ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT
KHOA HỆ THỐNG THÔNG TIN

NHÀ KHO DỮ LIỆU VÀ KHAI PHÁ DỮ LIỆU

Đề tài: Dự đoán ý định mua hàng của người tiêu dùng online khi ghé thăm các cửa hàng trực tuyến

Giảng viên hướng dẫn: **ThS. Nguyễn Thị Anh Thư**
Nhóm thực hiện: **Nhóm 4 - Lớp K18406C**

1. Trần Khánh Duy K184060780
2. Ngô Hữu Tài K184060801
3. Lâm Thị Hoài Thanh K184060802
4. Nguyễn Phạm Thủy Tiên K184060809

Thành phố Hồ Chí Minh, ngày 24 tháng 12 năm 2020

MỤC LỤC

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH ẢNH

LỜI CÁM ƠN

Lời đầu tiên, chúng em xin gửi lời tri ân sâu sắc đến cô Nguyễn Thị Anh Thư. Trong quá trình tìm hiểu và học tập môn Nhà kho dữ liệu và khai phá dữ liệu, chúng em đã nhận được sự giảng dạy và hướng dẫn rất tận tình, tâm huyết của cô. Cô đã giúp em tích lũy thêm nhiều kiến thức hay và bổ ích. Từ những kiến thức mà cô truyền đạt, chúng em xin trình bày lại những gì mình tích lũy được trong quá trình học.

Tuy nhiên, kiến thức về dữ liệu và khai phá dữ liệu của chúng em vẫn còn những hạn chế nhất định. Do đó, không tránh khỏi những thiếu sót trong quá trình hoàn thành đồ án này. Mong cô xem và góp ý để đồ án của chúng em được hoàn thiện hơn.

Kính chúc cô hạnh phúc và thành công hơn nữa trong sự nghiệp "trồng người". Kính chúc cô luôn dồi dào sức khỏe để tiếp tục dìu dắt nhiều thế hệ học trò đến những bến bờ tri thức.

Chúng em xin chân thành cảm ơn!

Nhóm 4

NHẬN XÉT CỦA GIẢNG VIÊN

……

……

……

……

……

……

……

……

……

……

……

……

I. Tổng quan

1. Giới thiệu

Internet đã và đang trở nên rất quan trọng trong đời sống xã hội nói chung và kinh doanh nói riêng. Internet tạo ra một khối lượng lớn các giao dịch kinh doanh trên toàn thế giới, góp phần thúc đẩy sự tăng trưởng của nền kinh tế số. Với việc chuyển đổi từ cửa hàng truyền thống sang trải nghiệm mua sắm trực tuyến, các ứng dụng thương mại điện tử cho phép người dùng tìm kiếm sản phẩm, so sánh giá và sau đó mua sản phẩm từ các nhà bán lẻ trực tuyến. Tuy nhiên, khách hàng có thể thoát ra khỏi trang web ở một bước nào đó trước khi họ hoàn tất giao dịch. Điều này làm ảnh hưởng đến lợi nhuận hoặc doanh thu bị mất của các nhà bán lẻ trực tuyến.

Do đó, việc nghiên cứu toàn bộ quá trình mua hàng để cải thiện tỷ lệ khách hàng quay lại trong lĩnh vực thương mại điện tử là điều cần thiết đối với các doanh nghiệp. Chẳng hạn, với trang web của một công ty bảo hiểm, các câu hỏi cần được đặt ra là: Có bao nhiêu khách hàng ghé thăm trang web? Có bao nhiêu khách hàng tìm hiểu các gói bảo hiểm khác nhau? Có bao nhiêu khách hàng "nhấp chuột" vào nút "Yêu cầu gửi báo giá"? Có bao nhiêu khách hàng khai báo thông tin cá nhân? Có bao nhiêu người trở thành khách hàng có mua bảo hiểm của công ty?

Vì vậy, nhóm chúng em quyết định lựa chọn đề tài ***"Xây dựng mô hình dự đoán ý định mua hàng của người tiêu dùng online khi ghé thăm các cửa hàng trực tuyến"*** để giải quyết những vấn đề trên.

2. Input và Output

Các thuộc tính được trích xuất từ thông tin tổng hợp và dữ liệu hoạt động trực tuyến. Từ đó, phân lớp các truy cập dựa theo mục đích của người tiêu dùng để xây dựng bài toán học máy có giám sát với mục tiêu hướng đến việc ước tính khuynh hướng của khách truy cập để hoàn thành giao dịch.

Như vậy, chúng ta sẽ xác định những người dùng truy cập vào trang web bán hàng online với mục đích mua hàng và cung cấp những nội dung dành riêng cho họ nếu họ thường hay rời khỏi trang web mà chưa hoàn thành bất kì giao dịch nào. Ta cũng sử dụng dữ liệu bán lẻ trực tuyến và kiểm tra hiệu suất của các phương pháp máy học dưới

những điều kiện khác nhau. Từ đó, xác định các nhân tố tách bạch nhất trong việc dự đoán ý định mua hàng bằng cách sử dụng các kỹ thuật lựa chọn tính năng cho bộ lọc.

Input:

+ Các dữ liệu về phiên truy cập của người mua hàng online trong bộ dữ liệu **Online Shopper Purchasing Intention**.
+ Số lượng thuộc tính: 18.
+ Tính chất dữ liệu: 10 thuộc tính Numeric và 8 thuộc tính Categorical.

Output:

+ Kết quả phân tích cho biết khách hàng có thực hiện giao dịch hay không. Với mỗi giao dịch được thực hiện sẽ đem lại lợi nhuận cho cửa hàng.
+ Tỷ lệ phần trăm hai lớp sau khi thực hiện các phương pháp phân lớp. Từ đó chọn ra phương pháp phân lớp phù hợp và có độ chính xác cao nhất.

3. Thách thức của bài toán

Một trong những khó khăn khi xây dựng mô hình này là lựa chọn phương pháp phân lớp có độ chính xác và hiệu quả tốt nhất phù hợp với thuật toán nhận dạng mẫu, cấu trúc dữ liệu thu được từ người dùng trong quá trình triển khai máy học *(machine learning)*. Một vấn đề khác có thể gặp phải trong quá trình phân tích ý định mua hàng của người tiêu dùng là sự chênh lệch giữa các lớp phân biệt. Có khả năng số trường hợp tích cực *(positive)* trong các hành động bỏ qua dữ liệu được thực hiện ít hơn số trường hợp tiêu cực *(negative)* đại diện cho tất cả các hành động khác. Trong trường hợp này, kết quả thu được sẽ được đánh giá bằng các thước đo thích hợp thay vì tỷ lệ chính xác và có thể dẫn đến đánh giá sai kết quả.

4. Đối tượng và phạm vi

Tập dữ liệu bao gồm các vectơ đặc trưng thuộc 12.330 phiên. Nhóm sử dụng dữ liệu luồng nhấp chuột *(clickstream data)* từ các cửa hàng trực tuyến và thông tin người dùng được trích xuất từ thông tin phiên của lượt truy cập trong khoảng thời gian một năm để tránh ảnh hưởng đến một chiến dịch cụ thể hay các dịp lễ nào.

5. Mục tiêu

Mục tiêu của đề tài là nghiên cứu và xây dựng mô dự đoán ý định mua hàng của người tiêu dùng online khi ghé thăm các cửa hàng trực tuyến. Trong bài toán này, các chuyển động chuột, liên kết và thông tin mà người dùng nhấp vào dùng để xác định các trang đã truy cập, số lần khách ghé thăm. Các hành động được thực hiện từ kết quả của những dữ liệu này sẽ được xác định và được sử dụng làm nhãn dữ liệu trong quá trình phân lớp bằng các thuật toán máy học có giám sát. Khi bất kỳ người dùng nào có các hành động phù hợp với mẫu được xác định trước, họ sẽ được phân lớp vào mẫu đó và chúng ta sẽ xác định người dùng có thực hiện giao dịch mua hàng khi truy cập vào trang web đó hay không.

II. Mô tả dữ liệu

1. Giới thiệu dữ liệu

+ Tên data: **Online Shopper Purchasing Intention**
+ Nguồn dữ liệu: UCI Machine Learning Repository
+ Link dữ liệu: https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataet
+ Bộ dữ liệu bao gồm nhiều thông tin khác nhau liên quan đến hành vi của khách hàng khi sử dụng các trang web mua sắm trực tuyến.
+ Có tất cả 12.330 mẫu dữ liệu có trong tập dữ liệu, không có dữ liệu bị thiếu.
+ Tính chất dữ liệu: 10 thuộc tính Numeric và 8 thuộc tính Categorical. Trong đó, thuộc tính Revenue đóng vai trò là nhãn dữ liệu dùng để phân lớp.

2. Các thuộc tính của dữ liệu

Tên thuộc tính	Ý nghĩa thuộc tính	Kiểu dữ liệu	Min Value	Max Value	Std. Dev.
Administrative	Số lượng trang khách truy cập về quản lý tài khoản	Numeric	0	27	3.32
Administrative_Duration	Tổng thời gian (tính bằng giây) mà khách truy cập đã dành để quản lý tài khoản các trang liên quan	Numeric	0	3398	176.70
Informational	Số lượng trang khách truy cập về trang web, thông tin liên lạc và địa chỉ của trang web mua sắm	Numeric	0	24	1.26
Informational_Duration	Tổng thời gian (tính bằng giây) khách truy cập trên các trang thông tin	Numeric	0	2549	140.64
ProductRelated	Số trang khách truy cập về các trang liên quan đến sản phẩm	Numeric	0	705	44.45

ProductRelated_Duration	Tổng thời gian (tính bằng giây) khách truy cập trên các trang liên quan đến sản phẩm	Numeric	0	63973	1912.25
BounceRates	Tỷ lệ trung bình khách truy cập vào trang web và sau đó rời khỏi trang mà không thực hiện bất kỳ yêu cầu nào khác đến máy chủ phân tích trong phiên đó	Numeric	0	0.2	0.04
ExitRates	Tỷ lệ phần trăm thoát khỏi trang web của người dùng	Numeric	0	0.2	0.05
PageValues	Thời gian trung bình truy cập vào trang web của người dùng trước khi hoàn tất giao dịch thương mại điện tử	Numeric	0	361	18.55

SpecialDay	Thời gian truy cập gần nhất vào trang web trong ngày các ngày lễ	Numeric	0	1.0	0.19

Bảng 1. Các thuộc tính số của dữ liệu

Bảng 1 thể hiện các thuộc tính số cùng với các tham số thống kê của từng thuộc tính. Trong đó, các thuộc tính "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" và "Product Related Duration" thể hiện số lượng các loại trang khác nhau mà khách truy cập trong phiên đó và tổng thời gian dành cho mỗi danh mục trang này. Giá trị của các thuộc tính này được lấy từ thông tin URL của các trang mà người dùng đã truy cập và được cập nhật theo thời gian thực khi người dùng thực hiện một hành động, ví dụ: chuyển từ trang này sang trang khác.

Các thuộc tính "Bounce Rate", "Exit Rate" và "Page Value" trong bảng 2.1 đại diện cho các số liệu được đo lường bởi Google Analytics cho mỗi trang trong trang web thương mại điện tử. Các giá trị này được lưu trữ trong hệ thống và được cập nhật tự động theo định kỳ.

Tên thuộc tính	**Ý nghĩa thuộc tính**	**Kiểu dữ liệu**	**Số giá trị của thuộc tính**
Month	Tháng tương ứng của ngày truy cập	Categorical	12
OperatingSystems	Hệ điều hành của khách truy cập	Categorical	8
Browser	Trình duyệt của khách truy cập	Categorical	13
Region	Khu vực địa lý nơi khách hàng truy cập vào trang web	Categorical	9
TrafficType	Mô tả số lượt người dùng truy cập và hoạt động trên website	Categorical	20

VisitorType	Loại khách truy cập là "Khách truy cập mới", "Khách truy cập quay lại" và "Khác"	Categorical	3
Weekend	Giá trị Boolean cho biết ngày của chuyến thăm có phải là cuối tuần hay không	Categorical	2
Revenue	Nhãn dữ liệu cho biết với mỗi lượt truy cập khách hàng có mua hàng hay không	Categorical	2

Bảng 2. Các thuộc tính phân loại của dữ liệu

Bảng 2 thể hiện các thuộc tính phân loại cùng với số giá trị phân loại của chúng. Các thuộc tính "Operating Systems", "Browser", "Traffic Type" và "Visitor Type" đại diện cho các số liệu được đo lường bởi Google Analytics cho mỗi trang trong trang web thương mại điện tử. Các thuộc tính "Weekend" và "Month" được lấy từ ngày khách truy cập trang web. Cuối cùng, thuộc tính "Revenue" đóng vai trò là nhãn dữ liệu cho biết với mỗi lượt truy cập có dẫn đến kết quả giao dịch hay không.

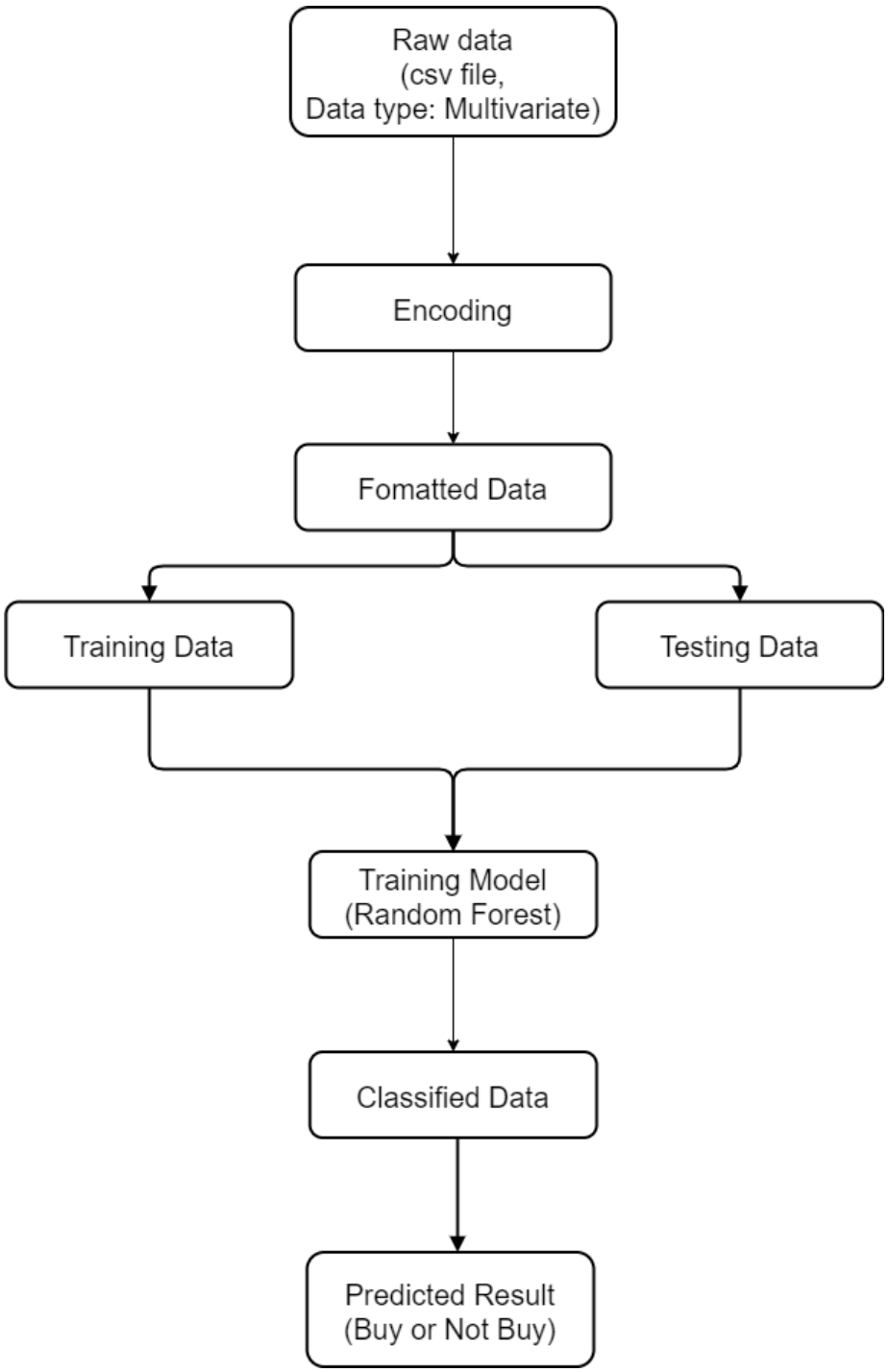
Administr	Administr	Informati	Informati	ProductRe	ProductRe	BounceRa	ExitRates	PageValu	SpecialDa	Month	Operating	Browser	Region	TrafficTyp	VisitorType	Weekend	Revenue
0	0	0	0	1	0	0,2	0,2	0	0	Feb	1	1	1	1	Returning_Visitc	FALSE	FALSE
0	0	0	0	2	64	0	0,1	0	0	Feb	2	2	1	2	Returning_Visitc	FALSE	FALSE
0	0	0	0	1	0	0,2	0,2	0	0	Feb	4	1	9	3	Returning_Visitc	FALSE	FALSE
0	0	0	0	2	2,666667	0,05	0,14	0	0	Feb	3	2	2	4	Returning_Visitc	FALSE	FALSE
0	0	0	0	10	627,5	0,02	0,05	0	0	Feb	3	3	1	4	Returning_Visitc	TRUE	FALSE
0	0	0	0	19	154,2167	0,015789	0,024561	0	0	Feb	2	2	1	3	Returning_Visitc	FALSE	FALSE
0	0	0	0	1	0	0,2	0,2	0	0,4	Feb	2	4	3	3	Returning_Visitc	FALSE	FALSE
1	0	0	0	0	0	0,2	0,2	0	0	Feb	1	2	1	5	Returning_Visitc	TRUE	FALSE
0	0	0	0	2	37	0	0,1	0	0,8	Feb	2	2	2	3	Returning_Visitc	FALSE	FALSE
0	0	0	0	3	738	0	0,022222	0	0,4	Feb	2	4	1	2	Returning_Visitc	FALSE	FALSE
0	0	0	0	3	395	0	0,066667	0	0	Feb	1	1	3	3	Returning_Visitc	FALSE	FALSE
0	0	0	0	16	407,75	0,01875	0,025833	0	0,4	Feb	1	1	4	3	Returning_Visitc	FALSE	FALSE
0	0	0	0	7	280,5	0	0,028571	0	0	Feb	1	1	1	3	Returning_Visitc	FALSE	FALSE
0	0	0	0	6	98	0	0,066667	0	0	Feb	2	5	1	3	Returning_Visitc	FALSE	FALSE
0	0	0	0	2	68	0	0,1	0	0	Feb	3	2	3	3	Returning_Visitc	FALSE	FALSE
2	53	0	0	23	1668,285	0,008333	0,016313	0	0	Feb	1	1	9	3	Returning_Visitc	FALSE	FALSE
0	0	0	0	1	0	0,2	0,2	0	0	Feb	1	1	4	3	Returning_Visitc	FALSE	FALSE
0	0	0	0	13	334,9667	0	0,007692	0	0	Feb	1	1	1	4	Returning_Visitc	TRUE	FALSE
0	0	0	0	2	32	0	0,1	0	0	Feb	2	2	1	3	Returning_Visitc	FALSE	FALSE
0	0	0	0	20	2981,167	0	0,01	0	0	Feb	2	4	4	4	Returning_Visitc	FALSE	FALSE
0	0	0	0	8	136,1667	0	0,008333	0	1	Feb	2	2	5	1	Returning_Visitc	TRUE	FALSE
0	0	0	0	2	0	0,2	0,2	0	0	Feb	3	3	1	3	Returning_Visitc	FALSE	FALSE
0	0	0	0	3	105	0	0,033333	0	0	Feb	3	2	1	5	Returning_Visitc	FALSE	FALSE

Hình 1. Minh họa bộ dữ liệu

III. Mô hình giải bài toán

Trong mô hình giải bài toán ở Hình 2, đầu tiên nhóm chúng em sẽ mã hóa dữ liệu của hai thuộc tính "Month" và "VisitorType". Sau đó dùng phương pháp Hold phân chia ngẫu nhiên tập dữ liệu ban đầu thành 2 tập dữ liệu độc lập là tập dữ liệu huấn luyện

và tập kiểm định mô hình. Mục đích nhằm kiểm tra độ hiệu quả của mô hình khi sử dụng nhiều tập dữ liệu khác nhau. Sau đó áp dụng các thuật toán phân lớp, cụ thể là thuật toán Random Forest để phân chia các mẫu vào hai lớp True, False của tập dữ liệu và xác định kết quả của bài toán.



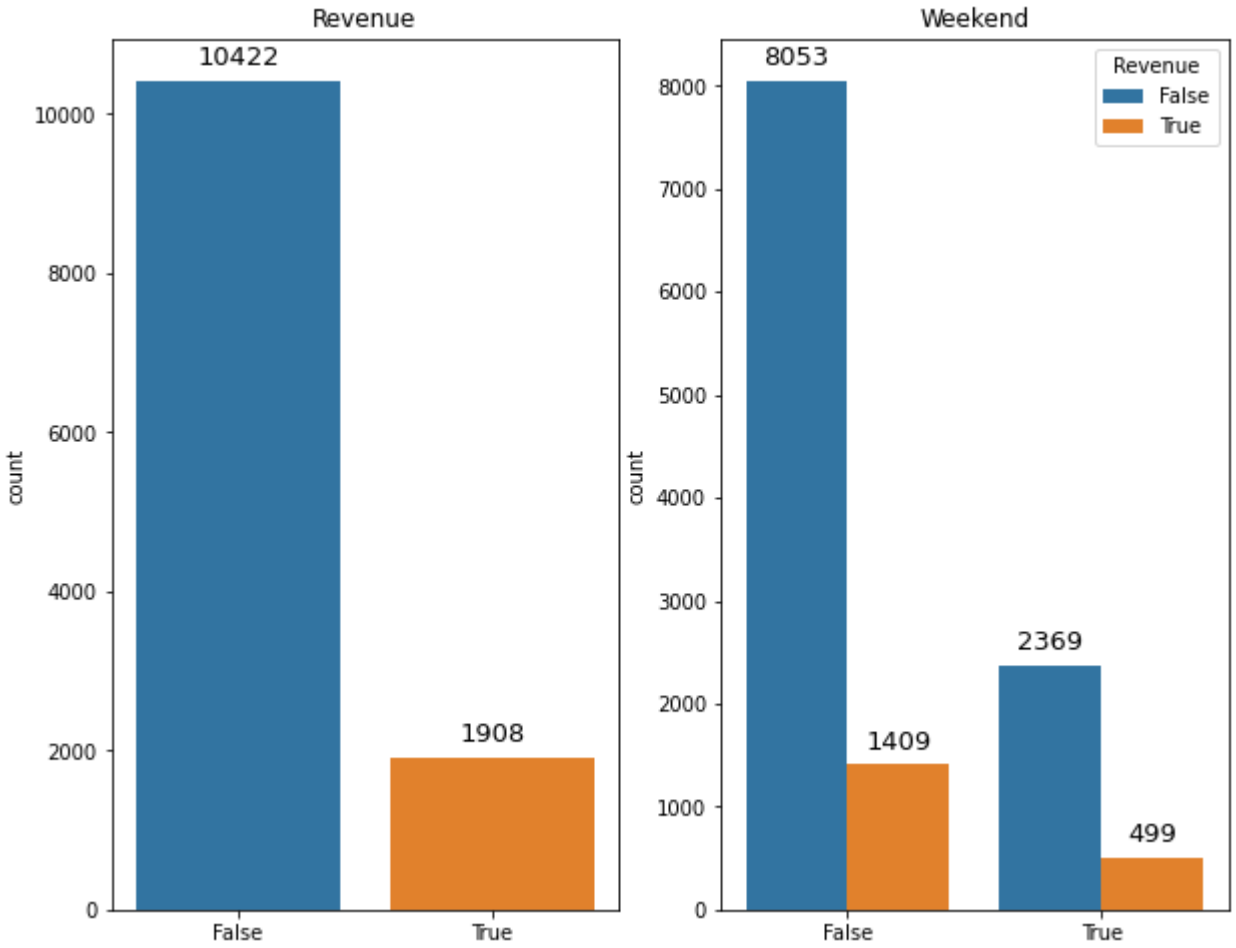
Hình 2. Framework bài toán

1. Tiền xử lý dữ liệu

Để thuận tiện cho việc phân lớp dữ liệu, chúng ta sẽ mã hóa các giá trị đặc trưng của hai thuộc tính "Month" và "VisitorType" từ dạng chuỗi thành hai giá trị 0 (False) và 1 (True) bằng phương pháp Mã hóa One Hot (One Hot Encoding).

One Hot Encoding là quá trình biến đổi từng giá trị thành các đặc trưng nhị phân chỉ chứa giá trị 1 hoặc 0. Mỗi mẫu trong đặc trưng phân loại sẽ được biến đổi thành một vecto có kích thước m chỉ với một trong các giá trị là 1 (biểu thị nó là **active**).

2. Phân tích dữ liệu

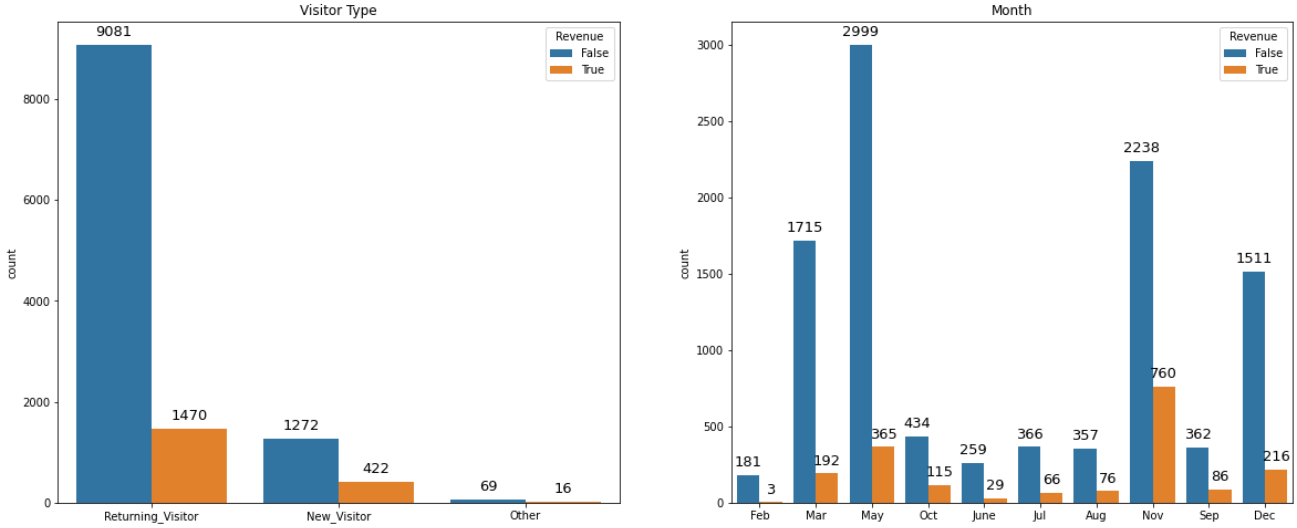


Hình 3. Biểu đồ phân chia hai lớp True - False trong Revenue và sự tương quan giữa Revenue – Weekend

Khi tiến hành kiểm tra độ tương quan của 2 cột là Revenue và cột Weekend, ta nhận thấy vào thời điểm cuối tuần, thì tỷ lệ không có doanh thu giảm 2.51% (từ 85.11% lên 82.6%). Điều đó, cho thấy cột Weekend có tác động tích cực lên cột Revenue.

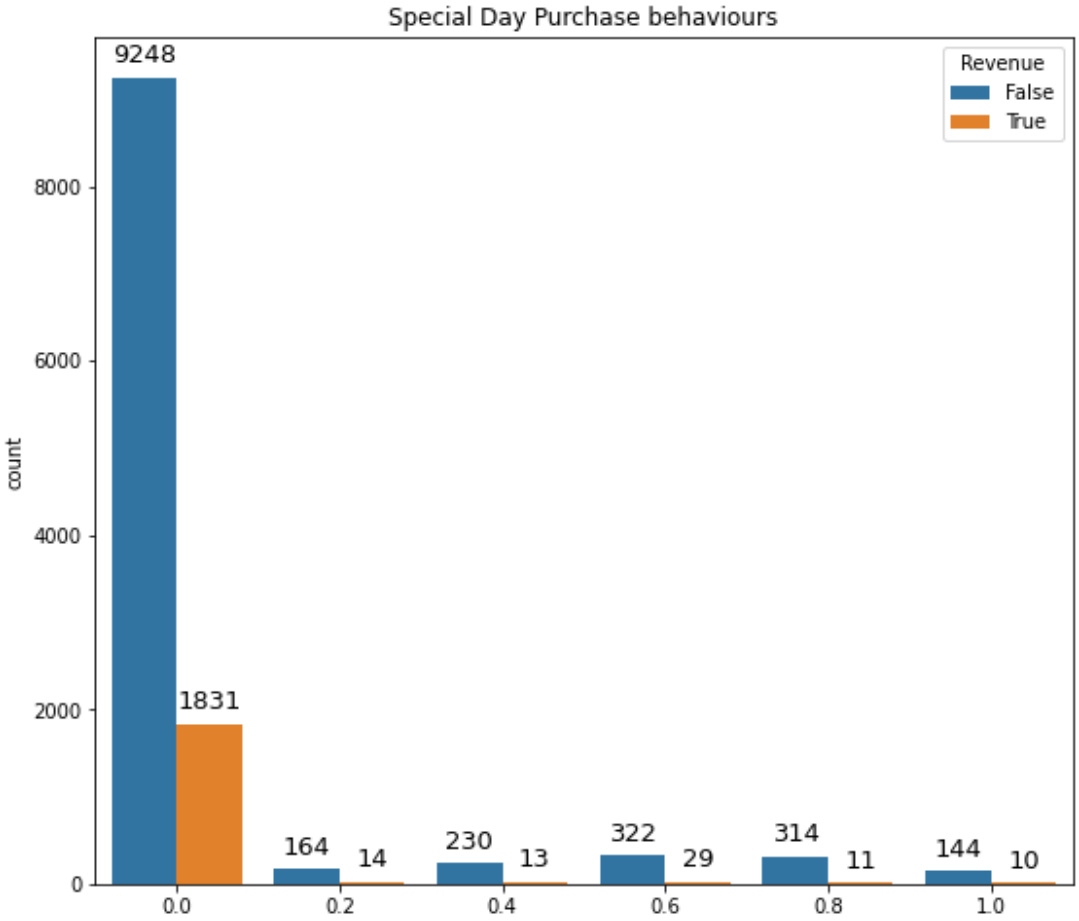
Tiếp tục sử dụng cột Visitor Type để kiểm tra độ tương quan với cột nhãn là cột Revenue. Từ biểu đồ ở bảng 4, ta thấy tỷ lệ khách hàng cũ không mang lại doanh thu chiếm 86.07%, điều này chỉ ra doanh nghiệp này không có khả năng giữ chân các khách hàng để mang lại lợi nhuận. Đồng thời, có 75.09% khách hàng mới cũng không mang lại doanh thu cho doanh nghiệp.

Kiểm tra tương quan với bảng Month, ta thu được các tháng 2 (98.37%), tháng 3 (89.91%), tháng 5 (89.15%), tháng 6 (89.93%) có tỷ lệ không mang lại doanh thu lớn hơn 85%.

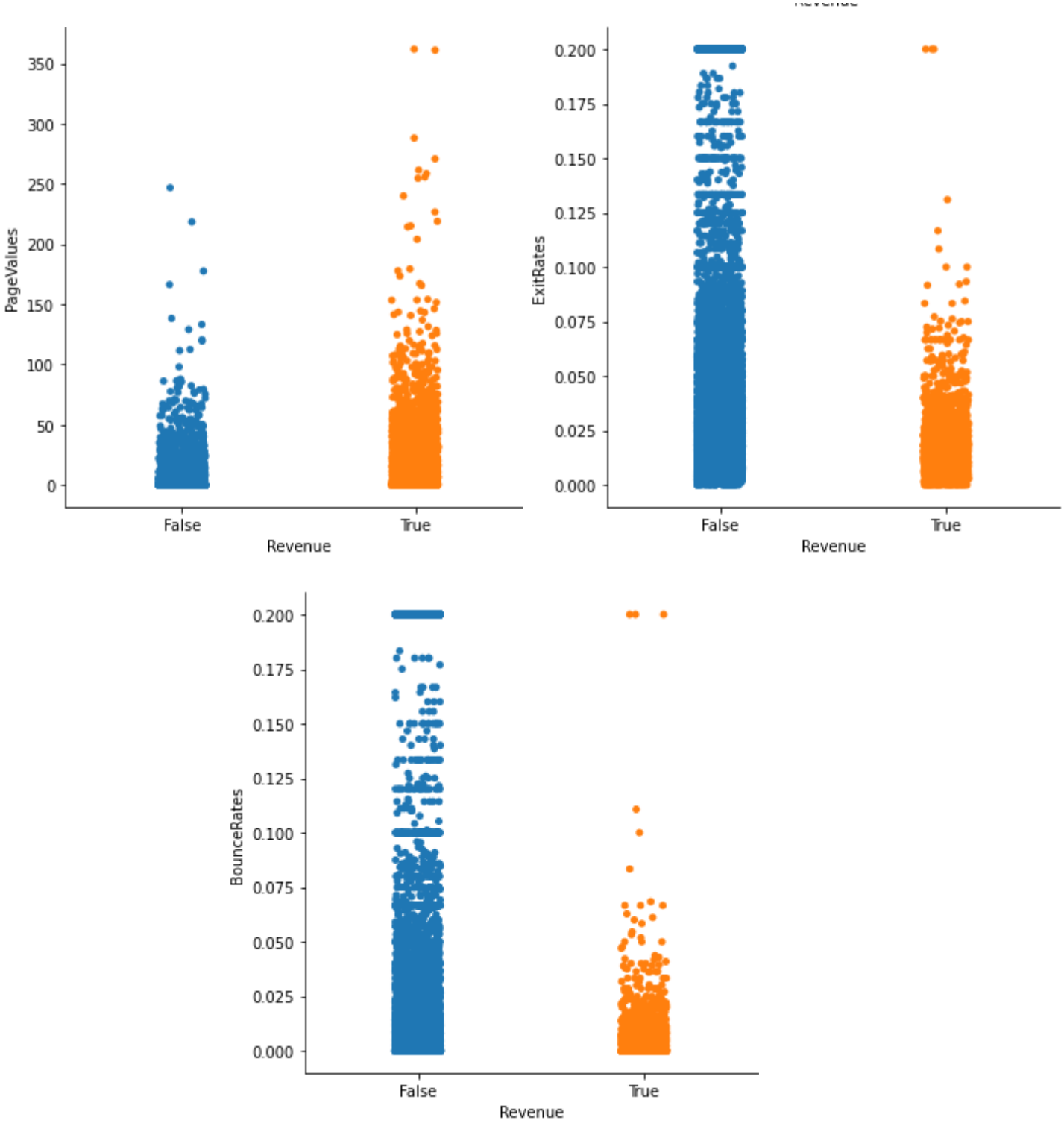


Hình 4. Biểu đồ tương quan giữa Revenue - VisitType – Month

Tiếp tục kiểm tra với bảng Special Day, tại đây ta nhận thấy vào những ngày sát với ngày lễ thì tỷ lệ không mang lại lợi nhuận chiếm cao nhất 96.61%.



Hình 5. Biểu đồ tương quan giữa Revenue – SpecialDay



Hình 6. Biểu đồ tương quan giữa Revenue và BounceRate, ExitRatse, PageValues

Cuối cùng, ta đưa cột nhãn để kiểm tra độ tương quan với 3 bảng PageValues, ExitRates. Đối với cột PageValues, ta nhận thấy nếu PageValues có giá trị thấp thì sẽ có tần suất không mang lại lợi nhuận cao hơn. Đồng thời, với cột ExitRate thì tần suất không mang lại lợi nhuận dàn trải đều, nhưng phần lớn đều tập trung ở mức ExitRate

nhỏ hơn 0.125. Còn với cột BounceRate thì tần suất không mang lại lợi nhuận lại tập trung nhiều ở mức BounceRate nhỏ hơn 0.100.

Như vậy, sau khi tiến hành tìm hiểu dữ liệu thông qua việc kiểm tra độ tương quan của cột nhãn Revenue với các bảng khác. Ta thu được kết quả là ngoài trừ thuộc tính Weekend thì các thuộc tính còn lại chưa phải một thông tin hoàn chỉnh để tiến hành phân lớp. Vì vậy, nhóm quyết định sử dụng 17 thuộc tính đã nêu trên để thực hiện phân lớp dữ liệu.

IV. Phương pháp đề xuất

Để xây dựng mô hình phân lớp dự đoán ý định mua hàng của người dùng khi truy cập các cửa hàng online, nhóm chúng em lựa chọn thực hiện mô hình phân lớp bằng thuật toán **Random Forest** và sử dụng hai phương pháp khác đó là Decision Tree và Naïve Bayes để so sánh với Random Forest.

1. Thuật toán Decision Tree

Decision tree - Cây quyết định là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Nói cách khác, cây quyết định là một danh sách tối thiểu các câu hỏi dạng yes/no mà người ta phải hỏi, để đánh giá xác suất đưa ra quyết định đúng đắn.

Thuật toán trong cây quyết định là giải thuật Iterative Dichotomiser 3 (gọi tắt là ID3). Nó biểu diễn các khái niệm *(concept)* ở dạng các cây quyết định *(Decision Tree)*. Biểu diễn này cho phép chúng ta xác định phân loại của một đối tượng bằng cách kiểm tra các giá trị của nó trên một số thuộc tính nào đó.

Giải thuật ID3 xây dựng cây quyết định được trình bày như sau:

Lặp:

+ Chọn A làm thuộc tính quyết định "tốt nhất" cho nút kế tiếp.
+ Gán A là thuộc tính quyết định cho nút.
+ Với mỗi giá trị của A, tạo nhánh con mới của nút.
+ Phân loại các mẫu huấn luyện cho các nút lá.

+ Nếu các mẫu huấn luyện được phân loại hoàn toàn thì NGƯNG. Ngược lại, lặp với các nút lá mới.

2. Thuật toán Naïve Bayes

Naïve Bayes là một thuật toán phân loại cho các vấn đề phân loại nhị phân (hai lớp) và đa lớp. Kỹ thuật này dễ hiểu nhất khi được mô tả bằng các giá trị đầu vào nhị phân hoặc phân loại.

Thuật toán Naïve Bayes sẽ tính xác suất cho các yếu tố, sau đó chọn kết quả với xác suất cao nhất. Lưu ý, giả định của thuật toán Naïve Bayes là các yếu tố đầu vào được cho là độc lập với nhau.

Công thức tổng quát của giải thuật Naïve Bayes về xác xuất xảy ra của y khi có X như sau:

$$\boldsymbol{P(y|X) = \frac{P(X|y)P(y)}{P(X)}}$$

Trong đó:

+ P(y|X) gọi là *posterior probability*: xác suất của mục tiêu y với điều kiện có đặc trưng X.
+ P(X|y) gọi là *likelihood*: xác suất của đặc trưng X khi đã biết mục tiêu y
+ P(y) gọi là xác suất trước *(prior probability)* của mục tiêu y.
+ P(X) gọi là xác suất trước *(prior probability)* của đặc trưng X.

3. Thuật toán Random Forest

3.1. Định nghĩa

Random Forest - Rừng ngẫu nhiên là một thuật toán học có giám sát. Như tên gọi của nó, Random Forest sử dụng các cây *(tree)* để làm nền tảng. Nó là một phương pháp tổng hợp (dựa trên cách tiếp cận phân chia và chinh phục) của các cây quyết định *(Decision Tree)* được tạo ra trên một tập dữ liệu được chia ngẫu nhiên. Bộ sưu tập phân loại cây quyết định này còn được gọi là rừng. Mỗi cây quyết định dự đoán một kết quả và kết quả nào được nhiều cây quyết định dự đoán thì đó là kết quả cuối cùng.

3.2. Thuật toán máy học

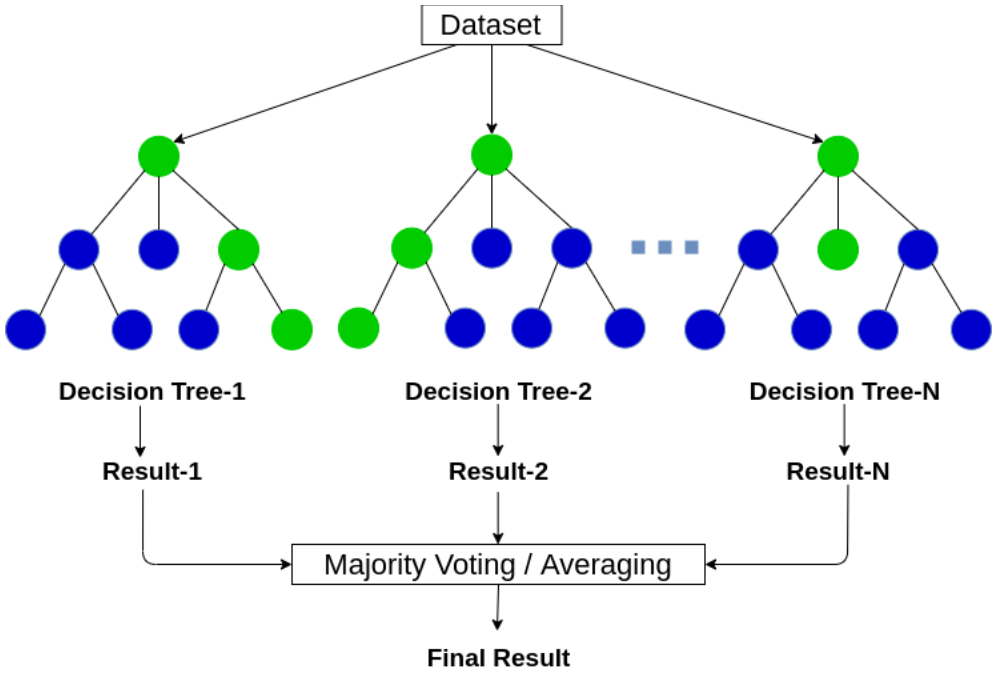
❖ **Mã giả cho hoạt động của Random Forest:**

1. Chọn ngẫu nhiên "k" features từ tập "m" features. (k << m)
2. Từ tập "k" features, tính toán ra node "d" là tốt nhất cho Node phân loại.
3. Chia các node con theo node tốt nhất vừa tìm được
4. Lặp lại bước 1-3 cho đến khi đạt đến k node
5. Lặp lại bước 1-4 để tạo ra "n" cây

❖ **Random Forest prediction:**

Để biểu diễn dự đoán sử dụng Random Forest đã huấn luyện, ta sử dụng các bước bên dưới:

1. Lấy các test features và sử dụng các Cây quyết định đã tạo ra để dự đoán kết quả, lưu nó vào một danh sách.
2. Tính toán số lượng vote trên toàn bộ Forest cho từng kết quả.
3. Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.



Hình 7. Sơ đồ thuật toán Random Forest

3.3. Đặc điểm của Random Forest

3.3.1. Ưu điểm

Thuật toán Random Forest giải quyết tốt các bài toán có nhiều dữ liệu nhiễu, thiếu giá trị. Do cách chọn ngẫu nhiên thuộc tính nên các giá trị nhiễu, thiếu ảnh hưởng không lớn đến kết quả. Khi rừng có nhiều cây hơn, chúng ta có thể tránh được việc overfitting với tập dữ liệu. Ngoài ra, thuật toán có thể được sử dụng trong cả bài toán phân loại và hồi quy. Random forests cũng có thể xử lý các giá trị còn thiếu.

3.3.2. Nhược điểm

Dữ liệu huấn luyện cần được đa dạng hóa và cân bằng về số nhãn lớp. Việc không cân bằng giữa các lớp khiến kết quả dự đoán của thuật toán có thể lệch về số nhãn lớp chiếm ưu thế. Thời gian huấn luyện của rừng có thể kéo dài tùy số cây và số thuộc tính phân chia.

3.4. Độ đo đặc trưng

Cho một tập dữ liệu huấn luyện (tập mẫu) chứa N mẫu dữ liệu, M thuộc tính X_j (j=1,2, ..., M) và $Y \in \{1, 2,.., C\}$ với $C \geq 2$ là biến phụ thuộc. Thuật toán Random Forest dùng chỉ số Gini để đo tính hỗn tạp của tập mẫu. Trong quá trình xây dựng các cây quyết định, Random Forest phát triển các nút con từ một nút cha dựa trên việc đánh giá chỉ số Gini của một không gian con mtry các thuộc tính được chọn ngẫu nhiên từ không gian thuộc tính ban đầu. Thuộc tính được chọn để tách nút t là thuộc tính làm cực tiểu độ hỗn tạp của các tập mẫu sau khi chia. Công thức tính chỉ số Gini cho nút t như sau:

$$\mathbf{Gini(t)} = \sum_{\mathbf{c-1}}^{\mathbf{c}} \mathbf{\Phi_c(t)[1 - \Phi_c(t)]}$$

trong đó $\Phi_c(t)$ là tần suất xuất hiện của lớp $c \in C$ trong nút t.

V. Cài đặt thực nghiệm

1. Chia dữ liệu thực nghiệm

Để tiến hành thực nghiệm, ta tiến hành chia bộ dữ liệu ban đầu thành 2 tập độc lập: tập huấn luyện *(training set)* chiếm 70%, tập thử nghiệm *(testing set)* chiếm 30%. Ta có bảng mô tả dữ liệu như sau:

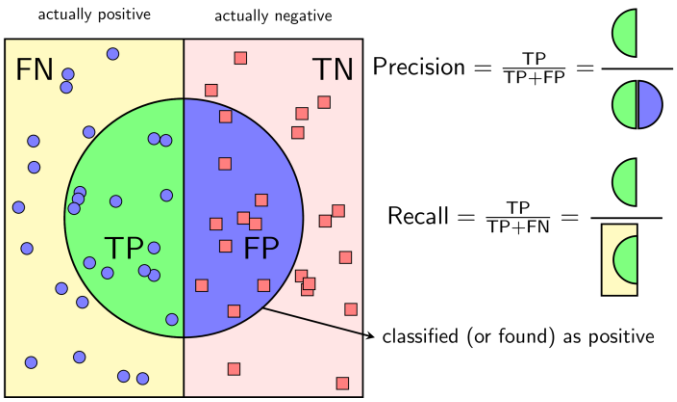
Tên dữ liệu: Online Shopper Purchasing Intention			
Số thuộc tính: 18 (10 thuộc tính Numeric và 8 thuộc tính Categorical)			
Nhãn dữ liệu: Revenue			
	Tổng số mẫu	**Tỷ lệ % các lớp**	
		False *(không mua hàng)*	**True** *(đã mua hàng)*
Tập dữ liệu ban đầu	12.330	84.5%	15.5%
Tập dữ liệu huấn luyện *(Training set)*	8.631	82%	18%
Tập dữ liệu thử nghiệm *(Testing set)*	3.699	90%	10%

Bảng 3. Tỷ lệ % hai lớp False – True trong tập dữ liệu ban đầu và tập dữ liệu thực nghiệm

2. Phương pháp đánh giá

2.1. Precision và Recall

Trước hết xét bài toán phân loại nhị phân. Ta cũng coi một trong hai lớp là *Positive*, lớp còn lại là *Negative*. Xét Hình 8 dưới đây:



Hình 8. Cách tính Precision và Recall

Khi đó, **Precision** được định nghĩa là tỉ lệ *số điểm Positive mô hình dự đoán đúng* trên *tổng số điểm mô hình dự đoán là Positive*. **Recall** được định nghĩa là tỉ lệ *số điểm Positive mô hình dự đoán đúng* trên *tổng số điểm thật sự là Positive* (hay *tổng số điểm được gán nhãn là Positive ban đầu*).

Precision và Recall theo công thức:

$$\boldsymbol{Precision} = \frac{\boldsymbol{TruePositive}}{\boldsymbol{TruePositive} + \boldsymbol{FalsePositive}}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TruePositive}}{\boldsymbol{TruePositive} + \boldsymbol{FalsePositive}}$$

Precision càng cao, tức là số điểm mô hình dự đoán là positive đều là positive càng nhiều. Precision = 1, tức là *tất cả số điểm mô hình dự doán là Positive đều đúng*, hay *không có điểm nào có nhãn là Negative mà mô hình dự đoán nhầm là Positive*.

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

2.2. F1 Score

F1-score là trung bình điều hòa *(harmonic mean)* của Precision và Recall (giả sử hai đại lượng này khác 0). F1-score được tinh theo công thức:

$$\boldsymbol{F1} = \boldsymbol{2}\frac{\boldsymbol{Precision} * \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}}$$

Căn cứ vào F1 ta sẽ chọn model, F1 càng cao thì càng tốt. Khi lý tưởng nhất thì F1 = 1 (khi Recall = Precision=1).

2.3. Accuracy

Khả năng mô hình phân loại dự báo chính xác, phân loại chính xác, hay xác định đúng class (nhóm, loại) cho dữ liệu cần phân loại. Khi ai đó nói rằng các phép đo là chính xác cao *(High Accuracy)* thì các phép đo đó rất gần với giá trị đích. Sự phân tán của các phép đo chính xác cao có thể tạo ra nhóm các kết quả đo xa nhau, hoặc dày đặc gần nhau.

3. Phương pháp thực nghiệm

+ Decision Tree
+ Naïve Bayes
+ Random Forest

4. Kết quả thực nghiệm

Trong bài toán này, việc khách hàng không thực hiện bất kỳ giao dịch nào trước khi rời khỏi trang web sẽ làm ảnh hưởng đến lợi nhuận hoặc doanh thu bị mất của các nhà bán lẻ trực tuyến. Vì vậy, chúng ta sẽ chỉ quan tâm đến lớp False của trong nhãn dữ liệu Revenue để xác định kết quả của mô hình.

Sau khi chia dữ liệu thực nghiệm, ta tiến hành phân lớp dữ liệu với 3 phương pháp lần lượt là Decision Tree, Naïve Bayes và Random Forest. Ta có kết quả thực nghiệm như sau:

4.1. Decision Tree

	Precision	Recall	F1-score	Accuracy
False	0.91	0.92	0.91	0.86

Bảng 4. Kết quả các độ đo của phương pháp Decision Tree

4.2. Naïve Bayes

	Precision	Recall	F1-score	Accuracy
False	0.91	0.85	0.88	0.81

Bảng 5. Kết quả các độ đo của phương pháp Naïve Bayes

4.3. Random Forest

	Precision	Recall	F1-score	Accuracy
False	0.92	0.96	0.94	0.90

Bảng 6. Kết quả các độ đo của phương pháp Random Forest

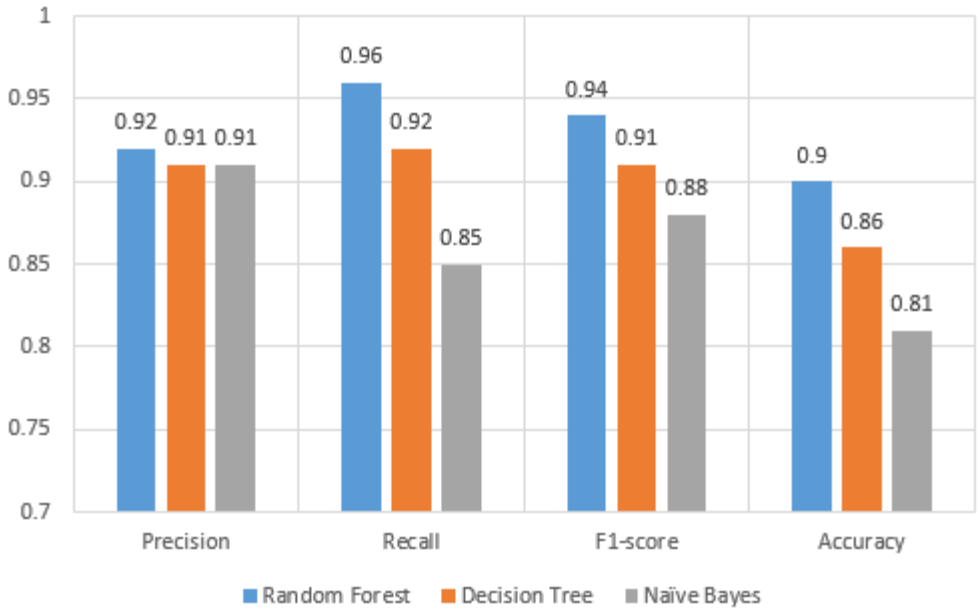
5. Đánh giá mô hình

Sau khi thu được bảng kết quả chạy thực nghiệm từ cả 3 phương pháp phân lớp. Ta có bảng tổng kết, đánh giá kết quả chạy thực nghiệm phương pháp Naïve Bayes, Random Forest và Decision Tree được phản ánh qua bảng sau:

	Precision	Recall	F1-score	Accuracy
Random Forest	0.92	0.96	0.94	0.90
Decision Tree	0.91	0.92	0.91	0.86
Naïve Bayes	0.91	0.85	0.88	0.81

Bảng 7. Kết quả các độ đo của 3 phương pháp

So sánh kết quả, ta nhận thấy phương pháp phân lớp Random Forest cho ra kết quả chia dữ liệu có độ đo Precision cao hơn hai phương pháp phân lớp Decision Tree và Naïve Bayes là 1%; độ đo Recall cao hơn phương pháp Decision Tree và Naïve Bayes lần lượt là 4% và 11%; ở độ đo F1 – score lần lượt là 3% và 6%; còn ở độ đo Accuracy lần lượt là 4% và 9%. Do đó, sau khi chạy thực nghiệm dữ liệu, nhóm đánh giá phương pháp phân lớp Random Forest có độ tin cậy cao và nhóm đề xuất sử dụng phương pháp này để thực hiện phân lớp đối với bộ dữ liệu trên.

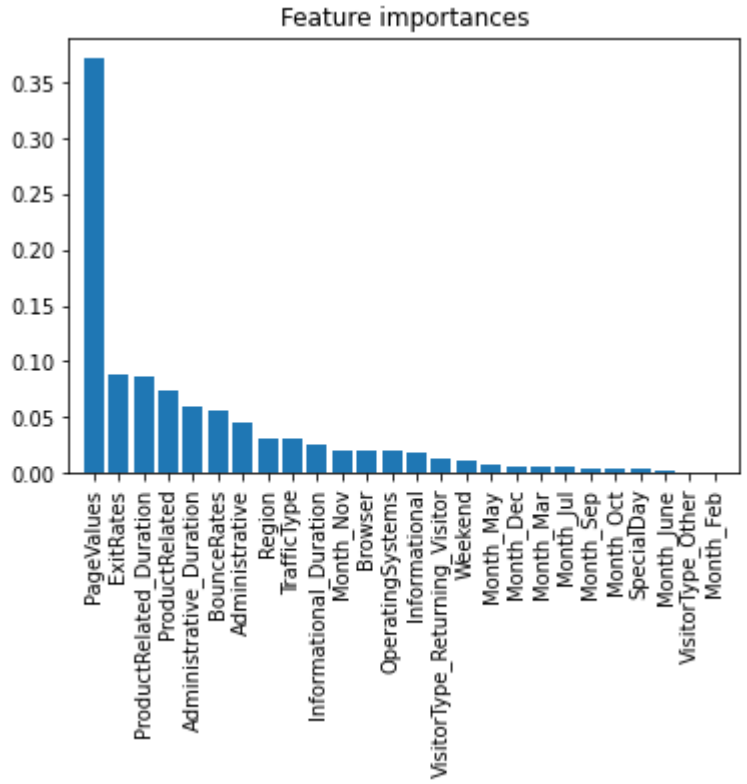


Hình 9. Biểu đồ so sánh các độ đo của 3 phương pháp

6. Xếp hạng các thuộc tính

Trong khai phá dữ liệu, phương pháp Random Forest thường dùng để xác định độ quan trọng của các thuộc tính trong tập dữ liệu. Các thuộc tính cao nhất trong cây (gần node gốc nhất) có thể được coi là quan trọng nhất. Việc lựa chọn thuộc tính có thể được thực hiện bằng cách lấy trung bình mức độ quan trọng trên tất cả các cây.

Để xếp hạng các thuộc tính, nhóm chúng em gọi "RandomForestClassifier ()" trong thư viện lớp Random Forest scikit-learning để lọc và sắp xếp các thuộc tính. Kết quả được hiển thị trong Hình 10.



Hình 10. Kết quả xếp hạng các thuộc tính từ phương pháp Random Forest

Ta có bảng sau:

Xếp hạng	Thuộc tính	Số đo độ quan trọng	Xếp hạng	Thuộc tính	Số đo độ quan trọng
1	PageValues	0.371793	14	Informational	0.018133
2	ExitRates	0.087309	15	VisitorType_Returning_Visitor	0.012881
3	ProductRelated_Duration	0.085934	16	Weekend	0.010316
4	ProductRelated	0.074531	17	Month_May	0.006726
5	Administrative_Duration	0.058947	18	Month_Dec	0.005901
6	BounceRates	0.055089	19	Month_Mar	0.004748
7	Administrative	0.044110	20	Month_Jul	0.004657
8	Region	0.030209	21	Month_Sep	0.003708
9	TrafficType	0.029920	22	Month_Oct	0.003704
10	Informational_Duration	0.025649	23	SpecialDay	0.003435
11	Month_Nov	0.020408	24	Month_June	0.002382
12	Browser	0.019479	25	VisitorType_Other	0.000551
13	OperatingSystems	0.019275	26	Month_Feb	0.000206

Bảng 8. Kết quả xếp hạng và số đo độ quan trọng của các thuộc tính

Từ Bảng 8 chúng ta có thể thấy rằng thuộc tính "PageValues" là thuộc tính quan trọng nhất trong tập dữ liệu. Các thứ hạng tiếp theo là những thuộc tính "ExitRates", "ProductRelated_Duration", "Product_Related", "Administrative_Duration".

VI. Kết luận và hướng phát triển

1. Kết luận

Trong báo cáo này, nhóm chúng em đã nghiên cứu hiệu suất của các thuật toán máy học có giám sát dựa trên dữ liệu thực nghiệm của người mua sắm trực tuyến. Mục tiêu của đề tài là xác định một mô hình thích hợp có thể dự đoán ý định mua của khách hàng khi truy cập trang web của một số cửa hàng trực tuyến được chính xác hơn. Nhóm chúng em đã sử dụng 3 phương pháp phân lớp khác nhau để giải quyết vấn đề này, cụ thể là Naïve Bayes, Decision Tree và Random Forest. Dựa trên kết quả thực nghiệm và so sánh, nhóm chúng em nhận thấy phương pháp Random Forest có các độ đo đánh giá và độ chính xác cao hơn hai thuật toán còn lại.

Vì vậy, ta kết luận rằng việc xây dựng mô hình phân lớp dựa trên thuật toán Random Forest phù hợp với các yêu cầu của bài toán và đem lại hiệu quả cao trong việc dự đoán ý định mua hàng của người tiêu dùng trực tuyến. Thông qua đó, các nhà bán lẻ trực tuyến có thể cải thiện tỷ lệ khách hàng quay lại và làm tăng lợi nhuận cho hoạt động kinh doanh trong lĩnh vực thương mại điện tử.

2. Hướng phát triển

Đối với những bài toán phân lớp, số lượng các thuộc tính thường rất lớn nhưng các thuộc tính không liên quan hoặc thừa có thể có những ảnh hưởng tiêu cực đối với các giải thuật phân lớp. Các thuộc tính, dữ liệu thừa hoặc không liên quan có thể là nguyên nhân dẫn đến việc học của giải thuật không được chính xác. Thêm vào đó, với sự có mặt của dữ liệu thừa hoặc dữ liệu không liên quan có thể làm cho bộ phân lớp trở lên phức tạp hơn. Điều này sẽ gây ra những khó khăn không cần thiết cho chúng ta trong việc diễn giải các kết quả học được từ tập huấn luyện. Do đó chúng ta cần giải quyết vấn đề này đối với các bài toán phân lớp.

Để cải thiện mô hình dự đoán ý định mua hàng của người tiêu dùng trực tuyến trong tương lai, nhóm chúng em đề xuất giải bài toán này bằng việc thu gọn kích thước dữ liệu thông qua những thuộc tính quan trọng đã được xếp hạng. Từ Bảng 8 – *Kết quả xếp hạng và số đo độ quan trọng của các thuộc tính*, nhóm chúng em lựa chọn ra 6 thuộc tính ảnh hưởng nhất đến việc phân lớp là “PageValues”, “ExitRates”,

“ProductRelated_Duration”, “ProductRelated”, “BounceRates”, “Administrative_Duration” nhằm đơn giản hóa mô hình, giảm thời gian chạy thuật toán và nâng cao độ chính xác của kết quả.

Bên cạnh đó, các mô hình máy học trong đề tài này được huấn luyện thông qua dữ liệu từ cùng một nguồn (chỉ trong một năm). Điều này có nghĩa là dữ liệu được sử dụng nhất quán nhưng ảnh hưởng đến tính tổng quát hóa của các mô hình. Vì vậy, nhóm chúng sẽ xem xét việc thu thập thêm dữ liệu hành vi của khách hàng trong vài năm tiếp theo nhằm tạo ra sự vượt trội cho dữ liệu.

TÀI LIỆU THAM KHẢO

[1] COUHP, "Random forest, thế nào là một rừng ngẫu nhiên," 01/24/2018. [Online]. Available: https://couhpcode.wordpress.com/2018/01/24/random-forest-the-nao-la-mot-rung-ngau-nhien/. [Accessed 15/12/2020].

[2] C. Okan Sakar, S. Olcay Polat, Mete Katircioglu & Yomi Kastro , "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," *Neural Computing and Applications volume,* p. 6893–6908(2019), 09 May 2018.

[3] Bamshad Mobasher, Honghua Dai, Tao Luo & Miki Nakagawa, "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization," *Data Mining and Knowledge Discovery,* p. 61–82(2002), January 2002.

[4] W. W. Moe, "Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream," *Journal of Consumer Psychology,* pp. 29-39, March 2003.

[5] G. Suchacka, M. Skolimowska-Kulig, A. Potempa, "Classification Of E-Customer Sessions Based On Support Vector Machine," ECMS 2015, p. 594–600 (2015).

[6] G. Suchacka, M. Skolimowska-Kulig, A. Potempa, "A k-Nearest Neighbors Method for Classifying User Sessions in E-Commerce Scenario," *Journal of telecommunications and information technology,* pp. 64-69, 2015.

[7] Grażyna Suchacka & Grzegorz Chodak , "Using association rules to assess purchase probability in online stores," *Information Systems and e-Business Management,* p. 751–780(2017).

[8] Karim Baaticorresponding author18 and Mouad Mohsil19, "Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest," *Artificial Intelligence Applications and Innovations,* p. 309–322(2015).

[9] Karim Baati, Tarek M Hamdani, Adel M Alimi, Ajith Abraham, "A new classifier for categorical data based on a possibilistic estimation and a novel

generalized minimum-based algorithm," *Journal of Intelligent & Fuzzy Systems*, pp. 1723-1731, 2017.

[10] Karim BaatiEmail authorTarek M. HamdaniAdel M. AlimiAjith Abraham, "A Modified Naïve Bayes Style Possibilistic Classifier for the Diagnosis of Lymphatic Diseases," in *Advances in Intelligent Systems and Computing*, 2017, pp. 479-488.

[11] Karim BaatiEmail authorTarek M. HamdaniAdel M. AlimiAjith Abraham, "A Modified Naïve Possibilistic Classifier for Numerical Data," in *Advances in Intelligent Systems and Computing*, 2017, pp. 417-426.

[12] R. F. Teixeira, "Using clickstream data to analyze online purchase intentions," 2015.

[13] W. L. YEUNG, "A review of data mining techniques for research in online shopping behaviour through frequent navigation paths," 2016. [Online]. Available: https://commons.ln.edu.hk/hkibswp/76/. [Accessed 2020].

[14] B. Clifton, Advanced Web Metrics with Google Analytics, 3rd Edition, 2012.

[15] L. Breiman, "Random Forests," *Machine Learning*, p. 5–32(2001).

[16] Karim BaatiEmail authorTarek M. HamdaniAdel M. AlimiAjith Abraham, "ecision Quality Enhancement in Minimum-Based Possibilistic Classification for Numerical Data," in *Advances in Intelligent Systems and Computing* , 2017, pp. 634-643.

[17] Chris Rygielski, J. Wang, D. C. Yen, "Data mining techniques for customer relationship management," pp. 483-502, 2002.

[18] Ramón Díaz-Uriarte & Sara Alvarez de Andrés , "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, pp. 3-20, 2006.

[19] F. Müller, "Will they Buy or just Browse? Predicting Purchase Intentions of Online Shoppers with Python," [Online]. Available: https://www.relataly.com/will-they-buy-or-just-visit-predicting-the-purchase-intention-of-online-shoppers/982/. [Accessed 12/12/2020].

[20] H. Sue. [Online]. Available: https://www.kaggle.com/henrysue/classifying-online-shopper-intention. [Accessed 15/12/2020].

[21] "Phương pháp đánh giá mô hình phân loại (classification model evalutation)," [Online]. Available: https://bigdatauni.com/vi/tin-tuc/phuong-phap-danh-gia-mo-hinh-phan-loai-classification-model-evalutation.html. [Accessed 12/01/2020].