ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH VÀ DỰ ĐOÁN NHU CẦU KHÁCH HÀNG SỬ DỤNG LẠI DỊCH VỤ TỪ ỨNG DỤNG ĐẶT ĐỒ ĂN TRỰC TUYẾN

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Bùi Nguyên Phương Linh	20521527
2	Nguyễn Thị Phương Thảo	20521936

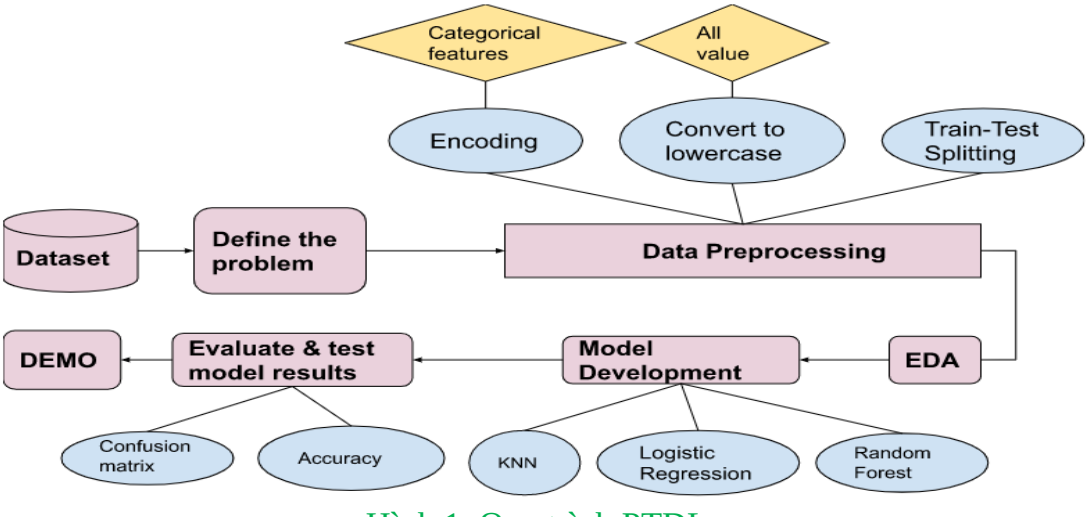
TP. HỒ CHÍ MINH – 12/2022

1. GIỚI THIỆU

Với sự bùng nổ của công nghệ 4.0 hiện nay, xu hướng đặt hàng, mua sắm trực tuyến diễn ra ngày càng phổ biến. Đặc biệt từ sau đại dịch COVID-19, dịch vụ giao đồ ăn trực tuyến đã nổi lên như một ngành công nghiệp toàn cầu, dẫn đến một "cuộc cách mạng giao hàng". Với ưu điểm là nhanh gọn, tiện lợi, người dùng chỉ cần một vài thao tác cơ bản trên ứng dụng thông minh mà không cần ra khỏi nhà những vẫn có thể dễ dàng mua được những món ăn mình muốn với mức giá không quá chênh lệch, thậm chí là rẻ hơn nhiều so với khi mua hàng tận nơi. Và để có thể nắm bắt tốt nhất nhu cầu cũng như thị yếu khách hàng, ứng dụng của máy học trở nên quan trọng và cần thiết hơn cả trong việc xây dựng các mô hình và đưa ra chiến lược phát triển cho doanh nghiệp kinh doanh bằng việc phân tích và dự đoán xem liệu khách hàng có nhu cầu sử dụng lại dịch vụ đặt đồ ăn trên ứng dụng trực tuyến hay không. Nghiên cứu này sử dụng bộ dữ liệu được thu thập từ 388 khách hàng sử dụng dịch vụ đặt đồ ăn trực tuyến tại Bangalore, Ấn Độ và ba mô hình máy học bao gồm Random Forest, Logistic Regression, K-Nearest Neighbors được xây dựng để phân loại và dự đoán nhu cầu khách hàng. Kết quả thu được cho thấy hầu hết các mô hình đều hoạt động cho ra kết quả phân loại và dự đoán có độ chính xác cao nhưng mô hình Random Forest vượt trội hơn tất cả với độ chính xác 96.15%.

Từ khóa: Random Forest, Logistic Regression, K-Nearest Neighbors.

2. NỘI DUNG



Hình 1. Quy trình PTDL

2.1. Khảo sát bộ dữ liệu

2.1.1. Giới thiệu bộ dữ liệu

Bộ dữ liệu sử dụng trong nghiên cứu có tên onlinedeliverydata, bao gồm những thông tin ảnh hưởng đến quyết định mua hàng của người dùng như nhân khẩu học, quyết định sử dụng dịch vụ đặt hàng, thời điểm giao hàng và xếp hạng những cửa hàng có ảnh hưởng. Tất cả dữ liệu đến từ trang web:

https://www.kaggle.com/datasets/benroshan/online-food-delivery-preferencesbangalore-region

Bộ dữ liệu onlinedeliverydata.csv gồm 55 thuộc tính và 388 dòng dữ liệu. Chúng tôi đã tiến hành chia bộ dữ liệu thành 2 tập là tập huấn luyện (train set) và tập kiểm thử (test set) với tỉ lệ lần lượt là 8:2.

2.1.2. Chi tiết bộ nhãn

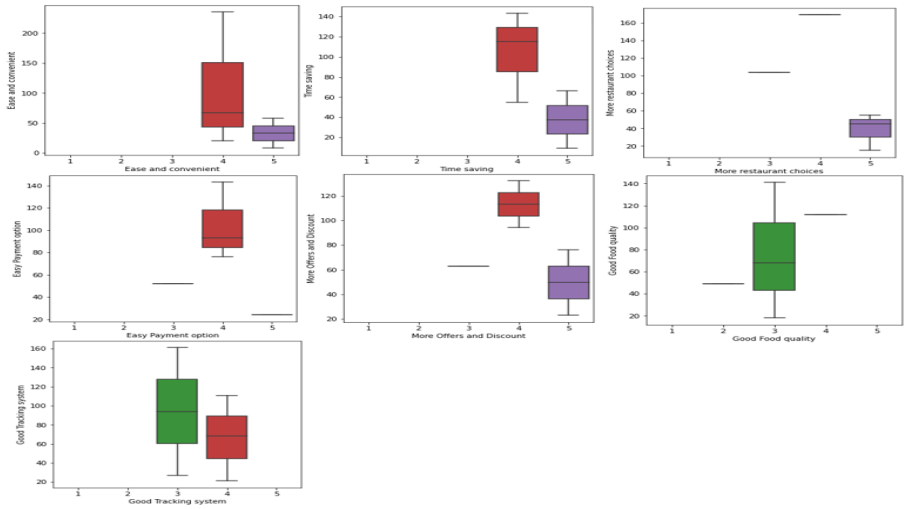
Dữ liệu bao gồm 55 thuộc tính nhưng để phục vụ cho việc chia dữ liệu, chúng tôi chỉ tập trung vào 40 thuộc tính được chia thành các biến như sau:

- Biến liên tục (Continuous variables):
  - + Age: Tuổi khách hàng (18-33).
  - + Family size: Số lượng thành viên trong gia đình (1-6).
- Biến phân loại (Categorical variables):
  - + Monthly Income: Thu nhập hằng tháng của khách hàng (0/ <10K/ 10K-25K/ 25K-50K/ >50K).
  - + Gender: Giới tính (Male/ Female).
  - + Marital Status: Tình trạng hôn nhân (Single/ Married/ Prefer not to say).
  - + Occupation: Nghề nghiệp (Student/ Employee/ Self Employeed/ House wife).
  - + Education Qualifications: Trình độ học vấn (Uneducated/ School/ Post Graduate/ Graduate/ Ph. D).
  - + Medium (P1): Hình thức đặt hàng (Food delivery apps/ Walk-in/ Direct call/ Web browser).
  - + Medium (P2): Hình thức đặt hàng (Web browser/ Direct call/ Walk-in).
  - + Meal(P1): Loại bữa ăn (Breakfast/ Snacks/ Lunch/ Dinner).
  - + Meal(P2): Loại bữa ăn (Lunch/ Dinner/ Snacks).
  - + Perference(P1): Loại thức ăn (Non veg foods (Lunch/ Dinner)/ Veg foods (Breakfast/ Lunch/ Dinner)/ Bakery items(snacks)/ Sweets).
  - + Perference(P2): Loại thức ăn (Bakery items/ Veg foods (Breakfast/ Lunch/ Dinner)/ Ice ream/ Cool drinks/ Sweets).
- Các biến ảnh hưởng đến việc khách hàng có nhu cầu đặt hàng lại đồ ăn như ‘Ease and convenient’, ‘Time saving’, ‘More restaurant choices, ‘Easy Payment option’, ‘More offers and Discount’, ‘Good Food quality’, ‘Good

Tracking system' với các nhãn tương ứng với quan điểm của khách hàng là 1-strongly disagree, 2-disagree, 3- neutral, 4-agree, 5-strongly agree.

- Các biến ảnh hưởng đến việc khách hàng hủy đơn: 'Long delivery time', 'Delay of delivery person getting assigned', 'Delay of delivery person picking up food', 'Wrong order delivered', 'Missing item', 'Order placed by mistake' với các nhãn tương ứng với quan điểm của khách hàng là 1-strongly disagree, 2-disagree, 3- neutral, 4-agree, 5-strongly agree.
- Các biến ảnh hưởng đến chất lượng thức ăn: 'High Quality of package', 'Freshnes', 'Temperature', 'Good Taste', 'Good Quantity' với các nhãn tương ứng với quan điểm của khách hàng là 1-strongly disagree, 2-disagree, 3- neutral, 4-agree, 5-strongly agree.
- Các biến liên quan đến việc giao hàng, vận chuyển: ' Influence of time', 'Order Time', 'Maximum wait time', 'Residence in busy location', 'Google Maps Accuracy', 'Good Road Condition', 'Low quantity low time', 'Delivery person ability', 'Less Delivery time', 'Number of calls', 'Politeness'.

2.1.3. Phân tích thăm dò dữ liệu (EDA)



Hình 2. Trực quan boxplot những thuộc tính ảnh hưởng đến quyết định sử dụng lại dịch vụ của khách hàng

Bộ dữ liệu có hơn 40 thuộc tính, do đó, sau khi tiến hành EDA, nhóm chỉ tập trung phân tích và đưa ra nhận xét về những thuộc tính có mức độ ảnh hưởng cao đối với nhu

cầu đặt lại đồ ăn của khách hàng. Với các nhãn tương ứng 1-strongly disagree, 2-disagree, 3-neutral, 4-agree, 5-strongly agree, các biến có sự ảnh hưởng cụ thể như sau.

- Ease and convenient:
  + Nhãn agree- phân bố từ khoảng 24 (giá trị nhỏ nhất) đến 280 (giá trị lớn nhất). Có mức trung vị ở khoảng 60 và độ phân tán rộng (xấp xỉ IQR= 150-35= 115). Dữ liệu có xu hướng nghiêng về phía dưới.
  + Nhãn strongly agree- phân bố từ khoảng 15 (giá trị nhỏ nhất) đến 50 (giá trị lớn nhất). Có mức trung vị ở khoảng 30 nhỏ hơn mức độ agree và độ phân tán bé hơn so với mức agree (xấp xỉ IQR= 40-20= 20). Dữ liệu có xu hướng đối xứng. Nhìn chung mức độ strongly agree còn khá thấp, phân bố rộng và có sự chênh lệch thấp hơn mức độ agree.
- Time saving:
  + Nhãn agree- phân bố từ khoảng 58 (giá trị nhỏ nhất) đến 145 (giá trị lớn nhất). Có mức trung vị ở khoảng 118 và độ phân phân tán rộng (xấp xỉ IQR= 128-85= 43). Dữ liệu có xu hướng nghiêng về phía trên. Nhìn chung mức độ agree khá cao, phân bố rộng.
  + Nhãn strongly agree- phân bố từ khoảng 8 (giá trị nhỏ nhất) đến 65 (giá trị lớn nhất). Có mức trung vị ở khoảng 38 nhỏ hơn mức độ agree và độ phân tán bé hơn so với mức agree (xấp xỉ IQR= 50-20= 30). Dữ liệu có xu hướng đối xứng. Nhìn chung mức độ strongly agree và sự chênh lệch còn thấp so với mức độ agree.
- More restaurant choices:
  + Nhãn neutral- nằm ở mức 100 và không có sự phân tán dữ liệu.
  + Nhãn agree- nằm ở mức 170 và không có sự phân tán dữ liệu.
  + Nhãn strongly agree- phân bố từ khoảng 10 (giá trị nhỏ nhất) đến 55 (giá trị lớn nhất). Có mức trung vị ở khoảng 40 và có độ phân tán xấp xỉ IQR= 50-30= 20. Dữ liệu có xu hướng nghiêng về phía trên. Mức độ strongly agree còn khá thấp so với mức agree, phân bố hẹp.
- Easy Payment option:

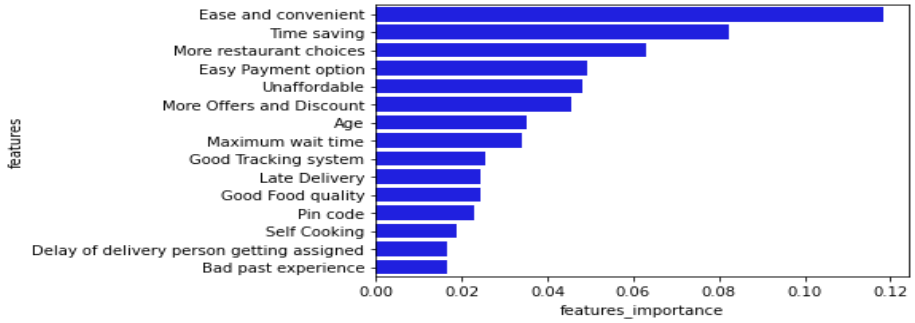
+ Nhãn neutral- nằm ở mức 55 và không có sự phân tán dữ liệu.
+ Nhãn agree- phân bố từ khoảng 77 (giá trị nhỏ nhất) đến 145 (giá trị lớn nhất). Có mức trung vị ở khoảng 90 và có độ phân tán xấp xỉ IQR= 120-82= 38. Mặc dù dữ liệu có xu hướng nghiêng về phía dưới nhưng mức độ agree khá cao.
+ Nhãn strongly agree- nằm ở mức 25 và không có sự phân tán dữ liệu.

– More Offers and Discount:

+ Nhãn neutral- nằm ở mức 62 và không có sự phân tán dữ liệu.
+ Nhãn agree- phân bố từ khoảng 95 (giá trị nhỏ nhất) đến 135 (giá trị lớn nhất). Có mức trung vị ở khoảng 110 và có độ phân phân tán xấp xỉ IQR= 122-100= 22. Dữ liệu có xu hướng đối xứng. Nhìn chung mức độ agree khá cao.
+ Nhãn strongly agree- phân bố từ khoảng 22 (giá trị nhỏ nhất) đến 78 (giá trị lớn nhất). Có mức trung vị ở khoảng 50 nhỏ hơn mức độ agree và độ phân tán lớn hơn so với mức agree (xấp xỉ IQR= 62-35= 27). Dữ liệu có xu hướng đối xứng. Nhìn chung mức độ strongly agree thấp và có sự chênh lệch cao so với mức độ agree.

– Good Food quality: Mặc dù có 3 mức độ được thể hiện là disagree, neutral và agree. Tuy nhiên mức độ disagree và agree không có sự phân tán dữ liệu. Nhãn neutral có mức trung vị khoảng 68 và dữ liệu có xu hướng nghiêng về phía dưới.

– Good Tracking system: Chỉ có 2 mức độ được thể hiện là neutral và agree. Trong đó, mức độ neutral khá cao, phân bố rộng. Mức độ agree có mức trung vị khoảng 70, nhỏ hơn mức độ neutral và độ phân tán cũng bé hơn.



Hình 3. Mức độ quan trọng của các thuộc tính

Biến "Ease and convenient" là biến quan trọng, ảnh hưởng nhiều nhất, tiếp theo sau là biến "Time saving" và "More restaurant choices" và ở vị trí thứ 4 là biến "Easy Payment option". Đây là 4 thuộc tính có mức độ ảnh hưởng nhiều nhất quyết định của khách hàng có muốn sử dụng lại dịch vụ hay không.

2.2. Phương pháp máy học

2.2.1. Mô hình máy học

- Logistic Regression: Là một mô hình hồi quy nhằm dự đoán giá trị đầu ra rời rạc y ứng với một véc-tơ đầu vào x. Mô hình thường được sử dụng nhiều cho các bài toán phân loại dùng để gán các đối tượng cho một tập hợp giá trị rời rạc. Một ví dụ điển hình là bài toán phân loại email, phân loại giao dịch trực tuyến an toàn hay không an toàn, khối u lành tính hay ác tính. Đầu ra dự đoán của Logistic Regression có dạng viết chung là:

$$\mathbf{f(x) = \theta(w^T x)}$$

- Random Forest: là một phương pháp máy học kết hợp các Decision Tree được huấn luyện theo kĩ thuật bagging (hoặc pasting). Thuật toán hoạt động bằng cách cho điểm dữ liệu mới được đánh giá/phân loại qua nhiều Decision Tree và lấy ra kết quả được đánh giá tốt nhất.
- K-Nearest Neighbors: Thuật toán K-Nearest Neighbors viết tắt là KNN. Đây là thuật toán sử dụng machine learning với phương pháp học có giám sát để phục vụ việc cho đưa ra quyết định hoặc dự đoán tương lai. KNN sử dụng công thức toán học để chọn ra K phần tử gần nhất từ tập dữ liệu data training để đưa ra quyết định.

2.2.2. Công cụ sử dụng

Nền tảng sử dụng: Google Colab.

Thư viện sử dụng:

- Sklearn: Thư viện cung cấp một tập các công cụ xử lý các bài toán machine learning và statistical modeling.
- Pandas: Thư viện dùng để thao tác, phân tích và dọn dẹp dữ liệu.
- Numpy: Thư viện dùng để xử lý mảng đa chiều, ma trận.
- Matplotlib: Thư viện dùng để vẽ đồ thị 2D.

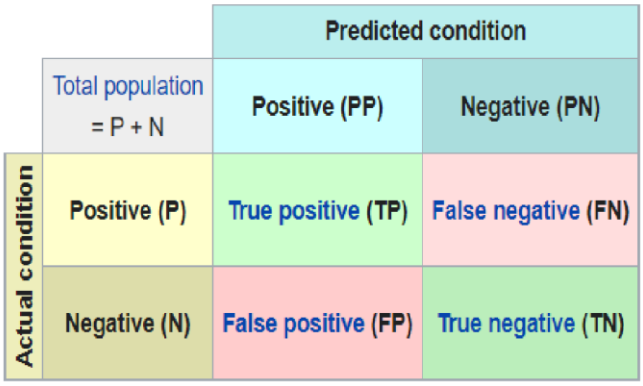
- Folium: Công cụ khai phá dữ liệu mạnh của python kế thừa các tính năng mạnh mẽ của thư viện rất nổi tiếng về map là `Leaflet.js` trong javascript.

Công cụ sử dụng vẽ DashBoard: Microsoft Power BI.

https://drive.google.com/file/d/1fyUhcBQwBPfP1OGRT5NlHS2VAGsXOx_a/view?usp=share_link

2.2.3. Các phương pháp đánh giá

- Ma trận nhầm lẫn: Giúp đánh giá được các giá trị cụ thể mỗi loại được phân loại như thế nào, lớp nào được phân loại đúng nhiều nhất, và dữ liệu thuộc lớp nào hay bị phân loại nhầm vào lớp khác.



Hình 4. Ma trận nhầm lẫn

- Ngoài ra nhóm còn sử dụng độ đo Accuracy để đánh giá mô hình bằng cách tính tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

2.3. Kết quả sau khi chạy mô hình

Sau khi chạy ba mô hình máy học bao gồm: Logistic Regression, Random Forest và K-Nearest Neighbors (KNN) với tham số mặc định, nhận thấy mô hình Random Forest (96.15%) đạt kết quả cao nhất trong ba mô hình.

Kết quả được thể hiện ở bảng 1

Mô hình	Kết quả độ đo Accuracy (%)
Logistic Regression	73.08

Random Forest	96.15
K-Nearest Neighbors	84.62

Bảng 1. Kết quả chạy mô hình máy học với tham số mặc định

2.4. Phân tích lỗi- Hướng phát triển

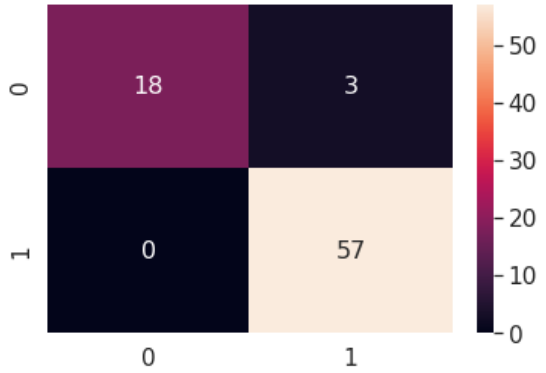
2.4.1. Phân tích lỗi

Từ ma trận nhầm lẫn, ta có thể tính được tỉ lệ dự đoán trên các nhãn như sau:

Nhãn	Chú thích
0	Người dùng không sử dụng lại dịch vụ đặt đồ ăn trực tuyến.
1	Người dùng sử dụng lại dịch vụ đặt đồ ăn trực tuyến.

Bảng 2. Bảng chú thích bộ nhãn

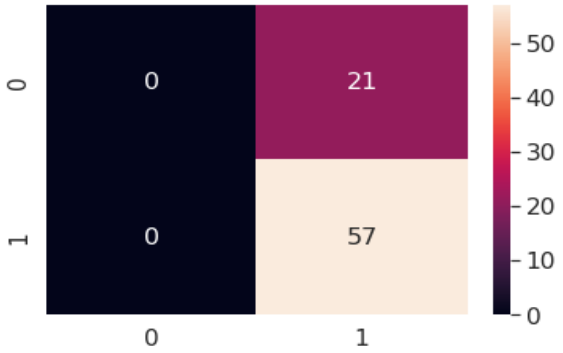
Mô hình Random Forest.



Hình 5. Ma trận nhầm lẫn trên tập kiểm tra với mô hình Random Forest

- Tỉ lệ dự đoán đúng với nhãn 0 đạt 18/21 = 85.71%.
- Tỉ lệ dự đoán đúng với nhãn 1 đạt 57/57= 100%.
- Trong việc dự đoán đúng nhãn 0 thì có 3 dự đoán nhầm thành nhãn 1 và không có sự dự đoán nhầm của nhãn 1.
- Sự chênh lệch giữa hai nhãn với nhau không quá lớn khi tỉ lệ dự đoán cho nhãn 0 đạt 85.71% và nhãn 1 đạt kết quả tuyệt đối 100%.

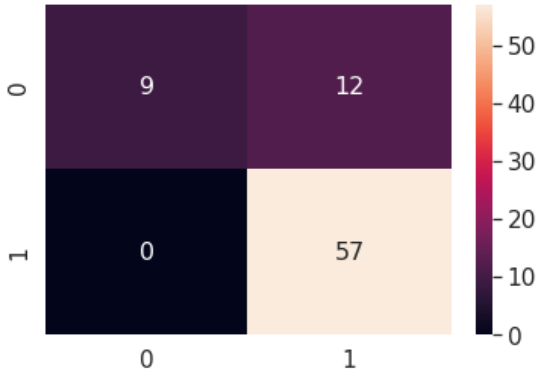
Mô hình Logistic Regression



Hình 6. Ma trận nhầm lẫn trên tập kiểm tra với mô hình Logistic Regression

- Tỉ lệ dự đoán đúng với nhãn 0 đạt 0/21= 0%.
- Tỉ lệ dự đoán đúng với nhãn 1 đạt 57/57= 100%.
- Trong việc dự nhãn 0 thì có 21 dự đoán nhầm thành nhãn 1 và không dự đoán nhầm của nhãn 1.
- Sự chênh lệch giữa hai nhãn với nhau rất lớn khi tỉ lệ dự đoán đúng của nhãn 0 là 0% và nhãn 1 lại đạt giá trị tuyệt đối 100%.

Mô hình K-Nearest Neighbors (KNN)



Hình 7. Ma trận nhầm lẫn trên tập kiểm tra với mô hình KNN

- Tỉ lệ dự đoán đúng với nhãn 0 đạt 9/21= 42.85%.
- Tỉ lệ dự đoán đúng với nhãn 1 đạt 57/57= 100%.
- Trong việc dự đoán nhãn 0 thì có 12 dự đoán nhầm nhãn 1 và không có sự đoán nhầm của nhãn 1.
- Sự chênh lệch giữa hai nhãn với nhau tương đối lớn khi tỉ lệ dự đoán đúng của nhãn 1 là tuyệt đối (100%) gần như gấp đôi nhãn 0 (42.85%).

Từ ma trận nhầm lẫn của ba mô hình trên ta thấy:

- Mô hình Random Forest có Accuracy cao nhất (96.15%) và kết quả của ma trận nhầm lẫn là tốt nhất. Nhãn 0 và nhãn 1 có tỉ lệ phân loại đúng đạt gần như là tuyệt đối (85.71% và 100%).
- Mô hình Logistic Regression có Accuracy thấp nhất (73.08%) và kết quả của ma trận nhầm lẫn là kém nhất. Mặc dù tỉ lệ phân loại đúng của nhãn 1 là tuyệt đối, đạt 100% nhưng nhãn 0 không có sự phân loại đúng.
- Mô hình K-Nearest Neighbors có Accuracy khá cao (84.62%) và kết quả của ma trận nhầm lẫn ở mức tương đối với tỉ lệ phân loại đúng của nhãn 0 là 42.85% và nhãn 1 là 100%.

Chúng tôi nhận thấy những nguyên nhân khiến mô hình cho ra kết quả không tốt có thể kể đến như do bộ dữ liệu chưa đủ lớn, dữ liệu ít dẫn đến mô hình hoạt động chưa tốt đặc biệt là mô hình Logistic Regression. Ngoài ra, do người dùng nhập các thông tin chưa hợp lí cũng dẫn đến sự sai sót trong việc chạy các mô hình.

2.4.2. Hướng phát triển

Từ những kết quả đã đạt được, chúng tôi đã vạch ra những bước phát triển tiếp cho việc nghiên cứu đề tài này:

- Trước tiên, chúng tôi sẽ xây dựng một bộ dữ liệu tốt hơn về mặt chất lượng lẫn số lượng, chú trọng hơn trong việc tiền xử lí dữ liệu và tiến hành tinh chỉnh mô hình bằng các siêu tham số để có thể đạt hiệu suất tốt hơn.
- Nghiên cứu và sử dụng các phương pháp phân loại phù hợp để tăng khả năng học cũng như dự đoán của mô hình. Đồng thời thử nghiệm và cài đặt thêm một vài mô hình máy học, mô hình học sâu khác để đưa ra kết quả tốt nhất.

3. KẾT LUẬN

Đồ án đã xây dựng được mô hình cho kết quả phân loại và dự đoán đạt độ chính xác trung bình khoảng 85% cụ thể là Random Forest 96.15%, Logistic Regression 73.08%, K-Nearest Neighbors 84.62%. Với độ chính xác khá cao và từ kết quả của ma trận nhầm lẫn, Random Forest là mô hình tốt nhất có thể ứng dụng để giải quyết bài toán đặt ra. Ngoài ra, qua nghiên cứu, nhóm đã tìm hiểu và sử dụng được một số phương pháp xử lí, phân tích thăm dò dữ liệu cũng như tiếp cận được nhiều phương pháp máy học áp dụng cho bài toán phân lớp đa lớp. Kết quả thu được từ ứng dụng máy học trên vận dụng vào thực tế giúp doanh nghiệp nắm bắt tốt nhất về tiềm năng, nhu cầu khách hàng, từ đó đưa ra chiến lược phát triển phù hợp.

TÀI LIỆU THAM KHẢO

[1] K Aditya Sobika, S.N Vivek Raj, A Study on Predicting Customer Willingness to Order Food Online During Covid-19 Pandemic Using Machine Learning Algorithms, 2021.

[2] Batool Madani, Hussam Alshraideh, Predicting Consumer Purchasing Decision in The Online Food Delivery Industry, 2021.

[3] Aurélien Géron, Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow, 2nd Edition, 2019.

[4] Yakob Utama Chandra, Cadelina Cassandra, Stimulus Factors of Order Online Food Delivery, 2019.

[5] Teck-Chai Lau, David ng, Online Food Delivery Services: Making Food Delivery the New Normal, 2019.

[6] Kathleen Griesbach, Adam Reich, Luke Elliott-Negri, Ruth Milkman, Algorithmic Control in Platform Food Delivery Work, 2019.

[7] Vicent Cheow Sern Yeo, See-Kwong Goh, Sajad Rezaei, Consumer experiences, attitude and behavioral intention toward online food delivery (OFD) services, 2017.

[8] Kaggle.com. https://www.kaggle.com/code/woojinhwang/simple-ml-prediction-predict-food-delivery-time (30/11/2022).

[9] Kaggle.com. https://www.kaggle.com/datasets/benroshan/online-food-delivery-preferencesbangalore-region (17/11/2022).

[10] Bultin.com. https://builtin.com/machine-learning/food-delivery-time-prediction (10/11/2022).

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Bùi Nguyên Phương Linh	Nhóm trưởng, Làm báo cáo Word, Thuyết trình, Làm slide, Hỗ trợ viết code, Tìm hiểu bộ dữ liệu.
2	Nguyễn Thị Phương Thảo	Viết code, Thuyết trình, Tìm hiểu bộ dữ liệu, Hỗ trợ làm báo cáo Word, Vẽ dashboard.