệu huấn luyện.

Giai đoạn kiểm chứng

Ở giai đoạn này, mô hình phân lớp ở bước đầu tiên sẽ được sử dụng để thực hiện phân lớp thử nghiệm và đánh giá mô hình phân lớp. Tập các dữ liệu test hay tập kiểm chứng được sử dụng trong giai đoạn. Do đó, tập dữ liệu kiểm chứng được sử dụng trong giai đoạn này phải độc lập với tập dữ liệu huấn luyện ở giai đoạn huấn luyện [4].

Quá trình thực hiện giai đoạn phân lớp thử nghiệm được mô tả trong hình 1.2.



Hình 1.2: Quá trình kiểm tra đánh giá mô hình phân lớp dữ liệu

Các kết quả phân lớp trong quá trình phân lớp thử nghiệm lại có thể sử dụng trong quá trình học tiếp theo.

Sau khi thực hiện xong hai giai đoạn trên, một mô hình phân lớp phù hợp nhất theo một ý nghĩa nào đó (thông qua việc đánh giá các độ đo của mô hình) sẽ được lựa

chọn để thực hiện việc phân lớp dữ liệu trong các bài toán ứng dụng khác nhau trong thực tế.

1.2.3. Các độ đo để đánh giá mô hình phân lớp dữ liệu

Sự phù hợp, mức độ hiệu quả của bất kỳ mô hình phân lớp dữ liệu nào cũng thường được xác định thông qua các độ đo được mô tả dưới đây.

Xét một lớp dữ liệu $c_i \in C = \{c_1, c_2, \ldots, c_m\}$ trong một bài toán phân lớp. Tập hợp các mẫu dữ liệu thuộc lớp c_i được gọi là các phần tử dương (positive). Tập hợp các mẫu dữ liệu không thuộc lớp c_i được gọi là các phần tử âm (negative). Kết quả phân lớp sau khi thực hiện phân lớp dữ liệu có thể xảy ra các trường hợp sau đây:

- True Positive (Trường hợp đúng dương): Phần tử dương được phân loại đúng là dương.
- False Positive (Trường hợp sai dương): Phần tử âm được phân loại sai thành dương.
- True Negative (Trường hợp đúng âm): Phần tử âm được phân loại đúng là âm.
- False Negative (Trường hợp sai âm): Phần tử dương được phân loại sai thành âm.

Ta gọi TP_i là số lượng các mẫu dữ liệu thuộc vào lớp c_i được phân loại đúng (chính xác) vào lớp c_i; gọi FP_i là số lượng các mẫu dữ liệu không thuộc lớp c_i nhưng bị phân loại sai vào lớp c_i; gọi TN_i là số lượng các mẫu dữ liệu không thuộc lớp c_i và được phân loại chính xác và gọi FN_i là số lượng các mẫu dữ liệu thuộc lớp c_i nhưng bị phân loại sai vào các lớp khác với lớp c_i.

Căn cứ vào các đại lượng trên, các khái niệm độ đo sau để đánh giá mức độ hiệu quả của mô hình phân lớp dữ liệu:

Độ đo Precision (Mức chính xác)

Định nghĩa: Precision = TP / (TP + FP).

Ý nghĩa: Giá trị Precision càng cao thể hiện khả năng để một kết quả phân lớp dữ liệu được đưa ra bởi bộ phân lớp là chính xác càng cao.

Độ đo Recall (Độ bao phủ, độ nhạy hoặc độ triệu hồi)

Định nghĩa: Recall = TP / (TP + FN).

Ý nghĩa: Giá trị Recall càng cao thể hiện khả năng kết quả đúng trong số các kết quả đưa ra của bộ phân lớp càng cao.

Độ đo Accuracy (Độ chính xác)

Định nghĩa: Accuracy = (TP + TN) / (TP + TN + FP + FN) * 100%.

Ý nghĩa: Accuracy phản ánh độ chính xác chung của bộ phân lớp dữ liệu.

Độ đo Specificity (Độ đặc hiệu)

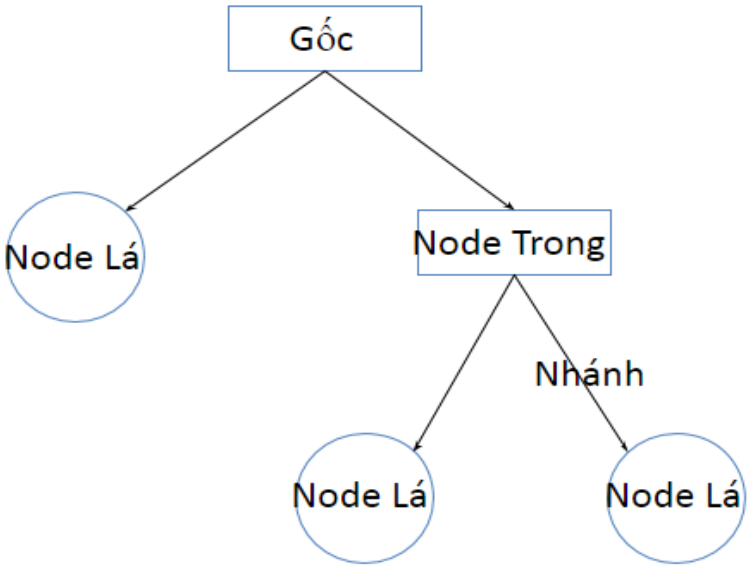
Định nghĩa: Specificity = TN/(TN+FP).

Ý nghĩa: Độ đo Specitivity đánh giá khả năng một dữ liệu là phần tử âm được bộ phân lớp cho ra kết quả chính xác.

1.3. Thuật toán Cây quyết định

1.3.1. Giới thiệu phương pháp

Cây quyết định [24] là một mô hình cấu trúc cây giống như một lưu đồ mà trong đó mỗi nút bên trong cây diễn tả cho việc kiểm tra một thuộc tính, mỗi nhánh trên cây sẽ đại diện cho một kết quả của quá trình kiểm tra và các nút lá sẽ đại diện cho các lớp hoặc phân phối lớp. Nút trên cùng sẽ là nút gốc. Quá trình xây dựng cây quyết định được thực hiện bằng việc phân tách các dữ liệu trong một nút, chia chúng thành các nút con. Quá trình tương tự được áp dụng cho từng các nút con một cách đệ quy cho đến khi không còn nút con nào có thể được tách ra nữa. Các nút không thể được chia nhỏ hơn nữa sẽ được phát triển thành các nút lá. Cây quyết định được biểu diễn dưới dạng một cấu trúc cây như trong hình 1.3 dưới đây.



(Nguồn: Internet)

Hình 1.3: Mô hình cây quyết định

Trong cây mô hình quyết định, mỗi nút trung gian [5], tức là nút khác với nút lá và nút gốc, sẽ tương ứng với một phép kiểm tra một thuộc tính. Mỗi nhánh phía dưới của nút đó sẽ tương ứng cho một giá trị của thuộc tính hay còn gọi là kết quả của phép thử. Khác với các nút trung gian, nút lá [5] không chứa thuộc tính cụ thể mà sẽ chứa các nhãn phân lớp. Để xác định nhãn phân lớp cho một dữ liệu mẫu bất kỳ, ta cho dữ liệu mẫu di chuyển từ gốc cây về phía nút lá. Tại mỗi nút trung gian, thuộc tính tương ứng với nút đó được kiểm tra, tùy vào giá trị của thuộc tính đó mà dữ liệu mẫu sẽ được chuyển xuống nhánh bên dưới tương ứng. Quá trình di chuyển này lặp lại cho đến khi dữ liệu mẫu đó tới được nút lá và được gán nhãn phân lớp là nhãn của nút lá tương ứng.

Quá trình xây dựng một cây quyết định thường được thực hiện như sau:

(1) Bắt đầu từ nút gốc nơi biểu diễn tất cả các mẫu của tập dữ liệu.

(2) Nếu tất cả các mẫu thuộc về cùng một lớp, nút đang xét sẽ trở thành nút lá và được gán nhãn chính bằng lớp đó.

(3) Ngược lại, dùng độ đo thuộc tính nào đó để chọn thuộc tính sẽ phân tách các mẫu tốt nhất vào các lớp tương ứng.

(4) Một nhánh được tạo ra cho từng giá trị của thuộc tính được chọn.

(5) Lặp lại quá trình trên để tạo cây quyết định.

(6) Tiến trình kết thúc chỉ khi bất kỳ điều kiện nào sau đây là đúng:

- Tất cả các mẫu của một nút cho trước đều thuộc về cùng một lớp.
- Không còn thuộc tính nào mà mẫu có thể dựa vào để phân hoạch xa hơn.
- Không còn mẫu nào cho nhánh.

Tuy nhiên, nếu chúng ta không lựa chọn được thuộc tính nào để phân loại hợp lý tại mỗi nút, cây quyết định sau khi xây dựng có thể rất phức tạp. Vì thế người ta thường sử dụng hai cách sau để xây dựng cây quyết định phù hợp:

- Dừng việc phát triển cây sớm hơn bình thường trước khi phân lớp hoàn toàn tập dữ liệu huấn luyện.
- Sử dụng một số kỹ thuật "cắt", "tỉa" cây phù hợp.

Xây dựng Cây quyết định dựa trên Entropy

Khái niệm Entropy [5] của một tập S được định nghĩa trong lý thuyết thông tin là số lượng mong đợi các bit cần thiết để mã hóa thông tin một thành phần rút ra một cách ngẫu nhiên từ tập S về lớp của nó. Đối với trường hợp tối ưu, mã sẽ có độ dài ngắn nhất. Theo lý thuyết thông tin, một mã có độ dài tối ưu sẽ được gán $-\log_2 p$ bits cho một thông điệp có xác suất là p.

Đối với trường hợp tập S là tập mẫu thì mỗi thành phần của tập S là một mẫu. Mỗi mẫu thuộc một lớp nào đó hay nói cách khác là có một giá trị phân loại. Giả sử các mẫu trong tập S thuộc về một lớp trong c lớp, trong đó lớp thứ i ($1 \leq i \leq c$) có tỉ lệ là p_i.

Độ đo Entropy của tập mẫu S được định nghĩa bởi công thức sau:

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

Về bản chất, độ đo Entropy sẽ phản ánh mức độ không đồng nhất của tập mẫu S. Entropy là một độ đo để đo độ pha trộn dữ liệu của một tập mẫu, Entropy càng nhỏ thì tập mẫu càng đồng nhất.

Từ đó, ta định nghĩa lượng thông tin thu thêm và ký hiệu là Gain cho một phép đo hiệu suất phân loại các mẫu của một thuộc tính. Cụ thể hơn, Gain(S, A) của thuộc tính A, trên tập S, được định nghĩa như sau:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Trong đó Values(A) là tập hợp các giá trị có thể có của thuộc tính A, và S_v là tập chứa các mẫu có thuộc tính A mang giá trị v trong tập S. Giá trị Gain(S, A) được sử dụng vào mục đích lựa chọn thuộc tính phân lớp dữ liệu tại mỗi nút trung gian và nút gốc trong quá trình xây dựng cây quyết định. Thuộc tính cho lượng thông tin thu thêm lớn nhất sẽ là thuộc tính được chọn.

Các thuật toán xây dựng cây quyết định dựa trên Entropy có thể tóm tắt như sau:

- Với mỗi thuộc tính bất kỳ A chưa được sử dụng trong quá trình xây dựng cây quyết định, tính Gain (S, A) theo công thức bên trên.
- Chọn một thuộc tính P sao cho giá trị Gain(S, P) có giá trị lớn nhất trong các thuộc tính A kể trên.
- Gán nút tương ứng với thuộc tính P có giá trị Gain lớn nhất.

Xây dựng cây quyết định dựa trên Gini index

Công thức Gini index thường được sử dụng phổ biến hơn Goodness of Split, là phương pháp hướng đến đo lường tần suất một đối tượng dữ liệu ngẫu nhiên trong tập dữ liệu ban đầu được phân loại không chính xác, trên cơ sở đối tượng dữ liệu đã nằm trong một tập con đã được phân ra từ dữ liệu ban đầu, có dán nhãn để thể hiện thuộc tính chung bất kỳ của các đối tường còn lại trong tập con này, giá trị phân loại chính là nhãn của tập con.

Gini index cũng chính là chỉ số đo lường mức độ đồng nhất hay mức độ nhiễu loạn của thông tin. Công thức Gini có thể áp dụng cho cả biến định tính và biến định lượng.

Gini index cho phép chúng ta đánh giá sự tối ưu của từng các phân nhánh thông qua xác định mức độ thuần khiết của từng node trong mô hình cây quyết định. Nếu tất cả các điểm dữ liệu nằm về cũng một lớp thì thể hiện sự đồng nhất không có nhiễu loạn ứng với Gini bằng 0, và sẽ càng lớn nếu các điểm dữ liệu khác biệt nhau và lớn nhất bằng 1.

$$GINI(t) = 1 - \sum_{j} [p(j \mid t)]^2$$

Do hệ số Gini dùng cho thuật toán CART mà CART chỉ giới hạn mỗi lần phân chỉ được 2 nhánh nên giả sử một biến có n thuộc tính thì sẽ có 2^n-2 tập con các cách phân nhánh trên cây quyết định.

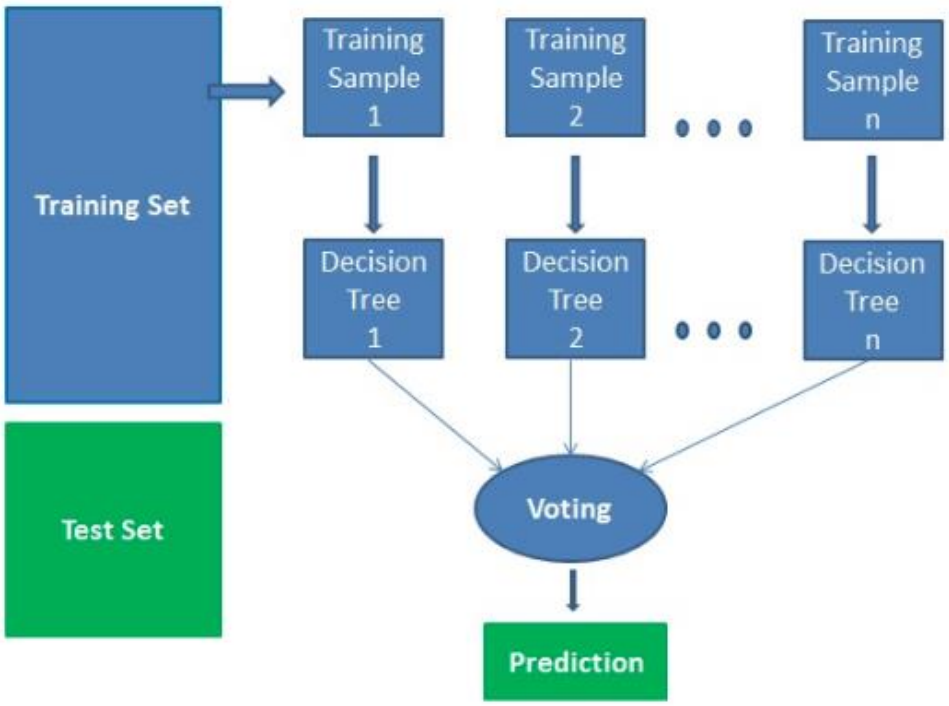
Công thức Gini để tính độ đồng nhất của một nút, vậy khi chúng ta có nhiều cách phân nhánh, mỗi cách có thể phân ra một số nút nhất định, tức có thể chia tập dữ liệu thành các tập con khác nhau theo các giá trị của biến dữ liệu, lúc này chúng ta có thêm công thức thứ 2 để tìm ra cách chia tối ưu nhất.

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Với n_i là số dữ liệu có trung nút con, n là số dữ liệu trong nút cha. Hệ số $GINI_{split}$ càng nhỏ tức cách phân nhánh càng tốt.

1.3.2. Thuật toán Rừng ngẫu nhiên

Rừng ngẫu nhiên [24] là một thuật toán học có giám sát. Như bạn có thể thấy từ tên của nó, nó tạo ra một khu rừng một cách ngẫu nhiên. “Khu rừng” mà ta tạo ra là một tập hợp các cây quyết định. Ý tưởng chính của phương pháp là sự kết hợp của các mô hình học tập làm tăng kết quả chung.



(Nguồn: Internet)

Hình 1.4: Thuật toán rừng ngẫu nhiên

Rừng ngẫu nhiên được đề xuất vào năm 2001 [2]. Đây là thuật toán phân loại có kiểm định dựa trên cây quyết định và kỹ thuật Bagging and Bootstrapping đã được cải tiến. Bootstrapping là một phương pháp rất nổi tiếng trong thống kê được giới thiệu bởi Efron vào năm 1979. Phương pháp này được thực hiện như sau: từ một quần thể ban đầu lấy ra một mẫu L = (x_1, x_2,..,x_n) gồm n thành phần để tính toán các tham số mong muốn. Trong các bước tiếp theo lặp lại b lần tạo ra mẫu L_b cũng gồm n phần bằng cách lấy lại mẫu với sự thay thế các thành phần trong mẫu ban đầu sau đó tính toán các tham số mong muốn. Phương pháp Bagging được xem như là một phương pháp tổng hợp kết quả có được từ các bootstraping sau đó huấn luyện mô hình từ các mẫu ngẫu nhiên này và cuối cùng đưa ra dự đoán phân loại dựa vào số phiếu bầu cao nhất của lớp phân loại. Cây quyết định là một sơ đồ phát triển có cấu trúc dạng cây phân nhánh đi từ gốc cho đến lá, giá trị các lớp phân loại của mẫu được đưa vào kiểm tra trên cây quyết định. Mỗi mẫu tương ứng có một đường đi từ gốc (tức là dữ liệu đầu vào) đến lá (tức

là các kết quả phân loại dự đoán đầu ra), đường đi này biểu diễn sự phân lớp của mẫu đó. Mỗi sơ đồ cây trong tập mẫu được tạo thành từ tập hợp các dữ liệu huấn luyện được lựa chọn ngẫu nhiên để huấn luyện mô hình phân loại Rừng ngẫu nhiên (mỗi tập mẫu bootstrap sẽ cho ra một cây và n cây tương ứng với n bootstrap). Khi một tập mẫu được rút ra từ tập huấn luyện (bootstrap) với sự thay thế có hoàn lại, thì thông thường có khoảng 1/3 các phần tử không nằm trong mẫu này và vì thế chúng không tham gia vào quá trình huấn. Điều này có nghĩa là chỉ có khoảng 2/3 các phần tử trong tập huấn luyện tham gia vào trong các tính toán để phân loại và 1/3 các phần tử này dùng để kiểm tra sai số. Dữ liệu kiểm tra được sử dụng để ước lượng sai số tạo ra từ việc kết hợp các kết quả phân loại riêng lẻ sau đó được tổng hợp trong mô hình Rừng ngẫu nhiên cũng như dùng để ước tính các biến quan trọng.

Rừng ngẫu nhiên chứa một lượng lớn các cây, mỗi cây được phát triển từ các tập huấn luyện được lựa chọn ngẫu nhiên. Hai tham số cần được xác định trong thuật toán phân loại này là ntree (số lượng cây được phát triển) và mtry (số lượng biến để phân chia tại mỗi node). Số ntree được lựa chọn phụ thuộc vào khoảng thời gian xử lý ngắn nhất để kết quả đạt được độ sai số thấp nhất và mtry biến động từ số biến độc lập tối thiểu (bằng 1) đến số biến độc lập tối đa được sử dụng trong phân loại.

Sau khi mô hình Random Forest được tạo thành, mỗi kết quả của các boostrap trong tập hợp sẽ bỏ phiếu cho lớp phổ biến nhất và cho ra một kết quả phân loại. Mô hình được tạo thành dựa vào phân loại có số phiếu bầu nhiều nhất của mỗi sơ đồ cây quyết định ntree.

1.4. Các công trình nghiên cứu liên quan

Dự đoán gói cước phù hợp vớinhu cầu sử dụng của khách hàng là chủ đề mà rất nhiều doanh nghiệp viễn thông quan tâm và có nhiều công trình công bố của nhiều tác giả. Nhìn chung, quy trình giải quyết bài toán gồm các công đoạn như: (i) Thu nhận dữ liệu; (ii) Tiền xử lý dữ liệu; (iii) Phân tích dữ liệu; (iv) Dự đoán bằng các mô hình phân lớp, một số công trình công bố như:

1.4.1.Model based collaborative filtering

Lọc cộng tác (CF) là thuật toán phổ biến cho các hệ thống khuyến nghị. Do đó, các mục được giới thiệu cho người dùng được xác định bằng cách khảo sát cộng đồng của họ. CF có quan điểm tốt vì nó có thể loại bỏ giới hạn của đề xuất bằng cách khám phá thêm các vật phẩm tiềm năng ẩn dưới cộng đồng. Những mặt hàng như vậy có khả năng phù hợp với người dùng và chúng nên được giới thiệu cho người dùng. Có hai cách tiếp cận chính cho CF: dựa trên bộ nhớ và dựa trên mô hình. Thuật toán dựa trên bộ nhớ tải toàn bộ cơ sở dữ liệu vào bộ nhớ hệ thống và đưa ra dự đoán cho đề xuất dựa trên cơ sở dữ liệu bộ nhớ nội tuyến đó. Nó đơn giản nhưng lại gặp phải vấn đề là dữ liệu khổng lồ. Thuật toán dựa trên mô hình cố gắng nén cơ sở dữ liệu khổng lồ vào một mô hình và thực hiện nhiệm vụ đề xuất bằng cách áp dụng cơ chế tham chiếu vào mô hình này. CF dựa trên mô hình có thể đáp ứng yêu cầu của người dùng ngay lập tức. Bài báo này khảo sát các kỹ thuật phổ biến để thực hiện các thuật toán dựa trên mô hình. Tác giả cũng đưa ra một ý tưởng mới cho cách tiếp cận dựa trên mô hình để đạt được độ chính xác cao và giải quyết vấn đề của ma trận thưa thớt bằng cách áp dụng các kỹ thuật suy luận dựa trên bằng chứng.

1.4.2. A Survey of Collaborative Filtering Techniques

Là một trong những cách tiếp cận thành công nhất để xây dựng hệ thống khuyến nghị, lọc cộng tác (CF) sử dụng các sở thích đã biết của một nhóm người dùng để đưa ra các đề xuất hoặc dự đoán về các sở thích chưa biết cho những người dùng khác. Trong bài báo này, trước tiên tác giả giới thiệu các nhiệm vụ CF và những thách thức chính của chúng, chẳng hạn như sự thưa thớt dữ liệu, khả năng mở rộng, từ đồng nghĩa, cừu xám, bảo vệ quyền riêng tư, v.v. và các giải pháp khả thi của chúng. Sau đó, tác giả trình bày ba danh mục chính của kỹ thuật CF: dựa trên bộ nhớ, dựa trên mô hình và thuật toán CF kết hợp (kết hợp CF với các kỹ thuật đề xuất khác), với các ví dụ cho các thuật toán đại diện của từng danh mục và phân tích hiệu suất dự đoán và khả năng giải quyết của chúng những thách thức. Từ các kỹ thuật cơ bản đến hiện đại, tác giả cố gắng trình bày một khảo sát toàn diện về các kỹ thuật CF, có thể được coi là lộ trình nghiên cứu và thực hành trong lĩnh vực này.

1.4.3. Collaborative Filtering for Multi-class Data Using Belief Nets **Algorithms**

Lưới tín ngưỡng Bayes, một trong những công cụ phân loại được sử dụng thường xuyên nhất, có thể được sử dụng cho các nhiệm vụ CF. Các công trình trước đây về việc áp dụng BN vào các nhiệm vụ CF chủ yếu tập trung vào dữ liệu lớp nhị phân và sử dụng các bộ phân loại Bayesian đơn giản hoặc cơ bản. Trong nghiên cứu này, tác giả áp dụng các mô hình BN nâng cao cho các tác vụ CF thay vì các tác vụ đơn giản và làm việc trên dữ liệu CF nhiều lớp trong thế giới thực thay vì dữ liệu lớp nhị phân tổng hợp. Kết quả thực nghiệm cho thấy rằng với khả năng xử lý dữ liệu không đầy đủ, hồi quy logistic mở rộng trên các mô hình Navie Bayes và Navie Bayes tăng cường trên cây (NB-ELR và TAN-ELR) luôn hoạt động tốt hơn so với thuật toán CF dựa trên tương quan Pearson hiện đại. Ngoài ra, các mô hình BNs CF được tối ưu hóa rất mạnh mẽ về khả năng đưa ra dự đoán, trong khi tính mạnh mẽ của thuật toán CF dựa trên tương quan Pearson giảm khi độ thưa thớt của dữ liệu tăng lên.

1.4.4. An intelligent decision support system for production planning based on machine learning

Bài báo này trình bày một phương pháp luận mới để giải quyết vấn đề quản lý Chuỗi cung ứng khép kín (CLSC) thông qua hệ thống ra quyết định dựa trên logic mờ được xây dựng trên máy học. Hệ thống sẽ đưa ra các quyết định để vận hành một nhà máy sản xuất được tích hợp trong CLSC nhằm đáp ứng các mục tiêu sản xuất khi có các yếu tố không chắc chắn. Một trong những đóng góp chính của đề xuất này là khả năng bác bỏ những ảnh hưởng mà sự mất cân đối trong phần còn lại của chuỗi gây ra đối với việc tồn kho nguyên vật liệu và thành phẩm. Đối với điều này, một thuật toán thông minh sẽ chịu trách nhiệm giám sát hoạt động của nhà máy và lập trình lại nhiệm vụ để đảm bảo đạt được các mục tiêu của quy trình. Kỹ thuật logic mờ và máy học được kết hợp để thiết kế công cụ. Phương pháp này đã được thử nghiệm tại bệnh viện công nghiệp với kết quả khả quan, do đó làm nổi bật tiềm năng của đề xuất này trong việc kết hợp nó vào khuôn khổ Công nghiệp 4.0.

1.4.5. Machine learning based decision support systems (DSS) for heart disease diagnosis

Đánh giá hiện tại đóng góp một cái nhìn tổng quan sâu rộng về các hệ thống hỗ trợ quyết định trong việc chẩn đoán bệnh tim trong các cơ sở lâm sàng. Các nhà điều tra đã sàng lọc và tóm tắt một cách độc lập các nghiên cứu liên quan đến hệ thống hỗ trợ quyết định lâm sàng dựa trên bệnh tim (DSS) được công bố cho đến ngày 8 tháng 6 năm 2015 trên PubMed. Dữ liệu được trích xuất từ hai mươi bài báo toàn văn đáp ứng các tiêu chí đưa vào được phân loại theo các trường sau; bệnh tim, phương pháp hình thành tập dữ liệu, thuật toán máy học, DSS dựa trên máy học, các loại so sánh, đánh giá kết quả và ý nghĩa lâm sàng của DSS được báo cáo. Trong tổng số 331 nghiên cứu, 20 nghiên cứu đáp ứng các tiêu chí thu nhận. Hầu hết các nghiên cứu liên quan đến bệnh tim thiếu máu cục bộ với mạng lưới thần kinh là kỹ thuật máy học (ML) phổ biến nhất. Trong số các kỹ thuật ML, ANN phân loại nhồi máu cơ tim với 97% và xạ hình tưới máu cơ tim với độ chính xác 87,5%, CART phân loại suy tim với 87,6%, mạng nơ-ron phân loại van tim với 97,4%, máy vector hỗ trợ phân loại sàng lọc rối loạn nhịp tim với 95,6%, hồi quy phân loại hội chứng mạch vành cấp với 72%, hệ thống nhận dạng miễn dịch nhân tạo phân loại bệnh mạch vành với 92,5% và các thuật toán di truyền và phân tích quyết định đa tiêu chí phân loại bệnh nhân đau ngực với độ chính xác 91%. Có 55% nghiên cứu xác nhận kết quả trong môi trường lâm sàng trong khi 25% xác nhận kết quả thông qua thiết lập thử nghiệm. Phần còn lại của các nghiên cứu (20%) không báo cáo khả năng áp dụng và tính khả thi của các phương pháp của họ trong các cơ sở lâm sàng. Nghiên cứu phân loại các kỹ thuật ML theo hiệu suất của chúng trong việc chẩn đoán các bệnh tim khác nhau. Nó phân loại, so sánh và đánh giá bộ so sánh dựa trên hiệu suất của bác sĩ, tiêu chuẩn vàng, các kỹ thuật ML khác, các mô hình khác nhau của cùng một kỹ thuật ML và các nghiên cứu không có sự so sánh. Nó cũng điều tra hiện tại, tương lai và không có ý nghĩa lâm sàng. Ngoài ra, các xu hướng của kỹ thuật học máy và thuật toán được sử dụng trong chẩn đoán bệnh tim cùng với việc xác định các lỗ hổng nghiên cứu được báo cáo trong nghiên cứu này. Các kết quả được báo cáo đề xuất các diễn giải

đáng tin cậy và các biểu diễn tự giải thích bằng đồ họa chi tiết của DSS. Nghiên cứu cho thấy nhu cầu thiết lập dữ liệu lâm sàng thời gian thực không mơ hồ để đào tạo DSS thích hợp trước khi nó có thể được sử dụng trong các cơ sở lâm sàng. Các hướng nghiên cứu trong tương lai của DSS dựa trên ML chủ yếu là hướng tới sự phát triển của các hệ thống tổng quát hóa có thể quyết định các phép đo lâm sàng dễ dàng truy cập và đánh giá trong thời gian thực.

1.5. Thư viện Scikit-learn

Là một thư viện mạnh mẽ có thể mang các thuật toán học máy vào trong một hệ thống thích hợp nhất. Thư viện này tích hợp rất nhiều thuật toán hiện đại và cổ điển hỗ trợ việc học và tiến hành đưa ra các giải pháp hữu ích cho bài toán học máy một cách đơn giản [27].

Scikit-learn (Sklearn) [30] là thư viện mạnh mẽ nhất dành cho các thuật toán học máy được viết trên ngôn ngữ Python. Thư viện cung cấp một tập các công cụ xử lý các bài toán học máy và mô hình thống kê gồm: phân loại, hồi quy.

Thư viện được cấp phép bản quyền chuẩn FreeBSD và chạy được trên nhiều nền tảng Linux. Scikit-learn được sử dụng như một tài liệu để học tập.

Để cài đặt scikit-learn trước tiên phải cài thư viện SciPy (Scientific Python). Những thành phần gồm:

- Numpy: Gói thư viện xử lý dãy số và ma trận nhiều chiều
- SciPy: Gói các hàm tính toán logic khoa học
- Matplotlib: Biểu diễn dữ liệu dưới dạng đồ thị 2 chiều, 3 chiều
- IPython: Sổ tay dùng để tương tác trực quan với Python
- SymPy: Gói thư viện các kí tự toán học
- Pandas: Xử lý, phân tích dữ liệu dưới dạng bảng

Những thư viện mở rộng của SciPy thường được đặt tên dạng SciKits. Như thư viện này là gói các lớp, hàm sử dụng trong thuật toán học máy thì được đặt tên là scikit-learn.

Scikit-learn hỗ trợ mạnh mẽ trong việc xây dựng các sản phẩm. Nghĩa là thư viện này tập trung sâu trong việc xây dựng các yếu tố: dễ sử dụng, dễ code, dễ tham khảo, dễ làm việc, hiệu quả cao.

Mặc dù được viết cho Python nhưng thực ra các thư viện nền tảng của scikit-learn lại được viết dưới các thư viện của C để tăng hiệu suất làm việc.

1.6. Pycharm

1.6.1. Giới thiệu

Pycharm là một nền tảng kết kết hợp được JetBrains phát triển như một IDE (Môi trường phát triển tích hợp) để phát triển các ứng dụng cho lập trình trong Python. Một số ứng dụng lớn như Tweeter, Facebook, Amazon và Pinterest sử dụng Pycharm để làm IDE Python của họ. Bài viết dưới đây sẽ giới thiệu chi tiết cho bạn về Pycharm cũng như hướng dẫn cách cài đặt và sử dụng Pycharm

1.6.2. Các tính năng của Pycharm

Pycharm có thể chạy trên Windows, Linux, hoặc Mac OS. Ngoài ra, nó cũng chứa các Mô đun và các gói giúp các lập trình viên phát triển phần mềm bằng Python trong thời gian ngắn với ít công sức hơn. Hơn nữa, nó cũng có khả năng tùy chỉnh theo yêu cầu của nhà phát triển.

Khi cài đặt Pycharm, LTV có thể sử dụng một số tính năng sau:

Trình chỉnh sửa mã thông minh:

- Giúp các lập trình viên viết mã chất lượng cao

- Bao gồm các lược đồ màu cho các từ khóa, lớp và hàm. Điều này giúp tăng khả năng đọc và hiểu mã

- Xác định lỗi một cách dễ dàng

- Cung cấp tính năng tự động hoàn thiện và hướng dẫn hoàn thiện mã

Điều hướng mã:

- Giúp các nhà phát triển trong việc chỉnh sửa và nâng cao mã với ít nỗ lực và thời gian hơn

- Với việc điều hướng mã, nhà phát triển có thể dễ dàng điều hướng một lớp,hàm hoặc tệp

- LTV có thể xác định vị trí của một phần tử, một ký hiệu hoặc một biến trong mã nguồn trong thời gian ngắn khi sử dụng Pycharm

- Bằng việc sử dụng chế độ thấu kính, nhà phát triển có thể kiểm tra và gỡ lỗi toàn bộ mã nguồn.

Tái cấu trúc:

- Sử dụng Pycharm có lợi thế là thực hiện các thay đổi hiệu quả và nhanh chóng đối với cả biến cục bộ và biến toàn cục

- Tái cấu trúc trong Pycharm cho phép các nhà phát triển cải thiện cấu trúc bên trong mà không thay đổi hiệu suất bên ngoài của mã

- Nó cũng cho phép phân chia các lớp với các chức năng mở rộng hơn

CHƯƠNG 2 – PHƯƠNG PHÁP KHUYẾN NGHỊ GÓI CƯỚC

2.1. Phân tích các yếu tố ảnh hưởng tới gói cước phù hợp với khách hàng

Việc chọn gói cước phù hợp với khách hàng phụ thuộc vào nhiều yếu tố, trong phần này luận văn sẽ đi sâu phân tích các yếu tố ảnh hưởng trực tiếp đến việc lựa chọn gói cước phù hợp cho khách hàng.

2.1.1. Các yếu tố về khách hàng

Các yếu tố phi chất lương là các yếu tố được hình thành gồm:

Tên Thành phố, Quận, Huyện: Như chúng ta đã biết khách hàng tuy sử dụng cùng 1 loại hình dịch vụ tuy nhiên do tập quán sinh hoạt văn hóa…. Mỗi vùng miền sẽ có nhưng đặc trưng riêng, điều kiện kinh tế khác nhau, do đó nhu cầu sử dụng dịch vụ cũng khác nhau, hành vi tiêu dùng cũng khác nhau.

Loại khách hàng: Doanh nghiệp, Tổ chức, Cá nhân…Những nhóm đối tượng khách hàng khác nhau cũng có nhưng đặc trưng khác nhau, yêu cầu về dịch vụ khác nhau, do đó chắc chắn ảnh hưởng đến nhu cầu sử dụng dịch vụ của khách hàng.

Độ tuổi khách hàng: Độ tuổi khách hàng phần nào đó thể hiện nhu cầu sử dụng dịch vụ của khách hàng. Ví dụ những người trẻ tuổi có nhu cầu sử dụng Internet tốc độ cao hơn để phục vụ cho các công việc online hoặc chơi game, xem phim trực tuyến. Những người lớn tuổi thì có xu hướng sử dụng dịch vụ MyTV để xem truyền hình, thời sự…

2.1.2. Các yếu tố về chất lượng dịch vụ

Tất cả mọi ngành nghề kinh doanh chất lượng sản phẩm dịch là linh hồn của doanh nghiệp, chất lượng càng cao thì sản phẩm được khách hàng ưu chuộng, doanh nghiệp bán được nhiều sản phẩm doanh thu mang về càng nhiều và cứ như thế doanh nghiệp ngày một phát triển, Viễn thông Tây Ninh cũng vậy, vì đã xác định chất lượng là mục tiêu hàng đầu để luôn cải thiện và hoàn chỉnh ngày một tốt hơn, từ đó có nhiều

giải pháp để thực hiện, chất lượng gồm chất lượng của dịch vụ và chất lượng phục vụ.

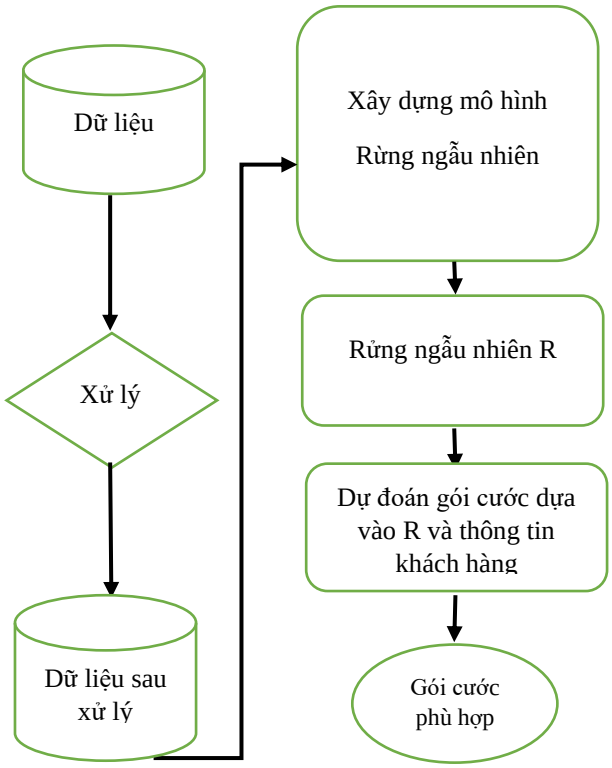
Các yếu tố chất lượng dịch vụ là chất lượng của từng dịch vụ cung cấp bao gồm:

- Băng thông: là băng thông tối đa của một gói cước khi cung cấp cho khách hàng.

- Loại IP tĩnh/động: tùy thuộc vào nhu cầu sử dụng của khách hàng mà gói cước khách hàng đăng ký có đi kèm với IP tĩnh hay không. Mặc định khách hàng sẽ được cung cấp IP WAN là IP động để sử dụng.

2.2. Mô hình dự đoán gói cước cho khách hàng

Để tiến hành dự đoán gói cước phù hợp với khách hàng ta sử dụng mô hình được mô tả như trong Hình 2.4 như sau:



Hình 2.1: Mô hình thực nghiệm dự đoán

Mô hình thực nghiệm dự báo được thực hiện thông qua các bước:

- Thu thập dữ liệu về khách hàng.

- Phân tích dữ liệu các yếu tố ảnh hưởng tới gói cước sử dụng của khách hàng.

- Tiền xử lý dữ liệu và cho ra tập dữ liệu phù hợp để dưa vào dự báo.

- Xây dựng mô hình rừng ngẫu nhiên thông qua việc xây dựng các cây quyết định thuộc rừng ngẫu nhiên dựa vào tập dữ liệu ban đầu.Đánh giá và lựa chọn cây quyết định có giá trị để đưa vào rừng ngẫu nhiên.

- Thực nghiệm và đánh giá, lựa chọn mô hình phù hợp dựa vào các độ đo. Trong quá trình thực nghiệm, luận văn lập lại nhiều lần để thay đổi tỉ lệ phân chia giữa tập huấn luyện và tập kiểm tra với các tập dữ liệu được chương trình chọn ngẫu nhiên và tiến hành kiểm tra chọn độ chính xác của các mô hình nào tốt nhất.

- Từ mô hình đề xuất chúng ta tiến hành thực nghiệm để đánh giá độ chính xác của mô hình, đồng thời lựa chọn mô hình tối ưu để đưa vào ứng dụng trong thực tiễn.

2.3. Sử dụng thuật toán phân lớp Rừng ngẫu nhiên thông qua bộ thư viện Scikit-learn

Để cài đặt scikit-learn trước tiên phải cài thư viện SciPy (Scientific Python). Những thành phần gồm:

- Numpy: Gói thư viện xử lý dãy số và ma trận nhiều chiều

- SciPy: Gói các hàm tính toán logic khoa học

- Matplotlib: Biểu diễn dữ liệu dưới dạng đồ thị 2 chiều, 3 chiều

- IPython: Notebook dùng để tương tác trực quan với Python

- SymPy: Gói thư viện các kí tự toán học

- Pandas: Xử lý, phân tích dữ liệu dưới dạng bảng

Những thư viện mở rộng của SciPy thường được đặt tên dạng SciKits. Như thư viện này là gói các lớp, hàm sử dụng trong thuật toán học máy thì được đặt tên là scikit-learn.

Scikit-learn hỗ trợ mạnh mẽ trong việc xây dựng các sản phẩm. Nghĩa là thư viện này tập trung sâu trong việc xây dựng các yếu tố: dễ sử dụng, dễ code, dễ tham khảo, dễ làm việc, hiệu quả cao.

Mặc dù được viết cho Python nhưng thực ra các thư viện nền tảng của scikit-learn lại được viết dưới các thư viện của C để tăng hiệu suất làm việc.

Các tham số trong Rừng Ngẫu nhiên được sử dụng để tăng khả năng dự đoán của mô hình hoặc để làm cho mô hình nhanh hơn. Luận văn sẽ thảo luận về các siêu tham số của hàm *random forest* gói *sklearn*[30].

Các tham số cần quan tâm trong khi xây dựng Rừng ngẫu nhiên như sau:

n_estimators: int, mặc định=100.

Số lượng cây trong Rừng ngẫu nhiên. Nói chung, số lượng cây cao làm tăng hiệu suất và làm cho các dự đoán ổn định hơn, nhưng nó cũng làm chậm quá trình tính toán.

criterion {“gini”, “entropy”}, mặc định=”gini”

Chức năng đo lường chất lượng của một cây quyết định được xây dựng. Các tiêu chí được hỗ trợ là “gini” và “entropy”. Lưu ý: tham số này dành riêng cho cây quyết định.

max_depth: int, mặc định=None

Chiều sâu tối đa của cây. Nếu không có, thì các nút được mở rộng cho đến khi tất cả các lá đều thuần túy hoặc cho đến khi tất cả các lá chứa ít hơn mẫu *min_samples_split*.

min_samples_split, kiểu dữ liệu là int hoặc float,mặc định=2

Số lượng mẫu tối thiểu cần thiết để tách một nút nội bộ một cây quyết định:

- Nếu kiểu dữ liệu là int, thì hãy coi *min_samples_split* là số mẫu nhỏ nhấtcho mỗi lần tách.

- Nếu kiểu dữ liệu là float, thì min_samples_split là một phân số và (*min_samples_split* * *n_samples*) là số lượng mẫu tối thiểu cho mỗi lần tách.

min_samples_leaf: kiểu dữ liệu là int hoặc float, mặc định=1

Số lượng mẫu tối thiểu cần thiết có ở một nút lá. Một điểm phân tách ở bất kỳ độ sâu nào sẽ chỉ được xem xét nếu nó để lại ít nhất các mẫu huấn luyện *min_samples_leaf* trong mỗi nhánh. Điều này có thể có tác dụng làm mịn mô hình.

Nếu kiểu dữ liệu là int, thì hãy coi *min_samples_leaf*là số mẫu nhỏ nhất tại một lá.

Nếu kiểu dữ liệu là float, thì min_samples_leaf là một phân số và (*min_samples_leaf* * *n_samples*) là số lượng mẫu tối thiểu cho mỗi lá.

max_features{“auto”, “sqrt”, “log2”}, kiểu dữ liệu là int hoặc float, mặc định=”auto”

Số lượng các thuộc tính cần xem xét khi tìm kiếm sự phân chia tốt nhất:

- Nếu kiểu dữ liệu là int, thì hãy xem xét max_features các thuộc tính tại mỗi lần phân chia.

- Nếu float, thì max_features là một phân số và (max_features * n_features) các thuộc tính được xem xét ở mỗi lần tách.

- Nếu “auto”, thì max_features = sqrt(n_features).

- Nếu “sqrt” thì max_features = sqrt (n_features) (giống như “auto”).

- Nếu “log2”, thì max_features = log2 (n_features).

- Nếu không, thì max_features = n_features.

max_leaf_nodes, mặc định=None

Xây dựng cây quyết định với max_leaf_nodes nút lá tối đa. Nếu Không thì không giới hạn số nút lá.

bootstrap, mặc định=True

Các mẫu bootstrap có được sử dụng khi xây dựng cây hay không. Nếu là False, toàn bộ tập dữ liệu được sử dụng để xây dựng từng cây.

oob_score, mặc định=False

Có sử dụng các mẫu ngoài túi để ước tính điểm tổng quát hay không. Chỉ khả dụng nếu bootstrap = True.

max_samples: kiểu dữ liệu là int hoặc float, mặc định=None

Nếu bootstrap là True, thì số lượng mẫu sẽ lấy từ tập mẫu X để xây dựng cây quyết định, trong đó:

- Nếu None (default), thì lấy X.shape [0] mẫu.

- Nếu int, thì lấy max_samples mẫu.

- Nếu float, thì lấy max_samples * X.shape [0] mẫu. Do đó, max_samples phải nằm trong khoảng (0.0, 1.0].

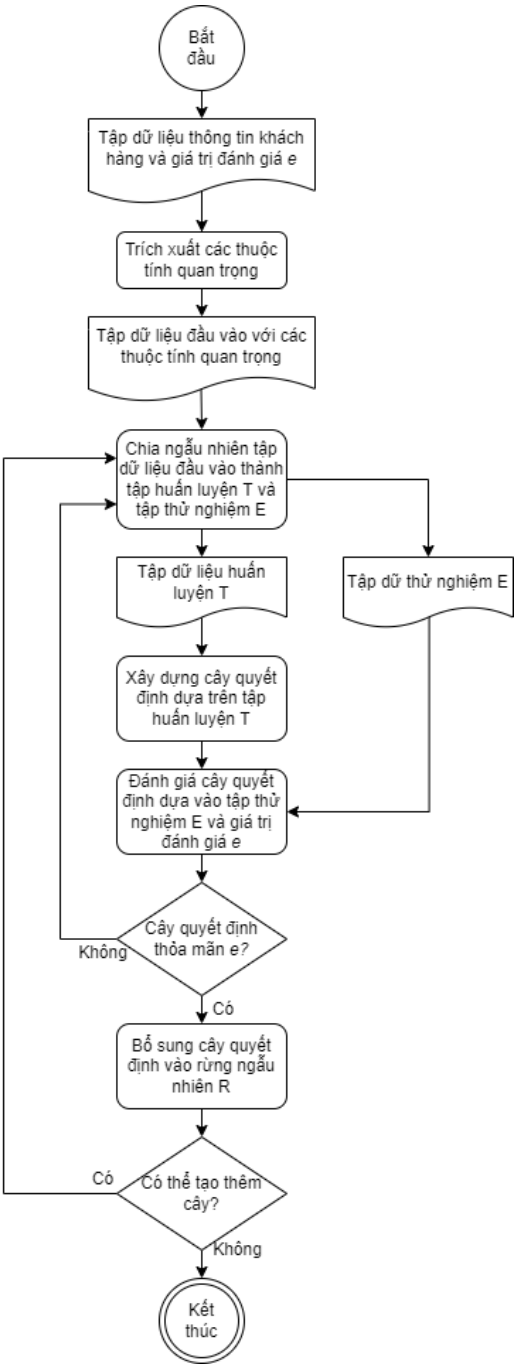
2.4. Sử dụng Pycharm để xây dựng ứng dụng web

Ứng dụng web được xây dựng bằng thư viện Flask trên ngôn ngữ Python. Server được xây dựng bằng ngôn ngữ Python để tiện cho việc truy xuất các model một cách dễ dàng hơn so với các ngôn ngữ khác.

Chức năng khuyến nghị gói cước: Sau khi nhập các trường dữ liệu trên chức năng này, kết quả sẽ hiển thị gói cước phù hợp nhất với khách hàng dựa vào mô hình học máy đã huấn luyện.

CHƯƠNG 3 - XÂY DỰNG MÔ HÌNH

Quá trình để xây dựng rừng ngẫu nhiên cho tập dữ liệu thông tin khách hàngđược biểu diễn qua lưu đồ giải thuật như sau.



Hình 3.1: Lưu đồ giải thuật xây dựng rừng ngẫu nhiên

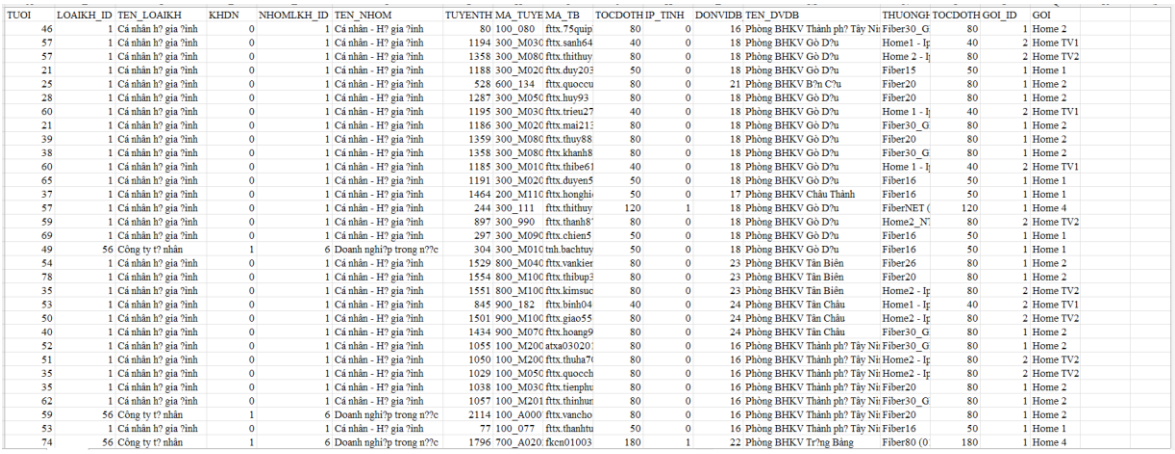
3.1. Dữ liệu

3.1.1. Thu thập dữ liệu

Hiện tại, các quy trình nghiệp vụ tại VNPT Tây Ninh đều được thao tác, thực hiện trên hệ thống thông tin Điều hành sản xuất kinh doanh (ĐHSXKD), đây là một hệ sinh thái lớn trong hệ thống quản lý của VNPT.

Hệ thống này cũng quản lý tất cả các việc thu thập thông tin khách hàng, quản lý thuê bao và các vấn đề liên quan. Vì vậy dữ liệu trong nghiên cứu này được trích xuất một phần từ cơ sở dữ liệu của hệ thống.

Sau khi thu thập dữ liệu, luận văn đã thu thập được tổng cộng hơn 140,000 mẫu và được lưu dưới dạng file .csv như ảnh minh họa.



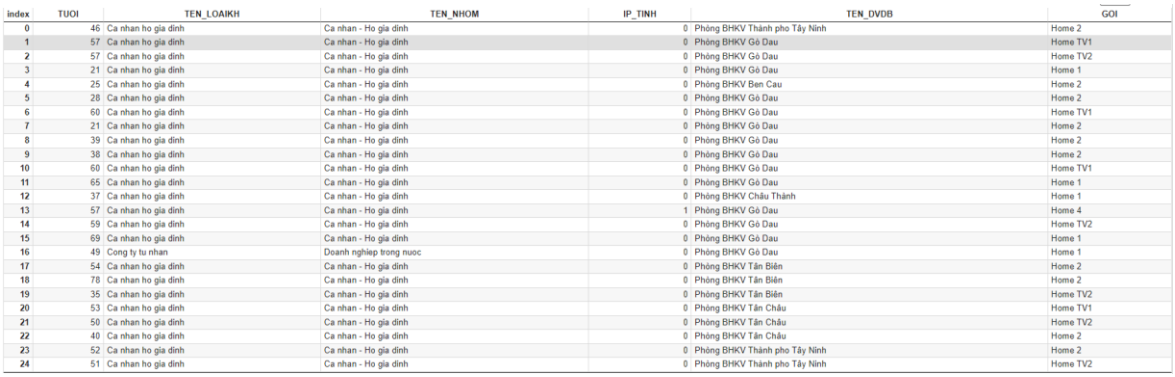
TUOI	LOAIKH_ID	TEN_LOAIKH	KHDN	NHOMLKH_ID	TEN_NHOM	TUYENTH	MA_TUYE	MA_TB	TOCDOTH	IP_TINH	DONVIDB	TEN_DVDB	THUONGI	TOCDOTH	GOI_ID	GOI
46	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	80	100_080	fttx.75quip	80	0	16	Phòng BHKV Thành ph? Tây Ni	Fiber30_G	80	1	Home 2
57	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1194	300_M03	fttx.sanh64	40	0	18	Phòng BHKV Gò D?u	Home1 - I	40	2	Home TV1
57	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1358	300_M08	fttx.thithuy	80	0	18	Phòng BHKV Gò D?u	Home 2 - I	80	2	Home TV2
21	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1188	300_M02	fttx.duy203	50	0	18	Phòng BHKV Gò D?u	Fiber15	50	1	Home 1
25	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	528	600_134	fttx.quoccu	80	0	21	Phòng BHKV B?n C?u	Fiber20	80	1	Home 2
28	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1287	300_M05	fttx.huy93	80	0	18	Phòng BHKV Gò D?u	Fiber20	80	1	Home 2
60	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1195	300_M03	fttx.trieu27	40	0	18	Phòng BHKV Gò D?u	Home 1 - I	40	2	Home TV1
21	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1186	300_M02	fttx.mai213	80	0	18	Phòng BHKV Gò D?u	Fiber30_G	80	1	Home 2
39	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1359	300_M08	fttx.thuy88	80	0	18	Phòng BHKV Gò D?u	Fiber20	80	1	Home 2
38	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1358	300_M08	fttx.khanh8	80	0	18	Phòng BHKV Gò D?u	Fiber30_G	80	1	Home 2
60	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1185	300_M01	fttx.thibe61	40	0	18	Phòng BHKV Gò D?u	Home 1 - I	40	2	Home TV1
65	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1191	300_M02	fttx.duyen5	50	0	18	Phòng BHKV Gò D?u	Fiber16	50	1	Home 1
37	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1464	200_M11	fttx.honghi	50	0	17	Phòng BHKV Châu Thành	Fiber16	50	1	Home 1
57	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	244	300_111	fttx.thithuy	120	1	18	Phòng BHKV Gò D?u	FiberNET (	120	1	Home 4
59	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	897	300_990	fttx.thanh8	80	0	18	Phòng BHKV Gò D?u	Home2_N	80	2	Home TV2
69	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	297	300_M09	fttx.chien5	50	0	18	Phòng BHKV Gò D?u	Fiber16	50	1	Home 1
49	56	Công ty t? nhân	1	6	Doanh nghi?p trong n??c	304	300_M01	tnh.bachtuy	50	0	18	Phòng BHKV Gò D?u	Fiber16	50	1	Home 1
54	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1529	800_M04	fttx.vankier	80	0	23	Phòng BHKV Tân Biên	Fiber26	80	1	Home 2
78	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1554	800_M10	fttx.thibup3	80	0	23	Phòng BHKV Tân Biên	Fiber20	80	1	Home 2
35	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1551	800_M10	fttx.kimsuc	80	0	23	Phòng BHKV Tân Biên	Home2 - I	80	2	Home TV2
53	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	845	900_182	fttx.binh04	40	0	24	Phòng BHKV Tân Châu	Home1 - I	40	2	Home TV1
50	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1501	900_M10	fttx.giao55	80	0	24	Phòng BHKV Tân Châu	Home2 - I	80	2	Home TV2
40	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1434	900_M07	fttx.hoang9	80	0	24	Phòng BHKV Tân Châu	Fiber30_G	80	1	Home 2
52	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1055	100_M20	atxa030201	80	0	16	Phòng BHKV Thành ph? Tây Ni	Fiber30_G	80	1	Home 2
51	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1050	100_M20	fttx.thuha7	80	0	16	Phòng BHKV Thành ph? Tây Ni	Home2 - I	80	2	Home TV2
35	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1029	100_M05	fttx.quocch	80	0	16	Phòng BHKV Thành ph? Tây Ni	Home2 - I	80	2	Home TV2
35	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1038	100_M03	fttx.tienphu	80	0	16	Phòng BHKV Thành ph? Tây Ni	Fiber20	80	1	Home 2
62	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	1057	100_M20	fttx.thinhur	80	0	16	Phòng BHKV Thành ph? Tây Ni	Fiber30_G	80	1	Home 2
59	56	Công ty t? nhân	1	6	Doanh nghi?p trong n??c	2114	100_A000	fttx.vancho	80	0	16	Phòng BHKV Thành ph? Tây Ni	Fiber20	80	1	Home 2
53	1	Cá nhân h? gia ?ình	0	1	Cá nhân - H? gia ?ình	77	100_077	fttx.thanhtu	50	0	16	Phòng BHKV Thành ph? Tây Ni	Fiber16	50	1	Home 1
74	56	Công ty t? nhân	1	6	Doanh nghi?p trong n??c	1796	700_A020	fkcn01003	180	1	22	Phòng BHKV Tr?ng Bàng	Fiber80 (0	180	1	Home 4

Hình 3.2: Dữ liệu thông tin khách hàng thu thập từ hệ thống ĐHSXKD

Dữ liệu thông tin khách hàng sau khi thu thập từ hệ thống ĐHSXKD cần thực hiện các bước tiền xử lý để loại bỏ các mẫu nhiễu trong tập dữ liệu như các dòng trống, các dòng không có giá trị.Các thông tin khách hàng từ tập dữ liệu sẽ được trích xuất để lấy các thuộc tính quan trọng với quá trình đề xuất gói cước, các thông tin được trích xuất cụ thể như sau:

Bảng 3.1 Bảng số trường và ý nghĩa từng trường dữ liệu

TT	Tên trường dữ liệu	Ý nghĩa	Kiểu dữ liệu
1	TEN_DVDB	Địa chỉ lắp đặt thuê bao	Liệt kê
2	TUOI	Độ tuổi khách hàng	Số
3	TEN_LOAIKH	Loại khách hàng là cá nhân hay doanh nghiệp	Liệt kê
4	TEN_NHOM	Tên nhóm khách hàng	Liệt kê
5	IP_TINH	Nhu cầu sử dụng IP tĩnh	Liệt kê
6	TOC_DO	Băng thông cần thiết cho khách hàng	Số
7	NAM_DK	Năm khách hàng đăng ký dịch vụ	Số
8	GOI	Gói cước khách hàng sử dụng	Liệt kê



index	TUOI	TEN_LOAIKH	TEN_NHOM	IP_TINH	TEN_DVDB	GOI
0	46	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Thành phố Tây Ninh	Home 2
1	57	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dầu	Home TV1
2	57	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dầu	Home TV2
3	21	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dầu	Home 1
4	25	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Bến Cầu	Home 2
5	28	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dầu	Home 2
6	60	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dầu	Home TV1
7	21	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dầu	Home 2
8	39	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dầu	Home 2
9	38	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dầu	Home 2
10	60	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dầu	Home TV1
11	65	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dầu	Home 1
12	37	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Châu Thành	Home 1
13	57	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	1	Phòng BHKV Gò Dầu	Home 4
14	59	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dầu	Home TV2
15	69	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dầu	Home 1
16	49	Cong ty tu nhan	Doanh nghiep trong nuoc	0	Phòng BHKV Gò Dầu	Home 1
17	54	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Tân Biên	Home 2
18	78	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Tân Biên	Home 2
19	35	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Tân Biên	Home TV2
20	53	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Tân Châu	Home TV1
21	50	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Tân Châu	Home TV2
22	40	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Tân Châu	Home 2
23	52	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Thành phố Tây Ninh	Home 2
24	51	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Thành phố Tây Ninh	Home TV2

Hình 3.3: Dữ liệu sau khi Import

Trường địa chỉ lắp đặt thuê bao: là nơi khách hàng đăng ký lắp đặt thuê bao, phản ánh vị trí địa lý nơi khách hàng sử dụng dịch vụ.

Trường độ tuổi: thông tin độ tuổi của khách hàng đăng ký thuê bao.

Trường tên loại khách hàng: thông tin loại khách hàng được phân nhóm theo nghề nghiệp của khách hàng.

Trường tên nhóm khách hàng: khách hàng được phân nhóm vào vào các nhóm khách hàng chính dựa vào mức độ ưu tiên.

Trường IP tĩnh: cho biết nhu cầu sử dụng IP tĩnh của khách hàng.

Trường năm đăng ký: cho biết năm khách hàng đăng ký bắt đầu sử dụng dịch vụ.

Trường tốc độ: cho biết băng thông cần thiết theo nhu cầu của khách hàng, được đo bằng Mbps.

Trường gói cước: cho biết gói cước mà khách hàng sử dụng.

3.1.2. Xử lý dữ liệu

Việc xử lý các dữ liệu ngoại lai sẽ giúp tăng cao độ chính xác cho các mô hình dự đoán hay các báo cáo doanh nghiệp một cách đáng kể. Từ kết quả của dữ liệu thu thập được, chúng ta có thể thấy rằng, dữ liệu vẫn một số trường hợp có thuộc tính có giá trị null khi không có thông tin.

Xứ lý dữ liệu ngoại lai/dữ liệu bất thường là một trong những thuật ngữ được sử dụng rất rộng rãi trong thế giới data và đặc biệt là data science. Xác định và loại bỏ dữ liệu ngoại lai là một bước cực kỳ quan trọng trong quá trình xử lý dữ liệu. Việc xử lý các dữ liệu ngoại lai sẽ giúp tăng cao độ chính xác cho các mô hình dự đoán hay các báo cáo doanh nghiệp một cách đáng kể.

Để xử lý vấn đề này, chúng ta sẽ điều giá trị 0 vào các giữ liệu bị null. Vì theo nhận định, các trường dữ liệu trước khi đưa vào mô hình dự doán đều được chuẩn

hóa sang dạng số bắt đầu từ 1 nên trường dữ liệu null tức giá trị bằng 0 sẽ là hợp lý nhất.

3.1.3. Mã hóa dữ liệu

Mã hóa dữ liệu là quá trình bắt buộc sử dụng các phương pháp mã hóa cụ thể để tạo ra các giá trị phân nhóm cụ thể cho từng đặc trưng.

Bước đầu tiên là mã hóa các dữ liệu của trường loại khách hàng, căn cứ vào sự phân bố của các loại khách hàng ta mã hóa như sau: Cá nhân hộ gia đình = 1, Công ty tư nhân = 2, Hành chính sự nghiệp = 3.

Từ kiểu dữ liệu có thể thấy, một số trường đang có kiểu dữ liệu chuỗi, vì vậy để có thể dễ dàng phân tích và đặc biệt là phục vụ cho các mô hình học máy ta sẽ chuyển chúng về dạng số. Phương pháp mã hóa dữ liệu sẽ dùng là phương pháp Label Encoder.

index	TEN_DVDB	GOI	TUOI	TEN_LOAIKH	TEN_NHOM	NAM_DK	IP_TINH	TOC_DO
0	4	0	50	1	1	17	0	50
1	2	3	47	1	1	17	0	40
2	5	1	34	1	1	18	0	80
3	5	1	68	1	1	18	0	80
4	5	1	29	1	1	18	0	80
5	3	1	45	1	1	18	0	80
6	8	1	49	1	1	18	0	80
7	7	3	29	1	1	20	0	40
8	4	3	29	1	1	20	0	40
9	8	4	43	1	1	20	0	80
10	6	4	27	1	1	19	0	80
11	7	4	45	1	1	19	0	80
12	5	1	37	56	6	20	0	100
13	1	4	72	1	1	19	0	80
14	0	1	41	1	1	18	0	80
15	0	1	35	1	1	18	0	80
16	3	1	40	1	1	19	0	80
17	8	4	56	1	1	18	0	80
18	6	2	39	56	6	18	1	150
19	8	1	35	1	1	18	0	80
20	8	1	50	1	1	18	0	80
21	3	4	44	1	1	20	0	80
22	5	4	43	3	5	20	1	80
23	3	4	54	1	1	20	0	80

Hình 3.4: Dữ liệu được mã hóa bằng phương pháp Label Encoder

3.2. Xây dựng mô hình khuyến nghị gói cước dựa vào thuật toán rừng ngẫu nhiên

Một khách hàng với các thông tin quan trọng sau khi được trích xuất và chuẩn hóa được chuyển về dạng véc tơ để làm đầu vào xây dựng Rừng ngẫu nhiên với đầu vào và đầu ra như sau:

Đầu vào: tập dataset Y là tập dữ liệu thông tin khách hàng với các thuộc tính độ tuổi, loại khách hàng, tên nhóm khách hàng, địa bàn khách hàng đăng ký dịch vụ, nhu cầu sử dụng IP tĩnh, năm đăng ký bắt đầu sử dụng dịch vu, băng thông cần thiết, gói cước khách hàng sử dụng và giá trị ngưỡng đánh giá σ (giá trị trong luận văn sử dụng là giá trị accuracy score và điều kiện là σ > 0.8).

Đầu ra: Rừng ngẫu nhiên với tập hợp các cây quyết định tối ưu R.

3.2.1. Lấy mẫu dữ liệu cho việc xây dựng cây quyết định trong rừng ngẫu nhiên

Chia tập dataset Y ngẫu nhiên thành 2 phần: tập dữ liệu để kiểm tra E (20%) và tập dữ liệu để huấn luyện T (80%). Sử dụng T để xây dựng Rừng ngẫu nhiên. Dữ liệu huấn luyện T với các thuộc tính độ tuổi, loại khách hàng, tên nhóm khách hàng, địa bàn khách hàng đăng ký dịch vụ, như cầu sử dụng IP tĩnh để xây dựng mô hình phân lớp. Tổng cộng có 5 thuộc tính dùng cho việc phân lớp dữ liệu. Giả sử có 1000 mẫu dữ liệu thông tin khách hàng, 800 mẫu sẽ được dùng để xây dựng một cây quyết định, 200 mẫu thông tin khách hàng còn lại dùng để đánh giá mô hình cây quyết định xây dựng được có thỏa mãn giá trị đánh giá σ hay không. Nếu thỏa mãn cây vừa xây dựng sẽ được thêm vào tập cây của rừng ngẫu nhiên.

```
df
```

	TUOI	TEN_LOAIKH	TEN_NHOM	IP_TINH	TEN_DVDB	GOI
0	46	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Thành pho Tây Ninh	Home 2
1	57	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dau	Home TV1
2	57	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dau	Home TV2
3	21	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dau	Home 1
4	25	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Ben Cau	Home 2
...	...	...	...	...	...	...
995	46	Cong ty tu nhan	Doanh nghiep trong nuoc	1	Phòng BHKV Trang Bàng	Home 4
996	79	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Trang Bàng	Home TV4
997	64	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Trang Bàng	Home 2
998	61	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Trang Bàng	Home 2
999	65	Cong ty tu nhan	Doanh nghiep trong nuoc	1	Phòng BHKV Trang Bàng	Home 4

1000 rows × 6 columns

Hình 3.5: Tập dữ liệu 1000 mẫu thông tin khách hàng

```
train
```

	TUOI	TEN_LOAIKH	TEN_NHOM	IP_TINH	TEN_DVDB	GOI
848	45	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Châu Thành	Home 2
961	66	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV D??ng Minh Châu	Home 2
179	58	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Tân Châu	Home TV1
812	34	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Châu Thành	Home 2
999	65	Cong ty tu nhan	Doanh nghiep trong nuoc	1	Phòng BHKV Trang Bàng	Home 4
...	...	...	...	...	...	...
662	52	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Tân Châu	Home TV1
231	45	Công ty nh? n??c	Doanh nghiep trong nuoc	0	Phòng BHKV Châu Thành	Home 2
741	39	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV D??ng Minh Châu	Home 2
339	21	Th??c Giáo d?c	Chính sách c??c	0	Phòng BHKV D??ng Minh Châu	Home TV2
682	37	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Tân Châu	Home TV2

800 rows × 6 columns

Hình 3.6: Tập huấn luyện cây quyết định với 800 mẫu được lấy ngẫu nhiên

```
test
```

	TUOI	TEN_LOAIKH	TEN_NHOM	IP_TINH	TEN_DVDB	GOI
396	51	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Trang Bàng	Home 2
760	53	Th??c Giáo d?c	Chính sách c??c	0	Phòng BHKV Trang Bàng	Home 4
670	48	Cong ty tu nhan	Doanh nghiep trong nuoc	0	Phòng BHKV Thành pho Tây Ninh	Home TV4
988	81	Cong ty tu nhan	Doanh nghiep trong nuoc	1	Phòng BHKV Trang Bàng	Home 4
67	45	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV D??ng Minh Châu	Home TV2
...	...	...	...	...	...	...
685	42	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Tân Châu	Home 2
276	56	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Tân Châu	Home 2
951	58	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV Gò Dau	Home TV2
939	61	Cong ty tu nhan	Doanh nghiep trong nuoc	1	Phòng BHKV Trang Bàng	Home 4
623	35	Ca nhan ho gia dinh	Ca nhan - Ho gia dinh	0	Phòng BHKV D??ng Minh Châu	Home 2

200 rows × 6 columns

Hình 3.7: Tập thử nghiệm với 200 mẫu còn lại để đánh giá cây quyết định

3.2.2. Xây dựng cây quyết định trong rừng ngẫu nhiên

Quá trình tạo cây quyết định với từng tập dữ liệu sau quá trình lấy mẫu ngẫu nhiên ở bước trước được thực hiện như sau:

Quá trình xây dựng cây bắt đầu bằng việc khởi động nút gốc thành nút nhị phân. Ban đầu, tất cả các mẫu dữ liệu là nằm trong nút gốc. CART triển khai một tổ hợp các thuật toán chuyên sâu tìm kiếm cách phân chia tốt nhất ở tất cả các điểm phân chia có thể có cho mỗi biến. Các phương pháp mà CART sử dụng để xây dựng cây quyết định được gọi là phân vùng đệ quy nhị phân. Thông qua chỉ số Gini như một quy tắc tách.

Bước 1: CART tách biến đầu tiên trong tất cả các biến tại các điểm phân tách có thể xảy ra, tại tất cả các giá trị mà biến giả định có trong tập mẫu. Tại mỗi điểm phân chia có thể có của một biến, mẫu phân tách thành hai nút con. Các trường hợp có câu trả lời "có" cho câu hỏi được đặt ra được gửi đến nút bên trái và những trường hợp phản hồi "không" được gửi đến đúng nút bên phải.

Bước 2: CART sau đó áp dụng tiêu chí phân tách dựa trên công thức GINIsplit tại mỗi điểm phân tách và đánh giá mức giảm tạp chất đạt được bằng cách sử dụng công thức:

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Bước 3: CART chọn cách tách tốt nhất cho các biến với sự phân chia mà sự giảm tạp chất là cao nhất. Ba bước trên được lặp lại cho mỗi biến còn lại ở nút gốc.

Bước 4: CART sau đó xếp hạng tất cả các cách chia tốt nhất trên mỗi biến theo sự giảm tạp chất đạt được bằng mỗi lần tách và chọn biến cùng điểm tách mà nó làm giảm tạm chất của nút gốc và nút trung gian nhiều nhất.

Bước 5: CART sau đó gán lớp cho các nút này theo quy tắc giảm thiểu phân loại. CART có một thuật toán tích hợp để cho phép người dùng xác định điểm gán

lớp trong quá trình tách. Mặc định là 1 đơn vị hoặc bằng giá trị phân loại. Bởi vì thủ tục CART là đệ quy, các bước 1 - 5 được áp dụng nhiều lần cho mỗi phần tử không phải nút lá ở mỗi giai đoạn kế tiếp.

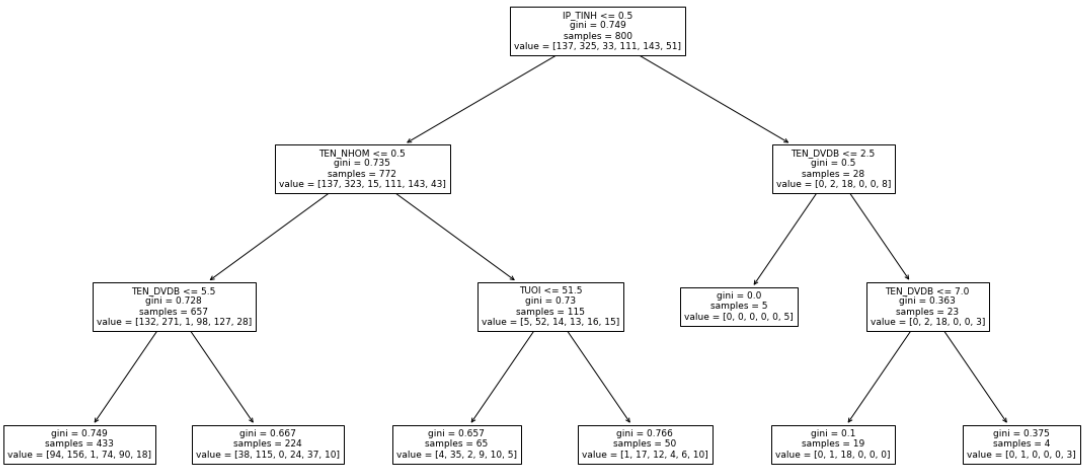
CART dừng quá trình phân tách khi:

- Chỉ có một mẫu trong mỗi nút.

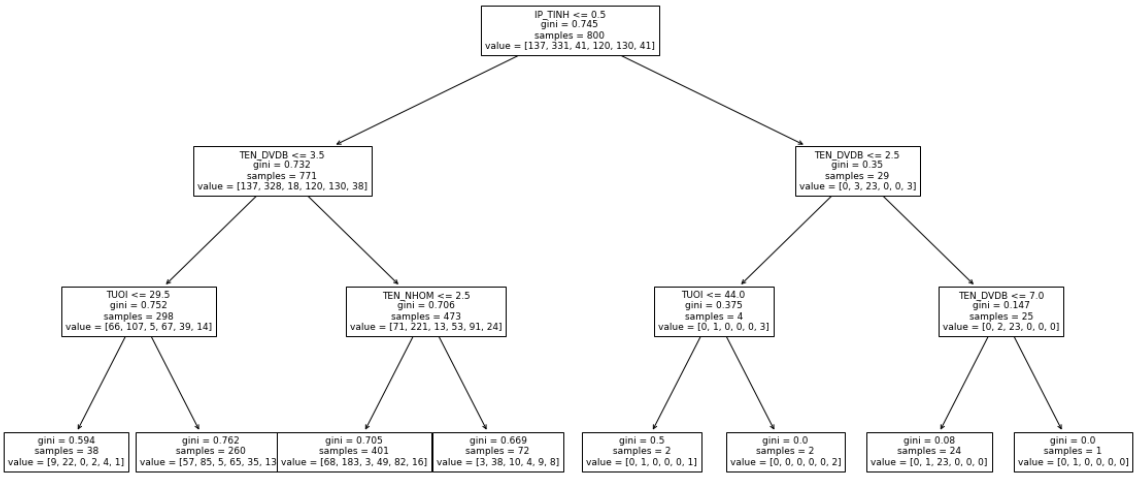
- Tất cả các mẫu trong mỗi nút con có phân loại giống hệt nhau của các biến phân lớp, tức là không thể tách.

- Cây đạt đủ độ sâu theo cài đặt (max_depth=10).

Một số ví dụ về cây quyết định được xây dựng với giá trị max_depth=3 được thể hiện qua các hình minh họa như sau:



Hình 3.8: Cây quyết định xây dựng trên mẫu huyến luyện ngẫu nhiên thứ nhất

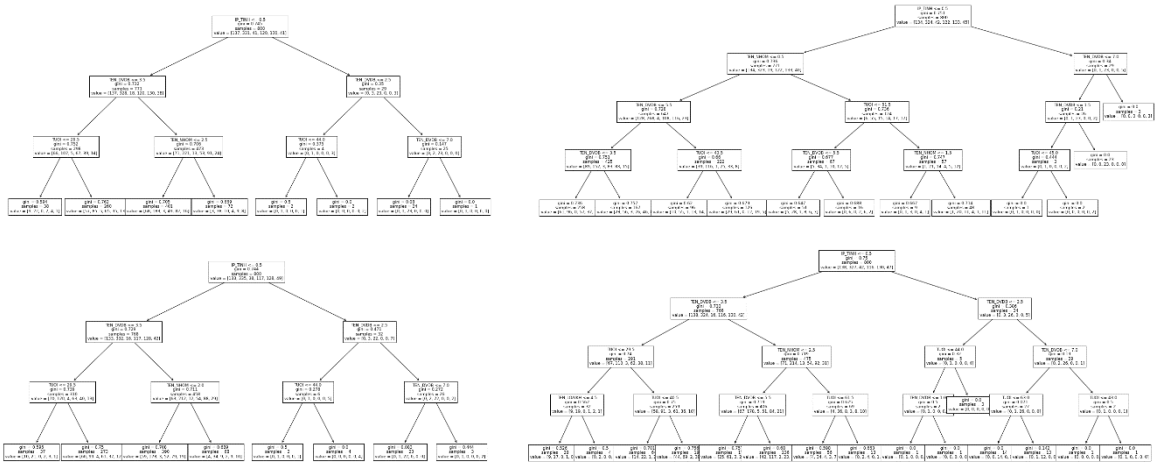


Hình 3.9: Cây quyết định xây dựng trên mẫu huyến luyện ngẫu nhiên thứ hai

3.2.3. Xây dựng rừng ngẫu nhiên

Cây quyết định sau khi xây dựng xong nếu thỏa giá trị đánh giá ban đầu sẽ được cập nhật vào rừng ngẫu nhiên R. Việc đánh giá dựa trên tập thử nghiệm E và giá trị đánh giá σ. Độ chính xác của cây quyết định, cụ thể là giá trị F1 Score của cây sẽ được tính dựa vào kết quả phân loại tập thử nghiệm E, nếu Accuracy Score > 0.8 thì cây sẽ được đưa vào rừng ngẫu nhiên và ngược lại.

Toàn bộ quá trình lấy mẫu ngẫu nhiên dữ liệu và xây dựng cây quyết định được thực hiện lặp lại cho đến khi không thể tạo thêm cây mới hoặc rừng ngẫu nhiên đạt đủ số lượng cây theo cài đặt ban đầu. Sau khi kết thúc ra sẽ thu được một mô hình rừng ngẫu nhiên gồm nhiều cây quyết định tối ưu được xây dựng trên tập mẫu được lấy ngẫu nhiên từ tập dataset ban đầu.



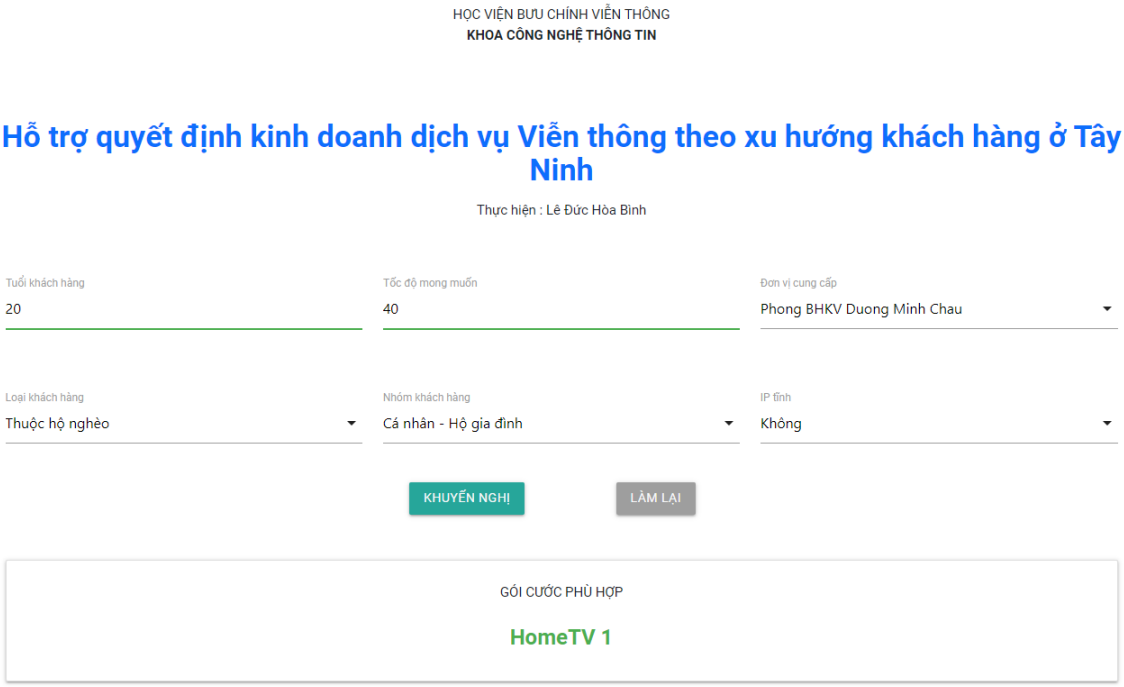
Hình 3.10: Một ví dụ rừng ngẫu nhiên với 4 cây quyết định

3.3. Xây dựng ứng dụng web

Sử dụng kết quả mô hình khuyến nghị gói cước dựa trên rừng ngẫu nhiên xây dựng được ở mục 3.2 để đưa vào ứng dụng web được xây dựng trên môi trường Pycharm.

Ứng dụng web được thiết kế để khách hàng hoặc nhân viên VNPT nhập các thông tin cần thiết là các thuộc tính phân lớp trong mô hình rừng ngẫu nhiên như: Tuổi khách hàng, Tốc độ mong muốn, Đơn vị cung cấp, Loại khách hàng, Nhóm khách hàng, nhu cầu dùng IP tĩnh.

Giao diện ứng dụng web được xây dựng như sau:



Hình 3.11: Giao diện ứng dụng web

CHƯƠNG 4 – PHÂN TÍCH VÀ ĐÁNH GIÁ

4.1. Phân tích độ chính xác của mô hình

Để đánh giá độ chính xác của mô hình đã xây dựng ta dựa vào kết quả phân lớp tập thử nghiệm là tập dữ liệu được trích xuất từ tập dataset ban đầu với 1000 mẫu thông tin khách hàng được xây dựng thủ công, trường thuộc tích gói cước được chọn từ kinh nghiệm của chuyên viên tư vấn bán hàng có kinh nghiệm để cho kết quả phù hợp với khách hàng nhất. Ta có ma trận hỗn loạn sau:

Bảng 4.1: Ma trận hỗn loạn

		Gói cước thực tế của khách hàng					
		Home 1	**Home 2**	**Home 3**	**Home TV 1**	**Home TV 2**	**Home TV 3**
Gói cước thông qua mô hình Khuyến nghị gói cước	**Home 1**	**T_1**	F_{21}	F_{31}	F_{41}	F_{51}	F_{61}
	Home 2	F_{12}	**T_2**	F_{32}	F_{42}	F_{52}	F_{62}
	Home 3	F_{13}	F_{23}	**T_3**	F_{43}	F_{53}	F_{63}
	Home TV 1	F_{14}	F_{24}	F_{34}	**T_4**	F_{54}	F_{64}
	Home TV 2	F_{15}	F_{25}	F_{35}	F_{45}	**T_5**	F_{65}
	Home TV 3	F_{16}	F_{26}	F_{36}	F_{46}	F_{56}	**T_6**

Trong đó:

- T_i là số lượng khách hàng có gói cước thông qua mô hình Khuyến nghị gói cước đúng với gói cước thực tế (i = 1, 2, 3, 4, 5, 6).

- F_{ij} là số lượng khách hàng có gói cước thông qua mô hình Khuyến nghị gói cước không đúng với gói cước thực tế (i, j = 1, 2, 3, 4, 5, 6).

Giá trị Accuracy Score dùng để đánh giá mô hình được tính như sau:

$$\textbf{Accuracy Score} = \frac{\Sigma \text{ Ti}}{\Sigma \text{ Fij} + \Sigma \text{ Ti}}$$

Với tập thử nghiệm 1000 mẫu thông tin khách hàng được lấy ngẫu nhiên từ tập dataset sau đó xây dựng gói cước thủ công ta có các thông tin sau:

- Tổng số khách hàng là 1000.

- Số khách hàng dùng gói Home 1 là 278.

- Số khách hàng dùng gói Home 2 là 153.

- Số khách hàng dùng gói Home 3 là 37.

- Số khách hàng dùng gói Home TV 1 là 369.

- Số khách hàng dùng gói Home TV 2 là 159.

- Số khách hàng dùng gói Home TV 3 là 4.

Thực hiện dự đoán gói cước cho 1000 khách hàng trên thông qua mô hình rừng ngẫu nhiên đã xây dựng, kết quả dự đoán gói cước khuyến nghị cho khách hàng trong 1 lần chạy thử nghiệm như sau:

- Số khách hàng thực tế dùng gói Home 1 được mô hình khuyến nghị đúng là 266. Số khách hàng thực tế dùng gói Home 1 bị mô hình khuyến nghị không đúng là 12.

- Số khách hàng thực tế dùng gói Home 2 được mô hình khuyến nghị đúng là 138. Số khách hàng thực tế dùng gói Home 2 bị mô hình khuyến nghị không đúng là 15.

- Số khách hàng thực tế dùng gói Home 3 được mô hình khuyến nghị đúng là 30. Số khách hàng thực tế dùng gói Home 3 bị mô hình khuyến nghị không đúng là 7.

- Số khách hàng thực tế dùng gói Home TV 1 được mô hình khuyến nghị đúng là 304. Số khách hàng thực tế dùng gói Home 1 bị mô hình khuyến nghị không đúng là 65.

- Số khách hàng thực tế dùng gói Home TV 2 được mô hình khuyến nghị đúng là 145. Số khách hàng thực tế dùng gói Home TV 2 bị mô hình khuyến nghị không đúng là 14.

- Số khách hàng thực tế dùng gói Home TV 3 được mô hình khuyến nghị đúng là 2. Số khách hàng thực tế dùng gói Home TV 3 bị mô hình khuyến nghị không đúng là 2.

Theo kết quả trên ta có:

Accuracy Score = $\frac{266+138+30+304+145+2}{1000} = 88.5\%$

Để đánh giá mức độ ảnh hưởng của các tham số quan trọng trong quá trình xây dựng rừng ngẫu nhiên là max_depth (độ sâu của cây quyết định) và n_ estimators (số lượng cây quyết định trong rừng ngẫu nhiên), các thử nghiệm được thực hiện 10 lần và các kết quả thu được dựa trên giá trị trung bình của các lần chạy như bảng kết quả đánh giá sau.

Bảng 4.2: Giá trị Accuracy Score với hai tham số quan trọng của rừng ngẫu nhiên

max_depth\n_estimators	20	30	50	100	200
5	88.28	88.28	88.39	88.31	88.3
10	88.42	**88.51**	88.43	88.47	88.39
15	88.23	88.25	88.32	88.31	88.27
20	87.74	87.79	87.64	87.56	87.46
25	87.32	87.45	87.54	87.36	87.31

Dựa vào các chỉ số đánh giá độ chính xác của mô hình trên có thể đưa ra các nhận định sau:

- Mô hình cho độ chính xác cao nhất với giá trị max_depth = 10, giá trị n_estimators = 30 với các giá trị Accuracy Score = 88.51%.

- Số lượng thuộc tính của khách hàng còn ít nên chưa tạo được sự bao quát cho mô hình nên độ chính xác của mô hình còn chưa thực sự cao. Việc sử dụng ít thuộc tính thông tin khách hàng dẫn đến trường hợp có thể một thuộc tính thực sự chưa ảnh hưởng đến việc lựa chọn gói cước cho khách hàng nhưng lại có giá trị ảnh hưởng cao trong mô hình khuyến nghị gói cước. Vì thế ta tiến hành đánh giá mức độ quan trọng của các thuộc tính trong mô hình để xác định các thuộc tính ảnh hưởng đến kết quả khuyến nghị gói cước có phải là yếu tố thực tế mang tính quyết định đến việc lựa chọn gói cước cho khách hàng hay không.

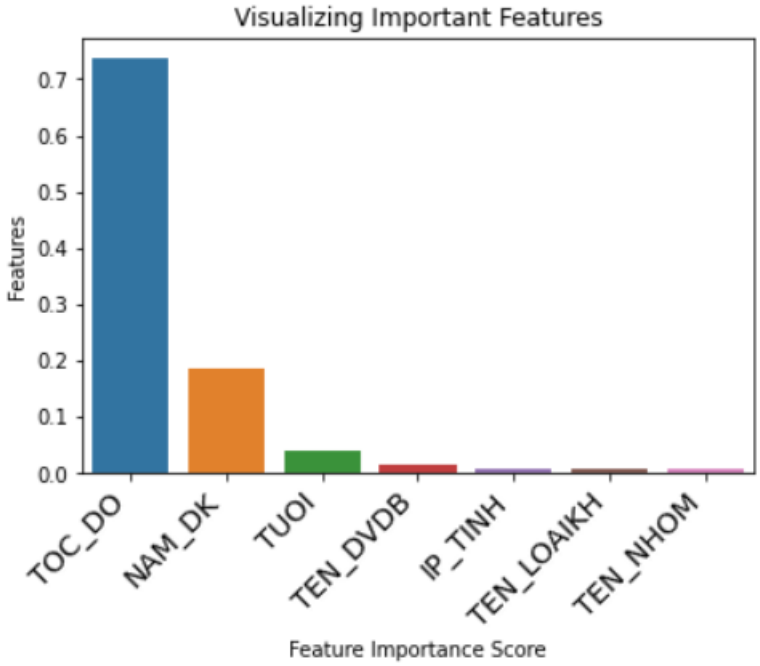
4.2. Xác định mức độ quan trọng của các thuộc tính

Mức độ quan trọng của các thuộc tính được xác định bằng độ giảm của chỉ số gini tại mỗi nút trong quá trình xây dựng cây quyết định. Độ giảm chỉ số gini càng nhiều ứng với mức độ quan trọng của thuộc tính càng cao.

Việc đánh giá mức độ quan trọng của các thuộc tính cho phép chúng ta phân tích được vai trò của mỗi thuộc tính trong việc xây dựng mô hình phân lớp. Trong luận văn mức độ quan trọng của các thuộc tính được thể hiện qua kết quả sau đây.

feature	importance
TOC_DO	0.735642
NAM_DK	0.187087
TUOI	0.040386
TEN_DVDB	0.014838
IP_TINH	0.007593
TEN_LOAIKH	0.007279
TEN_NHOM	0.007175

Hình 4.1: Kết quả mức độ quan trọng của các thuộc tính



Hình 4.2: Biểu đồ mức độ quan trọng của các thuộc tính

Như kết quả được thể hiện qua các hình ảnh bên trên, thuộc tính TOC_DO ảnh hưởng đến 73.56% kết quả gói cước sẽ được khuyến nghị cho khách hàng. Điều này hoàn toàn đúng với thực tế vì khi các chuyên gia tư vấn bán hàng sẽ ưu tiên gói cước nào có băng thông phù hợp với nhu cầu sử dụng của khách hàng nhất hơn là các thuộc tính khác được sử dụng trong mô hình khuyến nghị đã xây dựng.

Thuộc tính quan trọng thứ hai là NAM_DK với 18.71%, năm khách hàng đăng ký sử dụng dịch vụ đóng vai trò quan trọng trong việc quyết định gói cước mà khách hàng sử dụng cho thấy việc áp dụng các chính sách kinh doanh cho các gói cước ở các giai đoạn khác nhau là khác nhau. Trong một năm sẽ có một hoặc một số gói cước được ưa chuộng đặc biệt và phổ biến hơn các gói cước còn lại.

Các thuộc tính còn lại hầu như không ảnh hưởng nhiều đến gói cước được chọn để khuyến nghị cho khách hàng. Điều này có thể giải thích do tập dữ liệu thông tin khách hàng hiện tại ở các năm từ 2015-2021 chưa tập trung tư vấn và cung cấp các gói cước phù hợp với các đặc tính riêng biệt của từng khách hàng mà chỉ tập trung vào nhu cầu băng thông khách hàng cần sử dụng. Đây cũng là một điểm cần phải thay đổi trong việc tư vấn bán hàng để mang lại trải nghiệm tuyệt với nhất khi khách hàng sử dụng dịch vụ Internet do VNPT cung cấp.

CHƯƠNG 5 - KẾT LUẬN

5.1. Kết quả đạt được

5.1.1. Về mặt lý thuyết

Khai thác được tập dữ liệu thông tin khách hàng sử dụng Internet của VNPT Tây Ninh để xây dựng mô hình Khuyến nghị gói cước cho khách hàng.

Ứng dụng Trí tuệ nhân tạo (AI), Machine Learning, các thuật toán học máy vào việc khuyến nghị gói cước cho khách hàng.

Khai thác được các thuật toán phân lớp dữ liệu, cụ thể là mô hình cây quyết định và rừng ngẫu nhiên. Nắm bắt được quá trình xây dựng một cây quyết định dựa trên giá trị gini index hay entropy và quá trình xây dựng một rừng ngẫu nhiên dựa trên các cây quyết định.

Ứng dụng thư viện scikit-learn trên nền tảng python vào việc nghiên cứu các vấn đề học máy, sử dụng được các tham số để tối ưu mô hình rừng ngẫu nhiên xây dựng được.

5.1.2. Về mặt thực tiễn

Luận văn đã đưa ra được giải pháp khuyến nghị gói cước phù hợp với khách hàng sử dụng dịch vụ Internet của VNPT dựa vào việc phân tích tập dữ liệu khách hàng hiện có. Việc này sẽ là tiền đề để xây dựng một công cụ tư vấn bán hàng tự động thay thế cách tư vấn truyền thống nhân công mất thời gian và nhân lực nhưng đôi khi lại cho kết quả gói cước tư vấn cho khách hàng chưa thực sự phù hợp với khách hàng đối với các nhân viên chưa có kinh nghiệm.

Mô hình trên có thể hỗ trợ khách hàng chủ động tìm kiếm các gói cước phù hợp với bản thân khi cung cấp các thông tin cần thiết để chọn ra gói cước phù hợp nhất.

Xây dựng thành công mô hình khuyến nghị gói cước, phân tích và đánh giá mô hình xây dựng được để hiểu rõ hơn về các yếu tố ảnh hưởng đến việc lựa chọn một gói cước phù hợp cho khách hàng. Từ đó cũng rút ra được các điểm còn thiếu sót để có thể tiến hành thay đổi trong công tác thực hiện thực tế để mang lại trải nghiệm tốt nhất cho khách hàng sử dụng dịch vụ Internet của VNPT.

5.2. Hạn chế

Kết quả khuyến nghị gói cước đạt được chỉ ở mức tốt chứ chưa thật sự cao. Kết quả đạt được chưa bao quát được hết các trường hợp. Dữ liệu thông tin khách hàng cần bổ sung thêm các thuộc tính thực tế tác động đến việc lựa chọn gói cước của khách hàng như số lượng thiết bị cần sử dụng Internet, cơ sở hạ tầng lắp đặt (nhà cấp 4, nhà lầu, nhà trọ...), các mục đích sử dụng chính khi lắp đặt Internet...

Mô hình rừng ngẫu nhiên trong luận văn còn ở mức cơ bản, chưa phân tích sâu vào các tham số để phù hợp với mô hình dữ liệu thông tin khách hàng sử dụng Internet của VNPT.

Kết quả của giải pháp học máy phụ thuộc không chỉ giải thuật học máy mà còn do bộ dữ liệu sử dụng. Bộ dữ liệu hiện tại không chứa nhiều thuộc tính có giá trị cho việc khuyến nghị gói cước phù hợp đến khách hàng vì vậy cần phải tiếp nhận thêm thông các thông tin cần thiết khác của khách hàng để tư vấn gói cước phù hợp so với hiện tại.

5.3. Hướng phát triển

Điều chỉnh công tác tư vấn bán hàng sang hướng tối ưu hóa trải nghiệm cho khách hàng hơn, thu thập thêm các thông tin mang tính cá nhân hóa với từng khách hàng như số lượng thiết bị cần sử dụng Internet, cơ sở hạ tầng lắp đặt (nhà cấp 4, nhà lầu, nhà trọ...), các mục đích sử dụng chính khi lắp đặt Internet... để chọn ra gói cước phù hợp nhất. Sau đó sử dụng tập dữ liệu mới với các thuộc tính bổ sung để xây dựng lại một mô hình khuyến nghị gói cước phù hợp nhất với nhu cầu sử dụng của từng

khách hàng. Từ đó gia tăng trải nghiệm của khách hàng khi sử dụng dịch vụ của VNPT.

Tiến hành áp dụng giúp hỗ trợ việc tư vấn bán hàng cho nhân viên kinh doanh mới, bán hàng online trên các website và nền tảng thông tin số khác của Viễn thông Tây Ninh. Với mục đích thực hiện tư vấn gói cước phù hợp, nhanh chóng và có độ tin cậy cao với từng khách hàng riêng biệt.

Nghiên cứu bổ sung các yếu tố ảnh hưởng đến việc lựa chọn gói cước phù hợp cho khách hàng sau đó áp dụng, bổ sung vào các thuộc tính ban đầu để xây dựng lại mô hình khuyến nghị phù hợp và có độ chính xác cao hơn.

DANH MỤC TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Hoàng Ngọc Thanh, Trần Văn Lăng, Hoàng Tùng (2016), "*Một tiếp cận máy học để phân lớp các kiểu tấn công trong hệ thống phát hiện xâm nhập mạng*", Kỷ yếu Hội nghị khoa học Quốc gia FAIR'9, T. 502-507

[2] Nguyễn Thị Thanh Hương, Đoàn Minh Trung (2018), "*Áp dụng thuật toán phân loại Random Forest để xây dựng bản đồ sử dụng đất/thảm phủ tỉnh Đắk Lắk dựa vào ảnh vệ tinh Landsat 8 OLI*", Tạp chí Nông nghiệp & Phát triển nông thôn, T. 122-129

[3] Dang N. H. Thanh, Nguyen Quoc Hung, Tran Le Phuc Thinh (2020), "*Một góc nhìn từ bài toán phân lớp dữ liệu: Thang điểm đánh giá nào là quan trọng?*", Kỷ yếu hội thảo khoa học quốc gia Hệ thống thông tin trong kinh doanh và quản lý ISBM20, T. 276-279

[4] Đỗ Thị Lương (2019), "*Nghiên cứu một số thuật toán học máy để phân lớp dữ liệu và thử nghiệm*", Hà Nội, 62 trang

[5] Nguyễn Thị Thùy Linh (2005), "*Nghiên cứu các thuật toán phân lớp dữ liệu dựa trên cây quyết định*", Hà Nội, T. 21-27

[6] Đỗ Trung Tuấn (2010), "*Nhập môn trí tuệ nhân tạo*", Nhà xuất bản Đại học quốc gia Hà Nội.

[7] Nguyễn Ngọc Tuân (2016), "*Áp dụng khái dữ liệu dư báo thuê bao rời mạng trong mạng năm di động*", Đại học công nghệ-Đại học quốc Hà Nội.

[8] Nguyễn Thành Phúc (2019), "*Phân tích dự báo sản lượng các dịch vụ chuyển phát tại Bưu điện tỉnh Bình Dương*" năm 2019, Đại học Thủ Dầu Một-UBND tỉnh Bình Dương.

[9] Đoàn Văn Tâm (2019), "*Xây dựng mô hình dự đoán khách hàng tiềm năng cho các gói cước trong mạng di động"*, Đại học công nghệ - Đại học quốc gia Hà Nội.

[10] Trần Hữu Nam (2000), *"Nghiên cứu ứng dụng các phương pháp dự báo trong giáo dục - đào tạo"*, Viện Nghiên cứu phát triển giáo dục.

Tiếng Anh

[11] O'Sullivan, Dympna, et al. (2008), "*Using Secondary Knowledge to Support Decision Tree Classification of Retrospective Clinical Data*", Mining Complex Data (2008), pp. 238-251

[12] Christopher J.C. Burges (2000), "*A Tutorial on Support Vector Machines for Pattern Recognition*", Kluwer Academic Publishers, Boston

[13] Sunil Kumar, Saroj Ratnoo, Renu Bala (2020), "*Enhanced Decision Tree Algorithm for Discovery of Exceptions*", Department of Computer Science & Engineering, Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India, pp. 3-7.

[14] M. Uddin, R. Stadler, and A. Clemm (2013), "*A Query Language for Network Search*", Proceedings of the 13th IFIP/IEEE International Symposium on Integrated Network Management (IM'13). IEEE, pp. 109–117

[15] W. Zhou, L. Tang, C. Zeng, T. Li, L. Shwartz, and G. Y. Grabarnik (2016), "*Resolution recommendation for event tickets in service management*", IEEE Transactions on Network and Service Management, vol. 13, no. 4, pp. 954–967, Available: https://doi.org/10.1109/TNSM.2016.2587807

[16] D. Hausheer and C. Morariu (2008), "*Distributed Test-Lab: EMANICSLab*", University of Zurich, Switzerland, The 2nd International Summer School on Network and Service Management (ISSNSM'08)

[17] L. Breiman (2001), "*Random Forests*", Machine Learning, vol. 45, no. 1, pp. 5–32

[18] Gilles Louppe, "*Understanding Random Forest from theory to pratic*", University of Liège, Faculty of Applied Sciences, Department of Electrical Engineering & Computer Science, pp. 55-115

[19] S. Chatrchyan, V. Khachatryan, A. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan (2012) "*Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*", Physics Letters B, pp. 2

[20] A. Criminisi and J. Shotton (2013), "*Decision Forests for Computer Vision and Medical Image Analysis*". Springer, pp. 2, 39, 106 and 107

[21] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager (2010), "*Building Watson: An overview of the Deep QA project*", AI magazine, 31(3):59–79

[22] https://cdspninhthuan.edu.vn/, truy cập ngày 02/01/2022

[23] https://viblo.asia/, truy cập ngày 10/02/2022

[24] https://machinelearningcoban.com/, truy cập ngày 15/02/2022

[25] https://vi.wikipedia.org/, truy cập ngày 20/02/2022

[26] https://www.coursera.org/, truy cập ngày 20/02/2022

[27] https://scikit-learn.org/, truy cập ngày 27/02/2022