TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ TÀI

Xây dựng hệ thống nhận diện sự hài lòng của hành khách sau chuyến bay

Môn học: Khai thác dữ liệu và ứng dụng
Giảng viên: Nguyễn Thanh Phước

Họ và tên: Hà Lý Gia Bảo
MSSV: 3121410068
Lớp: DCT1223

Thành phố Hồ Chí Minh - Tháng 5/2025

TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ TÀI

Xây dựng hệ thống nhận diện sự hài lòng của hành khách sau chuyến bay

Thành phố Hồ Chí Minh - Tháng 5/2025

Mục lục

Chương 1

Giới thiệu

1.1 Bối cảnh của bài toán

Ngày nay, luôn có cạnh tranh khốc liệt trong ngành hàng không, sự hài lòng của hành khách đóng một vai trò then chốt quyết định sự thành công và phát triển bền vững của các hãng bay. Việc hiểu rõ những yếu tố nào ảnh hưởng đến sự hài lòng của hành khách và dự đoán mức độ hài lòng của họ là vô cùng quan trọng đối với các hãng hàng không. Hành khách hài lòng có xu hướng trở thành khách hàng trung thành, giới thiệu hãng bay cho người khác và ít có khả năng chuyển sang sử dụng dịch vụ của đối thủ cạnh tranh. Ngược lại, những hành khách không hài lòng có thể chia sẻ trải nghiệm tiêu cực của họ, gây ảnh hưởng xấu đến danh tiếng và doanh thu của hãng.

Việc hiểu rõ các yếu tố tác động đến sự hài lòng của hành khách và khả năng dự đoán mức độ hài lòng của họ trở nên vô cùng quan trọng. Điều này cho phép các hãng hàng không chủ động cải thiện chất lượng dịch vụ, tối ưu hóa trải nghiệm bay và xây dựng chiến lược kinh doanh hiệu quả hơn. Cùng với sự phát triển của các công nghệ thu thập dữ liệu, các hãng hàng không hiện nay có quyền truy cập vào một lượng lớn dữ liệu về trải nghiệm bay của hành khách. Các thông tin về hành khách và chuyến bay của họ, bao gồm thông tin cá nhân, đặc điểm chuyến bay và đánh giá các dịch vụ khác nhau, chứa đựng những thông tin giá trị có thể được khai thác để đạt được mục tiêu này. Tuy nhiên, việc khai thác hiệu quả nguồn dữ liệu này để đưa ra những hiểu biết sâu sắc và hành động cụ thể vẫn là một thách thức.

1.2 Định nghĩa bài toán

Bài toán được đặt ra là xây dựng một mô hình có khả năng dự đoán mức độ hài lòng của hành khách dựa trên các thông tin liên quan đến đặc điểm về chuyến bay và dịch vụ. Đây là một bài toán phân loại trong lĩnh vực khai phá dữ liệu.

- **Dữ liệu đầu vào:** Tập dữ liệu chứa thông tin chi tiết về từng hành khách và trải nghiệm chuyến bay của họ, bao gồm các thuộc tính:

  - Thông tin cá nhân của hành khách: Giới tính, Loại khách hàng, Tuổi, Loại hình du lịch.
  - Thông tin chuyến bay: Hạng ghế, Khoảng cách chuyến bay, Độ trễ khi khởi hành, Độ trễ khi đến.
  - Đánh giá của hành khách về các dịch vụ: Mức độ hài lòng với Wifi trên máy bay, Sự thuận tiện của thời gian khởi hành/đến, Vị trí cổng chờ, Dễ dàng đặt vé trực tuyến, Làm thủ tục lên máy bay trực tuyến, Sự thoải mái của ghế ngồi, Giải trí trên máy bay, Đồ ăn và thức uống, Dịch vụ trên máy bay, Chỗ để chân, Dịch vụ làm thủ tục, Xử lý hành lý, Dịch vụ trong chuyến bay, Mức độ sạch sẽ.

- **Dữ liệu đầu ra:** Biến mục tiêu là "satisfaction"với hai giá trị: hài lòng và trung lập hoặc không hài lòng.
- **Thuật toán:** Vì đây là bài toán phân loại nhị phân nên để giải quyết bài toán phân loại này, nhiều thuật toán học máy sẽ được xem xét và sử dụng cho huấn luyện và đánh giá, bao gồm:
  - Logistic Regression
  - Naive Bayes
  - K-Nearest Neighbors
  - Decision Tree
  - Multi Layer Perceptron
  - Random Forest
  - AdaBoost
  - Support Vector Machine

1.3 Các giải pháp thủ công để giải quyết bài toán

Trước khi có sự hỗ trợ của các kỹ thuật khai phá dữ liệu và học máy, các hãng hàng không thường dựa vào các phương pháp truyền thống và thủ công để đánh giá và cải thiện sự hài lòng của hành khách như:

- **Khảo sát thủ công:** Phát phiếu khảo sát giấy hoặc khảo sát trực tiếp hành khách sau chuyến bay. Phương pháp này tốn kém thời gian, công sức, chi phí và thường có tỷ lệ phản hồi thấp, dữ liệu thu được có thể không đại diện cho toàn bộ hành khách.
- **Phân tích phản hồi từ các kênh riêng lẻ:** Thu thập và phân tích phản hồi từ các kênh như email, tổng đài chăm sóc khách hàng, mạng xã hội. Việc tổng hợp và phân tích thủ công từ nhiều nguồn rời rạc thường khó khăn và không đưa ra được cái nhìn tổng thể.

- **Dựa trên kinh nghiệm và cảm tính:** Các nhà quản lý dựa trên kinh nghiệm và quan sát cá nhân để đưa ra quyết định cải thiện dịch vụ. Phương pháp này thiếu tính khách quan và có thể bỏ sót nhiều yếu tố quan trọng.

Những giải pháp này thường mang tính bị động, tốn kém và phản ứng chậm với các vấn đề phát sinh và khó có khả năng dự đoán trước mức độ hài lòng của hành khách ở quy mô lớn để có những can thiệp kịp thời.

1.4 Vai trò của khai phá dữ liệu trong việc giải quyết bài toán trên

Khai phá dữ liệu đóng vai trò quan trọng trong việc giải quyết bài toán dự đoán sự hài lòng của hành khách một cách hiệu quả và khoa học hơn bằng cách:

- **Phát hiện các mẫu ẩn (hay sâu hơn là khám phá ra các tri thức ẩn)**: Khai phá dữ liệu giúp phát hiện các mẫu tiềm ẩn, các mối quan hệ và xu hướng trong bộ dữ liệu lớn mà con người khó có thể nhận ra bằng các phương pháp thủ công.
  Ví dụ, khám phá xem yếu tố dịch vụ như wifi, ghế ngồi, đồ ăn có ảnh hưởng lớn nhất đến sự hài lòng của các nhóm hành khách khác nhau như khách hạng thương gia so với khách hạng phổ thông.

- **Xây dựng mô hình dự đoán**: Bằng cách sử dụng các thuật toán học máy, khai phá dữ liệu cho phép xây dựng các mô hình có khả năng dự đoán mức độ hài lòng của hành khách trong tương lai dựa trên các đặc điểm của họ và chuyến bay.

- **Phân khúc khách hàng**: Phân cụm hành khách thành các nhóm dựa trên mức độ hài lòng và các đặc điểm khác để đưa ra các chiến lược phục vụ phù hợp cho từng nhóm.

- **Hỗ trợ ra quyết định**: Cung cấp cái nhìn sâu sắc dựa trên dữ liệu để các nhà quản lý đưa ra quyết định chiến lược nhằm nâng cao trải nghiệm và sự hài lòng của hành khách.

- **Hỗ trợ ra quyết định dựa trên dữ liệu**: Kết quả từ các mô hình khai phá dữ liệu cung cấp cơ sở vững chắc để các hãng hàng không đưa ra các quyết định kinh doanh và vận hành một cách chủ động và hiệu quả hơn. Ví dụ: Nên tập trung cải thiện dịch vụ nào, cá nhân hóa trải nghiệm cho từng đối tượng khách hàng.

- **Tối ưu hóa nguồn lực**: Thay vì đầu tư dàn trải, các hãng bay có thể tập trung nguồn lực vào những yếu tố thực sự quan trọng ảnh hưởng đến sự hài lòng, từ đó tối ưu hóa chi phí và nâng cao hiệu quả hoạt động.

1.5 Kết quả ứng dụng của khai phá dữ liệu sau khi giải quyết bài toán trên

Việc giải quyết thành công bài toán phân loại mức độ hài lòng của hành khách bằng các kỹ thuật khai phá dữ liệu sẽ mang lại nhiều kết quả dự đoán và ứng dụng thực tiễn có giá trị cho các hãng hàng không:

- **Phân tích nguyên nhân gốc rễ:** Khi có sự không hài lòng, nhờ mô hình và việc phân tích dữ liệu có thể giúp xác định nguyên nhân cốt lõi.
- **Dự đoán sớm nguy cơ không hài lòng:** Mô hình có thể xác định những hành khách có khả năng cao không hài lòng trước hoặc trong chuyến bay, cho phép nhân viên chủ động can thiệp để cải thiện trải nghiệm của họ như cung cấp một ưu đãi nhỏ, hỏi thăm đặc biệt
- **Cá nhân hóa dịch vụ:** Hiểu được các yếu tố nào quan trọng đối với từng phân khúc khách hàng, hãng bay có thể điều chỉnh dịch vụ cho phù hợp.
  Ví dụ: khách đi công tác quan tâm đến wifi và sự đúng giờ, khách đi du lịch quan tâm đến giải trí và sự thoải mái
- **Định hướng cải tiến chất lượng dịch vụ:** Phân tích các yếu tố có trọng số cao trong mô hình dự đoán giúp hãng bay xác định chính xác những khía cạnh dịch vụ cần được ưu tiên cải thiện để mang lại tác động lớn nhất đến sự hài lòng chung.
- **Thiết kế chiến lược marketing hiệu quả:** Các chiến dịch quảng bá, chương trình khách hàng thân thiết, nhắm mục tiêu đến đúng đối tượng dựa trên mức độ hài lòng và các yếu tố ảnh hưởng đến họ.
- **Đánh giá hiệu quả của các thay đổi:** Khi hãng bay thực hiện các cải tiến dịch vụ, mô hình có thể được sử dụng để đánh giá xem những thay đổi đó có thực sự nâng cao sự hài lòng của hành khách hay không.
- **Tăng cường lợi thế cạnh tranh:** Bằng cách liên tục theo dõi và cải thiện sự hài lòng của hành khách dựa trên dữ liệu, các hãng hàng không có thể xây dựng được lòng trung thành của khách hàng và tạo ra lợi thế cạnh tranh bền vững.

Chương 2

Mô tả dữ liệu

2.1 Kích thước dữ liệu, chiều dữ liệu, trích dẫn nguồn dữ liệu:

Dữ liệu được sử dụng trong đồ án này có nguồn gốc từ cuộc khảo sát về mức độ hài lòng của hành khách hàng không, được công bố trên nền tảng Kaggle. Nguồn dữ liệu có thể được truy cập tại: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

Tập dữ liệu bao gồm thông tin về sự hài lòng của hành khách hàng không. Dữ liệu gốc bao gồm hai tệp CSV riêng biệt: một cho tập huấn luyện và một cho tập kiểm tra.

- Tập huấn luyện: Bao gồm 103,904 mẫu (hay là hành khách) và 25 cột (hay là thuộc tính).
- Tập kiểm tra: Bao gồm 25,976 mẫu và 25 cột tương tự.
- Tổng cộng: 129,880 mẫu.

Mỗi hàng đại diện cho một hành khách và có 25 cột. Sau khi loại bỏ các cột không cần thiết và thực hiện các bước tiền xử lý như mã hóa nhãn, tập dữ liệu được sử dụng để huấn luyện mô hình còn có 23 thuộc tính và 1 cột mục tiêu là satisfaction.

2.2 Các kiểu dữ liệu:

Tập dữ liệu bao gồm cả các thuộc tính hạng mục và số. Các kiểu dữ liệu chính trong tập dữ liệu gốc bao gồm:

- **Thuộc tính dạng số:**
  - Age: Tuổi của hành khách.
  - Flight Distance: Khoảng cách của chuyến bay.
  - Departure Delay in Minutes: Độ trễ khi khởi hành tính bằng phút.

  - Arrival Delay in Minutes: Độ trễ khi đến tính bằng phút.

- **Thuộc tính dạng hạng mục:**
  - Gender: Giới tính (Male, Female).
  - Customer Type: Loại khách hàng (Loyal Customer, disloyal Customer).
  - Type of Travel: Mục đích chuyến đi (Personal Travel, Business travel).
  - Class: Hạng ghế (Eco, Eco Plus, Business).
  - Inflight wifi service: Mức độ hài lòng với dịch vụ wifi (0-5).
  - Departure/Arrival time convenient: Mức độ hài lòng với sự thuận tiện của thời gian bay (0-5).
  - Ease of Online booking: Mức độ hài lòng với sự dễ dàng đặt vé trực tuyến (0-5).
  - Gate location: Mức độ hài lòng với vị trí cổng chờ (0-5).
  - Food and drink: Mức độ hài lòng với đồ ăn và thức uống (0-5).
  - Online boarding: Mức độ hài lòng với làm thủ tục trực tuyến (0-5).
  - Seat comfort: Mức độ hài lòng với sự thoải mái của ghế (0-5).
  - Inflight entertainment: Mức độ hài lòng với giải trí trên chuyến bay (0-5).
  - On-board service: Mức độ hài lòng với dịch vụ trên khoang (0-5).
  - Leg room service: Mức độ hài lòng với chỗ để chân (0-5).
  - Baggage handling: Mức độ hài lòng với xử lý hành lý (0-5).
  - Checkin service: Mức độ hài lòng với dịch vụ làm thủ tục (0-5).
  - Inflight service: Mức độ hài lòng với dịch vụ trong chuyến bay (0-5).
  - Cleanliness: Mức độ hài lòng với sự sạch sẽ (0-5).
  - satisfaction: Mức độ hài lòng chung của hành khách (neutral or dissatisfied, satisfied) và đây là biến mục tiêu để xây dựng mô hình.

Trong quá trình tiền xử lý, các thuộc tính hạng mục đã được chuyển đổi thành dạng số thông qua kỹ thuật Mã hóa Nhãn để phù hợp với yêu cầu đầu vào của các thuật toán học máy. Ví dụ, 'Female' được mã hóa thành 0 và 'Male' thành 1 hay 'neutral or dissatisfied' thành 0 và 'satisfied' thành 1.

2.3 Thống kê mô tả về dữ liệu:

2.3.1 Dữ liệu thiếu:

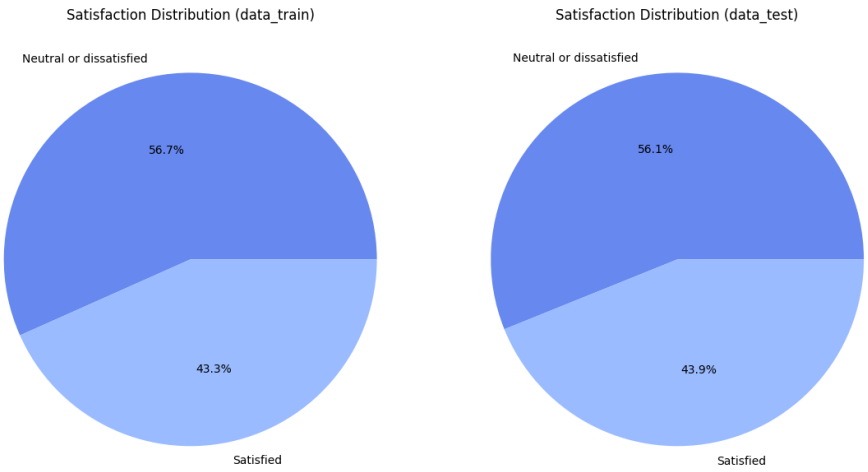
- Chỉ có một cột duy nhất có dữ liệu thiếu là Arrival Delay in Minutes với 310 (tập train) và 83 (tập test) bản ghi bị thiếu, chiếm khoảng 0.298352% (tập train) và 0.319526% (tập test) tổng số dữ liệu.

- Các cột khác không có giá trị thiếu.

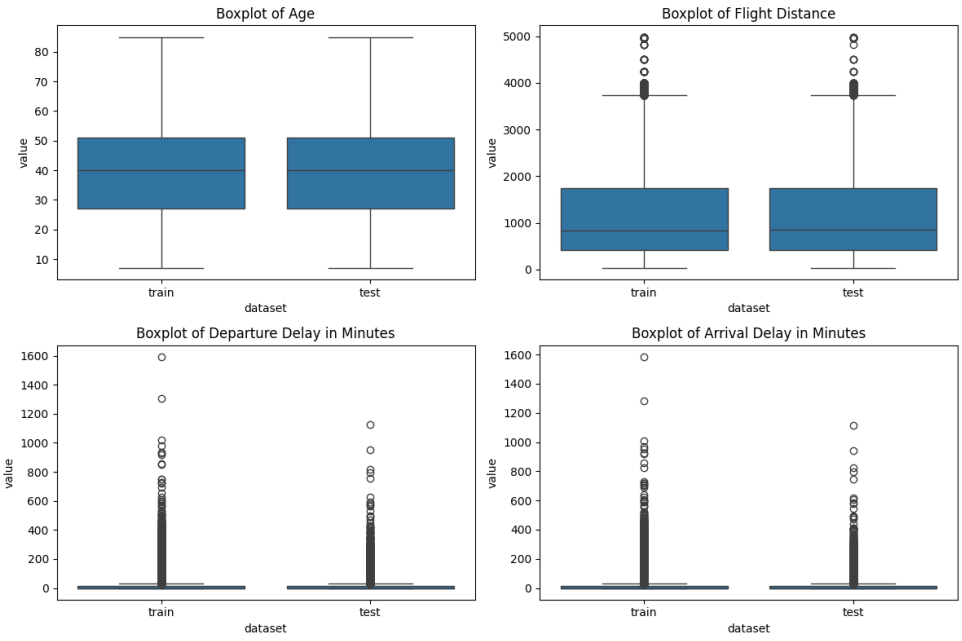
Đây là một tỷ lệ thiếu rất nhỏ và có thể được xử lý bằng các phương pháp đơn giản như điền giá trị trung bình, trung vị hoặc loại bỏ các hàng tương ứng. Vì vậy việc xử lý số lượng nhỏ dữ liệu thiếu này sẽ không ảnh hưởng nhiều đến tổng thể tập dữ liệu nên sẽ sử dụng điền giá trị trung vị để bổ sung vào những dữ liệu thiếu

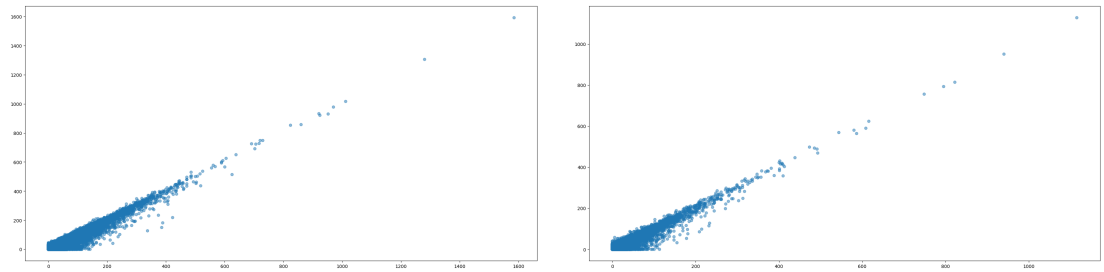
2.3.2 Mô tả thông tin về các biến

- **Phân phối của biến mục tiêu**:
  - Trong cả tập huấn luyện và tập kiểm tra, dữ liệu có sự hơi mất cân bằng nhẹ. Cụ thể, trong tập huấn luyện, khoảng 56.7% hành khách được phân loại là Trung lập hoặc không hài lòng và 43.3% là Hài lòng. Tỷ lệ này tương tự trong tập kiểm tra (56.1% và 43.9%).
  - Điều này cho thấy không có sự chênh lệch quá lớn giữa hai lớp nhưng vẫn cần lưu ý khi đánh giá mô hình, đặc biệt là khi sử dụng các chỉ số đánh giá như Precision và Recall cho từng lớp.



- **Phân phối của các thuộc tính dạng số:**
  - Age: Độ tuổi của hành khách phân bổ khá rộng, tập trung chủ yếu trong khoảng từ 20 đến 60 tuổi. Có một số giá trị ngoại lệ ở cả hai phía (trẻ hơn và lớn tuổi hơn).
  - Flight Distance: Hầu hết các chuyến bay có khoảng cách dưới 2000 đơn vị, nhưng có một số chuyến bay với khoảng cách rất xa, tạo ra một cái đuôi dài ở phía bên trên của phân phối (phân phối lệch phải nếu xoay theo chiều ngang).





Train Test

– Departure Delay in Minutes và Arrival Delay in Minutes: Cả hai thuộc tính này đều có phân phối lệch phải mạnh, với phần lớn các chuyến bay có độ trễ thấp (gần 0 phút). Tuy nhiên, có một số trường hợp trễ chuyến đáng kể, thể hiện qua các điểm ngoại lệ rất lớn.

  Cụ thể hơn: Dựa vào biểu đồ scatter plot giữa Departure Delay in Minutes và Arrival Delay in Minutes cho thấy một mối tương quan dương mạnh mẽ, điều này là hợp lý vì trễ khởi hành thường dẫn đến trễ khi hạ cánh.

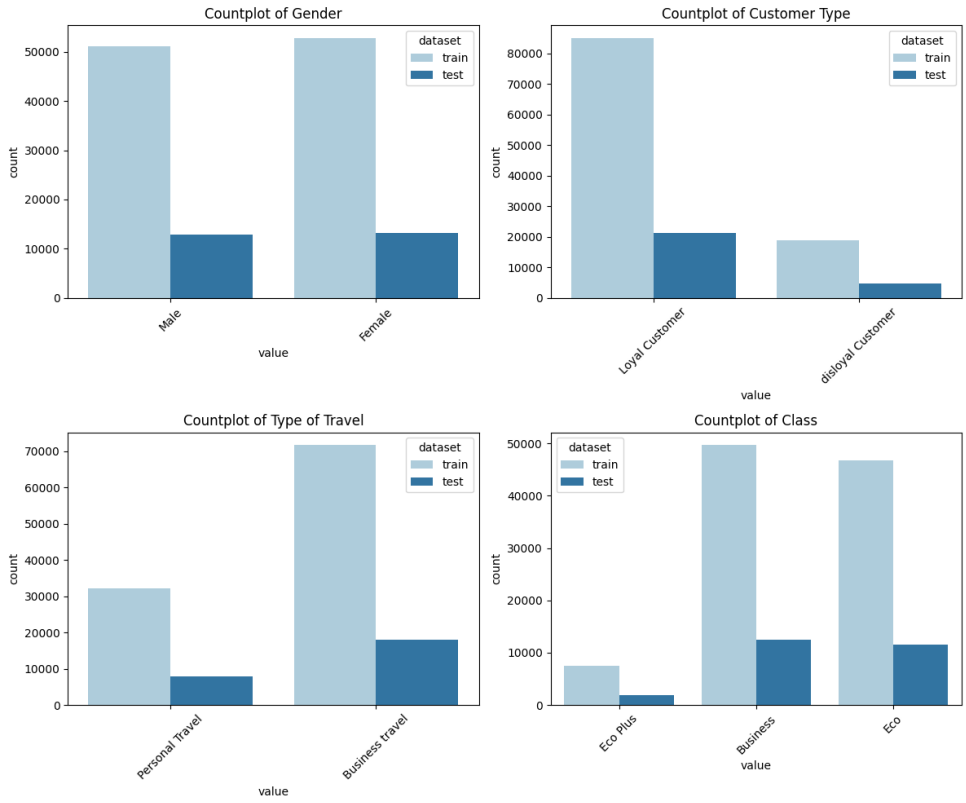
- **Phân phối của các thuộc tính dạng hạng mục:**

  – Gender: Số lượng hành khách nam và nữ khá tương đồng.

  – Customer Type: Phần lớn hành khách là Khách hàng trung thành so với Khách hàng không trung thành.

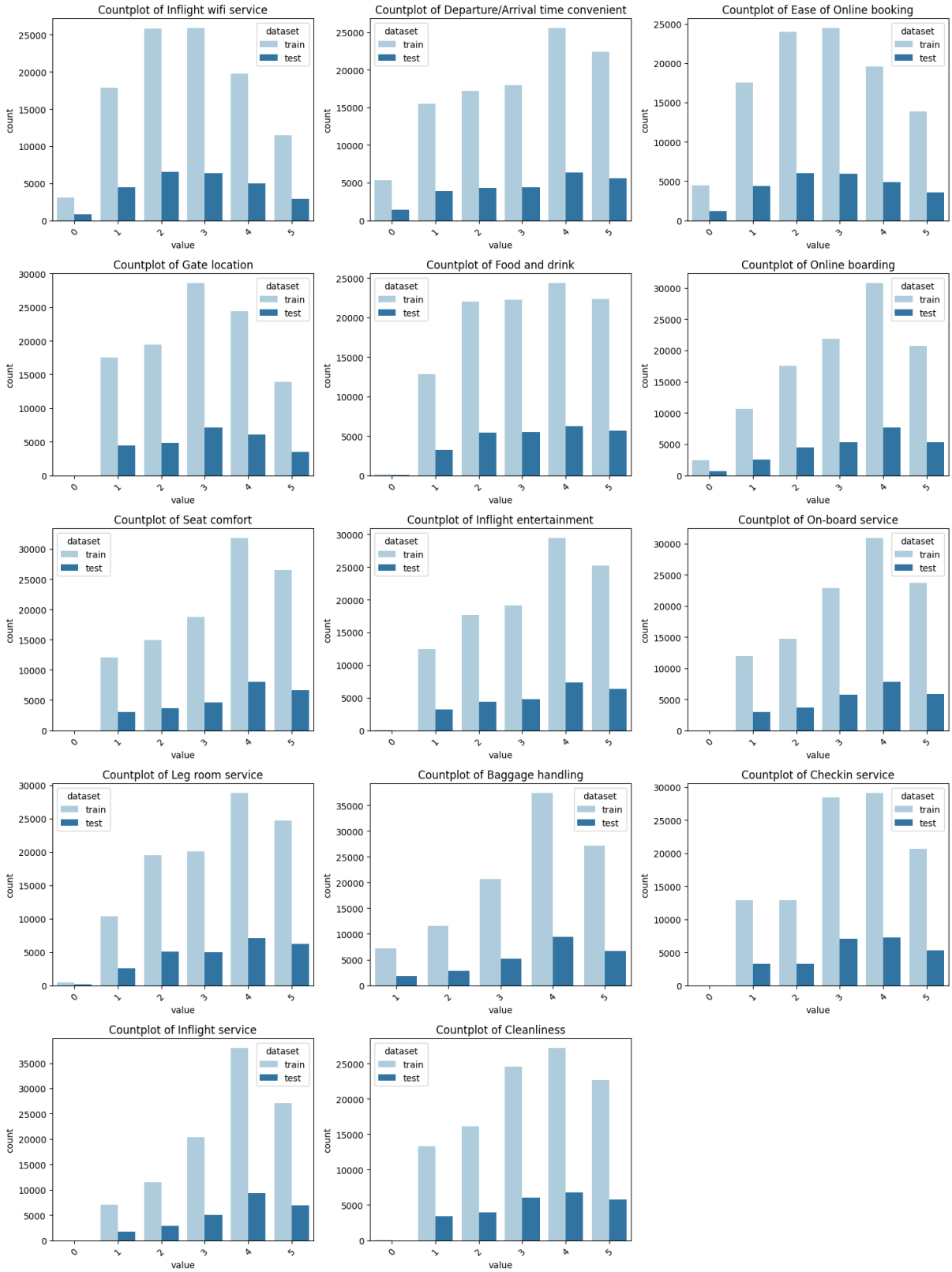
  – Type of Travel: Số lượng hành khách đi công tác nhiều hơn đáng kể so với hành khách đi du lịch cá nhân.

  – Class: Hạng ghế Business và Eco chiếm đa số, trong khi Eco Plus có số lượng ít hơn hẳn.



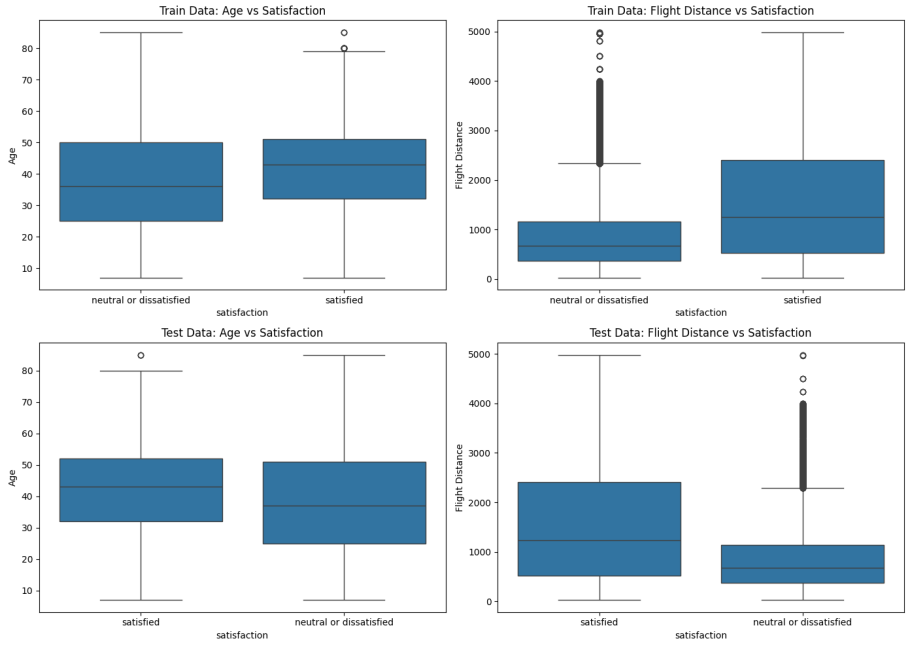
– Đối với các thuộc tính đánh giá dịch vụ:

  * Nhiều hành khách có xu hướng cho điểm cao (4 hoặc 5) cho các dịch vụ như Online boarding, Seat comfort, Cleanliness.
  * Các dịch vụ như Inflight wifi service và Ease of Online booking cũng nhận được nhiều đánh giá ở mức khá (3, 4, 5).
  * Một số dịch vụ có sự phân bổ điểm rộng hơn, cho thấy sự đa dạng trong cảm nhận của hành khách.


Countplot of Inflight wifi service
Countplot of Departure/Arrival time convenient
Countplot of Ease of Online booking
Countplot of Gate location
Countplot of Food and drink
Countplot of Online boarding
Countplot of Seat comfort
Countplot of Inflight entertainment
Countplot of On-board service
Countplot of Leg room service
Countplot of Baggage handling
Countplot of Checkin service
Countplot of Inflight service
Countplot of Cleanliness
dataset
train
test
count
value

2.3.3 Mối quan hệ giữa các thuộc tính và biến mục tiêu:

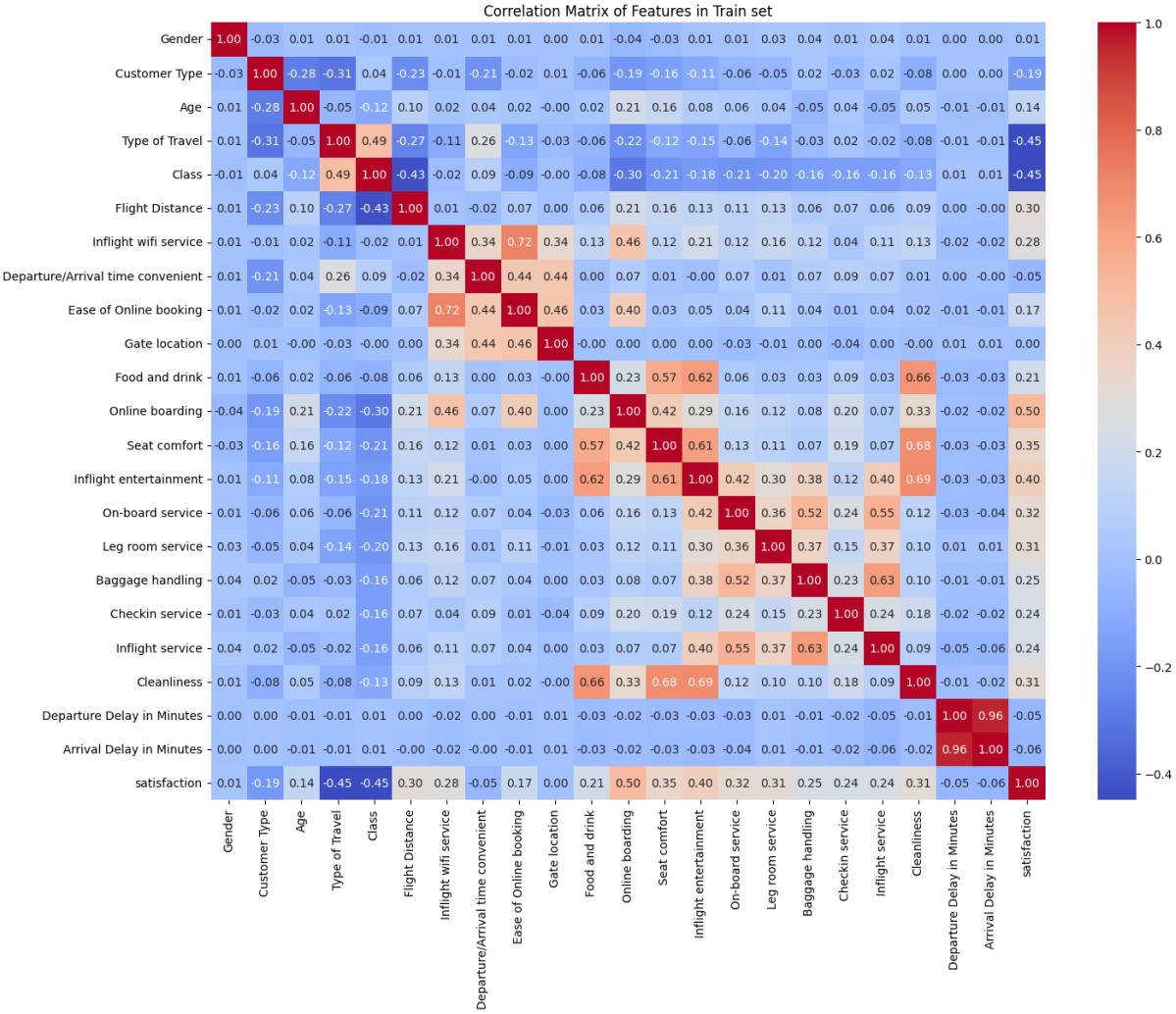
Các biểu đồ cột và hộp kết hợp với biến satisfaction cung cấp cái nhìn ban đầu về mối quan hệ:

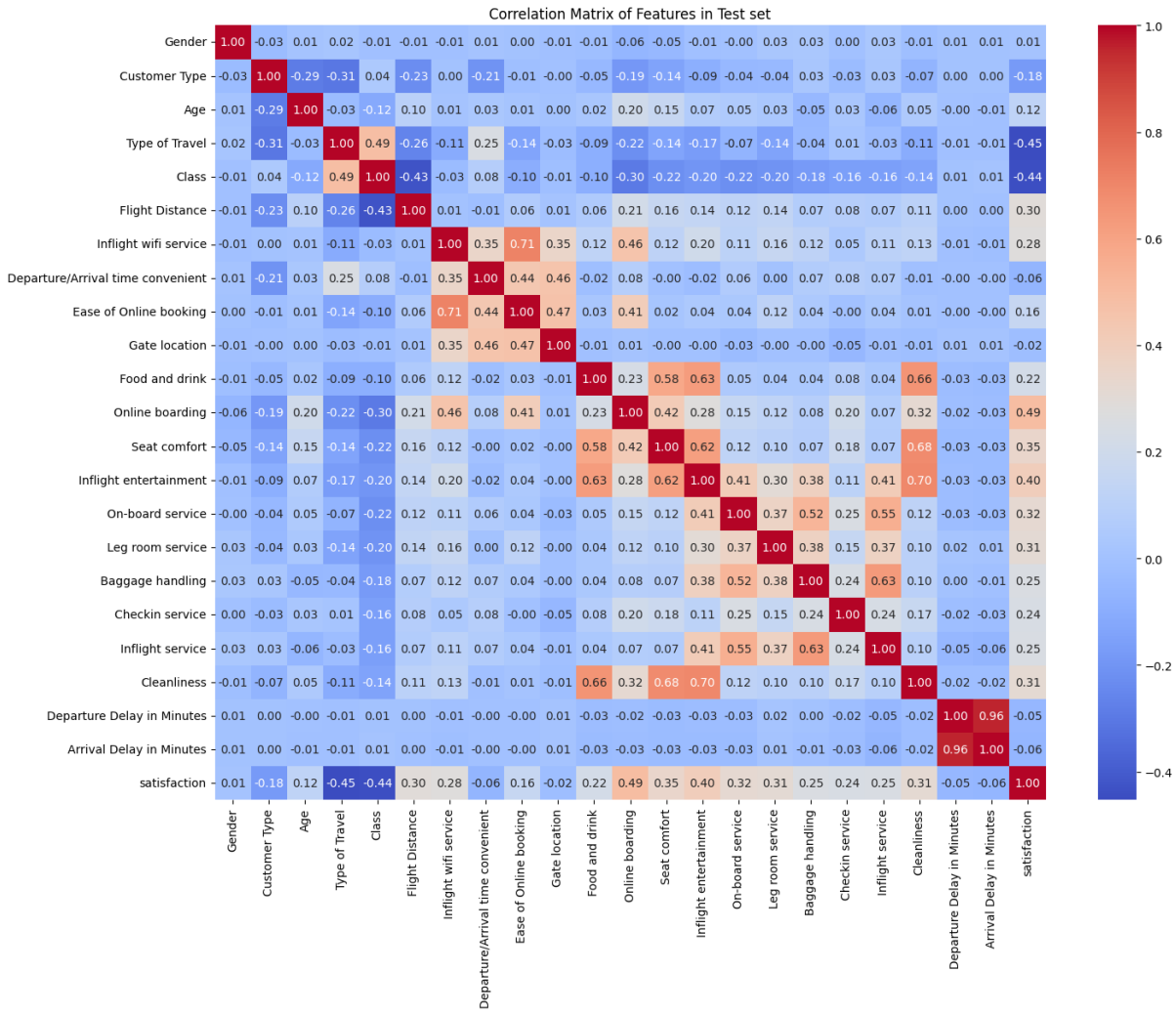


- Age vs Satisfaction: Có vẻ như hành khách ở độ tuổi trung niên (khoảng 40-60) có xu hướng hài lòng cao hơn.
- Flight Distance vs Satisfaction: Hành khách trên các chuyến bay dài hơn có xu hướng hài lòng hơn.
- Gender vs Satisfaction: Không có sự khác biệt quá rõ rệt về mức độ hài lòng giữa nam và nữ.
- Customer Type vs Satisfaction: Khách hàng trung thành có tỷ lệ không hài lòng cao hơn một chút so với khách hàng không trung thành, điều này có thể phản ánh kỳ vọng cao hơn từ họ. Tuy nhiên, số lượng khách hàng không trung thành hài lòng cũng đáng kể. Đây là một yếu tố quan trọng.
- Type of Travel vs Satisfaction: Hành khách đi công tác có tỷ lệ hài lòng cao hơn đáng kể so với hành khách đi du lịch cá nhân. Đây cũng là một yếu tố ảnh hưởng mạnh.
- Class vs Satisfaction: Hành khách hạng Business có tỷ lệ hài lòng cao nhất, tiếp theo là Eco Plus và cuối cùng là Eco.

2.3.4 Ma trận tương quan:

Ma trận tương quan cung cấp cái nhìn về mối quan hệ tuyến tính giữa tất cả các cặp biến trong tập dữ liệu (sau khi đã được mã hóa thành số). Các giá trị gần 1 hoặc -1 cho thấy mối tương quan mạnh (tương quan dương hoặc âm), trong khi các giá trị gần 0 cho thấy ít hoặc không có mối tương quan tuyến tính.





Dựa vào 2 hình trực quan ma trận tương quan, cho thấy sự tương đồng về các mối quan hệ giữa 2 tập train và test. Nên không cần lo đến sự sai lệch về mối quan hệ trong dữ liệu đã được chia sẵn của tác giả gây ảnh hưởng hiệu quả của mô hình máy học.

- Các dịch vụ trên chuyến bay thường có tương quan dương với nhau. Ví dụ: Inflight entertainment và Seat comfort, Food and drink và Cleanliness.
- Online boarding có tương quan dương mạnh với Inflight wifi service và Ease of Online booking.
- Biến mục tiêu satisfaction có tương quan dương đáng kể với các yếu tố như Online boarding, Class, Type of Travel, Inflight entertainment, Seat comfort.
- Departure Delay in Minutes và Arrival Delay in Minutes có tương quan dương rất cao (0.96), điều này là hiển nhiên vì thời gian chậm trễ khởi hành thường dẫn đến chậm trễ đến. Việc này cần được xem xét trong quá trình lựa chọn đặc trưng để tránh vấn đề đa cộng tuyến.

2.4 Những giả thiết khi thu thập dữ liệu:

Khi làm việc với tập dữ liệu này, một số giả thiết có thể được đặt ra (hoặc cần được xem xét) về quá trình thu thập dữ liệu:

- Tính đại diện của mẫu: Giả sử rằng mẫu khảo sát đủ lớn và đa dạng để đại diện cho tổng thể hành khách của các hãng hàng không liên quan. Nếu dữ liệu chỉ từ một hãng hoặc một số đường bay nhất định, kết quả có thể không khái quát được cho toàn ngành.

- Sự hiểu biết đồng nhất về thang đo: Giả sử tất cả hành khách hiểu và diễn giải các câu hỏi cũng như thang điểm đánh giá như điểm từ 0 đến 5 một cách tương tự nhau.

- Tính trung thực và khách quan của phản hồi: Giả sử hành khách cung cấp phản hồi một cách trung thực và dựa trên trải nghiệm thực tế của họ. Các yếu tố như tâm trạng tại thời điểm khảo sát, hoặc mong muốn làm hài lòng/gây khó dễ cho người khảo sát có thể ảnh hưởng đến câu trả lời.

- Không có yếu tố bên ngoài chi phối mạnh: Giả sử rằng các yếu tố không được thu thập trong bộ dữ liệu như thời tiết cụ thể, sự kiện đặc biệt xảy ra trên chuyến bay không được ghi nhận nên không gây ra những biến động lớn và ngẫu nhiên trong mức độ hài lòng.

- Không có thiên vị đáng kể trong việc lựa chọn mẫu: Quy trình chọn hành khách để khảo sát không có thiên vị đáng kể, đảm bảo rằng mẫu dữ liệu phản ánh đúng sự đa dạng của hành khách.

- Các biến độc lập với nhau với điều kiện trong một mức độ nào đó: Mặc dù có tương quan giữa một số biến, nhưng các biến đầu vào được giả định là cung cấp thông tin độc lập đủ để phân biệt các lớp hài lòng và không hài lòng.

- Dữ liệu được thu thập trong một khoảng thời gian nhất quán: Nếu dữ liệu được thu thập qua một khoảng thời gian rất dài, các thay đổi về chính sách của hãng bay, công nghệ, hoặc kỳ vọng của hành khách có thể làm thay đổi bản chất của các mối quan hệ trong dữ liệu.

Chương 3

Phương pháp khai phá dữ liệu

3.1 Quy trình khai phá dữ liệu

Quy trình khai phá dữ liệu để giải quyết bài toán dự đoán sự hài lòng của hành khách theo các bước chuẩn trong một quy trình thực hiện dự án của khoa học dữ liệu, mục đích để đảm bảo tính hệ thống và hiệu quả thì các bước bao gồm:

- **Hiểu biết về bài toán và dữ liệu:**
  - Xác định rõ mục tiêu của bài toán: dự đoán sự hài lòng của hành khách.
  - Hiểu rõ bối cảnh ngành hàng không và ý nghĩa của sự hài lòng hành khách.
  - Khám phá dữ liệu ban đầu, hiểu cấu trúc, các thuộc tính và ý nghĩa của chúng.
  - Xác định biến mục tiêu và các biến đặc trưng.
- **Thu thập dữ liệu:**
  - Tải tập dữ liệu Airline Passenger Satisfaction từ Kaggle.
  - Dữ liệu bao gồm hai tệp: train.csv và test.csv.
- **Tìm hiểu và khám phá dữ liệu:**
  - Kiểm tra sự mất cân bằng lớp của biến mục tiêu.
  - Xem xét cấu trúc, các thuộc tính, kiểu dữ liệu của từng thuộc tính.

- Thực hiện thống kê mô tả (giá trị trung bình, trung vị, độ lệch chuẩn, phân phối) cho các thuộc tính số.
- Phân tích tần suất cho các thuộc tính hạng mục.
- Trực quan hóa dữ liệu bằng biểu đồ (boxplot, histogram, countplot, correlation matrix) để hiểu rõ hơn về đặc điểm và mối quan hệ giữa các biến.
- Phân tích mối quan hệ giữa các biến đầu vào và biến mục tiêu thông qua các biểu đồ và ma trận tương quan.

- **Tiền xử lý dữ liệu:**
  - Xử lý dữ liệu thiếu: Dựa trên phân tích ở Chương 2, cột Arrival Delay in Minutes có một tỷ lệ nhỏ dữ liệu thiếu. Các giá trị thiếu này đã được xử lý bằng cách điền giá trị trung vị (do phân phối lệch).
  - Loại bỏ các thuộc tính không cần thiết: Các cột như Unnamed: 0 và id được loại bỏ vì chúng không cung cấp thông tin hữu ích cho việc dự đoán.
  - Mã hóa dữ liệu hạng mục: Các thuộc tính hạng mục như Gender, Customer Type, satisfaction được chuyển đổi thành dạng số sử dụng kỹ thuật Mã hóa Nhãn để các thuật toán học máy có thể học được.
  - Phân chia dữ liệu: Dữ liệu được cung cấp sẵn dưới dạng tập huấn luyện và tập kiểm tra nên không cần phải chia nữa. Tập huấn luyện được sử dụng để xây dựng mô hình và tập kiểm tra được sử dụng để đánh giá hiệu năng của mô hình trên dữ liệu chưa từng thấy.
  - Chuẩn hóa: Đối với một số thuật toán nhạy cảm với phạm vi giá trị của các thuộc tính như KNN, SVM, Neural Network, cần áp dụng các kỹ thuật chuẩn hóa để đưa dữ liệu về cùng một phạm vi.

- **Kỹ thuật đặc trưng:**
  - Mục tiêu là chọn ra tập hợp các thuộc tính quan trọng nhất giúp cải thiện hiệu suất mô hình và giảm độ phức tạp. Bằng cách sử dụng sử dụng các phương pháp như Recursive Feature Elimination (RFE) và Permutation Importance để xác định và

chọn ra các đặc trưng quan trọng nhất, giúp giảm chiều dữ liệu, tăng hiệu suất và cải thiện khả năng tổng quát của mô hình.

  - Ngoài ra, còn kỹ thuật khác là tạo ra các đặc trưng mới từ các đặc trưng hiện có để cung cấp thêm thông tin cho mô hình như kết hợp các đánh giá dịch vụ, tạo các biến tương tác.

- **Xây dựng và Huấn luyện mô hình:**
  Thử nghiệm nhiều thuật toán phân loại khác nhau phù hợp với bài toán nhị phân và huấn luyện trên tập dữ liệu huấn luyện đã qua tiền xử lý và lựa chọn đặc trưng.

- **Đánh giá mô hình:**
  Sử dụng các tiêu chí đánh giá phù hợp cho bài toán phân loại để đo lường hiệu suất của các mô hình trên và tập kiểm tra.

- **Tinh chỉnh tham số:**

  - Đối với mô hình có hiệu suất tốt, áp dụng các kỹ thuật tối ưu hóa siêu tham số để tìm ra bộ tham số tốt nhất, nhằm cải thiện hơn về hiệu suất của mô hình.

  - Kỹ thuật tinh chỉnh siêu tham số bao gồm:
    Các kỹ thuật thông dụng Grid Search, Random Search hoặc các phương pháp nâng cao và hiện đại hơn là Bayesian Optimization, Gradient-based Optimization, Asynchronous Successive Halving Algorithm (ASHA), Hyperband và Optuna để tìm ra bộ siêu tham số tối ưu cho mô hình tốt nhất.

- **Diễn giải kết quả và Kết luận:**
  Phân tích kết quả, so sánh hiệu suất giữa các mô hình, rút ra kết luận về mô hình tốt nhất và các yếu tố ảnh hưởng đến sự hài lòng của hành khách.

3.2 Nguyên lý hoạt động của thuật toán

Trong quá trình thực nghiệm, nhiều mô hình phân loại đã được thử nghiệm để tìm ra giải pháp tối ưu nhất cho bài toán. Dưới đây là nguyên lý hoạt động của từng mô hình đã được sử dụng:

3.2.1 Logistic Regression:

- **Nguyên lý hoạt động:** Mặc dù có tên là hồi quy, Logistic Regression là một thuật toán phân loại được sử dụng để dự đoán xác suất một sự kiện xảy ra. Nó sử dụng hàm sigmoid (hoặc logistic) để ánh xạ bất kỳ giá trị thực nào vào một giá trị trong khoảng [0,1], sau đó sử dụng một ngưỡng để phân loại.
- **Ưu điểm:** Đơn giản, dễ hiểu, hiệu quả cho các bài toán phân loại nhị phân, cung cấp xác suất dự đoán.
- **Nhược điểm:** Giả định mối quan hệ tuyến tính giữa các đặc trưng và log-odds của biến mục tiêu, có thể không hiệu quả với dữ liệu phức tạp, phi tuyến tính.

3.2.2 Naive Bayes:

- **Nguyên lý hoạt động:** Naive Bayes là một thuật toán phân loại dựa trên Định lý Bayes với giả định ngây thơ về sự độc lập có điều kiện giữa các đặc trưng khi biết lớp. Gaussian Naive Bayes là một biến thể phù hợp với các đặc trưng có phân phối liên tục, giả định rằng các đặc trưng tuân theo phân phối Gaussian (hay phân phối chuẩn).
  - Tính toán xác suất tiền nghiệm của mỗi lớp.
  - Tính toán xác suất có điều kiện của mỗi đặc trưng cho mỗi lớp (dựa trên phân phối Gaussian).
  - Sử dụng Định lý Bayes để tính xác suất hậu nghiệm của mỗi lớp và chọn lớp có xác suất cao nhất.
- **Ưu điểm:** Đơn giản, nhanh, hiệu quả với dữ liệu lớn, hoạt động tốt ngay cả với ít dữ liệu huấn luyện.
- **Nhược điểm:** Giả định độc lập giữa các đặc trưng hiếm khi đúng trong thực tế, có thể ảnh hưởng đến độ chính xác.

3.2.3 K-Nearest Neighbor:

- **Nguyên lý hoạt động:** KNN là một thuật toán học không tham số (non-parametric) dựa trên khoảng cách. Để phân loại một điểm dữ liệu mới, nó tìm K điểm dữ liệu gần nhất trong tập huấn luyện bằng cách dựa trên một độ đo khoảng cách như Euclidean. Đối với bài toán phân loại thì điểm dữ liệu mới được gán cho lớp chiếm đa số trong số K láng giềng gần nhất.
- **Ưu điểm:** Đơn giản, dễ cài đặt, không cần huấn luyện mô hình (lazy learner), có thể xử lý các đường biên quyết định phức tạp.
- **Nhược điểm:** Tốn kém về mặt tính toán khi tập dữ liệu lớn (cần tính khoảng cách đến tất cả các điểm), nhạy cảm với các đặc trưng không liên quan và thang đo dữ liệu.

3.2.4 Decision Tree:

- **Nguyên lý hoạt động:** Cây quyết định là một mô hình phân loại và hồi quy dạng cây. Nó phân chia dữ liệu thành các tập con nhỏ hơn dựa trên các quy tắc quyết định đơn giản (các câu hỏi về đặc trưng). Quá trình này được lặp lại cho đến khi các nút lá đạt được độ thuần khiết nhất định hoặc đạt đến một tiêu chí dừng. Tại mỗi nút, thuật toán chọn đặc trưng và ngưỡng phân tách tối ưu để tối đa hóa độ lợi thông tin (hay Information Gain) hoặc giảm thiểu độ không thuần khiết (hay hệ Gini Impurity).

- **Ưu điểm:** Dễ hiểu, dễ giải thích, không yêu cầu tiền xử lý dữ liệu phức tạp, có thể xử lý cả dữ liệu định tính và định lượng.

- **Nhược điểm:** Dễ bị quá khớp với dữ liệu huấn luyện, không ổn định (thay đổi nhỏ trong dữ liệu có thể tạo ra cây rất khác).

3.2.5 Neural Network - MLP:

- **Nguyên lý hoạt động:** Mạng nơ-ron đa lớp (Multi-layer Perceptron) là một loại mạng nơ-ron truyền thẳng (feedforward neural network) bao gồm ít nhất ba lớp: lớp đầu vào, một hoặc nhiều lớp ẩn và lớp đầu ra. Mỗi nơ-ron trong một lớp được kết nối với tất cả các nơ-ron trong lớp tiếp theo.

  – Dữ liệu được truyền qua các lớp, với mỗi nơ-ron thực hiện tổng trọng số của các đầu vào và áp dụng một hàm kích hoạt phi tuyến tính.

  – Quá trình huấn luyện sử dụng thuật toán lan truyền ngược (hay là backpropagation) để điều chỉnh trọng số của các kết nối, nhằm giảm thiểu sai số giữa đầu ra dự đoán và đầu ra thực tế.

- **Ưu điểm:** Khả năng học các mối quan hệ phức tạp, phi tuyến tính trong dữ liệu, hiệu quả cao với dữ liệu lớn và cấu trúc phức tạp.

- **Nhược điểm:** Được xem là Hộp đen vì khó giải thích, yêu cầu nhiều dữ liệu, tốn kém về mặt tính toán, nhạy cảm với việc khởi tạo trọng số và lựa chọn siêu tham số.

3.2.6 Random Forest:

- **Nguyên lý hoạt động:** Random Forest là một thuật toán học máy mạnh mẽ thuộc loại Ensemble Learning, cụ thể là Bootstrap Aggregating (Bagging). Nguyên lý hoạt động của nó dựa trên việc xây dựng một "rừng" gồm nhiều cây quyết định độc lập và sau đó tổng hợp kết quả của chúng để đưa ra dự đoán cuối cùng.

  – Lấy mẫu Bootstrap (Bootstrapping): Từ tập dữ liệu huấn luyện ban đầu, Random Forest tạo ra nhiều tập con khác nhau bằng cách lấy mẫu có hoàn lại (bootstrap samples). Mỗi tập con có cùng kích thước với tập dữ

liệu gốc nhưng có thể chứa các bản ghi lặp lại và bỏ sót một số bản ghi khác.

  - Xây dựng cây quyết định: Đối với mỗi tập con được lấy mẫu, một cây quyết định được xây dựng. Tuy nhiên, có một điểm khác biệt quan trọng so với cây quyết định thông thường: tại mỗi bước tách nút của cây, thuật toán chỉ xem xét một tập hợp con ngẫu nhiên các đặc trưng (features) thay vì tất cả các đặc trưng. Điều này giúp giảm thiểu sự tương quan giữa các cây và tăng tính đa dạng của rừng.
  - Tổng hợp kết quả (Aggregation): Đối với bài toán phân loại thì việc tổng hợp kết quả bằng cách mỗi cây trong rừng sẽ đưa ra một "phiếu bầu" cho lớp mà nó dự đoán. Lớp có số phiếu bầu cao nhất sẽ là dự đoán cuối cùng của Random Forest.
  - Giảm phương sai và chống quá khớp: Bằng cách kết hợp nhiều cây quyết định được xây dựng trên các tập dữ liệu con và tập đặc trưng con ngẫu nhiên, Random Forest giúp giảm đáng kể phương sai (variance) của mô hình và chống lại hiện tượng quá khớp (overfitting) so với một cây quyết định đơn lẻ.

- **Ưu điểm:** Độ chính xác cao, khả năng chống quá khớp tốt, có thể xử lý dữ liệu có nhiều đặc trưng, cung cấp tầm quan trọng của đặc trưng.

- **Nhược điểm:** Ít giải thích được hơn cây quyết định đơn lẻ, tốn kém hơn về mặt tính toán.

3.2.7 Adaptive Boosting:

- **Nguyên lý hoạt động:** AdaBoost là một thuật toán boosting, xây dựng một mô hình mạnh mẽ bằng cách kết hợp nhiều bộ phân loại "yếu" (weak learners), thường là các cây quyết định nông (stumps). Nó huấn luyện các bộ phân loại yếu một cách tuần tự, tập trung vào các mẫu mà các bộ phân loại trước đó đã phân loại sai.

  - Khởi tạo trọng số đều cho tất cả các mẫu.
  - Huấn luyện một bộ phân loại yếu trên dữ liệu.
  - Tăng trọng số của các mẫu bị phân loại sai để bộ phân loại tiếp theo tập trung vào chúng.
  - Giảm trọng số của các mẫu được phân loại đúng.
  - Gán trọng số cho mỗi bộ phân loại yếu dựa trên độ chính xác của nó.
  - Lặp lại quá trình cho đến khi đạt được số lượng bộ phân loại mong muốn hoặc hiệu suất không cải thiện.
  - Dự đoán cuối cùng là tổng trọng số của dự đoán từ tất cả các bộ phân loại yếu.

- **Ưu điểm:** Hiệu quả cao, khả năng chống quá khớp tốt, đơn giản để cài đặt.

- **Nhược điểm:** Nhạy cảm với dữ liệu nhiễu và các giá trị ngoại lai, có thể tốn thời gian nếu số lượng bộ phân loại yếu lớn.

3.2.8 Support Vector Machine:

- **Nguyên lý hoạt động:** SVM là một thuật toán phân loại mạnh mẽ tìm kiếm một siêu phẳng (hyperplane) tối ưu để phân tách các lớp dữ liệu. Siêu phẳng này được chọn sao cho khoảng cách (margin) đến các điểm dữ liệu gần nhất của mỗi lớp (gọi là support vectors) là lớn nhất. Đối với bài toán phân loại thì SVM sử dụng kỹ thuật "kernel trick" để ánh xạ dữ liệu vào một không gian chiều cao hơn, nơi dữ liệu có thể phân tách tuyến tính. Các hàm kernel phổ biến bao gồm Linear, Polynomial, Radial Basis Function (RBF).

- **Ưu điểm:** Hiệu quả cao trong không gian chiều cao, mạnh mẽ với dữ liệu có nhiều chiều, hiệu quả khi số lượng đặc trưng lớn hơn số lượng mẫu.

- **Nhược điểm:** Tốn kém về mặt tính toán với tập dữ liệu lớn, khó chọn hàm kernel và siêu tham số phù hợp, không cung cấp xác suất dự đoán trực tiếp.

3.3 Cách cài đặt mô hình, các tham số của mô hình

- **Môi trường cài đặt:** Các mô hình được cài đặt và huấn luyện trong môi trường Google Colaboratory.

- **Thư viện:** Chủ yếu sử dụng thư viện Scikit-learn trong Python cho việc triển khai các thuật toán học máy.

- **Tham số mô hình:**

  - Đối với hầu hết các mô hình ban đầu (Logistic Regression, Random Forest, Naive Bayes, KNN, Decision Tree, MLP, AdaBoost, SVM) đều được sử dụng các tham số mặc định của thư viện Scikit-learn để có một đánh giá cơ sở.
  - Riêng đối với mô hình Random Forest là mô hình tốt nhất cho bài toán này. Nên được tái sử dụng lại và tinh chỉnh bằng các tham số mới. Quá trình tinh chỉnh tìm ra siêu tham số tối đã được thực hiện và kết quả cho ra như sau:
    * criterion: entropy (Tiêu chí để đo chất lượng của một lần tách. Entropy đo độ hỗn loạn của thông tin).
    * max_depth: 20 (Độ sâu tối đa của cây. Giới hạn độ sâu giúp kiểm soát độ phức tạp của mô hình và tránh quá khớp).
    * max_features: 0.5 (Số lượng đặc trưng cần xem xét khi tìm kiếm sự phân tách tốt nhất. Ở đây là 50% tổng số đặc trưng).

    * min_samples_leaf: 1 (Số lượng mẫu tối thiểu cần có ở một nút lá. Đảm bảo mỗi lá có đủ dữ liệu để tránh quá khớp).
    * min_samples_split: 5 (Số lượng mẫu tối thiểu cần có để một nút có thể được tách. Đảm bảo rằng mỗi lần tách là có ý nghĩa).
    * n_estimators: 500 (Số lượng cây trong rừng. Số lượng cây càng lớn, mô hình càng mạnh và ổn định hơn nhưng cũng tốn thời gian tính toán hơn).

3.4 Tiêu chí đánh giá mô hình

Vì bài toán này là một bài toán phân loại nhị phân (hài lòng/không hài lòng). Các tiêu chí sau được sử dụng để đánh giá hiệu suất của các mô hình:

- **Accuracy (Độ chính xác):**
  - Công thức: (Số dự đoán đúng) / (Tổng số dự đoán)
  - Ý nghĩa: Tỷ lệ các mẫu được phân loại chính xác trên tổng số mẫu. Tuy nhiên, trong trường hợp dữ liệu mất cân bằng, accuracy có thể không phản ánh đầy đủ hiệu suất của mô hình.
- **Precision (Độ chuẩn xác):**
  - Công thức cho lớp dương: TP / (TP + FP)
  - Ý nghĩa: Trong số các trường hợp được dự đoán là "hài lòng", có bao nhiêu trường hợp thực sự "hài lòng". Precision cao cho thấy mô hình ít mắc lỗi dự đoán sai một trường hợp âm thành dương.
- **Recall (Độ phủ, hay Độ nhạy - Sensitivity):**
  - Công thức (cho lớp dương): TP / (TP + FN)
  - Ý nghĩa: Trong số tất cả các trường hợp thực sự "hài lòng", mô hình đã phát hiện (dự đoán đúng) được bao nhiêu trường hợp. Recall cao cho thấy mô hình ít bỏ sót các trường hợp dương.
- **F1-score:**
  - Công thức: 2 * (Precision * Recall) / (Precision + Recall)
  - Ý nghĩa: Là trung bình điều hòa của Precision và Recall. F1-score hữu ích khi cần cân bằng giữa Precision và Recall, đặc biệt khi có sự mất cân bằng dữ liệu.
- **AUC-ROC (Area Under the ROC Curve - Diện tích dưới đường cong ROC):**

- Đường cong ROC (Receiver Operating Characteristic) biểu diễn mối quan hệ giữa tỷ lệ True Positive Rate (Recall) và False Positive Rate (1 - Specificity) ở các ngưỡng phân loại khác nhau.
- AUC đo lường khả năng của mô hình trong việc phân biệt giữa các lớp. Giá trị AUC càng gần 1 thì mô hình càng tốt. AUC = 0.5 tương ứng với một mô hình dự đoán ngẫu nhiên. Đây là một độ đo dùng để đánh giá tốt, đặc biệt với dữ liệu mất cân bằng.

Trong đó:

- True Positives (TP): Số lượng mẫu dương được dự đoán đúng là dương.
- False Positives (FP): Số lượng mẫu âm bị dự đoán sai là dương.
- False Negatives (FN): Số lượng mẫu dương bị dự đoán sai là âm.
- True Negatives (TN): Số lượng mẫu âm được dự đoán đúng là âm.

Chương 4

Thực nghiệm

4.1 Yêu cầu về chương trình

Để thực hiện đề tài này, môi trường phát triển và các thư viện sau đây đã được sử dụng:

- **Ngôn ngữ lập trình:** Python 3.11.
- **Môi trường phát triển:** Google Colaboratory (Colab). Google Colab cung cấp một môi trường Jupyter Notebook dựa trên đám mây, cho phép chạy mã Python với quyền truy cập miễn phí vào GPU/TPU, rất phù hợp cho các tác vụ học máy và xử lý dữ liệu lớn.
- **Thư viện chính được sử dụng:**
  - pandas: Để xử lý và phân tích dữ liệu dạng bảng.
  - numpy: Để thực hiện các phép toán số học hiệu quả.
  - matplotlib.pyplot: Để tạo các biểu đồ tĩnh.
  - seaborn: Để tạo các biểu đồ thống kê hấp dẫn và thông tin.
  - Seaborn: Dựa trên Matplotlib, cung cấp giao diện cấp cao hơn để vẽ các biểu đồ thống kê hấp dẫn và nhiều thông tin.
  - sklearn (scikit-learn): Thư viện chính cho các thuật toán học máy. Ngoài chứa các thuật toán học máy thì chứa công cụ tiền xử lý, lựa chọn đặc trưng, đánh giá mô hình và tinh chỉnh siêu tham số. Các module cụ thể bao gồm:
    * sklearn.model_selection: Để chia dữ liệu, thực hiện kiểm định chéo và tìm kiếm siêu tham số (Halving Grid Search và Halving Random Search).
    * sklearn.preprocessing: Để tiền xử lý dữ liệu (Label Encoder, Standard Scaler).
    * sklearn.impute: Để xử lý dữ liệu thiếu (Simple Imputer).
    * sklearn.ensemble: Chứa các thuật toán ensemble như Random Forest Classifier, AdaBoost Classifier.

    * sklearn.linear_model: Chứa LogisticRegression.
    * sklearn.naive_bayes: Chứa GaussianNB.
    * sklearn.neighbors: Chứa KNeighborsClassifier.
    * sklearn.tree: Chứa DecisionTreeClassifier.
    * sklearn.neural_network: Chứa MLPClassifier.
    * sklearn.svm: Chứa SVC (Support Vector Classifier).
    * sklearn.metrics: Để đánh giá hiệu suất mô hình.
    * sklearn.feature_selection: Chứa RFE.
    * sklearn.inspection: Chứa permutation_importance.

- **Cấu hình máy tính:** Mặc dù Google Colab cung cấp tài nguyên đám mây nên cấu hình máy tính cục bộ không quá quan trọng. Tuy nhiên, để làm việc hiệu quả với các tập dữ liệu lớn và các mô hình phức tạp, việc có kết nối internet ổn định là cần thiết. Đối với các tác vụ nặng về tính toán như huấn luyện Neural Network hoặc SVM, việc sử dụng GPU hay TPU trên Colab là rất hữu ích.

4.2 Mô tả chi tiết các bước khai phá dữ liệu

Mô tả chi tiết quá trình khai phá dữ liệu dựa trên các bước thực nghiệm đã được thực hiện, cùng với đầu vào và đầu ra của mỗi bước. Chi tiết thực hiện như sau:

- **Thu thập và tải dữ liệu**

  Input: Hai tệp dữ liệu train.csv và test.csv từ bộ dữ liệu "Airline Passenger Satisfaction" trên Kaggle.

  Hoạt động: Tải dữ liệu vào môi trường Google Colab. Đọc dữ liệu từ các tệp CSV vào các đối tượng DataFrame của Pandas.

  Output: Hai DataFrame: data_train chứa dữ liệu huấn luyện và data_test chứa dữ liệu kiểm tra.

- **Khám phá và Tiền xử lý dữ liệu**

  Input: DataFrame data_train và data_test.

  Hoạt động:

  – Kiểm tra thông tin chung: Xem xét số lượng hàng, cột, kiểu dữ liệu của từng cột, thống kê mô tả cơ bản.

  – Xử lý dữ liệu thiếu:

    * Kiểm tra các giá trị thiếu trong mỗi cột. Dữ liệu này chỉ có cột Arrival Delay in Minutes giá trị thiếu.
    * Thực hiện điền giá trị thiếu. Một phương pháp phổ biến cho cột số có phân phối lệch là điền bằng giá trị trung vị.

- Loại bỏ cột không cần thiết: Loại bỏ cột Unnamed: 0 và id khỏi cả hai DataFrame vì chúng không mang giá trị dự đoán.
- Mã hóa biến mục tiêu (satisfaction): Chuyển đổi giá trị 'neutral or dissatisfied' thành 0 và 'satisfied' thành 1.
- Mã hóa các biến hạng mục (Categorical Feature Encoding): Sử dụng LabelEncoder từ Scikit-learn để chuyển đổi các cột hạng mục còn lại (Gender, Customer Type, Type of Travel, Class) thành dạng số.
- Phân tách thuộc tính và biến mục tiêu:
  * Trong data_train, tách riêng các thuộc tính đầu vào (X_train) và biến mục tiêu (y_train).
  * Trong data_test, tách riêng các thuộc tính đầu vào (X_test) và biến mục tiêu (y_test).
- Trực quan hóa dữ liệu: Tạo các biểu đồ để hiểu rõ hơn về phân phối dữ liệu và mối quan hệ giữa các biến.

Output:

- X_train, y_train: Dữ liệu huấn luyện đã được làm sạch và mã hóa.
- X_test, y_test: Dữ liệu kiểm tra đã được làm sạch và mã hóa.
- Các biểu đồ và thống kê mô tả cung cấp cái nhìn sâu sắc về dữ liệu.

- **Lựa chọn đặc trưng**

  Input: X_train, y_train.

  Hoạt động:

  - Recursive Feature Elimination (RFE):
    * Khởi tạo một mô hình ước lượng RandomForestClassifier.
    * Áp dụng RFE với mô hình ước lượng này để chọn ra số lượng đặc trưng mong muốn hoặc xếp hạng tất cả các đặc trưng.
  - Permutation Importance:
    * Huấn luyện một mô hình cơ sở
    * Tính toán Permutation Importance trên tập kiểm tra để đánh giá tầm quan trọng của từng đặc trưng.
  - Lựa chọn tập đặc trưng cuối cùng:
    Dựa vào kết quả từ RFE và Permutation Importance, quyết định tập hợp các đặc trưng sử dụng để huấn luyện các mô hình cuối cùng. Có thể chọn tất cả các đặc trưng gốc hoặc một tập con đặc trưng mới dựa trên xếp hạng.

  Output:

  - Danh sách các đặc trưng được xếp hạng theo tầm quan trọng.

- X_train_selected, X_test_selected: Dữ liệu chỉ chứa các đặc trưng đã được lựa chọn.

- **Huấn luyện và Đánh giá các mô hình phân loại**

  Input: X_train_selected, y_train, X_test_selected, y_test.

  Hoạt động:

  - Với mỗi thuật toán phân loại sau:
    * Logistic Regression
    * Naive Bayes (GaussianNB)
    * K-Nearest Neighbors
    * Decision Tree
    * Neural Network (MLPClassifier)
    * Random Forest
    * AdaBoost
    * Support Vector Machine (SVM)
  - Thực hiện các bước:
    * Khởi tạo mô hình với các tham số mặc định (hoặc các tham số cơ sở).
    * Huấn luyện mô hình trên X_train_selected và y_train. Ghi nhận thời gian huấn luyện.
    * Dự đoán nhãn trên X_test_selected.
    * Tính toán và ghi nhận các chỉ số đánh giá: Accuracy, AUC-ROC, Precision, Recall, F1 Score.

  Output:

  - Một bảng tổng hợp chứa các chỉ số đánh giá và thời gian huấn luyện cho từng mô hình.
  - Biểu đồ so sánh Accuray, AUC-ROC, thời gian huấn luyện giữa các mô hình.

- **Tinh chỉnh siêu tham số cho mô hình tốt nhất**

  Input: X_train_selected, y_train_selected và mô hình có hiệu suất tốt nhất từ Bước 4 (là Random Forest).

  Hoạt động:

  - Xác định không gian siêu tham số: Định nghĩa một lưới các giá trị siêu tham số tiềm năng cho Random Forest.
  - Sử dụng HalvingGridSearchCV:
    * Thực hiện tìm kiếm trên lưới siêu tham số bằng HalvingGridSearch. Kỹ thuật này bắt đầu với một lượng nhỏ tài nguyên (mẫu dữ liệu) và loại bỏ dần các ứng viên siêu tham số kém hiệu quả qua các vòng lặp, tập trung tài nguyên vào các ứng viên hứa hẹn hơn.

    - Đánh giá dựa trên roc_auc.
    - Ghi nhận thời gian thực hiện, bộ tham số tốt nhất và điểm CV tốt nhất.
  - Sử dụng HalvingRandomSearchCV:
    - Thực hiện tìm kiếm ngẫu nhiên trên không gian siêu tham số bằng HalvingRandomSearch. Tương tự như HalvingGridSearch nhưng chọn ngẫu nhiên các kết hợp tham số thay vì thử tất cả.
    - Đánh giá dựa trên roc_auc.
    - Ghi nhận thời gian thực hiện, bộ tham số tốt nhất và điểm CV tốt nhất.

  Output:

  - Thông tin chi tiết về quá trình chạy.
  - Bộ siêu tham số tốt nhất và điểm số CV tương ứng cho mỗi phương pháp tinh chỉnh bằng Best parameters và Best CV score.
  - Bảng tóm tắt so sánh kết quả của hai phương pháp.

- **Đánh giá mô hình sau khi tinh chỉnh**

  Input: Mô hình Random Forest với bộ siêu tham số tốt nhất tìm được, X_train_selected, y_train, X_test_selected, y_test.

  Hoạt động:

  - Huấn luyện lại mô hình Random Forest trên toàn bộ X_train_selected với bộ siêu tham số tối ưu.
  - Đánh giá hiệu suất của mô hình đã tinh chỉnh này trên X_test_selected bằng các chỉ số Accuracy, AUC-ROC, Precision, Recall, F1 Score.

  Output: Các chỉ số đánh giá cho mô hình Random Forest đã được tối ưu hóa.

Chương 5

Kết quả

Sau quá trình tiền xử lý dữ liệu, lựa chọn đặc trưng, huấn luyện và đánh giá các mô hình phân loại khác nhau và tinh chỉnh siêu tham số. Mô hình Random Forest đã được chọn làm mô hình cuối cùng do hiệu suất vượt trội và khả năng tổng quát tốt. Mô hình này được tinh chỉnh bằng HalvingGridSearchCV và cho thấy kết quả ấn tượng trên tập dữ liệu kiểm tra.

5.1 Kết quả khai phá dữ liệu:

5.1.1 Kết quả lựa chọn đặc trưng

- Recursive Feature Elimination (RFE):

  Các đặc trưng được RFE đánh giá là quan trọng nhất ở Ranking 1 bao gồm: Type of Travel, Age, Flight Distance, Class, Inflight wifi service, Inflight entertainment, Online boarding, Ease of Online booking, On-board service, Leg room service, Seat comfort.

- Permutation Importance:

  Các đặc trưng có Mean_Importance cao nhất bao gồm:

  - Type of Travel (0.128261)
  - Inflight wifi service (0.127749)
  - Customer Type (0.052348)
  - Online boarding (0.034436)
  - Checkin service (0.023845)
  - Class (0.019526)

  Đáng chú ý, Gender có Mean_Importance âm (-0.000050), cho thấy việc hoán vị đặc trưng này không làm giảm hiệu suất mô hình, ngụ ý rằng nó có thể không phải là một yếu tố dự đoán quan trọng trong mô hình cụ thể này.

- Sự nhất quán giữa các phương pháp lựa chọn đặc trưng:

  Có sự tương đồng giữa hai phương pháp. Type of Travel, Inflight wifi service, Online boarding, Class đều xuất hiện trong top các đặc trưng quan trọng ở cả RFE và Permutation Importance. Điều này củng cố tầm quan trọng của các yếu tố này đối với sự hài lòng của hành khách.

5.1.2 Hiệu suất của các mô hình phân loại (trước khi tinh chỉnh)

Bảng dưới đây tóm tắt hiệu suất của các mô hình phân loại đã được huấn luyện và đánh giá trên tập kiểm tra:

Bảng 5.1: Kết quả đánh giá các mô hình học máy

Mô hình	Thời gian HL (s)	Accuracy	AUC-ROC	Precision	Recall	F1 Score
Logistic Regression	0.3453	0.8668	0.9221	0.8614	0.8302	0.8455
Naive Bayes (GaussianNB)	0.0578	0.8703	0.9293	0.8812	0.8144	0.8465
K-Nearest Neighbor	0.3817	0.9358	0.9736	0.9600	0.8908	0.9241
Decision Tree	0.2795	0.9459	0.9509	0.9429	0.9333	0.9380
Neural Network (MLP)	89.0628	0.9578	0.9925	0.9701	0.9326	0.9510
Random Forest	8.2405	0.9564	0.9893	0.9609	0.9390	0.9498
AdaBoost	2.1670	0.9033	0.9676	0.8881	0.8920	0.8901
Support Vector Machine (SVM)	1364.3680	0.9537	0.9881	0.9599	0.9335	0.9465

Phân tích kết quả:

- Mô hình có độ chính xác (Accuracy) và AUC-ROC cao nhất:
  - Neural Network (MLP) đạt được Accuracy cao nhất (0.9578) và AUC-ROC cao nhất (0.9925). Điều này cho thấy MLP có khả năng phân biệt rất tốt giữa hai lớp "hài lòng" và "không hài lòng". Precision của MLP cũng rất cao (0.9701), nghĩa là khi mô hình dự đoán một hành khách hài lòng, khả năng cao là dự đoán đó đúng.
  - Random Forest cũng cho kết quả rất tốt, với Accuracy là 0.9564 và AUC-ROC là 0.9893. Mô hình này có Recall cao nhất (0.9390), cho thấy nó có khả năng phát hiện tốt các trường hợp hành khách thực sự hài lòng.
  - Support Vector Machine (SVM) cũng đạt hiệu suất cao tương tự (với Accuracy 0.9537, AUC-ROC 0.9881), tuy nhiên thời gian huấn luyện của SVM là lâu nhất (1364.3680 giây), đây là một nhược điểm lớn.
- Các mô hình khác:
  - K-Nearest Neighbor và Decision Tree cũng cho kết quả khá tốt, với Accuracy lần lượt là 0.9358 và 0.9459.
  - Logistic Regression, Naive Bayes và AdaBoost có hiệu suất thấp hơn so với các mô hình còn lại, nhưng vẫn ở mức chấp nhận được (Accuracy $>$ 0.86).

- Thời gian huấn luyện:
  - Naive Bayes có thời gian huấn luyện nhanh nhất (0.0578 giây).
  - MLP và đặc biệt là SVM yêu cầu thời gian huấn luyện đáng kể. Random Forest có thời gian huấn luyện trung bình (8.2405 giây).

Kết luận về mô hình chính xác nhất (trước tinh chỉnh): Dựa trên các tiêu chí đánh giá, Neural Network (MLP) và Random Forest là hai mô hình nổi bật nhất, với MLP nhỉnh hơn một chút về Accuracy và AUC-ROC, trong khi Random Forest có Recall tốt hơn và thời gian huấn luyện nhanh hơn đáng kể so với MLP.

5.1.3 Kết quả tinh chỉnh siêu tham số cho Random Forest

Quá trình tinh chỉnh siêu tham số cho mô hình Random Forest mang lại kết quả sau:

Bảng 5.2: Kết quả tinh chỉnh siêu tham số

Phương pháp	Thời gian tinh chỉnh (s)	Best CV Score (roc_auc)	Best Params
HalvingGridSearchCV	4842.002971	0.992751	{'criterion': 'entropy', 'max_depth': 20, 'max_features': 0.5, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 500}
HalvingRandomSearchCV	3125.234851	0.988711	{'n_estimators': 500, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 20, 'criterion': 'gini'}

- HalvingGridSearchCV đã tìm ra một bộ tham số cho Random Forest với điểm AUC-ROC trên tập validation (CV score) là 0.9928, cao hơn một chút so với điểm AUC-ROC ban đầu của Random Forest (0.9893) trên tập test. Điều này cho thấy tiềm năng cải thiện hiệu suất của Random Forest sau khi được tinh chỉnh.
- Thời gian tinh chỉnh khá lớn, đặc biệt với HalvingGridSearchCV.

5.1.4 Hiệu suất của mô hình phân loại tốt nhất Random Forest (sau khi tinh chỉnh)

Mô hình Random Forest cuối cùng đã được đánh giá trên tập kiểm tra độc lập để xác định khả năng tổng quát hóa của nó. Các tiêu chí đánh giá được sử dụng là Accuracy, AUC-ROC, Precision, Recall và F1-Score.

Các chỉ số này cho thấy mô hình Random Forest đạt độ chính xác rất cao trong việc phân loại sự hài lòng của hành khách. Đặc biệt, giá trị AUC-ROC gần 1 (0.9931) khẳng định khả năng phân biệt xuất sắc giữa hai lớp "hài lòng" và "không hài lòng" của mô hình.

Bảng 5.3: Hiệu suất của mô hình cuối cùng trên tập kiểm tra

Chỉ số	**Giá trị**	
Accuracy	0.9591	(95.91%)
AUC-ROC	0.9931	
Precision	0.9680	
Recall	0.9377	
F1 Score	0.9526	

Bảng 5.4: Báo cáo phân loại (Classification Report) của mô hình cuối cùng

Lớp	**precision**	**recall**	**f1-score**	**support**
0	0.95	0.98	0.96	14573
1	0.97	0.94	0.95	11403
accuracy		0.96		25976
macro avg	0.96	0.96	0.96	25976
weighted avg	0.96	0.96	0.96	25976

- Lớp 0 (Neutral or dissatisfied):
  - Precision: 0.95 (95% số hành khách được dự đoán là không hài lòng thực sự là không hài lòng).
  - Recall: 0.98 (98% số hành khách không hài lòng thực sự đã được mô hình nhận diện đúng).
  - F1-Score: 0.96.
- Lớp 1 (Satisfied):
  - Precision: 0.97 (97% số hành khách được dự đoán là hài lòng thực sự là hài lòng).
  - Recall: 0.94 (94% số hành khách hài lòng thực sự đã được mô hình nhận diện đúng).
  - F1-Score: 0.95.

Báo cáo phân loại cho thấy mô hình hoạt động rất tốt trên cả hai lớp, với độ chính xác cao và khả năng nhận diện tốt cả hai loại hành khách.

Confusion Matrix của mô hình cuối cùng, trong đó:

- **True Negative (TN)** là Số hành khách thực sự không hài lòng và được dự đoán là không hài lòng.

Bảng 5.5: Ma trận nhầm lẫn của mô hình dự đoán sự hài lòng của hành khách

	Dự đoán: Không hài lòng	**Dự đoán: Hài lòng**
Thực tế: Không hài lòng	TN: 14220	FP: 353
Thực tế: Hài lòng	FN: 710	TP: 10693

- **False Positive (FP)** là số hành khách thực sự hài lòng nhưng bị dự đoán sai là không hài lòng.
- **False Negative (FN)** là số hành khách thực sự không hài lòng nhưng bị dự đoán sai là hài lòng).
- **True Positive (TP)** là số hành khách thực sự hài lòng và được dự đoán là hài lòng).

5.1.5 Tính giải thích được của mô hình

- Mô hình dễ giải thích:
  - Decision Tree: Mặc dù không phải là mô hình tốt nhất về độ chính xác, Cây quyết định cung cấp các quy tắc rõ ràng, dễ hiểu về cách đưa ra dự đoán.
  - Logistic Regression: Cung cấp các trọng số cho từng đặc trưng, cho biết hướng và mức độ ảnh hưởng của đặc trưng đó đến xác suất hài lòng.
- Mô hình có khả năng giải thích thông qua tầm quan trọng đặc trưng:
  - Random Forest: Mặc dù bản thân là một "hộp đen" do cấu trúc gồm nhiều cây, Random Forest có thể cung cấp thông tin về tầm quan trọng của các đặc trưng, tương tự như kết quả từ Permutation Importance. Các đặc trưng như Type of Travel, Inflight wifi service, Customer Type, Online boarding được xác định là quan trọng, giúp hiểu được yếu tố nào tác động mạnh đến sự hài lòng.
- Mô hình khó giải thích:
  - Neural Network (MLP): Thường rất khó để hiểu chính xác tại sao MLP đưa ra một dự đoán cụ thể do cấu trúc phức tạp với nhiều lớp và kết nối phi tuyến.
  - SVM (với kernel phi tuyến): Cũng có thể khó giải thích trực tiếp.

Các mô hình đạt độ chính xác cao (MLP, Random Forest) có khả năng học được các mối quan hệ phức tạp và phi tuyến giữa các đặc trưng đầu vào và biến mục tiêu satisfaction. Chúng có thể nhận diện được các mẫu tinh vi trong dữ liệu mà các mô hình đơn giản hơn có thể bỏ qua. Việc các đặc trưng liên quan trực tiếp đến trải nghiệm dịch vụ (ví dụ: Inflight wifi service, Online boarding, Seat comfort)

và loại hình chuyến đi/khách hàng (Type of Travel, Class, Customer Type) được đánh giá là quan trọng cho thấy các mô hình đã học được những yếu tố thực sự ảnh hưởng đến cảm nhận của hành khách. Ví dụ, hành khách đi công tác (Type of Travel = Business travel) thường có kỳ vọng cao hơn về các dịch vụ tiện ích và hiệu quả, và mô hình có thể đã nắm bắt được điều này. Tương tự, chất lượng wifi (Inflight wifi service) là một yếu tố ngày càng quan trọng, đặc biệt với khách đi công tác hoặc những người muốn duy trì kết nối.

5.2 So sánh kết quả thực tế với kết quả dự đoán:

5.2.1 Chất lượng của dữ liệu:

- **Ưu điểm:**
  - **Dữ liệu tương đối đầy đủ:** Chỉ có một lượng rất nhỏ dữ liệu thiếu ở cột Arrival Delay in Minutes, giúp giảm thiểu sự phức tạp trong quá trình tiền xử lý và đảm bảo mô hình được huấn luyện trên một tập dữ liệu gần như hoàn chỉnh.
  - **Đa dạng đặc trưng:** Dữ liệu bao gồm nhiều loại đặc trưng khác nhau, bao quát nhiều khía cạnh của trải nghiệm chuyến bay, từ thông tin cá nhân đến các đánh giá dịch vụ. Điều này cung cấp đủ thông tin để mô hình học các mối quan hệ phức tạp.
  - **Phân bố biến mục tiêu tương đối cân bằng:** Tỷ lệ giữa "hài lòng" và "không hài lòng" không quá chênh lệch, giúp mô hình không bị thiên vị quá mức về một lớp nào đó.
- **Nhược điểm tiềm ẩn:**
  - Dữ liệu khảo sát có thể chứa đựng sự chủ quan.
  - Thang điểm đánh giá (0-5) có thể được hiểu khác nhau bởi những người khác nhau.
  - Có thể còn các yếu tố khác ảnh hưởng đến sự hài lòng chưa được thu thập. Ví dụ: thái độ cụ thể của tiếp viên, sự cố nhỏ trên chuyến bay không được ghi nhận.
  - Sự mất cân bằng nhẹ của biến mục tiêu như 56.7% không hài lòng/trung lập vs 43.3% hài lòng trong tập huấn luyện cần được lưu ý, nhưng các mô hình hàng đầu (MLP, Random Forest) vẫn đạt được Recall tốt cho lớp "hài lòng".

5.2.2 Nguyên lý hoạt động của thuật toán có giải quyết được vấn đề hay không?

Có. Các thuật toán phân loại được sử dụng đều là những công cụ mạnh mẽ và phù hợp cho bài toán dự đoán một biến mục tiêu hạng mục dựa trên các đặc trưng

đầu vào. Việc các mô hình như MLP và Random Forest đạt được độ chính xác và AUC-ROC trên 95% và 0.98-0.99 tương ứng trên tập kiểm tra cho thấy chúng đã học được các quy luật cơ bản từ dữ liệu huấn luyện và có khả năng tổng quát hóa tốt trên dữ liệu mới. Sự thành công này đến từ khả năng của các thuật toán này trong việc:

- Xử lý các mối quan hệ phi tuyến (MLP, Random Forest, SVM với kernel).
- Tương tác giữa các đặc trưng (Random Forest, MLP).
- Giảm thiểu overfitting (Random Forest thông qua bagging và chọn đặc trưng ngẫu nhiên; MLP có thể dùng các kỹ thuật điều chuẩn).

Chương 6

Kết luận

Đồ án đã giải quyết thành công bài toán phân loại mức độ hài lòng của hành khách bằng cách áp dụng quy trình khai phá dữ liệu một cách có hệ thống dựa trên tập dữ liệu "Airline Passenger Satisfaction". Nhiều thuật toán học máy đã được thử nghiệm. Các mô hình như Neural Network (MLP) và Random Forest đã chứng tỏ hiệu suất vượt trội trong việc dự đoán sự hài lòng. Các yếu tố quan trọng ảnh hưởng đến sự hài lòng cũng đã được xác định, cung cấp những hiểu biết giá trị cho các hãng hàng không.

Mặc dù có một số hạn chế về tính giải thích của các mô hình phức tạp và sự phụ thuộc vào dữ liệu, giải pháp này mang lại tiềm năng ứng dụng to lớn trong việc nâng cao trải nghiệm khách hàng và tối ưu hóa hoạt động kinh doanh của các hãng hàng không. Các đề xuất cải tiến và phát triển trong tương lai hứa hẹn sẽ làm cho giải pháp ngày càng hoàn thiện và hiệu quả hơn.

6.1 Khả năng ứng dụng của giải pháp:

Các mô hình phân loại mức độ hài lòng của hành khách được phát triển và các giải pháp dự đoán sự hài lòng của hành khách dựa trên khai phá dữ liệu có tiềm năng ứng dụng rộng rãi trong ngành hàng không:

- **Hệ thống cảnh báo sớm và can thiệp chủ động:**
  - Tích hợp mô hình vào hệ thống đặt vé hoặc làm thủ tục để dự đoán những hành khách có khả năng cao sẽ không hài lòng.
  - Thông tin này cho phép nhân viên mặt đất hoặc phi hành đoàn có những hành động can thiệp kịp thời (ví dụ: một lời chào đặc biệt, một ưu đãi nhỏ, hỏi thăm về nhu cầu) nhằm cải thiện trải nghiệm của họ trước hoặc trong chuyến bay.
- **Cá nhân hóa dịch vụ và trải nghiệm khách hàng:**

  Phân tích các yếu tố quan trọng đối với từng phân khúc khách hàng để cung cấp các dịch vụ hoặc ưu đãi phù hợp hơn. Ví dụ, ưu tiên chất lượng wifi cho khách đi công tác hoặc cung cấp thêm lựa chọn giải trí cho khách đi du lịch cá nhân.

- **Định hướng chiến lược cải tiến chất lượng dịch vụ:**

  Các đặc trưng có trọng số cao trong mô hình (ví dụ: Inflight wifi service, Online boarding, Seat comfort) chỉ ra những lĩnh vực mà việc cải thiện sẽ có tác động lớn nhất đến sự hài lòng chung. Hãng bay có thể ưu tiên đầu tư nguồn lực vào các khía cạnh này.

- **Tối ưu hóa chiến dịch marketing và chương trình khách hàng thân thiết:**

  - Xác định các nhóm khách hàng dễ hài lòng hoặc khó hài lòng để thiết kế các thông điệp marketing phù hợp.
  - Điều chỉnh các quyền lợi trong chương trình khách hàng thân thiết dựa trên những yếu tố mà khách hàng trung thành thực sự đánh giá cao.

- **Đo lường và theo dõi hiệu quả của các sáng kiến mới:**

  Khi hãng bay triển khai một dịch vụ mới hoặc cải tiến một quy trình, mô hình có thể được sử dụng để đánh giá tác động của những thay đổi đó lên mức độ hài lòng của hành khách theo thời gian.

- **Tối ưu hóa hoạt động và phân bổ nguồn lực:**

  Giúp hãng hàng không tập trung nguồn lực (nhân sự, tài chính) vào những yếu tố then chốt mang lại sự hài lòng cao nhất, thay vì đầu tư dàn trải.

- **Nâng cao lợi thế cạnh tranh:**

  Trong một thị trường cạnh tranh, việc chủ động thấu hiểu và đáp ứng vượt trội nhu cầu của hành khách sẽ giúp xây dựng lòng trung thành, thu hút khách hàng mới và tạo dựng danh tiếng thương hiệu vững chắc.

6.2 Ưu điểm – nhược điểm của giải pháp:

- **Ưu điểm:**

  - Độ chính xác cao: Các mô hình hàng đầu như Neural Network (MLP) và Random Forest đã đạt được độ chính xác và AUC-ROC rất cao (trên 95% và 0.99 tương ứng sau tinh chỉnh cho Random Forest), cho thấy khả năng dự đoán tốt.
  - Khả năng chống quá khớp tốt: Là một thuật toán ensemble, Random Forest ít bị quá khớp hơn so với các mô hình đơn lẻ như Decision Tree, đảm bảo khả năng tổng quát hóa tốt trên dữ liệu mới.
  - Xử lý dữ liệu đa dạng: Mô hình có khả năng xử lý cả dữ liệu định tính và định lượng, cũng như các mối quan hệ phi tuyến tính phức tạp giữa các đặc trưng.

- Xác định được các yếu tố quan trọng: Quá trình lựa chọn đặc trưng (RFE, Permutation Importance) và khả năng diễn giải của một số mô hình (ví dụ: feature importances từ Random Forest) đã giúp xác định các yếu tố chính ảnh hưởng đến sự hài lòng như Type of Travel, Inflight wifi service, Customer Type, Online boarding, Class.
- Mạnh mẽ với dữ liệu thiếu và ngoại lai: Random Forest khá mạnh mẽ trước các giá trị thiếu (đã được xử lý) và các giá trị ngoại lai (đặc biệt trong các cột chậm trễ), giúp mô hình ổn định hơn.
- Quy trình rõ ràng: Các bước khai phá dữ liệu được thực hiện theo một quy trình chuẩn, từ tiền xử lý đến tinh chỉnh siêu tham số, đảm bảo tính khoa học và có thể tái tạo.
- Khả năng tự động hóa: Sau khi được huấn luyện và triển khai, mô hình có thể tự động đưa ra dự đoán cho lượng lớn hành khách, giúp tiết kiệm thời gian và công sức so với các phương pháp thủ công.
- Quyết định dựa trên dữ liệu: Cung cấp một phương pháp tiếp cận khoa học, dựa trên bằng chứng từ dữ liệu để cải thiện dịch vụ, thay vì dựa trên cảm tính hay kinh nghiệm chủ quan.
- Sử dụng các công cụ và kỹ thuật tiên tiến: Việc áp dụng các thuật toán học máy hiện đại và các kỹ thuật tinh chỉnh siêu tham số cho thấy sự cập nhật với các phương pháp tiên tiến trong lĩnh vực.

- **Nhược điểm:**
  - Tính giải thích hạn chế của một số mô hình: Neural Network (MLP) là một mô hình "hộp đen", rất khó để giải thích cặn kẽ tại sao nó lại đưa ra một quyết định cụ thể. Điều này có thể gây khó khăn cho việc thuyết phục và triển khai trong một số môi trường doanh nghiệp. SVM với kernel phi tuyến cũng tương tự.
  - Thời gian huấn luyện và tinh chỉnh: Một số mô hình hiệu quả (MLP, SVM) yêu cầu thời gian huấn luyện đáng kể. Quá trình tinh chỉnh siêu tham số cũng rất tốn thời gian và tài nguyên tính toán.
  - Phụ thuộc vào chất lượng dữ liệu: Hiệu suất của mô hình phụ thuộc lớn vào chất lượng và tính đầy đủ của dữ liệu đầu vào. Nếu dữ liệu khảo sát có sai lệch, không đại diện, hoặc thiếu các yếu tố quan trọng, mô hình có thể không phản ánh đúng thực tế.
  - Dữ liệu tĩnh: Mô hình được xây dựng dựa trên một bộ dữ liệu tại một thời điểm nhất định. Kỳ vọng và hành vi của hành khách có thể thay đổi theo thời gian, đòi hỏi mô hình cần được cập nhật và huấn luyện lại định kỳ với dữ liệu mới.
  - Yêu cầu dữ liệu được mã hóa: Các biến định tính cần được mã hóa thành dạng số (Label Encoding) trước khi đưa vào mô hình. Mặc dù Label Encoding đơn giản, nó có thể áp đặt một thứ tự không có thật cho các danh mục, điều này có thể là một nhược điểm nếu các danh mục không có thứ tự tự nhiên.

- Nhạy cảm với dữ liệu huấn luyện (đối với Permutation Importance): Kết quả Permutation Importance có thể thay đổi một chút nếu mô hình được huấn luyện lại hoặc nếu có sự thay đổi nhỏ trong dữ liệu.
- Khó nắm bắt các yếu tố cảm tính và tình huống đặc biệt: Các mô hình dựa trên dữ liệu có cấu trúc có thể khó nắm bắt được các yếu tố cảm tính tinh tế hoặc các sự cố bất thường, đơn lẻ có thể ảnh hưởng lớn đến trải nghiệm của một hành khách cụ thể.

6.3 Bài học kinh nghiệm:

Để phát triển và ứng dụng giải pháp này một cách hiệu quả hơn trong tương lai với một số đề xuất sau:

- Xử lý dữ liệu nâng cao:
  - One-Hot Encoding: Thay vì Label Encoding, xem xét sử dụng One-Hot Encoding cho các biến định tính không có thứ tự tự nhiên (như Gender, Customer Type, Type of Travel, Class) để tránh việc mô hình hiểu sai mối quan hệ thứ tự.
  - Xử lý ngoại lai: Mặc dù Random Forest khá mạnh mẽ, việc xử lý các giá trị ngoại lai cực đoan trong các cột chậm trễ (Departure Delay in Minutes, Arrival Delay in Minutes) bằng các kỹ thuật như Winsorization hoặc log transformation có thể cải thiện hiệu suất của một số mô hình khác.
- Thu thập và tích hợp thêm dữ liệu đa dạng:
  - Dữ liệu văn bản từ phản hồi mở: Thu thập và phân tích ý kiến phản hồi dạng văn bản của hành khách (từ khảo sát, mạng xã hội, email) bằng các kỹ thuật Xử lý Ngôn ngữ Tự nhiên (NLP) để có cái nhìn sâu sắc hơn về lý do hài lòng/không hài lòng.
  - Dữ liệu vận hành chi tiết hơn: Tích hợp thêm dữ liệu về chuyến bay (ví dụ: loại máy bay, tuổi máy bay, tình trạng thời tiết cụ thể, thông tin về phi hành đoàn) để khám phá thêm các yếu tố tiềm ẩn.
  - Dữ liệu lịch sử tương tác của khách hàng: Nếu có, sử dụng lịch sử bay, các yêu cầu đặc biệt trước đó của khách hàng để cá nhân hóa dự đoán.
- Khám phá các thuật toán và kỹ thuật nâng cao hơn:
  - Các mô hình Ensemble phức tạp hơn: Thử nghiệm với các thuật toán vốn thường cho hiệu suất rất cao trong các bài toán phân loại dạng bảng.
  - Deep Learning cho dữ liệu bảng: Khám phá các kiến trúc mạng neural sâu hơn hoặc các mô hình chuyên biệt cho dữ liệu bảng nếu có đủ dữ liệu và tài nguyên tính toán.
  - Giải thích mô hình (Explainable AI - XAI): Áp dụng các kỹ thuật XAI để hiểu rõ hơn về quyết định của các mô hình "hộp đen" như MLP.

  - Phân tích tương tác giữa các đặc trưng để hiểu cách các yếu tố kết hợp với nhau ảnh hưởng đến sự hài lòng.

- Xây dựng hệ thống triển khai và giám sát mô hình (MLOps):
  - Thiết lập một quy trình để triển khai mô hình vào môi trường sản xuất.
  - Giám sát hiệu suất của mô hình theo thời gian và tự động cảnh báo khi hiệu suất suy giảm.
  - Xây dựng cơ chế để huấn luyện lại mô hình định kỳ với dữ liệu mới nhất.

- Thử nghiệm A/B Testing trong thực tế:

  Sau khi triển khai, thực hiện các thử nghiệm A/B để đánh giá tác động thực sự của các can thiệp dựa trên dự đoán của mô hình (ví dụ: so sánh mức độ hài lòng của nhóm khách hàng được can thiệp với nhóm đối chứng).

- Tập trung vào phân khúc khách hàng cụ thể:

  Xây dựng các mô hình riêng biệt cho từng phân khúc khách hàng quan trọng (ví dụ: khách hàng hạng thương gia, khách hàng thường xuyên) vì các yếu tố ảnh hưởng đến sự hài lòng của họ có thể khác nhau.

- Triển khai mô hình:

  Xây dựng một ứng dụng web hoặc API đơn giản để triển khai mô hình, cho phép hãng hàng không nhập dữ liệu hành khách mới và nhận được dự đoán về mức độ hài lòng một cách trực quan.

Tài liệu tham khảo

[1] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[2] A. Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2 edition, 2019.

[3] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3 edition, 2011.

[4] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, T. E. Oliphant, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

[5] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009.

[6] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(03):90–95, 2007.

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[8] Matplotlib development team. Matplotlib Documentation, Accessed 2025. URL https://matplotlib.org/.

[9] W. McKinney. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, 2010.

[10] Michael Waskom. Seaborn: statistical data visualization – Documentation, Accessed 2025. URL https://seaborn.pydata.org/.

[11] NumPy community. NumPy Documentation, Accessed 2025. URL https://numpy.org/.

[12] Pandas development team. Pandas: Python Data Analysis Library – Documentation, Accessed 2025. URL https://pandas.pydata.org/.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Joly, M. Perrot, and É. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] scikit-learn developers. Scikit-learn: Machine Learning in Python – Documentation, Accessed 2025. URL https://scikit-learn.org/.

[15] scikit-learn developers. 1.11. Ensemble methods, Accessed 2025. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.ensemble.

[16] scikit-learn developers. 3.13. Feature selection, Accessed 2025. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection.

[17] scikit-learn developers. 1.1. Linear Models, Accessed 2025. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model.

[18] scikit-learn developers. 3.3. Metrics and scoring: quantifying the quality of predictions, Accessed 2025. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics.

[19] scikit-learn developers. 3.2. Model selection and evaluation, Accessed 2025. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.model_selection.

[20] scikit-learn developers. 1.8. Naive Bayes, Accessed 2025. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.naive_bayes.

[21] scikit-learn developers. 1.6. Nearest Neighbors, Accessed 2025. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.neighbors.

[22] scikit-learn developers. 1.17. Neural network models (supervised), Accessed 2025. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.neural_network.

[23] scikit-learn developers. 3.3. Preprocessing data, Accessed 2025. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing.

[24] scikit-learn developers. 1.4. Support Vector Machines, Accessed 2025. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.svm.

[25] scikit-learn developers. 1.9. Decision Trees, Accessed 2025. URL https://scikit-learn.org/stable/modules/classes.html#module-sklearn.tree.

[26] P. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson Education, 2005.

[27] Teejmahal. Airline Passenger Satisfaction. Kaggle, 2022. URL https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction/data. Truy cập ngày 24 tháng 5, 2025.

[28] M. L. Waskom. Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.