

TRƯỜNG ĐẠI HỌC SÀI GÒN  
KHOA TOÁN - ỨNG DỤNG



ĐỀ TÀI

**Phân tích dữ liệu đặt phòng khách sạn bằng các  
kỹ thuật khai phá dữ liệu**

HỌ VÀ TÊN : NGUYỄN ĐĂNG TIẾN                  MSSV: 3123580050

HỌC PHẦN : KHAI PHÁ DỮ LIỆU

**GIẢNG VIÊN:**  
**TS. ĐỖ NHU TÀI**

Thành phố Hồ Chí Minh - 2025

### **Lời cảm ơn**

Lời nói đầu tiên cho em xin gửi lời cảm ơn chân thành đến **TS. Đỗ Như Tài** giảng viên bộ môn *Khai Phá Dữ Liệu* lớp DDU1231. Khoa Toán – Úng Dụng, Đại học Sài Gòn, người đã tận tình hướng dẫn, hỗ trợ và cung cấp nền tảng kiến thức vững chắc giúp em hoàn thành bài tiểu luận cuối kỳ này. Những góp ý và định hướng chuyên môn của thầy là yếu tố quan trọng giúp em hoàn thiện trong suốt quá trình nghiên cứu.

Em cũng không quên gửi lời cảm ơn tới sự giúp đỡ của các thành viên trong lớp, công nhân cán bộ viên chức của nhà trường đã tạo điều kiện, cung cấp cơ sở vật chất để em có thể học hỏi nghiên cứu. Một lần nữa xin gửi lời cảm ơn chân thành cũng như lời chúc sức khỏe đến quý thầy cô và các bạn đọc.

Do hạn chế về kiến thức và phạm vi nghiên cứu, bài tiểu luận khó tránh khỏi những thiếu sót. Em rất mong nhận được sự góp ý nhiệt tình từ quý thầy cô và các độc giả để em có thể rút kinh nghiệm và hoàn thiện hơn trong những nghiên cứu tiếp theo.

**Sinh viên thực hiện**

**Nguyễn Đăng Tiên**

## Tóm tắt

Trong bối cảnh ngành du lịch – khách sạn ngày càng cạnh tranh, việc dự báo chính xác nhu cầu đặt phòng và hiểu rõ hành vi khách hàng đóng vai trò then chốt trong tối ưu hóa vận hành, lập kế hoạch nhân sự và định giá chiến lược. Tuy nhiên, để đưa ra quyết định hiệu quả, các nhà quản trị cần trả lời những câu hỏi quan trọng sau: (1) **Nhu cầu và hành vi của khách hàng là gì?** (2) **Làm thế nào để phân khúc khách hàng hiệu quả?** (3) **Nhóm khách hàng nào có nguy cơ hủy phòng cao?** (4) **Dịch vụ nào thường được đặt kèm với từng loại phòng?** và (5) **Xu hướng nhu cầu đặt phòng trong 30–90 ngày tới sẽ như thế nào?**

Để giải quyết các vấn đề trên, nghiên cứu này nhằm dự báo nhu cầu đặt phòng khách sạn và phân tích hành vi khách hàng thông qua việc tích hợp bốn nhóm kỹ thuật khai phá dữ liệu: phân cụm, phân loại, luật kết hợp và chuỗi thời gian.

Dữ liệu lịch sử đặt phòng được tiền xử lý và phân tích mô tả để nhận diện đặc trưng theo mùa vụ cũng như các yếu tố ảnh hưởng đến nhu cầu lưu trú. Bốn kỹ thuật chính được áp dụng như sau: (1) K-means phân nhóm khách hàng theo hành vi đặt phòng; (2) Random Forest dự đoán khả năng hủy phòng; (3) Apriori xác định mẫu dịch vụ thường được đặt kèm; và (4) các mô hình chuỗi thời gian SARIMAX dự báo nhu cầu đặt phòng trong 30–90 ngày tới.

Kết quả cho thấy SARIMAX là mô hình dự báo hiệu quả nhất trong khi phân cụm K-means xác định được 6 nhóm khách hàng chính và luật kết hợp phát hiện 7 mẫu dịch vụ có liên quan mạnh. Nghiên cứu chứng minh rằng việc kết hợp đa mô hình cung cấp cái nhìn toàn diện về hoạt động khách sạn, từ đó hỗ trợ tối ưu hóa vận hành và hoạch định doanh thu hiệu quả.

## Mục lục

Lời cảm ơn .....	2
Tóm tắt .....	3
<b>CHƯƠNG I: GIỚI THIỆU .....</b>	<b>7</b>
1.1.    Bối cảnh nghiên cứu .....	7
1.2.    Lý do chọn đề tài.....	7
1.3.    Tính cấp thiết của đề tài .....	8
1.4.    Mục tiêu nghiên cứu.....	9
1.4.1.    Mục tiêu tổng quát .....	9
1.4.2.    Mục tiêu cụ thể .....	9
1.5.    Các câu hỏi nghiên cứu.....	10
1.6.    Đối tượng và phạm vi nghiên cứu.....	11
1.6.1.    Đối tượng nghiên cứu .....	11
1.6.2.    Phạm vi nghiên cứu .....	12
1.7.    Phương pháp nghiên cứu.....	13
1.8.    Cấu trúc tiểu luận .....	15
<b>CHƯƠNG II: CƠ SỞ LÝ THUYẾT .....</b>	<b>17</b>
2.1. Các khái niệm liên quan.....	17
2.1.1. Khai phá dữ liệu (Data Mining).....	17
2.1.2. Khái niệm cơ bản trong học máy .....	17
2.1.3. Kiểu dữ liệu .....	18
2.2. Lý thuyết về tiền xử lý dữ liệu .....	18
2.2.1. Khái niệm tiền xử lý dữ liệu .....	18
2.2.2. Các bước trong tiền xử lý dữ liệu .....	19
2.3. Lý thuyết về các thuật toán .....	21
2.3.1. Phân loại (Classification) .....	21
2.3.2. Phân cụm (Clustering).....	23
2.3.3. Luật kết hợp (Association Rules) .....	24
2.3.4. Chuỗi thời gian (Time Series) .....	25

<b>2.4. Nghiên cứu liên quan .....</b>	<b>26</b>
<b>CHƯƠNG III: DỮ LIỆU &amp; PHƯƠNG PHÁP ĐỀ XUẤT .....</b>	<b>28</b>
<b>3.1. Mô tả bảng dữ liệu .....</b>	<b>28</b>
<b>3.1.1. Tổng quan dữ liệu sử dụng.....</b>	<b>28</b>
<b>3.1.2. Quy mô và cấu trúc dữ liệu .....</b>	<b>28</b>
<b>3.1.3. Các thuộc tính sử dụng trong nghiên cứu .....</b>	<b>30</b>
<b>3.1.4. Vai trò của dữ liệu đối với các bài toán khai phá.....</b>	<b>33</b>
<b>3.2. Quy trình tiền xử lý dữ liệu .....</b>	<b>34</b>
<b>3.2.1. Kiểm tra và làm sạch dữ liệu .....</b>	<b>34</b>
<b>3.2.2. Chuyển đổi và chuẩn hóa dữ liệu .....</b>	<b>36</b>
<b>3.2.3. Tạo thuộc tính mới .....</b>	<b>37</b>
<b>3.2.4. Chọn lọc thuộc tính .....</b>	<b>38</b>
<b>3.2.5. Chuẩn bị dữ liệu cho mô hình hóa .....</b>	<b>39</b>
<b>3.3. Khám phá dữ liệu (EDA) .....</b>	<b>39</b>
<b>3.3.1. Thống kê mô tả.....</b>	<b>39</b>
<b>3.3.2. Phân bố dữ liệu của các thuộc tính chính.....</b>	<b>41</b>
<b>3.3.3. Phân tích mối quan hệ với biến mục tiêu .....</b>	<b>44</b>
<b>3.3.4. Phân tích tương quan giữa các thuộc tính .....</b>	<b>46</b>
<b>3.3.5. Phân tích theo thời gian .....</b>	<b>48</b>
<b>3.3.5. Nhận xét và ý nghĩa rút ra từ Khám phá dữ liệu .....</b>	<b>50</b>
<b>3.4. Lựa chọn mô hình .....</b>	<b>50</b>
<b>3.4.1. Mô hình phân lớp (Classification) .....</b>	<b>51</b>
<b>3.4.2. Mô hình phân cụm (Clustering) .....</b>	<b>51</b>
<b>3.4.3. Khai phá luật kết hợp (Association Rules Mining).....</b>	<b>52</b>
<b>3.4.4. Phân tích chuỗi thời gian (Time Series Analysis) .....</b>	<b>53</b>
<b>3.4.5. Tổng kết lựa chọn mô hình .....</b>	<b>53</b>
<b>3.5. Quy trình mô hình hóa đề xuất.....</b>	<b>53</b>
<b>3.5.1. Xác định bài toán khai phá dữ liệu.....</b>	<b>54</b>
<b>3.5.2. Chuẩn bị dữ liệu cho mô hình hóa.....</b>	<b>54</b>

<b>3.5.3. Huấn luyện mô hình.....</b>	<b>55</b>
<b>3.5.4. Đánh giá và so sánh mô hình .....</b>	<b>55</b>
<b>3.5.5. Trực quan hóa về diễn giải kết quả .....</b>	<b>56</b>
<b>3.5.6. Tổng kết quy trình mô hình hóa.....</b>	<b>56</b>
<b>CHƯƠNG IV: THỰC NGHIỆM.....</b>	<b>58</b>
<b>4.1. Thiết lập thực nghiệm.....</b>	<b>58</b>
<b>4.2. Kết quả mô hình phân loại .....</b>	<b>58</b>
<b>4.3. Kết quả mô hình phân cụm .....</b>	<b>62</b>
<b>CHƯƠNG V: KẾT LUẬN.....</b>	<b>68</b>
<b>5.1. Tóm tắt kết quả.....</b>	<b>68</b>
<b>5.2. Trả lời câu hỏi nghiên cứu .....</b>	<b>68</b>
<b>5.3. Hạn chế của nghiên cứu .....</b>	<b>69</b>
<b>5.4. Hướng mở rộng.....</b>	<b>69</b>
<b>Tài liệu tham khảo .....</b>	<b>70</b>

## CHƯƠNG I: GIỚI THIỆU

### 1.1. Bối cảnh nghiên cứu

Trong những năm gần đây, ngành du lịch và khách sạn toàn cầu chứng kiến sự phục hồi mạnh mẽ với tốc độ tăng trưởng trung bình từ 5% đến 8% mỗi năm (**UNWTO, 2025**) sau đại dịch COVID-19, nhưng cũng phải đối mặt với mức độ cạnh tranh ngày càng cao do sự bão hòa của thị trường và sự xuất hiện của các mô hình kinh doanh mới. Các yếu tố như biến động mùa vụ, thay đổi chính sách thị trường, sự phát triển của các nền tảng đặt phòng trực tuyến (OTA) và sự đa dạng hóa trong hành vi tiêu dùng khiến nhu cầu lưu trú ngày càng phức tạp và khó dự báo.

Cùng với xu hướng chuyển đổi số, các hệ thống quản lý khách sạn hiện đại (PMS) ngày càng tạo ra khối lượng dữ liệu khổng lồ về hành vi đặt phòng, tỷ lệ hủy phòng, lựa chọn dịch vụ và mô hình tiêu dùng. Tuy nhiên, nhiều nghiên cứu cho thấy phần lớn các khách sạn vừa và nhỏ chưa tận dụng hiệu quả nguồn dữ liệu này để hỗ trợ các quyết định quản trị chiến lược (**Kim et al., 2020**). Các phương pháp truyền thống dựa trên kinh nghiệm hoặc mô hình thống kê đơn giản thường không đủ khả năng xử lý tính phi tuyến và đa chiều của dữ liệu khách sạn đương đại.

Sự biến động cao của nhu cầu—đặc biệt trong các dịp lễ, mùa cao điểm du lịch hoặc giai đoạn thấp điểm—đặt ra thách thức nghiêm trọng đối với việc phân bổ nguồn lực, lập kế hoạch nhân sự và tối ưu hóa tỷ lệ công suất phòng. Nghiên cứu của **Chiang et al., 2017** chỉ ra rằng việc thiếu khả năng dự báo chính xác dẫn đến hai hậu quả tiêu cực: dư thừa công suất gây lãng phí chi phí vận hành, và tình trạng quá tải làm giảm chất lượng dịch vụ cũng như uy tín thương hiệu. Do đó, các phương pháp phân tích dữ liệu tiên tiến và mô hình dự báo nhu cầu đã trở thành yếu tố then chốt trong chiến lược quản trị doanh thu của ngành khách sạn.

Mặc dù đã có nhiều nghiên cứu áp dụng kỹ thuật khai phá dữ liệu trong lĩnh vực khách sạn, phần lớn tập trung vào một hoặc hai kỹ thuật đơn lẻ như dự báo chuỗi thời gian (**Law et al., 2018**) hoặc phân loại khách hàng (**Dolnicar & Le, 2017**). Các nghiên cứu tích hợp đa kỹ thuật—kết hợp phân cụm, phân loại, luật kết hợp và dự báo chuỗi thời gian—vẫn còn hạn chế, đặc biệt trong bối cảnh khách sạn tại Việt Nam. Nghiên cứu này nhằm lắp đầy khoảng trống đó bằng cách đề xuất một khung phân tích tích hợp, không chỉ dự báo nhu cầu mà còn đồng thời phân tích hành vi khách hàng, dự đoán rủi ro hủy phòng và nhận diện cơ hội bán kèm dịch vụ. Kết quả nghiên cứu kỳ vọng sẽ đóng góp cả về mặt lý thuyết (chứng minh hiệu quả của phương pháp tích hợp) và thực tiễn (cung cấp hệ thống hỗ trợ quyết định cho các khách sạn vừa và nhỏ).

### 1.2. Lý do chọn đề tài

*Thứ nhất, từ góc độ thực tiễn:* Ngành khách sạn đang trải qua quá trình chuyển đổi số mạnh mẽ, trong đó việc ra quyết định dựa trên dữ liệu trở thành yếu tố then chốt để duy trì lợi thế cạnh tranh. Các phương pháp quản trị truyền thống dựa trên kinh nghiệm đã bộc lộ nhiều hạn chế khi đối mặt với sự thay đổi nhanh chóng của hành vi khách hàng, sự gia tăng về khối lượng và độ phức tạp của dữ liệu, cũng như xu hướng cá nhân hóa dịch vụ trong ngành khách sạn hiện đại. Điều này tạo ra nhu cầu cấp thiết về các giải pháp phân tích dữ liệu tiên tiến để giảm thiểu rủi ro và tối ưu hóa hiệu quả kinh doanh.

*Thứ hai, từ góc độ nghiên cứu:* Các kỹ thuật khai phá dữ liệu (Data Mining) đã chứng minh tiềm năng trong việc giải quyết nhiều bài toán trong ngành khách sạn, bao gồm phân khúc khách hàng, dự đoán rủi ro hủy phòng, phân tích các mẫu tiêu dùng dịch vụ và dự báo nhu cầu theo thời gian. Tuy nhiên, qua khảo sát tài liệu, nhận thấy rằng phần lớn các nghiên cứu hiện có chỉ có thể tập trung vào một hoặc hai kỹ thuật riêng lẻ để giải quyết từng bài toán cụ thể, trong khi còn thiếu các công trình tích hợp nhiều kỹ thuật vào một khung phân tích thống nhất để khai thác toàn diện giá trị của dữ liệu khách sạn.

*Thứ ba, từ góc độ học tập:* Đề tài này cung cấp cơ hội để ứng dụng các kiến thức về khai phá dữ liệu vào một bài toán thực tiễn có tính phức hợp cao. Qua đó, bản thân em có thể trải nghiệm toàn bộ quy trình khai phá dữ liệu - từ tiền xử lý, khám phá dữ liệu, xây dựng và đánh giá mô hình, đến việc diễn giải kết quả trong bối cảnh kinh doanh cụ thể. Đây là cơ hội quý giá để nâng cao năng lực phân tích dữ liệu và rèn luyện tư duy giải quyết vấn đề thực tế.

Từ những lý do trên, em lựa chọn nghiên cứu đề tài này nhằm xây dựng một khung mô hình phân tích toàn diện, góp phần tăng tính thực tiễn và tính hệ thống trong việc khai thác dữ liệu ngành khách sạn cũng như phần nào đó có thể luyện tập việc ứng dụng các kỹ thuật lên bộ dữ liệu thực tế phù hợp vần đề, yêu cầu thực tiễn của doanh nghiệp.

### 1.3. Tính cấp thiết của đề tài

Từ góc độ khai phá dữ liệu, nhu cầu đặt phòng khách sạn là một hiện tượng có mức độ biến động cao dưới tác động của mùa vụ, thời điểm trong năm, các sự kiện đặc biệt và nhiều yếu tố ngoại sinh khác. Theo **STR Global (2023)** – thông tin công khai trên *Data Insights Blog* – công suất phòng tại nhiều thị trường lớn thường dao động **khoảng 50–60% trong mùa thấp điểm** và tăng lên **trên 70–80% trong mùa cao điểm**, phản ánh chênh lệch theo mùa có thể lên tới **20–30%** tùy khu vực. Báo cáo **Deloitte Hospitality Outlook Asia Pacific (2022)** cũng ghi nhận biến động công suất đáng kể theo mùa tại nhiều thị trường trong khu vực châu Á – Thái Bình Dương.

Tại Việt Nam, xu hướng biến động này thể hiện rõ rệt hơn. Theo **Tổng cục Thống kê Việt Nam (2023)**, công suất bình quân của các cơ sở lưu trú trong một số tháng thấp điểm chỉ ở mức **khoảng 40%**. Các điểm du lịch biển như Đà Nẵng và Nha Trang có thời điểm thấp hơn mức trung bình này theo các báo cáo tình hình du lịch địa phương được công bố công khai. Trong khi đó, theo **Hiệp hội Khách sạn Việt Nam (VHA)**, nhiều khách sạn cần duy trì **tối thiểu 55–65% công suất** để đạt mức hòa vốn. Điều này cho thấy tình trạng dư thừa công suất trong mùa thấp điểm không chỉ làm giảm doanh thu mà còn làm tăng chi phí cố định tính trên mỗi phòng được sử dụng.

Ngược lại, vào mùa cao điểm hoặc các dịp lễ đặc biệt, công suất phòng thường tăng vọt. Các báo cáo công khai từ **UN Tourism (2023)** và **STR (2022–2023)** ghi nhận công suất của nhiều thị trường quốc tế như Nhật Bản (mùa hoa anh đào), châu Âu (mùa hè du lịch) hay Mỹ (các sự kiện thể thao lớn) đều có thời điểm đạt **trên 90%**. Tại Việt Nam, dữ liệu từ **Tổng cục Thống kê (2023)** cũng cho thấy công suất phòng trong các dịp lễ lớn như 30/4–1/5 và Tết Nguyên đán thường tăng mạnh, tiệm cận mức tối đa tại nhiều điểm du lịch. Việc công suất tăng đột biến gây áp lực lớn lên hệ thống vận hành và nhân sự, làm giảm chất lượng dịch vụ và ảnh

hướng tiêu cực đến trải nghiệm khách hàng. Điều này cho thấy khả năng dự báo công suất chính xác là yêu cầu then chốt để tối ưu hóa vận hành khách sạn.

Ngoài biến động theo mùa vụ, độ chính xác của dự báo công suất còn bị ảnh hưởng nghiêm trọng bởi tỷ lệ hủy phòng cao trong thời đại OTA. Báo cáo của OTA Insight (2023) và Expedia Group (2022) cho thấy tỷ lệ hủy phòng toàn cầu thường dao động 20–30%, cao hơn đáng kể ở các khách sạn áp dụng chính sách hủy linh hoạt. Nghiên cứu của Antonio et al. (2019), công bố trên *Data in Brief*, ghi nhận tỷ lệ hủy phòng trung bình khoảng 27,5% tại 31 khách sạn ở Bồ Đào Nha, với sự biến động lớn tùy theo kênh đặt phòng và loại hình khách sạn. Tại Việt Nam, mặc dù chưa có số liệu chính thức toàn ngành, nhiều khách sạn ghi nhận chênh lệch đáng kể giữa công suất dự kiến và thực tế do tỷ lệ hủy gia tăng khi khách đặt qua OTA.

Khi kết hợp cả hai yếu tố—biến động công suất theo mùa và tỷ lệ hủy phòng cao—độ phức tạp của bài toán dự báo nhu cầu tăng lên đáng kể. Ví dụ, trong mùa cao điểm, nếu 25–30% phòng bị hủy vào phút chót, khách sạn không chỉ mất doanh thu từ những phòng đã đặt mà còn mất cơ hội bán phòng cho khách hàng khác do thời gian phản ứng hạn chế. Ngược lại, trong mùa thấp điểm, việc dự báo sai về tỷ lệ hủy có thể dẫn đến việc từ chối đặt phòng hợp lệ hoặc không tối ưu được chiến lược giá.

Mặc dù các khách sạn hiện đại ngày nay sở hữu lượng dữ liệu lớn, đa dạng và liên tục tăng, phần lớn các nghiên cứu hiện hành lại chỉ tập trung vào từng kỹ thuật đơn lẻ như dự báo chuỗi thời gian hoặc phân loại hủy phòng. Sự thiếu vắng các mô hình tích hợp đa kỹ thuật—kết hợp dự báo công suất, phân tích hành vi hủy phòng và tối ưu hóa vận hành—khiến tiềm năng khai thác dữ liệu chưa được tận dụng đầy đủ.

Vì vậy, để tài mang tính cấp thiết cả về **khoa học** (bổ sung mô hình tích hợp) lẫn **thực tiễn** (hỗ trợ quyết định cho các khách sạn).

## 1.4. Mục tiêu nghiên cứu

### 1.4.1. Mục tiêu tổng quát

Nghiên cứu nhằm phát triển mô hình khai phá dữ liệu tích hợp để phân tích hành vi đặt phòng với , dự đoán khả năng hủy phòng và dự báo nhu cầu lưu trú, qua đó hỗ trợ khách sạn nâng cao hiệu quả vận hành và quản trị doanh thu.

### 1.4.2. Mục tiêu cụ thể

#### (1) Mục tiêu mô tả

- Phân tích các đặc trưng hành vi đặt phòng theo thời gian, mùa vụ, loại phòng và đặc điểm khách hàng.
- Khám phá các yếu tố ảnh hưởng đến giá phòng, thời điểm đặt và tỷ lệ hủy phòng.

#### (2) Mục tiêu phân tích – dự đoán

- Phân khúc khách hàng dựa trên hành vi đặt phòng bằng các kỹ thuật phân cụm.
- Dự đoán khả năng hủy phòng bằng mô hình phân loại để hỗ trợ kiểm soát rủi ro và nâng cao độ chính xác dự báo.
- Khai phá các mẫu dịch vụ hoặc loại phòng thường xuất hiện cùng nhau bằng kỹ thuật luật kết hợp.
- Dự báo nhu cầu đặt phòng trong 30–90 ngày tới bằng mô hình chuỗi thời gian.

### (3) Mục tiêu ứng dụng

- Rút ra ý nghĩa từ kết quả mô hình và đề xuất giải pháp hỗ trợ khách sạn tối ưu vận hành, hoạch định nguồn lực và quản trị doanh thu.

## 1.5. Các câu hỏi nghiên cứu

Từ những mục tiêu trên, đề tài đặt ra các câu hỏi nghiên cứu sau:

### A. Câu hỏi mô tả (EDA – Exploratory Data Analysis)

- **Mục tiêu:** Hiểu bức tranh tổng quan về nhu cầu đặt phòng, hành vi khách hàng và biến động vận hành để định hướng cho các mô hình khai phá dữ liệu.
  1. Xu hướng đặt phòng thay đổi như thế nào theo thời gian và mùa vụ?
  2. Những phân khúc khách hàng nào có tỷ lệ hủy phòng cao hoặc hành vi đặt phòng đặc thù?
  3. Loại phòng hoặc dịch vụ nào được đặt phổ biến nhất trong từng giai đoạn?
  4. Những yếu tố nào ảnh hưởng đến biến động giá phòng và tỷ lệ hủy phòng?

### B. Câu hỏi dự đoán (Classification – Hủy phòng)

- **Mục tiêu:** Xây dựng mô hình phân loại để dự đoán khả năng hủy phòng dựa trên thông tin đặt phòng ban đầu, giúp tối ưu doanh thu và quản lý rủi ro.
  5. Liệu khả năng hủy phòng có thể được dự đoán dựa trên thông tin đặt phòng ban đầu hay không?
  6. Các mô hình phân loại (Decision Tree, Random Forest, Logistic Regression) cho kết quả như thế nào theo các thước đo Accuracy, Precision, Recall và F1-score?

### C. Câu hỏi phân cụm (Clustering – Phân khúc khách hàng)

7. Có thể phân nhóm khách hàng dựa trên hành vi đặt phòng (thời gian đặt, ADR, số đêm lưu trú, loại phòng...) hay không?
8. Các nhóm khách hàng được hình thành có sự khác biệt đáng kể về hành vi chi tiêu hoặc khả năng hủy phòng không?

### C. Câu hỏi luật kết hợp (Association Rule Mining)

- **Mục tiêu:** Khám phá các nhóm khách hàng tự nhiên dựa trên hành vi đặt phòng nhằm hỗ trợ chiến lược marketing và cá nhân hóa dịch vụ.

9. Những dịch vụ hoặc loại phòng nào có xu hướng được đặt cùng nhau và có thể tạo thành các luật kết hợp mạnh?

#### D. Câu hỏi ứng dụng – Chiều sâu kinh doanh

- **Mục tiêu:** Kết nối kết quả khai phá dữ liệu với quyết định kinh doanh, mang lại giá trị thực sự cho doanh nghiệp.

10. Các yếu tố nào ảnh hưởng mạnh nhất đến khả năng hủy phòng và có thể được sử dụng để giảm thiểu rủi ro?

11. Kết quả dự báo nhu cầu trong 30–90 ngày tới có thể hỗ trợ khách sạn tối ưu giá phòng, nhân sự và quản trị doanh thu như thế nào?

### 1.6. Đối tượng và phạm vi nghiên cứu

#### 1.6.1. Đối tượng nghiên cứu

Đề tài lấy đối tượng nghiên cứu chính là tập dữ liệu đặt phòng khách sạn ([Hotel Booking Demand Dataset](#)) được công bố trong bài báo “Hotel booking demand datasets” (Antonio et al., 2019) bao gồm hơn 119.000 bản ghi, mô tả hành vi đặt phòng tại hai loại khách sạn: City Hotel và Resort Hotel. Các đặc trưng của bộ dữ liệu này được phân tích bằng các kỹ thuật khai phá dữ liệu như EDA, phân loại, phân cụm và luật kết hợp, Cụ thể như sau:

##### (1) Đối tượng dữ liệu

- Thông tin khách hàng: quốc gia, số lượng người lớn, trẻ em, em bé

- Thông tin đặt phòng: lead time, số đêm lưu trú, loại phòng đặt trước và loại phòng được giao

- Thông tin giá và giao dịch: ADR (Average Daily Rate), loại bữa ăn, kênh đặt phòng

- Thông tin hành vi: khách có hủy phòng hay không (is\_canceled), số lần thay đổi đặt phòng, yêu cầu đặc biệt

- Thông tin thời gian: tháng đặt phòng, năm đặt phòng, thời điểm đến khách sạn

##### (2) Đối tượng phân tích

- Xu hướng và đặc điểm hành vi đặt phòng theo thời gian, mùa vụ và phân khúc khách hàng.

- Hành vi hủy phòng và các yếu tố ảnh hưởng.

- Mối quan hệ giữa các loại phòng, dịch vụ và phân khúc khách hàng.

- Cấu trúc phân nhóm khách hàng dựa trên hành vi tiêu dùng và đặt phòng.

##### (3) Đối tượng mô hình hóa (Data Mining Techniques)

- Mô hình phân loại (Classification) dự đoán khả năng hủy phòng.

- Mô hình **phân cụm (Clustering)** nhằm phân khúc khách hàng.
- **Luật kết hợp (Association Rule Mining)** nhằm khai phá các mẫu hành vi đặt phòng.
- **EDA – Exploratory Data Analysis** nhằm hiểu đặc trưng dữ liệu ban đầu.

### **1.6.2. Phạm vi nghiên cứu**

**Phạm vi nghiên cứu** giới hạn trong dữ liệu đặt phòng theo giai đoạn được cung cấp (2015 - 2017), tập trung vào phân tích hành vi, dự đoán hủy phòng, phân khúc khách hàng và khai phá quy luật đặt dịch vụ, không đi sâu vào mô hình deep learning hoặc dự báo dài hạn.

#### **(1) Phạm vi nội dung**

- Đề tài tập trung vào:

- Phân tích mô tả dữ liệu đặt phòng.
- Dự đoán khả năng hủy phòng bằng các mô hình phân loại.
- Phân nhóm khách hàng dựa trên đặc trưng hành vi.
- Khai phá luật kết hợp giữa loại phòng, dịch vụ và đặc điểm khách hàng.
- Nghiên cứu áp dụng mô hình để dự báo tổng lượng đặt phòng theo chuỗi thời gian.
- Ứng dụng kết quả để đề xuất giải pháp quản trị doanh thu và giảm rủi ro hủy phòng.

- Đề tài **không** đi sâu vào:

- Hành vi của khách sạn ngoài bộ dữ liệu được cung cấp.
- Quy trình vận hành nội bộ hoặc chiến lược marketing ngoài dữ liệu.

#### **(2) Phạm vi dữ liệu**

Dữ liệu lấy từ tập dữ liệu khách sạn (Hotel Booking Demand Dataset) công bố vào năm 2019 và được lưu trữ tại Kaggle

- Nguồn dữ liệu trên Kaggle: [Hotel booking demand](#)
- Khoảng thời gian dữ liệu được thu thập: **từ năm 2015 đến 2017**
- Nghiên cứu chỉ phân tích các biến có độ đầy đủ và chất lượng đảm bảo sau tiền xử lý.

#### **(3) Phạm vi kỹ thuật được sử dụng**

- Các thuật toán được sử dụng:
  - **Phân loại:** Decision Tree, Random Forest, Logistic Regression.
  - **Phân cụm:** K-means

- **Luật kết hợp:** Apriori

- **Chuỗi thời gian:** SARIMAX

- Các kỹ thuật Deep Learning nâng cao **không nằm trong phạm vi nghiên cứu.**

## 1.7. Phương pháp nghiên cứu

Phương pháp nghiên cứu trong đề tài được xây dựng dựa trên quy trình khai phá dữ liệu CRISP-DM (Cross Industry Standard Process for Data Mining) gồm sáu giai đoạn: (1) Hiểu bài toán kinh doanh, (2) Hiểu dữ liệu, (3) Chuẩn bị dữ liệu, (4) Mô hình hóa, (5) Đánh giá mô hình và (6) Triển khai ứng dụng. Các kỹ thuật phân tích được sử dụng trong nghiên cứu bao gồm EDA, phân loại, phân cụm và luật kết hợp. Cụ thể như sau:

### 1. Hiểu bài toán kinh doanh

Mục tiêu của nghiên cứu này chính là dự đoán hủy phòng, phân khúc khách hàng và khai phá mẫu dịch vụ

### 2. Phương pháp hiểu và mô tả dữ liệu (EDA – Exploratory Data Analysis)

Các hoạt động EDA được thực hiện gồm:

- Thống kê mô tả các biến định tính và định lượng (giá trị trung bình, tứ phân vị, .. )
- Phân tích phân phối dữ liệu, nhận diện ngoại lai và giá trị thiếu (null)
- Phân tích xu hướng thời gian theo năm, tháng, mùa du lịch
- Khám phá mối quan hệ giữa các biến như giá phòng trung bình, quốc gia, loại phòng, kênh đặt phòng
- Trả lời các câu hỏi nghiên cứu trong nhóm A (EDA) được đề cập trong phần Mục tiêu nghiên cứu

### 3. Phương pháp chuẩn bị dữ liệu

Quy trình chuẩn bị dữ liệu bao gồm:

Làm sạch dữ liệu: xử lý giá trị thiếu, nhiễu hoặc không hợp lệ, Xử lý dữ liệu mất cân bằng bằng các phương pháp (SMOTE, Class weights)

- Mã hóa biến định tính bằng **One-Hot Encoding** hoặc **Ordinal Encoding**
- Chuẩn hóa biến định lượng khi cần thiết bằng **StandardScaler** hoặc **MinMaxScaler**
- Tạo chuỗi và gom nhóm dữ liệu theo tần suất, xử lý khoảng trống
- Xây dựng các biến mới (feature engineering), bao gồm:
  - o Nhóm lead time
  - o Tổng số đêm lưu trú
  - o Mùa du lịch

- Chênh lệch giữa phòng đặt và phòng được giao
- Chia dữ liệu thành tập huấn luyện với tỷ lệ train/test split là 80/20 và tập kiểm tra cho mô hình phân loại

#### **4. Phương pháp mô hình hóa**

*Mô hình phân loại:*

- Các thuật toán được sử dụng: Decision Tree, Random Forest, Logistic Regression
- Đánh giá mô hình theo các thước đo: Accuracy, Precision, Recall, F1-score và ROC-AUC
- Lựa chọn mô hình tối ưu dựa trên sự cân bằng giữa Precision – Recall và độ chính xác tổng thể

*Mô hình phân cụm*

- Áp dụng thuật toán K-means
- Xác định số cụm tối ưu bằng phương pháp Elbow hoặc Silhouette Score
- Phân tích và diễn giải đặc điểm từng cụm dựa trên: ADR, lead\_time, số đêm lưu trú, loại phòng, số lượng khách

*Luật kết hợp*

- Sử dụng thuật toán Apriori
- Lựa chọn các luật mạnh dựa trên các chỉ số: Support, Confidence và Lift
- Khai thác các mối quan hệ kết hợp giữa loại phòng, dịch vụ, quốc gia, mùa du lịch và kênh đặt phòng

*Chuỗi thời gian*

- Sử dụng mô hình SARIMAX
- Xác định mô hình nào hiệu quả dựa vào RMSE, MAPE, ...
- Nhằm đưa ra phân tích dự đoán xem khách sẽ đến khi nào và bao nhiêu

#### **5. Phương pháp đánh giá mô hình**

Việc đánh giá mô hình được thực hiện theo các nội dung sau:

- So sánh hiệu quả các mô hình phân loại dựa trên bảng tổng hợp chỉ số
- Đánh giá mức độ tách biệt và ổn định của các cụm khách hàng
- Phân tích ý nghĩa thực tiễn của các luật kết hợp được khai phá
- Đối chiếu kết quả với hệ thống câu hỏi nghiên cứu nhằm đảm bảo mức độ phù hợp và độ tin cậy

#### **6. Ứng dụng kết quả**

Kết quả phân tích được diễn giải theo góc độ quản trị khách sạn, từ đó đề xuất các giải pháp:

- Giảm thiểu rủi ro hủy phòng: điều chỉnh chính sách giữ phòng, yêu cầu đặt cọc, phân loại nhóm khách có nguy cơ hủy cao
- Tối ưu hóa giá phòng theo mùa vụ và theo phân khúc khách hàng
- Xây dựng chiến lược marketing cá nhân hóa dựa trên phân cụm khách hàng
- Gợi ý gói dịch vụ phù hợp dựa trên các luật kết hợp
- Đề xuất khung ứng dụng kết quả phân tích trong hoạt động vận hành và quản trị doanh thu

## 1.8. Cấu trúc tiểu luận

### Chương 1: Giới thiệu

- Trình bày bối cảnh nghiên cứu, lý do chọn đề tài và tính cấp thiết trong ngành khách sạn.
- Xác định mục tiêu và hệ thống câu hỏi nghiên cứu cụ thể.
- Phác thảo phương pháp nghiên cứu và hướng tiếp cận đề tài.

### Chương 2: Cơ sở lý thuyết

- Tổng quan các khái niệm cơ bản liên quan đến dữ liệu đặt phòng khách sạn.
- Trình bày cơ sở lý thuyết về các thuật toán khai phá dữ liệu được sử dụng: Phân loại, Phân cụm, Luật kết hợp và Chuỗi thời gian.
- Tổng hợp các nghiên cứu liên quan đã ứng dụng Khai phá dữ liệu trong lĩnh vực khách sạn.

### Chương 3: Dữ liệu & Phương pháp đề xuất

- Mô tả chi tiết tập dữ liệu (kích thước, thuộc tính) và nguồn gốc dữ liệu.
- Trình bày quy trình tiền xử lý dữ liệu, bao gồm làm sạch dữ liệu, xử lý giá trị thiếu/ngoại lệ, và xây dựng biến mới (feature engineering).
- Phân tích dữ liệu mô tả ban đầu (EDA) để khám phá các đặc trưng dữ liệu.
- Mô tả quy trình mô hình hóa tích hợp được đề xuất.

### Chương 4: Thực nghiệm & Kết quả & Thảo luận

- Trình bày thiết lập thực nghiệm và các độ đo đánh giá mô hình đã sử dụng.
- Trình bày kết quả chi tiết của các mô hình: Phân cụm (K-means), Phân loại (Random Forest), Luật kết hợp (Apriori), và Chuỗi thời gian (SARIMAX).
- Tiến hành đánh giá, so sánh hiệu quả giữa các mô hình và thảo luận ý nghĩa kết quả dưới góc độ kinh doanh.

## **Chương 5: Kết luận**

- Tóm tắt các kết quả chính đã đạt được và trả lời toàn bộ các câu hỏi nghiên cứu ban đầu.
- Rút ra những chiêu sâu có ý nghĩa thực tiễn cho hoạt động quản trị khách sạn.
- Nêu rõ các hạn chế của nghiên cứu và đề xuất hướng mở rộng cho các nghiên cứu tiếp theo.

## CHƯƠNG II: CƠ SỞ LÝ THUYẾT

### 2.1. Các khái niệm liên quan

#### 2.1.1. Khai phá dữ liệu (Data Mining)

Khai phá dữ liệu là quá trình trích xuất tri thức tiềm ẩn từ các tập dữ liệu lớn thông qua việc áp dụng các phương pháp thống kê, học máy và mô hình hóa. Theo quan điểm hiện đại, Data Mining là một bước quan trọng trong quy trình phát hiện tri thức (KDD – Knowledge Discovery in Databases), bao gồm các hoạt động như tiền xử lý, xây dựng mô hình, đánh giá và diễn giải kết quả.

Trong lĩnh vực khách sạn, khai phá dữ liệu giúp dự đoán hành vi đặt phòng, nhận diện nhóm khách hàng, phát hiện các mẫu dịch vụ thường đi cùng nhau và dự báo nhu cầu trong tương lai. Đây là nền tảng cho các ứng dụng quản trị doanh thu, tối ưu vận hành và chiến lược market

#### 2.1.2. Khái niệm cơ bản trong học máy

*Học máy (Machine Learning)* là một lĩnh vực thuộc trí tuệ nhân tạo, tập trung nghiên cứu và phát triển các thuật toán cho phép máy tính tự động học hỏi từ dữ liệu và cải thiện hiệu suất thực hiện nhiệm vụ mà không cần được lập trình một cách tường minh. Thay vì xây dựng các quy tắc cố định, học máy khai thác dữ liệu lịch sử để phát hiện các mẫu và mối quan hệ tiềm ẩn, từ đó đưa ra dự đoán hoặc hỗ trợ ra quyết định trong các bài toán thực tế.

*Dữ liệu (Data)* đóng vai trò trung tâm trong học máy và thường được biểu diễn dưới dạng bảng, trong đó mỗi hàng tương ứng với một đối tượng (mẫu dữ liệu) và mỗi cột tương ứng với một thuộc tính (đặc trưng). Đối với các bài toán có giám sát, dữ liệu còn bao gồm nhãn – là giá trị đầu ra cần dự đoán. Chất lượng, độ dày dặn và tính nhất quán của dữ liệu ảnh hưởng trực tiếp đến độ chính xác và khả năng tổng quát của mô hình học máy.

- *Mỗi hàng*: Tương ứng với một đối tượng quan sát
- *Mỗi cột*: Tương ứng với một thuộc tính mô tả đối tượng
- Với bài toán học có giám sát, dữ liệu còn bao gồm **biến mục tiêu** – giá trị cần dự đoán

#### Các phương pháp học máy:

*Học có giám sát (Supervised Learning)* là phương pháp học sử dụng tập dữ liệu đã được gán nhãn. Mục tiêu của phương pháp này là xây dựng mô hình học được mối quan hệ giữa dữ liệu đầu vào và nhãn đầu ra.

#### *Các loại bài toán:*

- **Phân loại (Classification)**: dự đoán nhãn phân loại cho dữ liệu đầu vào
- **Hồi quy (Regression)**: dự đoán giá trị liên tục

*Học không giám sát* (*Unsupervised Learning*) sử dụng dữ liệu không có nhãn nhằm khám phá cấu trúc ẩn và các mẫu trong dữ liệu. Phân cụm và giảm chiều dữ liệu là hai bài toán điển hình của phương pháp này.

- **Phân cụm (Clustering):** nhóm các đối tượng có đặc điểm tương đồng
- **Giảm chiều (Dimensionality Reduction):** Giảm số biến giữ lại thông tin quan trọng
- **Phát hiện bất thường (Anomaly Detection):** Tìm các điểm dữ liệu khác biệt

*Học bán giám sát* kết hợp cả dữ liệu có nhãn và không có nhãn để nâng cao hiệu quả học trong trường hợp dữ liệu gán nhãn hạn chế. Học tăng cường là phương pháp trong đó mô hình học thông qua quá trình tương tác với môi trường và nhận phản hồi dưới dạng phần thưởng hoặc hình phạt.

### 2.1.3. Kiểu dữ liệu

Kiểu dữ liệu (*Data Types*) trong học máy và khai phá dữ liệu dùng để mô tả bản chất và cách biểu diễn của dữ liệu. Việc xác định đúng kiểu dữ liệu là cơ sở quan trọng cho quá trình tiền xử lý, lựa chọn thuật toán và xây dựng mô hình học máy. Mỗi kiểu dữ liệu yêu cầu các phương pháp xử lý và kỹ thuật phân tích khác nhau.

*Phân loại dữ liệu theo giá trị:*

- **Biến định lượng:** Là dữ liệu có giá trị số và có thể thực hiện các phép toán số học
  - **Dữ liệu rời rạc:** Giá trị nguyên đếm được
  - **Dữ liệu liên tục:** Giá trị thực trong một khoảng, thường là kết quả của việc đo lường
- **Biến định tính:** Là dữ liệu biểu diễn các nhóm hoặc danh mục, không mang ý nghĩa số học
  - **Dữ liệu không thứ tự:** Các giá trị phân biệt
  - **Dữ liệu có thứ tự:** Các giá trị có mức độ rõ ràng
- **Biến thời gian:** Dữ liệu biểu diễn thời điểm hoặc khoảng thời gian
- **Dữ liệu nhị phân:** Là trường hợp đặc biệt của dữ liệu định tính, chỉ nhận hai giá trị

Việc nhận diện và xử lý đúng từng loại dữ liệu là bước quan trọng đầu tiên trong quy trình khai phá dữ liệu. Chi tiết việc các kỹ thuật xử lý cụ thể sẽ được trình bày ở mục 2.2 (Lý thuyết về tiền xử lý dữ liệu)

## 2.2. Lý thuyết về tiền xử lý dữ liệu

### 2.2.1. Khái niệm tiền xử lý dữ liệu

*Tiền xử lý dữ liệu* (*Data Preprocessing*) là giai đoạn chuẩn bị và cải thiện chất lượng dữ liệu trước khi tiến hành các phương pháp phân tích, khai phá dữ liệu và xây dựng mô hình dự báo. Dữ liệu thô thu thập từ thực tế thường tồn tại các vấn đề như thiếu dữ liệu, sai lệch, nhiễu,

trùng lặp hoặc không đồng nhất về định dạng. Nếu không được xử lý phù hợp, các vấn đề này có thể làm sai lệch kết quả phân tích và làm giảm hiệu quả của các mô hình khai thác dữ liệu.

### **Tiền xử lý dữ liệu nhằm:**

- Nâng cao chất lượng và độ tin cậy của dữ liệu đặt phòng.
- Loại bỏ hoặc giảm thiểu các sai sót, dữ liệu thiếu và dữ liệu nhiễu.
- Chuẩn hóa và biến đổi dữ liệu về dạng phù hợp cho các phương pháp phân tích.
- Tạo điều kiện cho các mô hình phân loại, phân nhóm, khai phá luật kết hợp và dự báo chuỗi thời gian hoạt động hiệu quả.

Trong nghiên cứu dữ liệu đặt phòng khách sạn, các phương pháp xử lý đóng vai trò nền tảng và mang tính quyết định đối với toàn bộ quy trình phân tích. Chất lượng dữ liệu đầu vào ảnh hưởng trực tiếp đến độ tin cậy của kết quả nghiên cứu. Cụ thể, quá trình giúp đảm bảo các mô hình dự đoán hủy phòng phản ánh đúng hành vi của khách hàng, hỗ trợ phân nhóm khách hàng dựa trên đặc trưng hành vi một cách rõ ràng và ổn định, nâng cao độ tin cậy của các luật kết hợp giữa loại phòng, dịch vụ và đặc điểm khách hàng, đồng thời cải thiện độ chính xác của các mô hình dự báo tổng lượng đặt phòng theo chuỗi thời gian.

Như vậy, tiền xử lý dữ liệu không chỉ là bước chuẩn bị ban đầu mà còn là nền tảng quyết định chất lượng của toàn bộ quá trình khai phá dữ liệu

#### **2.2.2. Các bước trong tiền xử lý dữ liệu**

##### **2.2.2.1. Khảo sát và hiểu dữ liệu**

Khảo sát và hiểu dữ liệu là bước đầu tiên nhằm nắm bắt cấu trúc, quy mô và các đặc điểm của tập dữ liệu. Bước này bao gồm việc xác định nguồn dữ liệu, số lượng bản ghi, số lượng thuộc tính và phân loại các thuộc tính theo các kiểu dữ liệu như dữ liệu số, dữ liệu danh mục, dữ liệu nhị phân và dữ liệu thời gian. Ngoài ra, các thống kê mô tả cơ bản như giá trị trung bình, giá trị nhỏ nhất, giá trị lớn nhất và phân bố dữ liệu thường được sử dụng để phát hiện các bất thường ban đầu.

##### **2.2.2.2. Làm sạch dữ liệu**

Làm sạch dữ liệu là bước nhằm loại bỏ hoặc hiệu chỉnh các lỗi tồn tại trong dữ liệu thô. Các bản ghi trùng lặp được phát hiện và loại bỏ để tránh gây sai lệch kết quả phân tích. Bên cạnh đó, các giá trị không hợp lệ hoặc không hợp lý được kiểm tra và xử lý phù hợp. Việc chuẩn hóa định dạng dữ liệu cũng được thực hiện nhằm đảm bảo tính nhất quán trong toàn bộ tập dữ liệu.

##### **2.2.2.3. Xử lý giá trị thiếu và ngoại lai**

Xử lý giá trị thiếu và ngoại lai là bước quan trọng nhằm nâng cao chất lượng dữ liệu. Giá trị thiếu là hiện tượng phổ biến trong các tập dữ liệu thực tế và có thể ảnh hưởng đến độ chính xác của quá trình khai phá. Trước hết, cần phân tích tỷ lệ và nguyên nhân của dữ liệu thiếu trong từng thuộc tính. Đối với các thuộc tính có tỷ lệ thiếu nhỏ, các phương pháp thay thế như giá trị

trung bình, trung vị hoặc giá trị xuất hiện nhiều nhất thường được áp dụng. Trong trường hợp tỷ lệ thiểu lớn hoặc thuộc tính không mang nhiều ý nghĩa phân tích, việc loại bỏ thuộc tính có thể được cân nhắc.

Dữ liệu nhiễu và ngoại lai có thể làm giảm chất lượng kết quả phân tích và độ chính xác của mô hình. Các giá trị ngoại lai thường được phát hiện thông qua các phương pháp thống kê hoặc trực quan hóa. Sau khi xác định, dữ liệu ngoại lai có thể được loại bỏ hoặc điều chỉnh về ngưỡng hợp lý nhằm giảm ảnh hưởng tiêu cực đến quá trình khai phá dữ liệu.

#### **2.2.2.4. Chuẩn hóa và biến đổi dữ liệu**

Chuẩn hóa và biến đổi dữ liệu được thực hiện do các thuộc tính trong tập dữ liệu có thể nằm trên các thang đo khác nhau. Việc chuẩn hóa giúp đảm bảo sự công bằng giữa các thuộc tính trong quá trình phân tích. Ngoài ra, các phép biến đổi dữ liệu có thể được áp dụng để cải thiện phân bố của dữ liệu, từ đó nâng cao hiệu quả của các thuật toán khai phá.

#### **2.2.2.5. Mã hóa dữ liệu danh mục**

Các thuật toán khai phá dữ liệu và học máy thường yêu cầu dữ liệu đầu vào ở dạng số. Do đó, các thuộc tính danh mục cần được mã hóa bằng các phương pháp phù hợp. Việc lựa chọn phương pháp mã hóa phụ thuộc vào đặc điểm của từng thuộc tính, đặc biệt là việc thuộc tính đó có mang tính thứ tự hay không.

#### **2.2.2.6. Lựa chọn và trích chọn đặc trưng**

Lựa chọn và trích chọn đặc trưng nhằm giảm thiểu dữ liệu và loại bỏ các thuộc tính không liên quan hoặc trùng lặp về mặt thông tin. Đồng thời, các đặc trưng mới có thể được xây dựng từ dữ liệu ban đầu để phản ánh rõ hơn bản chất và cấu trúc tiềm ẩn của dữ liệu. Bước này giúp mô hình trở nên đơn giản hơn, dễ diễn giải và đạt hiệu quả cao hơn.

#### **2.2.2.7. Chuẩn bị dữ liệu cho từng bài toán khai phá**

Tùy theo mục tiêu khai phá, dữ liệu có thể được chuẩn bị theo những cách khác nhau. Đối với bài toán phân loại, dữ liệu cần có nhãn rõ ràng; đối với phân cụm, dữ liệu không sử dụng nhãn; trong khai phá luật kết hợp, dữ liệu thường được biểu diễn dưới dạng giao dịch; còn trong bài toán chuỗi thời gian, yếu tố thứ tự thời gian đóng vai trò then chốt. Việc chuẩn bị dữ liệu phù hợp với từng bài toán giúp đảm bảo các thuật toán khai phá hoạt động hiệu quả và chính xác.

#### **2.2.2.8. Xử lý dữ liệu mất cân bằng**

Dữ liệu mất cân bằng là hiện tượng phổ biến trong các bài toán phân loại, khi số lượng mẫu giữa các lớp có sự chênh lệch đáng kể. Trong trường hợp này, các mô hình học máy có xu hướng thiên lệch về lớp chiếm ưu thế, dẫn đến kết quả dự đoán không phản ánh đúng đặc điểm của lớp thiểu số. Do đó, việc xử lý dữ liệu mất cân bằng là cần thiết nhằm nâng cao khả năng học và độ tin cậy của mô hình.

Các phương pháp xử lý dữ liệu mất cân bằng thường tập trung vào việc điều chỉnh phân bổ dữ liệu giữa các lớp. Một số hướng tiếp cận phổ biến bao gồm lấy mẫu lại dữ liệu, chẳng hạn như giảm số lượng mẫu của lớp chiếm đa số hoặc tăng cường số lượng mẫu của lớp thiểu số. Bên cạnh đó, việc điều chỉnh trọng số cho các lớp trong quá trình huấn luyện cũng được sử dụng nhằm giảm ảnh hưởng của sự mất cân bằng dữ liệu.

Việc lựa chọn phương pháp xử lý dữ liệu mất cân bằng cần được cân nhắc dựa trên đặc điểm của tập dữ liệu và mục tiêu phân tích. Xử lý phù hợp hiện tượng mất cân bằng dữ liệu góp phần cải thiện khả năng phát hiện các trường hợp quan trọng thuộc lớp thiểu số, từ đó nâng cao hiệu quả của các mô hình phân loại.

Nhìn chung, các bước tiền xử lý dữ liệu đóng vai trò quan trọng trong việc đảm bảo chất lượng dữ liệu đầu vào, từ đó nâng cao hiệu quả và độ tin cậy của các phương pháp khai phá dữ liệu.

### 2.3. Lý thuyết về các thuật toán

Trong nghiên cứu phân tích dữ liệu đặt phòng khách sạn, các thuật toán khai phá dữ liệu và học máy được lựa chọn nhằm phục vụ nhiều mục tiêu khác nhau, bao gồm dự đoán khả năng hủy phòng, phân nhóm khách hàng, khai phá mối quan hệ giữa các yếu tố đặt phòng và dự báo xu hướng đặt phòng theo thời gian. Tùy theo đặc điểm của từng bài toán, các nhóm thuật toán phân loại, phân cụm, luật kết hợp và chuỗi thời gian được áp dụng tương ứng.

#### 2.3.1. Phân loại (Classification)

Bài toán phân loại được sử dụng chủ yếu nhằm dự đoán **khả năng hủy phòng** của khách hàng dựa trên các đặc trưng liên quan đến hành vi đặt phòng ban đầu.

Các thuật toán phân loại được sử dụng bao gồm:

##### 2.3.1.1. Logistic Regression

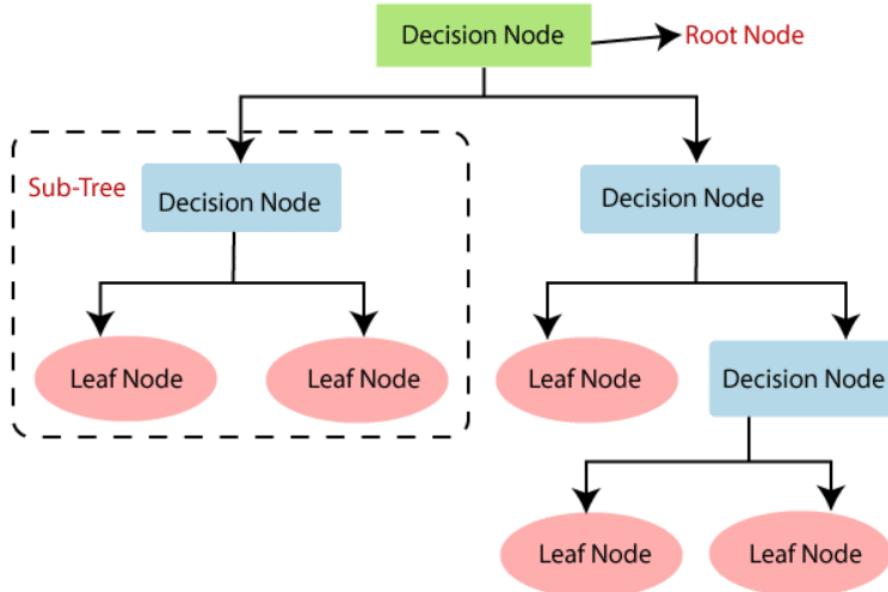
Logistic Regression là một mô hình thống kê được sử dụng rộng rãi để ước lượng xác suất xảy ra của một biến phụ thuộc nhị phân. Mô hình mô tả mối quan hệ giữa các biến độc lập và xác suất xảy ra của sự kiện thông qua hàm logistic:

$$P(x) = \frac{1}{1 + e^{-z}}, \quad z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

Trong bài toán dự đoán hủy phòng, Logistic Regression cho phép đánh giá mức độ ảnh hưởng của từng yếu tố như thời gian đặt trước, loại phòng, giá phòng hoặc kênh đặt phòng đến xác suất khách hàng hủy phòng. Nhờ tính đơn giản, khả năng diễn giải cao và chi phí tính toán thấp, mô hình này thường được sử dụng như **mô hình cơ sở (baseline)** để so sánh với các phương pháp phức tạp hơn.

##### 2.3.1.2. Cây quyết định (Decision Tree)

Cây quyết định là thuật toán phân loại dựa trên cấu trúc cây, trong đó dữ liệu được phân tách tuần tự thông qua các điều kiện rẽ nhánh tại các nút. Mỗi nút trong cây đại diện cho một thuộc tính, mỗi nhánh biểu diễn một điều kiện phân tách, và mỗi nút lá tương ứng với kết quả phân loại cuối cùng.



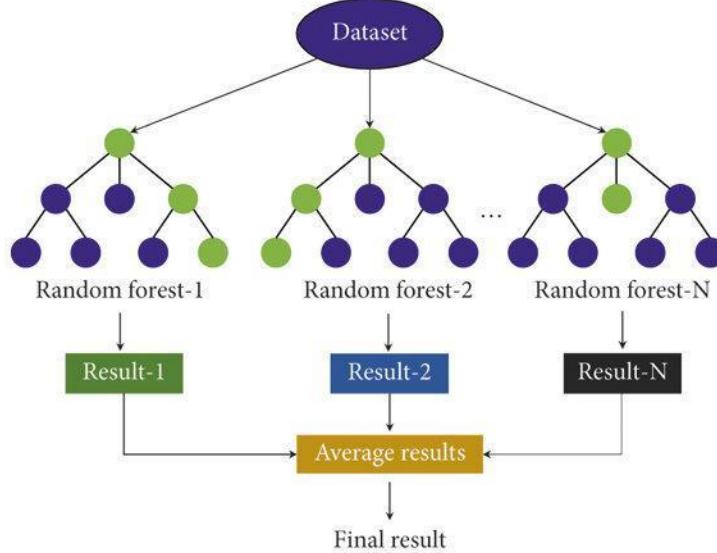
Hình 1: Minh họa Decision Tree

Hình minh họa cho thấy cấu trúc tổng quát của một cây quyết định, trong đó nút gốc (**Root Node**) đại diện cho thuộc tính được chọn để phân tách dữ liệu ban đầu. Tại mỗi nút quyết định (**Decision Node**), dữ liệu được chia thành các nhánh dựa trên một điều kiện xác định, chẳng hạn như so sánh giá trị của một thuộc tính với một ngưỡng cho trước. Quá trình phân tách này tiếp tục lặp lại cho đến khi đạt tới các nút lá (**Leaf Node**), nơi mỗi nút tương ứng với một nhãn phân loại cuối cùng. Một phần của cây có thể được xem như một cây con (**Sub-tree**), đại diện cho các quyết định cục bộ trong không gian dữ liệu. Cấu trúc phân cấp này cho phép mô hình biểu diễn các quy tắc phân loại dưới dạng các chuỗi điều kiện logic, giúp việc diễn giải kết quả trở nên trực quan và dễ hiểu.

Ưu điểm của Decision Tree là khả năng mô hình hóa các mối quan hệ phi tuyến giữa các biến đầu vào và biến mục tiêu, đồng thời dễ dàng diễn giải thông qua các quy tắc quyết định trực quan. Trong bối cảnh phân tích hành vi hủy phòng, cây quyết định giúp xác định rõ các yếu tố quan trọng và các ngưỡng giá trị có ảnh hưởng mạnh đến quyết định hủy của khách hàng. Tuy nhiên, mô hình này có thể gặp hiện tượng quá khớp nếu không được kiểm soát độ sâu hoặc các tham số phù hợp.

### 2.3.1.2. Rừng cây ngẫu nhiên (Random Forest)

Random Forest là một phương pháp học tổ hợp (ensemble learning), trong đó nhiều cây quyết định được xây dựng độc lập dựa trên các tập con ngẫu nhiên của dữ liệu và tập thuộc tính. Kết quả dự đoán cuối cùng được xác định thông qua cơ chế bỏ phiếu đa số:



Hình 2: Minh họa Random Forest

Hình minh họa mô tả cơ chế hoạt động của Random Forest, trong đó tập dữ liệu ban đầu được sử dụng để xây dựng nhiều cây quyết định độc lập thông qua quá trình lấy mẫu ngẫu nhiên (bootstrap sampling). Mỗi cây trong rừng được huấn luyện trên một tập con khác nhau của dữ liệu và chỉ xem xét một tập con ngẫu nhiên các thuộc tính tại mỗi nút phân tách. Các cây quyết định này đưa ra các dự đoán riêng lẻ, sau đó được tổng hợp thông qua cơ chế bỏ phiếu đa số (đối với bài toán phân loại) để tạo ra kết quả dự đoán cuối cùng. Cách tiếp cận học tổ hợp này giúp mô hình giảm độ nhạy với nhiễu trong dữ liệu và cải thiện khả năng tổng quát hóa so với việc sử dụng một cây quyết định đơn lẻ.

Việc kết hợp nhiều cây quyết định giúp Random Forest giảm thiểu hiện tượng quá khóp và nâng cao độ ổn định của mô hình so với một cây đơn lẻ. Trong nghiên cứu này, Random Forest được kỳ vọng mang lại hiệu quả dự đoán cao hơn trong trường hợp dữ liệu có nhiều biến và mối quan hệ phức tạp giữa các yếu tố ảnh hưởng đến hành vi hủy phòng của khách hàng.

### 2.3.2. Phân cụm (Clustering)

Dựa trên phương pháp phân cụm, xem xét khả năng phân nhóm khách hàng dựa trên hành vi đặt phòng, cũng như sự khác biệt giữa các nhóm khách hàng được hình thành. Kết quả phân cụm cho phép nhận diện các nhóm hành vi tương đồng, từ đó hỗ trợ xây dựng các chính sách quản trị doanh thu, thiết kế ưu đãi phù hợp và góp phần giảm thiểu rủi ro hủy phòng cho từng nhóm khách hàng. Trên cơ sở đó, phương pháp phân cụm được sử dụng nhằm xem xét khả năng hình thành các nhóm khách hàng khác biệt dựa trên hành vi đặt phòng và đặc điểm chi tiêu.

### 2.3.2.1. Thuật toán K-Means

K-Means là thuật toán phân cụm phổ biến, hoạt động bằng cách chia dữ liệu thành **K** cụm sao cho tổng bình phương khoảng cách từ các điểm dữ liệu đến tâm cụm là nhỏ nhất. Thuật toán có ưu điểm là đơn giản, dễ triển khai và hiệu quả với tập dữ liệu lớn.

$$\min \sum_{i=1}^K \sum_{x \in C_i} ||x - \mu_i||^2$$

Trong đó  $C_i$  là cụm thứ  $i$  và  $\mu_i$  là tâm của cụm tương ứng. Thuật toán K-Means có ưu điểm là đơn giản, dễ triển khai và hiệu quả đối với các tập dữ liệu có kích thước lớn. Trong nghiên cứu này, K-Means được sử dụng để phân nhóm khách hàng dựa trên các đặc trưng hành vi đặt phòng, nhằm khám phá cấu trúc tiềm ẩn trong dữ liệu và làm cơ sở cho các phân tích ở các chương tiếp theo.

### 2.3.3. Luật kết hợp (Association Rules)

**Khai phá luật kết hợp** (Association Rule Mining) là phương pháp phân tích dữ liệu nhằm phát hiện các mối quan hệ đồng thời giữa các thuộc tính trong tập dữ liệu, thông qua việc tìm ra các mẫu hành vi thường xuyên xuất hiện cùng nhau. Phương pháp này thường được áp dụng để khám phá thói quen hoặc xu hướng lựa chọn của người dùng dựa trên các giao dịch quan sát được.

Một luật kết hợp có dạng  $A \rightarrow B$  được đánh giá thông qua ba độ đo chính:

- Support: Phản ánh mức độ phổ biến của luật trong tập dữ liệu

$$Support(A \rightarrow B) = P(A \cap B)$$

- Confidence: Thể hiện xác suất xảy ra của B khi A xảy ra

$$Confidence(A \rightarrow B) = \frac{P(A \cap B)}{P(A)}$$

- Lift: Thể hiện mức độ phụ thuộc A và B

$$Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{P(B)}$$

Trong phạm vi nghiên cứu này, khai phá luật kết hợp được thực hiện bằng các thuật toán như **Apriori** nhằm khám phá mối quan hệ giữa loại phòng, các dịch vụ đi kèm và đặc điểm khách hàng. Các luật thu được giúp làm rõ xu hướng lựa chọn dịch vụ của khách hàng, từ đó hỗ trợ để xuất các gói dịch vụ phù hợp và nâng cao hiệu quả quản trị doanh thu. Các kết quả cụ thể của quá trình khai phá luật kết hợp sẽ được trình bày và phân tích chi tiết ở các chương tiếp theo.

#### 2.3.3.1. Apriori

Apriori là một trong những thuật toán kinh điển được sử dụng trong khai phá luật kết hợp, nhằm tìm ra các tập mục phổ biến (frequent itemsets) và từ đó sinh ra các luật kết hợp thỏa mãn các ngưỡng đánh giá nhất định. Thuật toán hoạt động dựa trên nguyên lý Apriori, theo đó: *nếu một tập mục là phổ biến thì mọi tập con của nó cũng phải là phổ biến.*

Quy trình hoạt động của Apriori bao gồm hai bước chính. Trước hết, thuật toán tiến hành tìm kiếm các tập mục phổ biến bằng cách lặp lại quá trình sinh tập ứng viên và loại bỏ các tập không đạt ngưỡng support tối thiểu. Quá trình này bắt đầu từ các tập mục đơn lẻ và mở rộng dần sang các tập mục có kích thước lớn hơn. Sau khi xác định được các tập mục phổ biến, các luật kết hợp được sinh ra và đánh giá dựa trên các độ đo như confidence và lift, nhằm giữ lại các luật có ý nghĩa.

Trong bối cảnh dữ liệu đặt phòng khách sạn, thuật toán Apriori cho phép khám phá các mối quan hệ thường xuyên xuất hiện giữa loại phòng, các dịch vụ đi kèm và đặc điểm khách hàng. Các luật kết hợp thu được giúp làm rõ xu hướng lựa chọn dịch vụ của khách hàng, từ đó hỗ trợ việc thiết kế các gói dịch vụ phù hợp và nâng cao hiệu quả quản trị doanh thu. Tuy nhiên, Apriori có hạn chế là chi phí tính toán cao khi số lượng thuộc tính lớn hoặc ngưỡng support thấp, do đó cần lựa chọn tham số phù hợp khi áp dụng vào thực tế.

### 2.3.4. Chuỗi thời gian (Time Series)

**Phân tích chuỗi thời gian** là phương pháp nghiên cứu dữ liệu được thu thập theo trình tự thời gian nhằm nhận diện xu hướng, tính mùa vụ và các biến động theo thời gian. Trong nghiên cứu này, phân tích chuỗi thời gian được áp dụng để dự báo tổng lượng đặt phòng trong tương lai dựa trên dữ liệu lịch sử.

Việc dự báo chính xác nhu cầu đặt phòng theo thời gian giúp khách sạn chủ động trong công tác lập kế hoạch kinh doanh, phân bổ nguồn lực và xây dựng các biện pháp quản trị phù hợp trong các giai đoạn cao điểm và thấp điểm. Trên cơ sở đó, nghiên cứu xem xét khả năng ứng dụng các mô hình chuỗi thời gian nhằm hỗ trợ tối ưu hóa giá phòng, nhân sự và quản trị doanh thu trong ngắn và trung hạn.

#### 2.3.4.1. SARIMAX

SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) là mô hình thống kê mở rộng từ ARIMA, cho phép mô hình hóa đồng thời xu hướng, tính mùa vụ và ảnh hưởng của các biến ngoại sinh. Mô hình được ký hiệu dưới dạng:

$$\text{SARIMAX}(p, d, q)(P, D, Q, s)$$

Trong đó ( $q$ ) là các tham số của thành phần không mùa vụ và ( $s$ ) biểu diễn thành phần mùa vụ với chu kỳ  $s$ . SARIMAX đặc biệt phù hợp với dữ liệu đặt phòng có tính chu kỳ theo tháng hoặc theo mùa, đồng thời cho phép đưa thêm các yếu tố bên ngoài như ngày lễ hoặc sự kiện đặc biệt vào quá trình dự báo.

## 2.4. Nghiên cứu liên quan

Trong những năm gần đây, việc ứng dụng các kỹ thuật khai phá dữ liệu và học máy trong lĩnh vực du lịch – khách sạn đã thu hút sự quan tâm đáng kể từ cộng đồng nghiên cứu. Dữ liệu đặt phòng khách sạn được xem là nguồn dữ liệu quan trọng, phản ánh hành vi khách hàng, xu hướng tiêu dùng và rủi ro hủy phòng, qua đó hỗ trợ hiệu quả cho công tác quản trị và ra quyết định.

Nhiều nghiên cứu tập trung vào **phân tích mô tả và phân loại dữ liệu đặt phòng**, nhằm làm rõ các đặc điểm hành vi của khách hàng và dự đoán khả năng hủy phòng. Các mô hình phân loại như Logistic Regression, Decision Tree, Random Forest và SVM đã được áp dụng rộng rãi để ước lượng xác suất hủy phòng của từng đơn đặt (Chen et al., 2023; Moro et al., 2019). Kết quả cho thấy các mô hình phi tuyến, đặc biệt là Random Forest, thường mang lại độ chính xác dự báo cao hơn trong bối cảnh dữ liệu có mối quan hệ phức tạp giữa các biến.

Bên cạnh đó, **phân cụm khách hàng** là một hướng nghiên cứu quan trọng nhằm nhận diện các nhóm khách hàng có hành vi tương đồng dựa trên tần suất đặt phòng, thời gian lưu trú và mức chi tiêu. Deldadehasl (2025) đã áp dụng thuật toán K-Means kết hợp với các kỹ thuật phân tích đa tiêu chí để phân đoạn khách hàng trong ngành khách sạn, cho thấy phân cụm giúp cải thiện hiệu quả xây dựng chiến lược marketing và tối ưu doanh thu. Tương tự, Eibl (2024) sử dụng phân cụm để phân tích hành vi người dùng trong các tình huống đặt phòng, góp phần hỗ trợ cá nhân hóa dịch vụ và nâng cao trải nghiệm khách hàng.

Ngoài phân cụm, **khai phá luật kết hợp** cũng được sử dụng để khám phá các mối quan hệ đồng thời giữa loại phòng, dịch vụ đi kèm và đặc điểm khách hàng. Các nghiên cứu như Arreeras et al. (2019) cho thấy luật kết hợp có thể làm rõ các mẫu hành vi thường xuyên xuất hiện cùng nhau, từ đó hỗ trợ thiết kế gói dịch vụ, bán chéo (cross-selling) và nâng cao giá trị khách hàng. Tại Việt Nam, Nguyen Van Chuc và Dao Thi Giang (2015) đã ứng dụng kết hợp phân cụm và luật kết hợp trong phân tích dữ liệu khách hàng khách sạn, cho thấy tính khả thi của các kỹ thuật khai phá dữ liệu trong bối cảnh thực tế.

Bên cạnh các bài toán trên, **phân tích chuỗi thời gian** cũng được áp dụng rộng rãi để dự báo nhu cầu du lịch và lượng đặt phòng theo thời gian. Song và Li (2008) cung cấp một tổng quan toàn diện về các mô hình dự báo nhu cầu du lịch, nhấn mạnh vai trò của các mô hình chuỗi thời gian truyền thống và các phương pháp học máy trong quản trị doanh thu và lập kế hoạch nguồn lực. Các kết quả nghiên cứu cho thấy việc kết hợp dữ liệu lịch sử với mô hình dự báo giúp doanh nghiệp chủ động hơn trong việc điều chỉnh giá và chính sách kinh doanh.

Tuy nhiên, phần lớn các nghiên cứu trước đây thường tập trung vào **từng bài toán riêng lẻ**, chẳng hạn như chỉ phân loại hủy phòng hoặc chỉ phân cụm khách hàng, trong khi còn hạn chế các nghiên cứu tích hợp đồng thời nhiều kỹ thuật phân tích trên cùng một bộ dữ liệu. Do đó, đề tài này kế thừa các nghiên cứu trước và mở rộng bằng cách kết hợp phân tích mô tả, phân loại,

phân cụm, khai phá luật kết hợp và phân tích chuỗi thời gian nhằm hỗ trợ toàn diện cho công tác quản trị doanh thu và giảm thiểu rủi ro hủy phòng trong lĩnh vực khách sạn.

### CHƯƠNG III: DỮ LIỆU & PHƯƠNG PHÁP ĐỀ XUẤT

#### 3.1. Mô tả bảng dữ liệu

##### 3.1.1. Tổng quan dữ liệu sử dụng

Bộ dữ liệu *Hotel Booking Demand* đã được giới thiệu tổng quan trong **Chương 1 – Đôi tượng nghiên cứu**.

Trong phạm vi Chương này, dữ liệu được xem xét dưới góc độ **đầu vào cho quy trình khai phá dữ liệu**, nhằm phục vụ cho các bài toán **phân lớp, phân cụm, khai phá luật kết hợp và phân tích chuỗi thời gian**.

Dataset ghi nhận thông tin chi tiết về các lượt đặt phòng tại hai loại khách sạn là **City Hotel** và **Resort Hotel**, phản ánh hành vi đặt phòng, đặc điểm lưu trú và trạng thái hủy đặt phòng của khách hàng trong **giai đoạn 2015–2017**.

Trên cơ sở đó, các bước tiền xử lý và mô hình hóa dữ liệu sẽ được trình bày chi tiết trong các mục tiếp theo của chương này.

##### 3.1.2. Quy mô và cấu trúc dữ liệu

Bộ dữ liệu bao gồm khoảng **119.000 bản ghi** với **32 thuộc tính**, trong đó mỗi bản ghi tương ứng với một lần đặt phòng.

Các thuộc tính trong dataset có thể được chia thành các nhóm chính như sau:

**BẢNG ĐẶC TRUNG CỦA BỘ DỮ LIỆU HOTEL BOOKING DEMAND**

Nhóm đặc trưng	Tên đặc trưng	Mô tả
Khách sạn	Hotel	Loại khách sạn (City Hotel hoặc Resort Hotel )
Thời gian đặt phòng	Is_cancel	Đặt phòng bị hủy = 1 Đặt phòng không hủy = 0
	Lead_time	Số ngày từ lúc đặt phòng đến lúc nhận phòng
	Arrival_date_year	Năm nhận phòng
	Arrival_date_month	Tháng nhận phòng (January, February, ..)

Thời gian đặt phòng	Arrival_date_week_number	Tuần trong năm
	Arrival_date_day_of_month	Ngày trong tháng
	Stay_in_weekend_night	Số đêm lưu trú vào cuối tuần vào thứ 6 thứ 7
	Stay_in_week_nights	Số đêm lưu trú vào các ngày trong tuần (Mon - Thu)
Khách hàng	Adults	Số lượng người lớn
	Children	Số lượng trẻ em
	Babies	Số trẻ sơ sinh
	Meal	Loại suất ăn (BB, HB, FB, SC, ...)
	Country	Quốc gia của khách (mã ISO)
	Market_segment	Kênh phân phối (Online TA, Direct, Corporate, ...)
	Distribution_channel	Kênh phân phối (Direct, TA/TO)
	Is_repeated_guest	Khách quay lại = 1; Khách mới = 0
	Previous_cancellations	Số lần đặt trước đó bị hủy
	Previous_bookings_not_canceled	Số lần đặt trước đó không bị hủy

	Reserved_room_type	Loại phòng khách đặt
	Assigned_room_type	Loại phòng được giao khi nhận phòng
Tài chính và giá phòng	Booking_changes	Số lần khách thay đổi đặt phòng
	Deposit_type	Loại đặt phòng (No Deposite, Non Refund, Refundable)
	Agent	ID đại lý du lịch
	Company	ID công ty (nếu theo đoàn công ty)
	Adr	Giá phòng trung bình mỗi năm
Điều kiện đặt phòng	Required_car_parking_spaces	Số chỗ đỗ xe khách yêu cầu
	Total_of_special_requests	Số yêu cầu đặc biệt (tầng cao, giường đôi, ...)
Trạng thái đặt phòng	Reservation_status	Trạng thái (Cancel, Check-Out, No-Show)
	Reservation_status_date	Ngày ghi nhận trạng thái

Bảng 1: Các đặc trưng trong bộ dữ liệu

### 3.1.3. Các thuộc tính sử dụng trong nghiên cứu

Trong nghiên cứu này, không sử dụng toàn bộ 32 thuộc tính của bộ dữ liệu *Hotel Booking Demand*. Thay vào đó, các thuộc tính được **chọn lọc có chủ đích** dựa trên mục tiêu phân tích, bản chất của từng bài toán khai phá dữ liệu và ý nghĩa thực tiễn trong bối cảnh quản lý đặt phòng khách sạn.

#### 3.1.3.1. Thuộc tính cho bài toán phân lớp

Trong bài toán phân loại, biến mục tiêu (*target variable*) được xác định là **is canceled**, đại diện cho trạng thái hủy hay không hủy của mỗi đơn đặt phòng, từ đó hình thành một bài toán phân loại nhị phân.

Các biến đầu vào (*independent variables*) được lựa chọn dựa trên các nghiên cứu trước đây về hành vi đặt phòng khách sạn, đồng thời phản ánh đầy đủ các yếu tố thời gian, hành vi, đặc điểm lưu trú và mức độ cam kết của khách hàng, bao gồm:

**Nhóm biến thời gian** (*lead\_time, arrival\_date\_month, arrival\_date\_year, arrival\_date\_week\_number, ...*): phản ánh khoảng cách giữa thời điểm đặt và thời điểm lưu trú, cũng như tính mùa vụ, vốn được xem là yếu tố quan trọng ảnh hưởng đến xác suất hủy đặt phòng.

**Nhóm biến hành vi đặt phòng** (*booking\_changes, deposit\_type, ...*): thể hiện mức độ thay đổi và cam kết của khách hàng đối với đơn đặt phòng; các nghiên cứu chỉ ra rằng khách hàng có nhiều thay đổi hoặc không đặt cọc thường có xu hướng hủy cao hơn.

**Nhóm biến lưu trú** (*stays\_in\_week\_nights, stays\_in\_weekend\_nights, ...*): mô tả đặc điểm thời gian lưu trú, qua đó phản ánh mục đích chuyến đi (nghỉ dưỡng hay công tác), yếu tố có liên quan trực tiếp đến hành vi hủy.

**Nhóm biến khách hàng** (*adults, children, ...*): biểu thị quy mô và cấu trúc nhóm khách, ảnh hưởng đến sự linh hoạt và khả năng thay đổi kế hoạch lưu trú.

**Nhóm biến giá** (*adr*): đại diện cho mức chi phí trung bình mỗi ngày, có tác động đến quyết định duy trì hay hủy đặt phòng trong bối cảnh biến động giá và thu nhập.

Việc lựa chọn các nhóm biến trên nhằm giúp mô hình học máy nắm bắt toàn diện đặc điểm hành vi đặt phòng, mức độ cam kết cũng như bối cảnh kinh tế – thời gian của từng đơn đặt, từ đó nâng cao khả năng dự báo trạng thái hủy đặt phòng.

### 3.1.3.2. Thuộc tính cho bài toán phân cụm

Đối với bài toán phân cụm, mục tiêu nghiên cứu không nhằm dự đoán biến đích cụ thể mà tập trung vào việc **khám phá các nhóm khách hàng có đặc điểm và hành vi đặt phòng tương đồng**, từ đó hỗ trợ phân tích hành vi và xây dựng chiến lược kinh doanh phù hợp.

Các biến được lựa chọn cho bài toán phân cụm chủ yếu là các biến định lượng và có ý nghĩa hành vi, bao gồm:

**Hành vi đặt phòng** (*lead\_time, booking\_changes, deposit\_type*): phản ánh mức độ chủ động, linh hoạt và cam kết của khách hàng.

**Đặc điểm lưu trú** (*stays\_in\_week\_nights, stays\_in\_weekend\_nights*): giúp phân biệt các nhóm khách có mục đích lưu trú khác nhau như công tác, nghỉ dưỡng ngắn hạn hoặc dài hạn.

**Đặc điểm khách hàng** (*adults, children*): thể hiện quy mô và cấu trúc nhóm khách, yếu tố ảnh hưởng đến nhu cầu và hành vi tiêu dùng.

**Giá** (*adr*): đại diện cho mức chi tiêu trung bình, cho phép phân biệt các phân khúc khách hàng theo khả năng chi trả.

Trước khi tiến hành phân cụm, dữ liệu được tiền xử lý thông qua chuẩn hóa nhằm giảm ảnh hưởng của sự khác biệt về thang đo giữa các biến. Kết quả phân cụm kỳ vọng sẽ giúp nhận diện các nhóm khách hàng đặc trưng, làm cơ sở cho việc cá nhân hóa dịch vụ và hỗ trợ các phân tích tiếp theo như luật kết hợp hoặc phân loại.

### 3.1.3.3. Thuộc tính cho khai phá luật kết hợp

Bài toán luật kết hợp được sử dụng nhằm **khai phá các mối quan hệ tiềm ẩn giữa các đặc trưng của đơn đặt phòng**, đặc biệt là các mẫu hành vi thường xuyên xuất hiện đồng thời trong dữ liệu.

Các biến được sử dụng trong khai phá luật kết hợp chủ yếu là các biến rời rạc hoặc đã được rời rạc hóa (*discretization*), bao gồm:

**Thời gian đặt phòng** (*lead\_time* – phân nhóm ngắn, trung bình, dài)

**Hành vi và cam kết** (*deposit\_type, booking\_changes*)

**Đặc điểm lưu trú** (*stays\_in\_week\_nights, stays\_in\_weekend\_nights*)

**Trạng thái hủy** (*is\_canceled*)

Việc rời rạc hóa các biến liên tục cho phép chuyển đổi dữ liệu sang dạng giao dịch, phù hợp với các thuật toán khai phá luật kết hợp như Apriori

Các luật kết hợp được đánh giá dựa trên các chỉ số **support, confidence và lift**, qua đó giúp xác định các mẫu hành vi phổ biến, chẳng hạn như mối liên hệ giữa thời gian đặt phòng dài, không đặt cọc và khả năng hủy cao. Kết quả của bài toán này cung cấp góc nhìn diễn giải (interpretability), hỗ trợ các nhà quản lý trong việc xây dựng chính sách đặt cọc hoặc điều chỉnh điều khoản hủy phòng.

### 3.1.3.4. Thuộc tính cho phân tích chuỗi thời gian

Phân tích chuỗi thời gian được áp dụng nhằm **khám phá xu hướng, tính mùa vụ và biến động theo thời gian của hoạt động đặt phòng và hủy phòng**, từ đó hỗ trợ công tác dự báo và hoạch định nguồn lực.

Dữ liệu chuỗi thời gian được xây dựng bằng cách tổng hợp các đơn đặt phòng theo đơn vị thời gian (ngày, tuần hoặc tháng), với các biến quan tâm bao gồm:

**Số lượng đơn đặt phòng**

## Số lượng đơn bị hủy

### Tỷ lệ hủy đặt phòng theo thời gian

Việc phân tích tập trung vào việc xác định các thành phần chính của chuỗi thời gian như **xu hướng dài hạn (trend)**, **tính mùa vụ (seasonality)** và **nhiều ngẫu nhiên (noise)**. Các mô hình chuỗi thời gian như SARIMAX hoặc các phương pháp làm mịn có thể được áp dụng để dự báo nhu cầu và tỷ lệ hủy trong các giai đoạn tương lai.

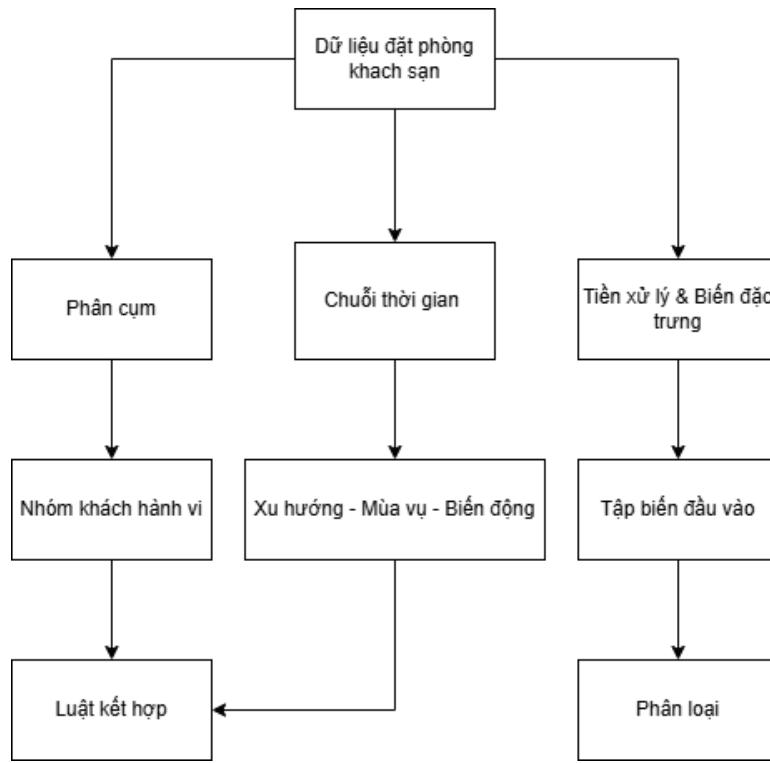
Kết quả phân tích chuỗi thời gian giúp các nhà quản lý khách sạn chủ động trong việc điều chỉnh chiến lược giá, chính sách hủy phòng và phân bổ nguồn lực theo từng giai đoạn cao điểm hoặc thấp điểm trong năm.

#### 3.1.4. Vai trò của dữ liệu đối với các bài toán khai phá

Dữ liệu sau khi được tiền xử lý được sử dụng làm đầu vào cho các bài toán khai phá dữ liệu khác nhau nhằm phục vụ các mục tiêu nghiên cứu của đế tài, bao gồm:

- **Phân lớp:** dự đoán khả năng hủy đặt phòng của khách hàng.
- **Phân cụm:** phân nhóm khách hàng dựa trên các đặc trưng và hành vi đặt phòng tương đồng.
- **Luật kết hợp:** khai phá các mẫu hành vi đặt phòng phổ biến và các mối quan hệ tiềm ẩn giữa các thuộc tính.
- **Chuỗi thời gian:** phân tích xu hướng đặt phòng theo thời gian và dự đoán tần suất đặt phòng trong tương lai.

Nhờ quy mô lớn cùng với tính đa dạng của các thuộc tính, bộ dữ liệu *Hotel Booking Demand* đáp ứng tốt yêu cầu thực nghiệm của đế tài, đồng thời phù hợp với mục tiêu và nội dung của học phần Khai phá dữ liệu.



Hình 3: Minh họa mối quan hệ dữ liệu và các bài toán

Mối liên hệ giữa dữ liệu và các bài toán khai phá được minh họa trong Hình 3

### 3.2. Quy trình tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước then chốt trong quy trình khai phá dữ liệu, đóng vai trò quan trọng trong việc đảm bảo chất lượng và hiệu quả của các mô hình học máy. Đối với bộ dữ liệu *Hotel Booking Demand* có quy mô lớn và cấu trúc thuộc tính đa dạng, một quy trình tiền xử lý có hệ thống đã được xây dựng nhằm chuẩn bị dữ liệu đầu vào cho các bài toán phân tích và mô hình hóa.

Cụ thể, quy trình này bao gồm việc xử lý các giá trị thiếu, loại bỏ các bản ghi không hợp lệ, chuẩn hóa dữ liệu số và lựa chọn các thuộc tính phù hợp, qua đó giúp dữ liệu đầu vào đáp ứng yêu cầu của các bài toán phân lớp, phân cụm, luật kết hợp và phân tích chuỗi thời gian.

Chi tiết các bước sẽ được trình bày ở các mục sau

#### 3.2.1. Kiểm tra và làm sạch dữ liệu

##### 3.2.1.1. Xử lý giá trị thiếu

Qua quá trình kiểm tra dữ liệu, một số thuộc tính xuất hiện giá trị thiếu, chủ yếu bao gồm *children*, *country*, *agent* và *company*. Các phương pháp xử lý giá trị thiếu được lựa chọn dựa trên bản chất và vai trò của từng thuộc tính trong bộ dữ liệu:

**Đối với thuộc tính *children* (biến số):** các giá trị thiếu được thay thế bằng **giá trị trung vị**, do trong thực tế phần lớn các đơn đặt phòng không có trẻ em, đồng thời phương pháp này giúp giảm ảnh hưởng của các giá trị ngoại lệ.

**Đối với thuộc tính *country* (biến phân loại):** các giá trị thiếu được gán bằng **giá trị xuất hiện nhiều nhất (mode)** nhằm giữ lại các bản ghi và hạn chế làm sai lệch phân bố dữ liệu.

**Đối với các thuộc tính *agent* và *company*:** các giá trị thiếu được thay thế bằng **0**, biểu thị các trường hợp đặt phòng không thông qua đại lý hoặc công ty, phù hợp với ý nghĩa thực tế của hai thuộc tính này.

Cách xử lý trên giúp hạn chế mất mát dữ liệu, đảm bảo tính nhất quán của tập dữ liệu và tạo điều kiện thuận lợi cho các bước phân tích và mô hình hóa tiếp theo.

### 3.2.1.2. Loại bỏ dữ liệu không hợp lệ

Qua quá trình kiểm tra dữ liệu, một số bản ghi có giá trị bất hợp lý đã được loại bỏ khỏi tập dữ liệu. Cụ thể:

**Các đơn đặt phòng có tổng số khách (adults + children + babies) bằng 0,** không phản ánh một giao dịch đặt phòng hợp lệ trong thực tế.

**Các bản ghi có giá trị *adr* (giá phòng trung bình mỗi ngày) âm,** trái với ý nghĩa kinh tế của biến và có khả năng là lỗi trong quá trình thu thập hoặc ghi nhận dữ liệu.

Việc loại bỏ các bản ghi này nhằm đảm bảo tính hợp lệ của dữ liệu, đồng thời hạn chế nhiễu và sai lệch trong quá trình huấn luyện và đánh giá các mô hình học máy.

### 3.2.1.3. Xử lý giá nhiễu và ngoại lai

Trong quá trình khám phá dữ liệu, một số thuộc tính số xuất hiện các giá trị ngoại lai có độ lệch lớn so với phân bố chung của dữ liệu. Các giá trị này có thể phát sinh do sai sót trong quá trình thu thập dữ liệu hoặc phản ánh các trường hợp hiếm gặp, từ đó có khả năng gây ảnh hưởng tiêu cực đến hiệu quả và độ ổn định của các mô hình học máy.

Việc phát hiện ngoại lai được thực hiện thông qua phân tích thống kê mô tả và trực quan hóa dữ liệu (boxplot), tập trung vào các thuộc tính số quan trọng như *lead\_time*, *adr*, *children*, *babies*, *adults*, *stays\_in\_week\_nights* và *stays\_in\_weekend\_nights*. Dựa trên kết quả phân tích, các giá trị nằm ngoài khoảng phân vị hợp lý được xem là ngoại lai.

Đối với các giá trị ngoại lai, thay vì loại bỏ hoàn toàn bản ghi, phương pháp **giới hạn giá trị (capping)** được áp dụng nhằm giảm ảnh hưởng của các giá trị cực đoan nhưng vẫn giữ lại

thông tin tổng thể của dữ liệu. Cụ thể, các giá trị vượt quá ngưỡng trên và dưới được điều chỉnh về mức giới hạn tương ứng.

Cách tiếp cận này giúp hạn chế tác động của ngoại lai đến quá trình huấn luyện mô hình, đồng thời đảm bảo không làm mất đi các quan sát hợp lệ, qua đó nâng cao tính ổn định và khả năng tổng quát hóa của các mô hình phân tích tiếp theo.

Việc xử lý ngoại lai được thực hiện trước bước chuẩn hóa dữ liệu nhằm đảm bảo phân bố dữ liệu đầu vào phù hợp với yêu cầu của các thuật toán học máy.

### 3.2.2. Chuyển đổi và chuẩn hóa dữ liệu

#### 3.2.2.1 Chuyển đổi kiểu dữ liệu

Các thuộc tính thời gian như *arrival\_date\_month* được chuyển đổi từ dạng chuỗi sang dạng số thứ tự của tháng nhằm thuận tiện cho việc phân tích và mô hình hóa. Thuộc tính sau khi chuyển đổi được lưu với tên *arrival\_month\_num* trong bộ dữ liệu.

Bên cạnh đó, các thuộc tính phân loại được chuẩn hóa về cùng định dạng nhằm đảm bảo tính nhất quán của dữ liệu, tạo điều kiện thuận lợi cho các bước xử lý tiếp theo như mã hóa biến và huấn luyện các mô hình học máy.

#### 3.2.2.2. Mã hóa biến phân loại

Để sử dụng cho các thuật toán học máy, các biến phân loại trong bộ dữ liệu được mã hóa phù hợp với bản chất của từng thuộc tính và cơ chế hoạt động của từng mô hình. Trong nghiên cứu này, **One-Hot Encoding** được áp dụng cho mô hình **Logistic Regression**, trong khi **Ordinal Encoding** được sử dụng cho các mô hình **Decision Tree** và **Random Forest**.

Đối với **Logistic Regression**, mô hình giả định mối quan hệ tuyến tính giữa các biến đầu vào và log-odds của biến mục tiêu. Việc sử dụng One-Hot Encoding giúp biểu diễn các biến phân loại dưới dạng các biến nhị phân độc lập, tránh việc áp đặt mối quan hệ thứ tự không tồn tại giữa các nhãn. Nhờ đó, mô hình có thể học được ảnh hưởng riêng biệt của từng giá trị phân loại và đảm bảo tính đúng đắn của giả định tuyến tính.

Ngược lại, đối với các mô hình dựa trên cây quyết định như **Decision Tree** và **Random Forest**, Ordinal Encoding có thể được áp dụng mà không gây ra sai lệch đáng kể. Các mô hình này không dựa trên khoảng cách hay giả định tuyến tính, mà học các quy tắc phân tách dựa trên ngưỡng giá trị. Do đó, việc gán các nhãn phân loại thành giá trị số nguyên không làm ảnh hưởng đến hiệu quả học của mô hình, đồng thời giúp giảm số chiều dữ liệu và cải thiện hiệu quả tính toán so với One-Hot Encoding.

**Label Encoding** không được sử dụng trực tiếp trong nghiên cứu này do phương pháp này có thể tạo ra mối quan hệ thứ tự giả giữa các nhãn phân loại vốn không có thứ tự tự nhiên. Điều này đặc biệt không phù hợp đối với các mô hình tuyến tính như Logistic Regression, khi mô hình có thể diễn giải sai sự khác biệt giữa các nhãn như một xu hướng tăng hoặc giảm. Mặc

dù Label Encoding về mặt kỹ thuật có thể áp dụng cho các mô hình cây quyết định, nghiên cứu này vẫn ưu tiên Ordinal Encoding nhằm đảm bảo tính nhất quán, khả năng kiểm soát ý nghĩa của giá trị mã hóa và tính minh bạch của pipeline tiền xử lý.

Việc lựa chọn phương pháp mã hóa theo từng nhóm mô hình được thực hiện nhất quán nhằm đảm bảo tính công bằng trong so sánh kết quả thực nghiệm.

### 3.2.2.3. Chuẩn hóa dữ liệu số

Các thuộc tính số có thang đo khác nhau như *lead\_time*, *adr* và *total\_nights* được chuẩn hóa bằng phương pháp **Standardization**, nhằm đưa các biến về cùng thang đo và đảm bảo chúng đóng góp công bằng vào mô hình làm giảm ảnh hưởng của các giá trị ngoại lai còn sót lại và phù hợp với các mô hình học máy được sử dụng trong nghiên cứu. Bước chuẩn hóa này đặc biệt quan trọng đối với các thuật toán dựa trên khoảng cách như **K-Means**, đồng thời giúp cải thiện độ ổn định của các mô hình học máy được sử dụng trong nghiên cứu.

## 3.2.3. Tạo thuộc tính mới

Nhằm phản ánh rõ hơn hành vi đặt phòng của khách hàng và cung cấp các đặc trưng tổng hợp có ý nghĩa cho mô hình, một số thuộc tính mới đã được xây dựng từ các biến gốc trong bộ dữ liệu. Cụ thể:

### Tổng số khách:

$$total\_guests = adults + children + babies$$

Biến này phản ánh quy mô nhóm khách trong mỗi đơn đặt phòng.

### - Tổng số đêm lưu trú:

$$total\_stay = stays\_in\_week\_nights + stays\_in\_weekend\_nights$$

Thuộc tính này thể hiện thời gian lưu trú tổng thể của khách hàng.

### - Chuyến đi gia đình (*is\_family\_trip*):

Biến nhị phân được tạo dựa trên hai thuộc tính *children* và *babies*. Nếu tổng số trẻ em lớn hơn 0 thì biến nhận giá trị 1 (chuyến đi gia đình), ngược lại nhận giá trị 0.

### - Khách quay lại (*has\_previous*):

Biến nhị phân được xây dựng dựa trên thuộc tính *previous\_cancellations*. Nếu giá trị lớn hơn 0, biến nhận giá trị 1, biểu thị khách hàng đã từng có lịch sử đặt phòng trước đó.

### - Giá phòng trung bình trên mỗi khách (*adr\_per\_guest*):

$$adr\_per\_guest = adr / total\_quest$$

Thuộc tính này phản ánh mức chi tiêu trung bình trên mỗi khách, giúp chuẩn hóa thông tin giá phòng theo quy mô nhóm khác

- **Phân nhóm thời gian đặt trước (*lead\_time\_bucket*):**

*lead\_time* được rời rạc hóa thành 5 nhóm: 0–30 ngày, 30–90 ngày, 90–180 ngày, 180–365 ngày và trên 365 ngày, nhằm phục vụ phân tích hành vi và khai phá luật kết hợp.

- **Mùa du lịch (season):**

Thời điểm lưu trú được phân chia theo mùa dựa trên tháng nhận phòng: mùa đông (12–2), mùa xuân (3–5), mùa hè (6–8) và mùa thu (9–11).

- **Thời điểm nhận phòng (*arrival\_date*):**

Thuộc tính này được tổng hợp từ *arrival\_date\_year*, *arrival\_month\_num* và *arrival\_date\_day\_of\_month* nhằm biểu diễn thời điểm nhận phòng một cách thống nhất.

Các thuộc tính được xây dựng giúp giảm số lượng biến đầu vào, tăng khả năng diễn giải và cung cấp thông tin tổng hợp có ý nghĩa hơn cho các mô hình phân tích và dự đoán.

Việc xây dựng các đặc trưng mới này góp phần nâng cao hiệu quả dự báo, đồng thời hỗ trợ tốt hơn cho các bài toán phân cụm và khai phá luật kết hợp.

Do một số đặc trưng được xây dựng từ các biến gốc, tồn tại khả năng trùng thông tin giữa các thuộc tính. Vì vậy, trong quá trình mô hình hóa, các biến có quan hệ phụ thuộc được lựa chọn có điều kiện tùy theo từng bài toán và thuật toán nhằm hạn chế đa cộng tuyến, giảm nhiễu và nâng cao hiệu quả mô hình.

#### 3.2.4. Chọn lọc thuộc tính

Không phải tất cả các thuộc tính trong bộ dữ liệu đều phù hợp cho việc khai phá và mô hình hóa. Do đó, quá trình chọn lọc thuộc tính được thực hiện nhằm giảm nhiễu, hạn chế trùng thông tin và nâng cao hiệu quả của các mô hình học máy.

Cụ thể, việc chọn lọc thuộc tính được dựa trên ba tiêu chí chính:

**Phân tích tương quan giữa các biến số**, nhằm phát hiện và loại bỏ các thuộc tính có mối quan hệ phụ thuộc mạnh, từ đó hạn chế hiện tượng đa cộng tuyến.

**Ý nghĩa thực tiễn của thuộc tính**, đảm bảo các biến được giữ lại có khả năng phản ánh hành vi đặt phòng và đặc điểm của khách hàng.

**Đánh giá mức độ quan trọng của thuộc tính (feature importance)** thông qua mô hình Random Forest, được sử dụng như một công cụ hỗ trợ nhằm xác định các biến có đóng góp lớn vào khả năng dự đoán.

Bên cạnh đó, các thuộc tính mang tính định danh hoặc không có giá trị dự đoán, chẳng hạn như *reservation\_status\_date*, *reservation\_status*, *assigned\_room\_type*, *company*, *agent*, được loại bỏ khỏi tập dữ liệu để tránh rò rỉ thông tin và đảm bảo tính khách quan trong quá trình huấn luyện mô hình.

### 3.2.5. Chuẩn bị dữ liệu cho mô hình hóa

Sau khi hoàn tất các bước tiền xử lý và lựa chọn thuộc tính, tập dữ liệu được chia thành hai phần độc lập: **tập huấn luyện (training set)** và **tập kiểm tra (test set)**.

Tập huấn luyện được sử dụng để xây dựng và điều chỉnh các mô hình khai phá dữ liệu, trong khi tập kiểm tra được giữ lại để đánh giá hiệu quả dự đoán của mô hình trên dữ liệu chưa từng được sử dụng trong quá trình huấn luyện. Việc chia dữ liệu theo cách này giúp đảm bảo tính khách quan và hạn chế hiện tượng quá khớp (overfitting) trong quá trình đánh giá các mô hình được đề xuất ở các phần tiếp theo.

## 3.3. Khám phá dữ liệu (EDA)

Phân tích khám phá dữ liệu (Exploratory Data Analysis – EDA) được thực hiện nhằm hiểu rõ cấu trúc của tập dữ liệu, phân bố của các thuộc tính quan trọng và mối quan hệ giữa các biến. Thông qua EDA, các đặc điểm nổi bật của dữ liệu như xu hướng, sự mất cân bằng, cũng như các mối tương quan tiềm ẩn được nhận diện.

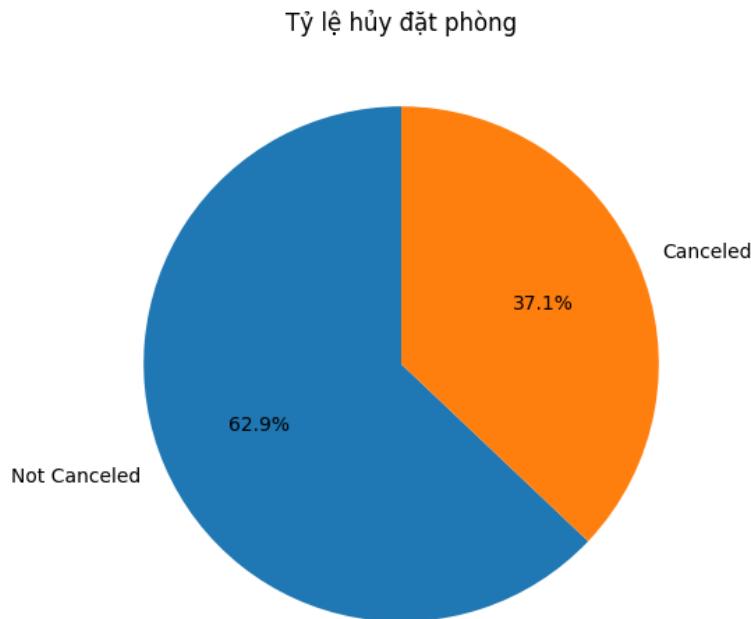
Các kết quả thu được từ EDA đóng vai trò nền tảng cho việc lựa chọn đặc trưng, xử lý dữ liệu và đề xuất các mô hình khai phá dữ liệu phù hợp, đồng thời định hướng thiết kế các thí nghiệm và chiến lược mô hình hóa ở các phần tiếp theo của nghiên cứu.

EDA cũng giúp phát hiện các vấn đề tiềm ẩn trong dữ liệu, từ đó hỗ trợ điều chỉnh pipeline tiền xử lý và nâng cao độ tin cậy của kết quả thực nghiệm.

### 3.3.1. Thống kê mô tả

Phân tích thống kê mô tả cho thấy một số đặc điểm đáng chú ý của bộ dữ liệu. Tỷ lệ các đơn đặt phòng bị hủy chiếm một tỷ lệ đáng kể trong tổng số booking, phản ánh tính không ổn

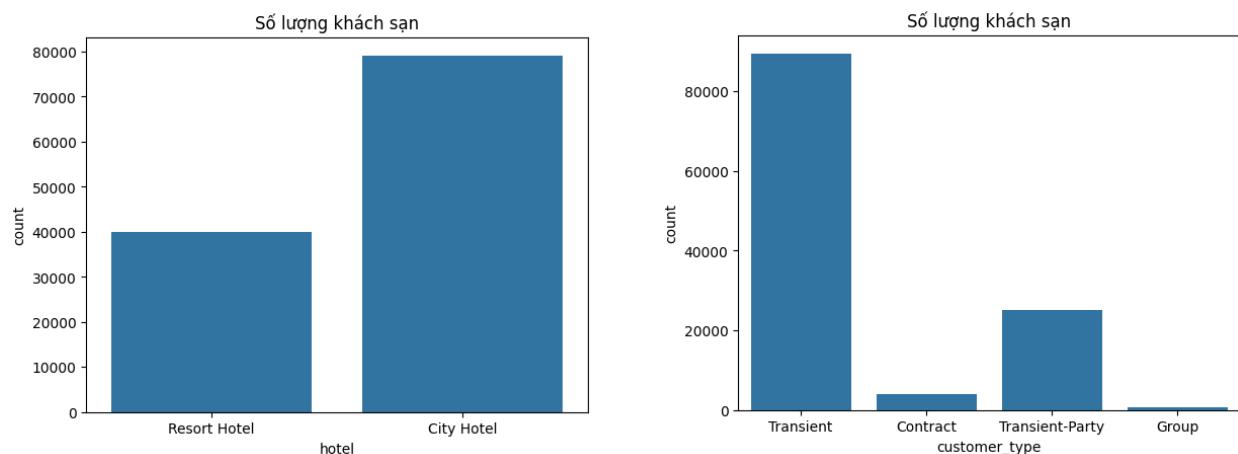
định trong hành vi đặt phòng của khách hàng và cho thấy tính phù hợp của dữ liệu đối với bài toán phân lớp.



Hình 4.1: Minh họa phân bố tỷ lệ hủy phòng

Quan sát biểu đồ phân bố theo loại khách sạn cho thấy số lượng booking tập trung chủ yếu ở **City Hotel**, trong khi **Resort Hotel** chiếm tỷ trọng nhỏ hơn. Điều này phản ánh sự mất cân đối về phân bố quan sát giữa hai nhóm khách sạn và cần được lưu ý khi so sánh hành vi đặt phòng giữa các loại hình.

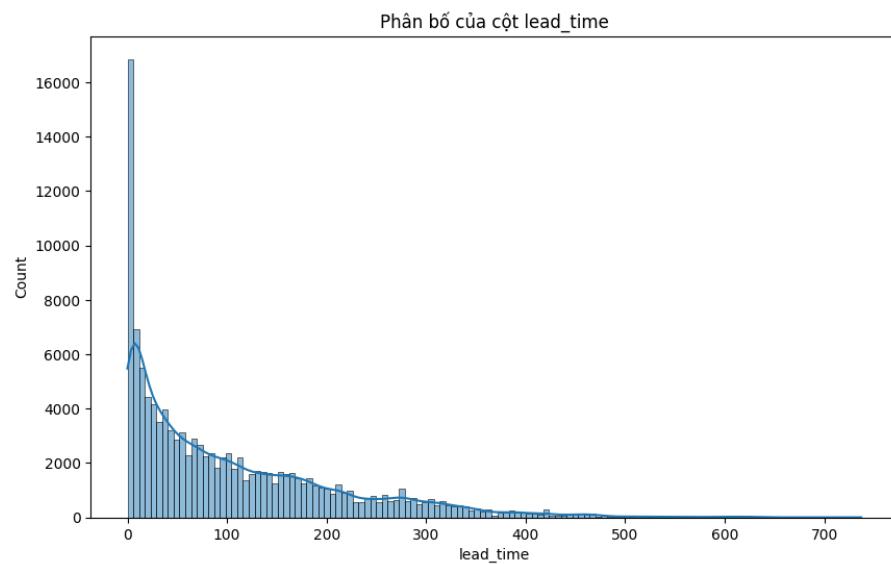
Đối với biến *customer\_type*, nhóm khách hàng **Transient** chiếm tỷ lệ lớn nhất, cho thấy phần lớn đơn đặt phòng đến từ khách lẻ/khách đặt ngắn hạn. Các nhóm **Contract**, **Transient-Party** và **Group** xuất hiện với tần suất thấp hơn, gợi ý rằng dữ liệu có sự chênh lệch về quy mô giữa các phân khúc khách hàng; vì vậy khi phân tích theo phân khúc hoặc mô hình hóa, cần cân nhắc ảnh hưởng của sự mất cân bằng này đến kết luận.



Hình 4.2: Minh họa phân khúc của dữ liệu:

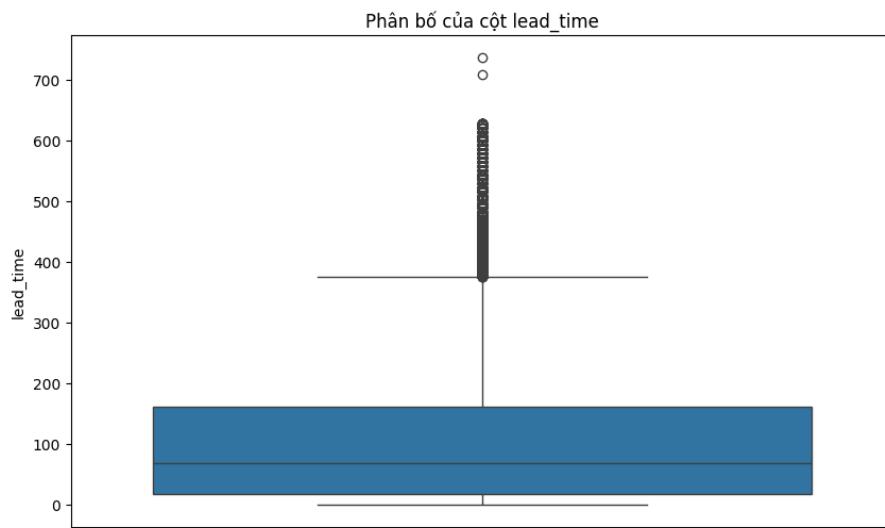
### 3.3.2. Phân bố dữ liệu của các thuộc tính chính

Quan sát histogram cho thấy *lead\_time* có phân bố lệch phải với mật độ tập trung lớn ở các giá trị nhỏ, trong khi vẫn tồn tại một số trường hợp đặt phòng rất sớm tạo thành **đuôi dài**. Điều này phản ánh sự khác biệt đáng kể giữa các nhóm khách hàng về thời điểm đặt trước ngày lưu trú. Do đó, *lead\_time* là một đặc trưng quan trọng trong quá trình mô hình hóa dự đoán hành vi đặt phòng/hủy phòng; đồng thời cần cân nhắc xử lý các giá trị cực đoan và chuẩn hóa thang đo để nâng cao độ ổn định của mô hình.



Hình 4.2: Minh họa phân khúc của dữ liệu:

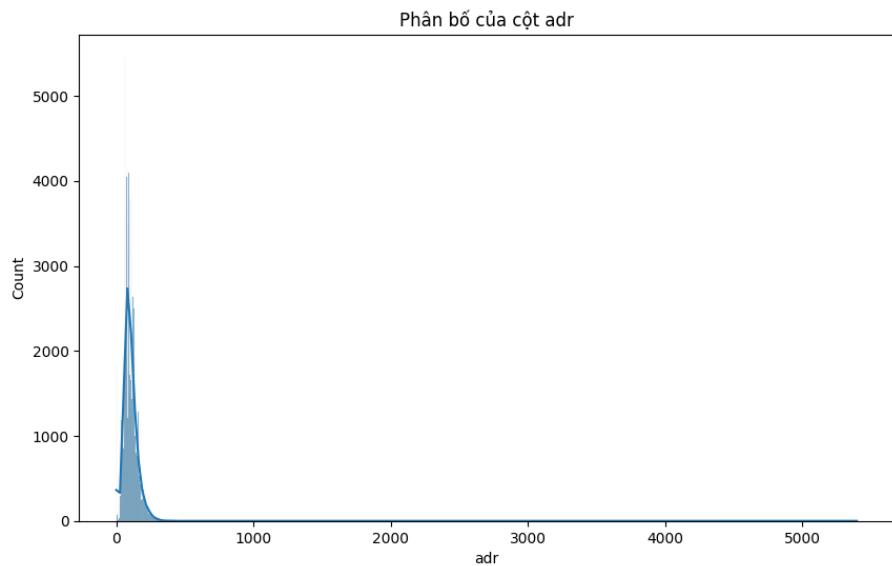
Boxplot của biến *lead\_time* cho thấy phân bố có **nhiều giá trị ngoại lai ở phía trên**, thể hiện một số đơn đặt phòng được thực hiện rất sớm so với phần lớn quan sát còn lại. Đồng thời,



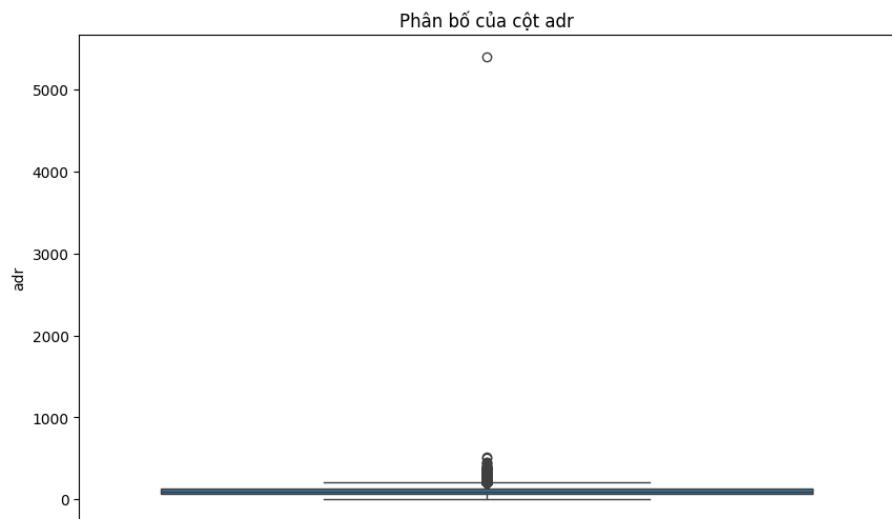
Hình 4.4: Minh họa Boxplot *lead\_time*

phản hộp (IQR) tập trung ở vùng giá trị thấp cho thấy đa số khách hàng có thời gian đặt trước tương đối ngắn. Sự hiện diện của các ngoại lai này có thể làm sai lệch quá trình huấn luyện đối với các mô hình nhạy với thang đo; do đó cần cân nhắc áp dụng các biện pháp xử lý ngoại lai (ví dụ: giới hạn giá trị/capping) hoặc biến đổi phù hợp trước khi mô hình hóa.

Quan sát histogram cho thấy biến *adr* có phân bố lệch phải mạnh, với phần lớn giá trị tập trung ở mức thấp đến trung bình, trong khi tồn tại một số giá trị rất lớn tạo thành **đuôi dài**. Điều này phản ánh mức giá phòng có độ biến thiên cao và có thể chịu ảnh hưởng bởi các trường hợp đặc biệt, do đó *adr* là một đặc trưng quan trọng nhưng cần cân nhắc xử lý ngoại lai và chuẩn hóa trước khi đưa vào mô hình.

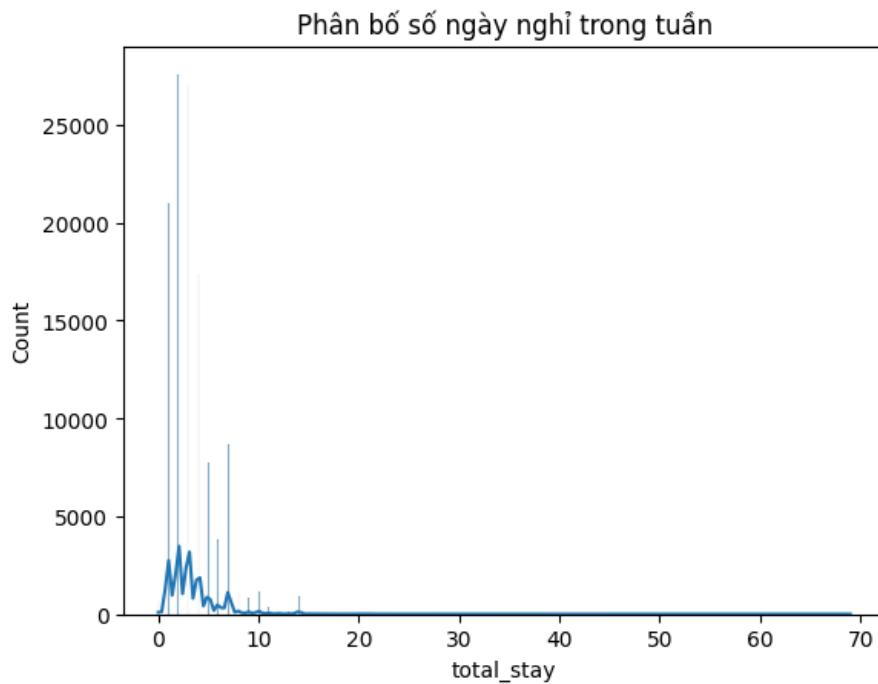


Hình 4.5: Minh họa phân bố của *adr*



Hình 4.6: Minh họa Boxplot *adr*

Boxplot của biến *adr* cho thấy phân bố tập trung chủ yếu ở vùng giá trị thấp đến trung bình, đồng thời xuất hiện **nhiều điểm ngoại lai ở phía trên**, bao gồm một số trường hợp có giá trị **rất lớn** so với phần lớn quan sát. Điều này cho thấy mức giá phòng có độ biến thiên cao và các trường hợp đặc biệt có thể gây ảnh hưởng đáng kể đến các mô hình học máy nhạy với thang đo. Do đó, biến *adr* cần được cân nhắc **xử lý ngoại lai** (ví dụ: giới hạn giá trị/capping) và **chuẩn hóa** trước khi đưa vào mô hình hóa nhằm đảm bảo tính ổn định của kết quả.

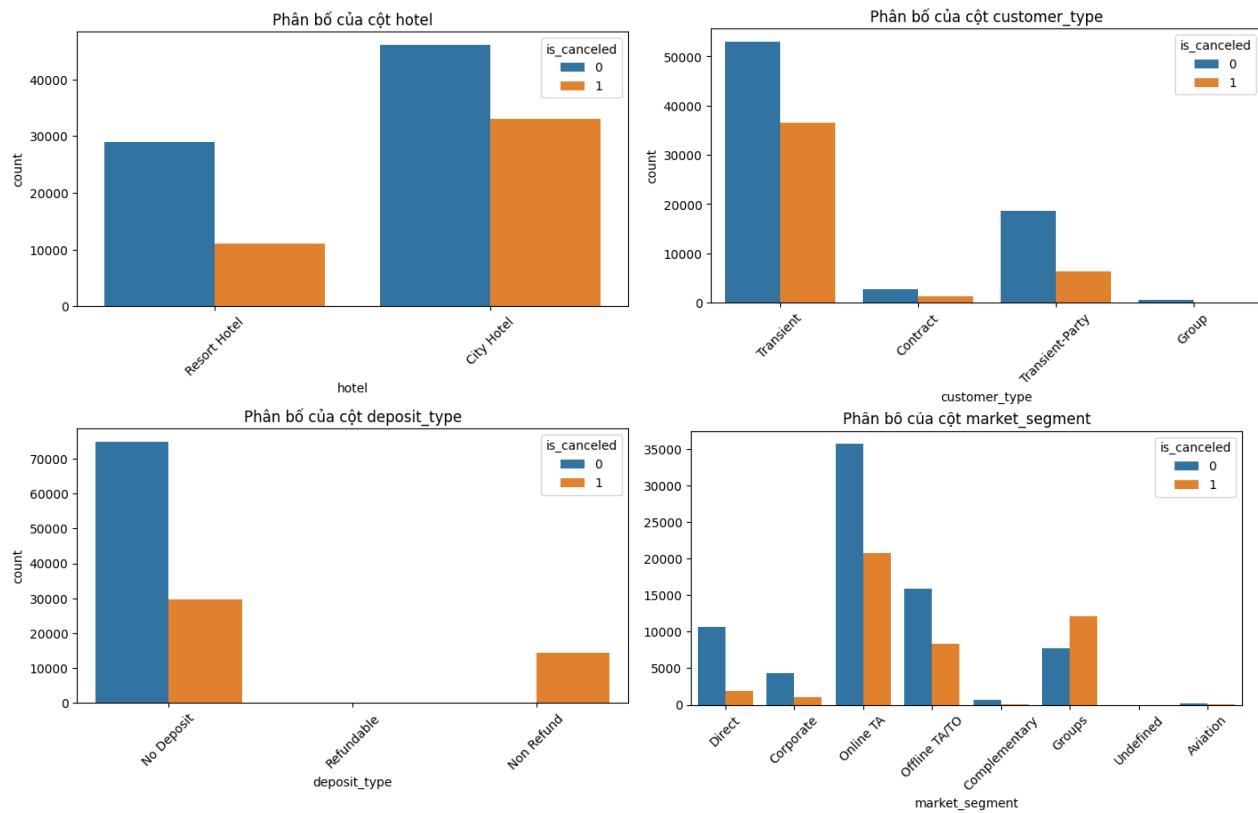


Hình 4.7: Minh họa phân bố của *total\_stay*

Biểu đồ phân bố của *total\_stay* cho thấy phần lớn các đơn đặt phòng có thời gian lưu trú **ngắn**, tập trung chủ yếu ở một vài ngày đầu, trong khi số lượng trường hợp lưu trú dài ngày giảm nhanh và tạo thành **đuôi dài** về phía các giá trị lớn. Điều này phản ánh hành vi đặt phòng phổ biến là lưu trú ngắn hạn, đồng thời tồn tại một số trường hợp đặc biệt có thời gian lưu trú rất dài. Do đó, *total\_stay* là một đặc trưng có ý nghĩa để mô tả hành vi lưu trú và có thể liên quan đến tỷ lệ hủy đặt phòng; tuy nhiên cần cân nhắc xử lý các giá trị cực đoan và chuẩn hóa thang đo trước khi mô hình hóa.

### 3.3.3. Phân tích mối quan hệ với biến mục tiêu

Các biểu đồ phân bố theo biến mục tiêu *is\_canceled* cho thấy tỷ lệ hủy đặt phòng có sự khác biệt rõ rệt giữa các nhóm thuộc tính phân loại, gợi ý rằng các biến này có khả năng phân biệt tốt giữa hai lớp và phù hợp để đưa vào mô hình phân lớp.



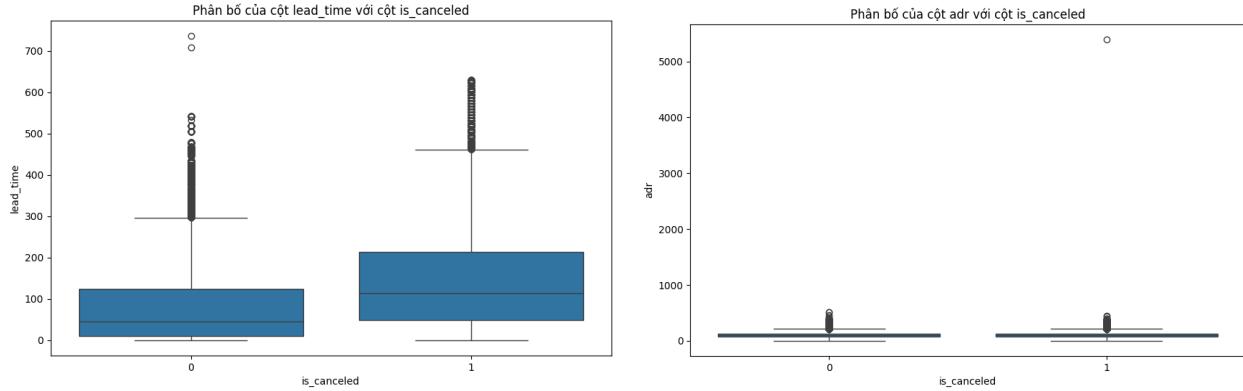
Hình 4.8: Các yếu tố phân loại ảnh hưởng đến biến mục tiêu

Đối với thuộc tính *hotel*, số lượng booking và số lượng hủy đều tập trung nhiều hơn ở **City Hotel** do quy mô quan sát lớn hơn. Tuy nhiên, để đánh giá mức độ rủi ro hủy giữa hai loại khách sạn một cách khách quan, cần so sánh theo **tỷ lệ hủy (cancellation rate)** thay vì chỉ dựa trên số lượng tuyệt đối.

Với *customer\_type*, nhóm **Transient** chiếm ưu thế về số lượng ở cả hai trạng thái hủy và không hủy, phản ánh rằng phần lớn booking đến từ khách lẻ. Các nhóm **Contract**, **Transient-Party** và **Group** có quy mô nhỏ hơn; do đó khi phân tích tỷ lệ hủy theo nhóm cần lưu ý yếu tố mất cân bằng về số lượng mẫu.

Đáng chú ý, biến *deposit\_type* thể hiện sự khác biệt mạnh theo trạng thái hủy: nhóm **No Deposit** chiếm tỷ trọng lớn nhất trong dữ liệu, trong khi nhóm **Non Refund** có cấu trúc phân bố khác biệt, gợi ý mức độ cam kết thanh toán có thể liên quan đến hành vi hủy đặt phòng. Điều này cho thấy *deposit\_type* là một đặc trưng quan trọng cần được xem xét trong mô hình dự đoán.

Ngoài ra, phân bố theo *market\_segment* cho thấy hành vi hủy không đồng nhất giữa các phân khúc thị trường; một số phân khúc có số lượng hủy tương đối cao so với số lượng đặt phòng, gợi ý đây là biến có giá trị trong việc phân khúc khách hàng và giải thích rủi ro hủy phòng.



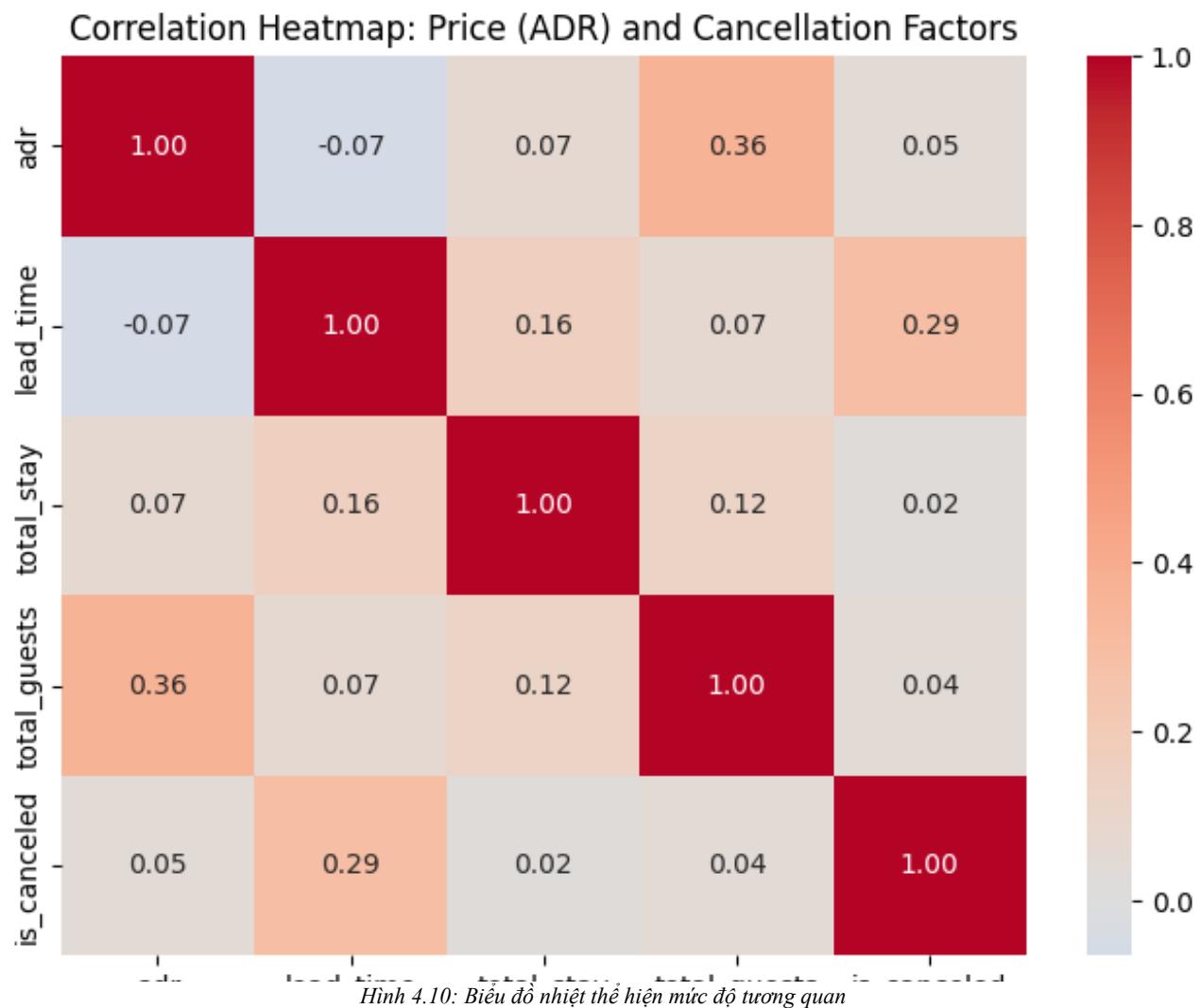
Hình 4.9: Các yếu tố biến số ảnh hưởng đến biến mục tiêu

Boxplot cho thấy biến *lead\_time* có sự khác biệt rõ rệt giữa hai nhóm **không hủy** và **hủy**. Nhóm bị hủy xu hướng có *thời gian chờ cao hơn*, thể hiện khách hàng đặt phòng càng sớm thì rủi ro hủy có thể tăng. Đồng thời, cả hai nhóm đều xuất hiện nhiều điểm ngoại lai ở phía giá trị lớn, cho thấy hành vi đặt trước rất sớm tồn tại nhưng không phổ biến; điều này cũng cống có nhu cầu xử lý ngoại lai và chuẩn hóa trước khi huấn luyện mô hình.

Đối với biến *adr*, phân bố giữa hai nhóm hủy và không hủy có sự chênh lệch nhưng không thể hiện rõ ràng như *lead\_time* do ảnh hưởng của các giá trị ngoại lai. Tuy vậy, sự khác biệt về trung vị và độ phân tán của *adr* vẫn gợi ý rằng mức giá có thể liên quan đến hành vi hủy đặt phòng, và cần được kiểm chứng định lượng thông qua mô hình phân lớp/đánh giá mức độ quan trọng của đặc trưng trong các phần tiếp theo.

Nhìn chung, *lead\_time* cho tín hiệu phân tách tốt hơn so với *adr* trong EDA, do đó được kỳ vọng là đặc trưng quan trọng trong bài toán dự đoán hủy đặt phòng.

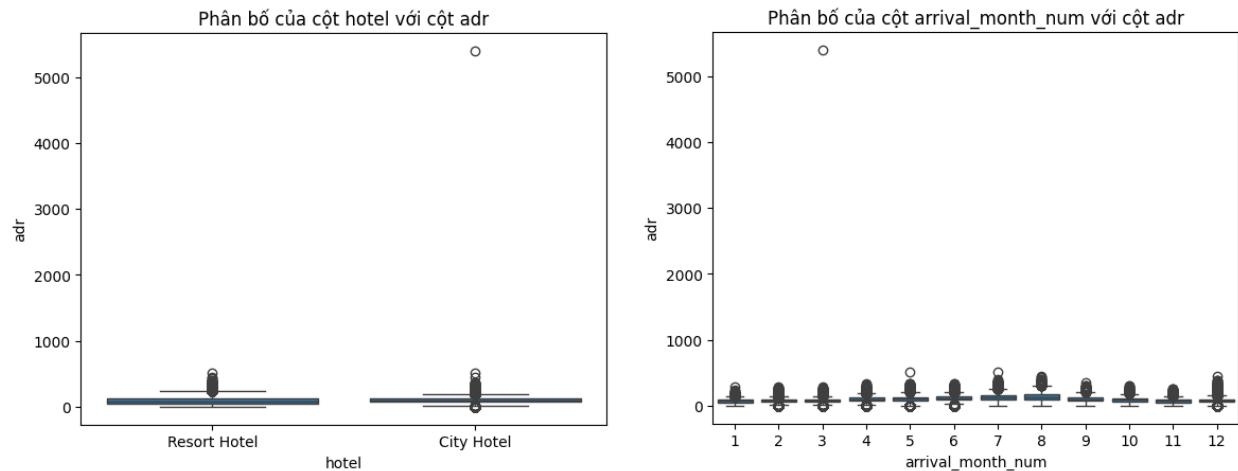
### 3.3.4. Phân tích tương quan giữa các thuộc tính



Ma trận tương quan cho thấy mức độ liên hệ tuyến tính giữa một số biến số chính và biến mục tiêu *is\_canceled*. Đáng chú ý, *lead\_time* có **tương quan dương mức vừa với *is\_canceled*** ( $r \approx 0.29$ ), gợi ý rằng thời gian đặt trước càng dài thì khả năng hủy đặt phòng có xu hướng tăng. Ngược lại, *adr* có **tương quan rất thấp với *is\_canceled*** ( $r \approx 0.05$ ) và *total\_stay* gần như không **tương quan** ( $r \approx 0.02$ ), cho thấy mối liên hệ tuyến tính trực tiếp giữa các biến này với trạng thái hủy là không rõ rệt trong phạm vi các biến được xét.

Ngoài ra, heatmap cũng cho thấy *adr* có **tương quan dương với *total\_guests*** ( $r \approx 0.36$ ), phản ánh xu hướng giá phòng trung bình mỗi ngày tăng theo quy mô nhóm khách (hoặc các booking có nhiều khách thường gắn với mức giá cao hơn). Các kết quả này cung cấp cơ sở ban đầu cho việc lựa chọn đặc trưng và thiết kế mô hình, tuy nhiên cần lưu ý rằng tương quan Pearson chỉ phản ánh quan hệ tuyến tính; do đó các mối quan hệ phi tuyến và tương tác giữa biến

sẽ được kiểm chứng thêm thông qua các mô hình học máy (Decision Tree/Random Forest) ở các phần tiếp theo.



Hình 4.11: Minh họa boxplot tương quan của các đặc trưng liên quan

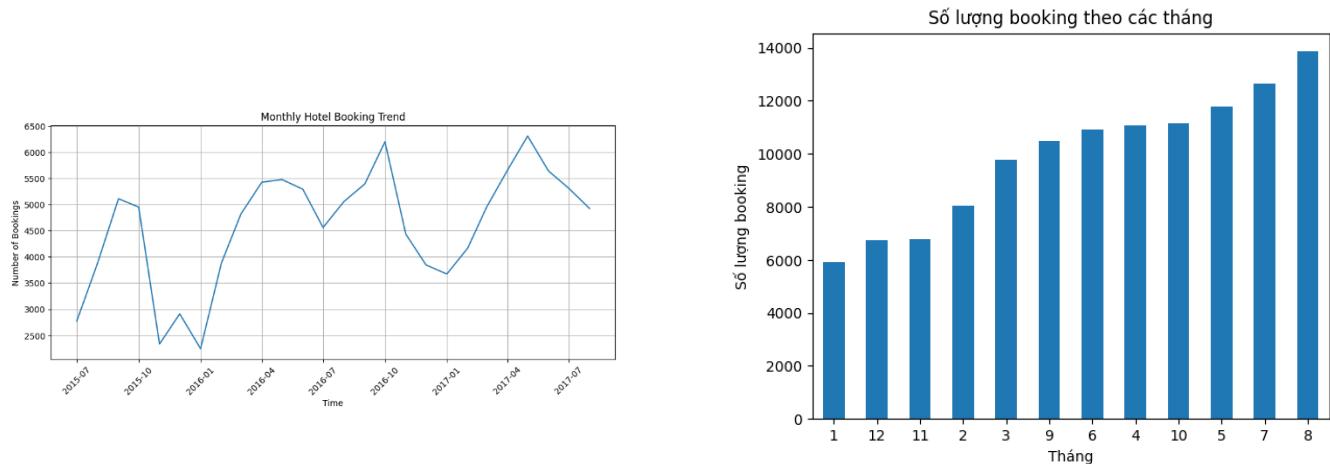
Boxplot theo **loại khách sạn** cho thấy phân bố *adr* khác nhau giữa **Resort Hotel** và **City Hotel**, thể hiện rằng phân khúc khách sạn có ảnh hưởng đến mức giá phòng trung bình mỗi ngày. Bên cạnh sự khác biệt về mức giá trung tâm, cả hai nhóm đều xuất hiện các điểm ngoại lai, cho thấy tồn tại một số trường hợp có mức giá đặc biệt cao so với phần lớn quan sát.

Khi xét theo **thời điểm nhận phòng** (*arrival\_month\_num*), phân bố *adr* thay đổi theo từng tháng, gợi ý **tác động của yếu tố mùa vụ** đến mức giá. Một số tháng có mức giá trung tâm và độ phân tán cao hơn, phản ánh nhu cầu cao điểm hoặc đặc thù thị trường theo thời gian. Kết quả này cung cấp cơ sở để đưa các biến thời gian (tháng/mùa) vào mô hình nhằm giải thích biến động giá và hỗ trợ phân tích xu hướng ở các phần tiếp theo.

Ngoài ra, sự xuất hiện của các giá trị ngoại lai rất lớn trong cả hai biểu đồ cho thấy cần cân nhắc xử lý ngoại lai (ví dụ: capping) trước khi mô hình hóa, nhằm đảm bảo tính ổn định và giảm ảnh hưởng của các trường hợp cực đoan.

Nhìn chung, *hotel* và *arrival\_month\_num* đều cho thấy mối liên hệ với biến động *adr*, đồng thời dữ liệu tồn tại ngoại lai lớn cần được xử lý trước khi huấn luyện mô hình.

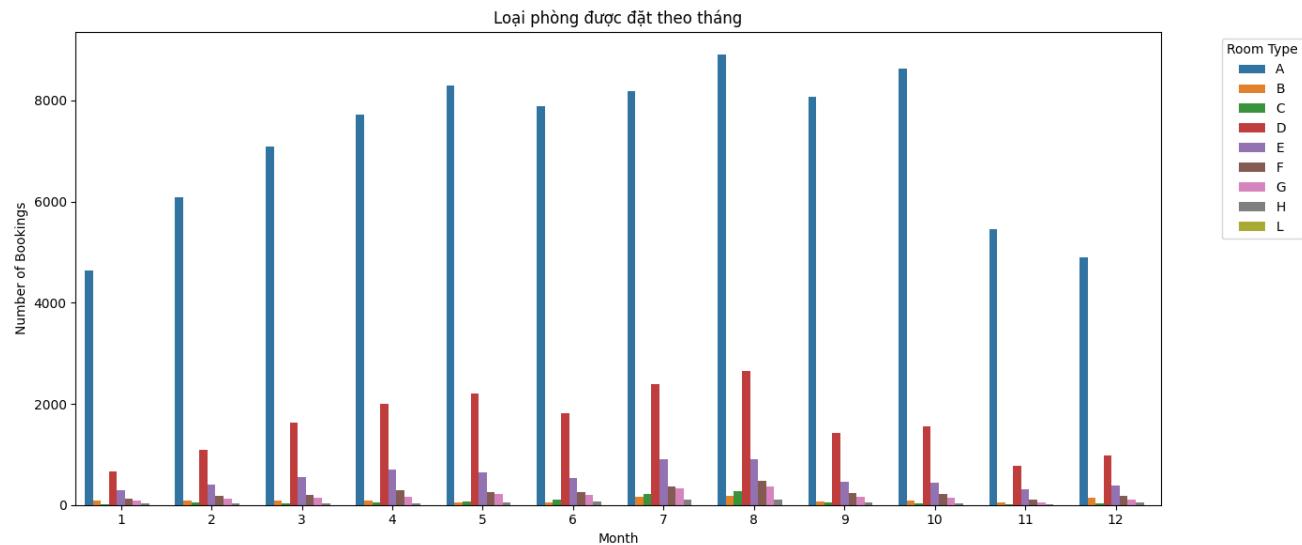
### 3.3.5. Phân tích theo thời gian



Hình 4.12: Minh họa số lượng đặt phòng qua thời gian

Biểu đồ chuỗi thời gian theo tháng cho thấy số lượng booking **biến động theo thời gian** và xuất hiện các giai đoạn tăng–giảm rõ rệt, phản ánh tính không ổn định của nhu cầu đặt phòng theo từng thời kỳ. Bên cạnh xu hướng chung, chuỗi dữ liệu còn cho thấy dấu hiệu **dao động mang tính mùa vụ**, khi số lượng booking thay đổi theo các tháng trong năm.

Kết quả từ biểu đồ cột theo tháng cũng có nhận định về **tính mùa vụ**: lượng booking không phân bố đồng đều giữa các tháng, trong đó một số tháng có số lượng booking cao hơn đáng kể so với các tháng còn lại. Điều này gợi ý rằng yếu tố thời gian (tháng/mùa) có vai trò quan trọng trong việc giải thích biến động nhu cầu, đồng thời là cơ sở để xây dựng các đặc trưng thời gian (*arrival\_month\_num/season*) và triển khai các mô hình dự báo chuỗi thời gian ở các phần tiếp theo.

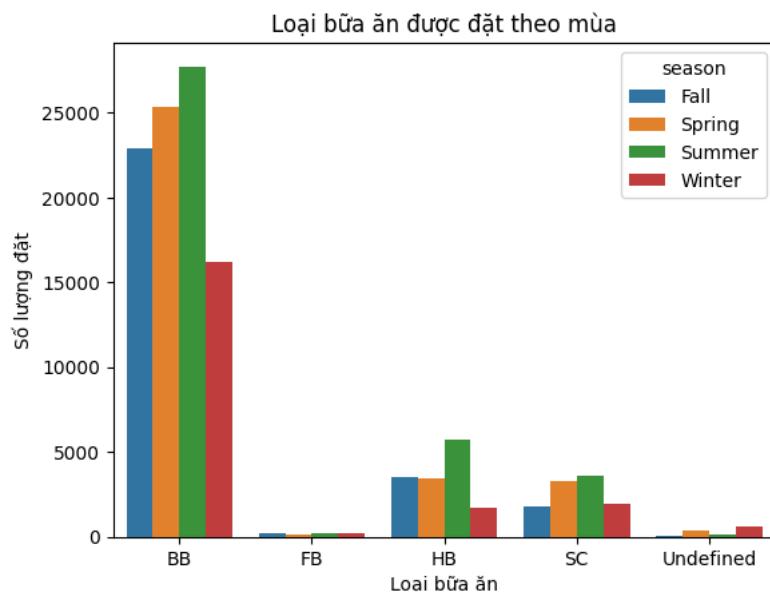


Hình 4.13: Minh họa Loại phòng được đặt theo tháng

Biểu đồ cho thấy số lượng booking giữa các loại phòng có sự chênh lệch rõ rệt theo thời gian. Nhìn chung, **một số loại phòng chiếm ưu thế về tần suất đặt phòng** trong hầu hết các tháng (đặc biệt là room type A), trong khi các loại phòng còn lại có số lượng booking thấp hơn đáng kể. Điều này phản ánh cấu trúc nhu cầu tập trung vào một vài loại phòng phổ biến.

Khi xét theo từng tháng, số lượng booking của các loại phòng cũng **biến động theo mùa vụ**, với xu hướng tăng trong một số giai đoạn cao điểm và giảm vào các tháng thấp điểm. Bên cạnh loại phòng chủ đạo, một số room type khác (ví dụ nhóm có mức đặt phòng trung bình như D/E) cũng thể hiện sự thay đổi theo tháng, gợi ý sự dịch chuyển trong cơ cấu lựa chọn phòng theo thời gian.

Kết quả này cung cấp cơ sở để xác định các loại phòng “chủ lực” theo từng giai đoạn, đồng thời hỗ trợ lập kế hoạch phân bổ công suất và chiến lược giá theo mùa trong các phân tích tiếp theo.



Hình 4.14: Minh họa loại bữa ăn được đặt theo mùa

Biểu đồ cho thấy lựa chọn **loại bữa ăn (meal)** có sự khác biệt theo **mùa (season)**. Nhìn chung, gói **BB (Bed & Breakfast)** chiếm ưu thế rõ rệt ở tất cả các mùa, phản ánh đây là lựa chọn phổ biến nhất trong dữ liệu. Các lựa chọn **HB** và **SC** xuất hiện với tần suất thấp hơn nhưng vẫn thể hiện sự biến động theo mùa, trong đó mức đặt có xu hướng tăng ở một số mùa cao điểm (đặc biệt là mùa hè).

Ngược lại, các nhóm **FB** và **Undefined** chiếm tỷ trọng rất nhỏ, cho thấy nhu cầu đối với các gói này không phổ biến hoặc dữ liệu ghi nhận không đầy đủ ở một số booking. Kết quả này gợi ý rằng yếu tố mùa vụ không chỉ tác động đến số lượng booking mà còn ảnh hưởng đến cơ cấu dịch vụ đi kèm (meal plan), qua đó có thể hỗ trợ hoạch định chính sách dịch vụ và marketing theo mùa.

BB là lựa chọn chủ đạo quanh năm, trong khi HB và SC biến động theo mùa, phản ánh sự thay đổi nhu cầu dịch vụ theo giai đoạn.

### 3.3.5. Nhận xét và ý nghĩa rút ra từ Khám phá dữ liệu

Kết quả phân tích khám phá dữ liệu (EDA) cho thấy bộ dữ liệu *Hotel Booking Demand* có tính đa dạng cao về hành vi đặt phòng và phù hợp để triển khai đồng thời các bài toán phân lớp, phân cụm, luật kết hợp và chuỗi thời gian.

*Thứ nhất*, phân bố của các biến số quan trọng như **lead\_time**, **adr** và **total\_stay** đều có xu hướng **lịch phải và xuất hiện ngoại lai**, phản ánh sự khác biệt đáng kể giữa các nhóm khách hàng và tồn tại các trường hợp đặc biệt (đặt sớm, giá rất cao hoặc lưu trú dài). Điều này nhấn mạnh vai trò của các bước tiền xử lý như xử lý ngoại lai và chuẩn hóa thang đo nhằm đảm bảo độ ổn định cho quá trình mô hình hóa.

*Thứ hai*, khi phân tích theo biến mục tiêu **is\_canceled**, EDA cho thấy một số đặc trưng có khả năng phân biệt hai nhóm hủy và không hủy, đặc biệt là **lead\_time** (nhóm bị hủy có xu hướng đặt trước dài hơn). Bên cạnh đó, các biến phân loại như **deposit\_type**, **market\_segment**, **hotel** và **customer\_type** thể hiện sự khác biệt về phân bố theo trạng thái hủy, gợi ý rằng mức độ cam kết đặt phòng và kênh/nhóm khách hàng có liên hệ đến rủi ro hủy. Đây là cơ sở để lựa chọn tập đặc trưng và triển khai các mô hình phân lớp ở chương thực nghiệm.

*Thứ ba*, phân tích theo thời gian cho thấy số lượng booking biến động theo tháng và có dấu hiệu **mùa vụ**, đồng thời cơ cấu lựa chọn **loại phòng (room type)** và **gói bữa ăn (meal)** cũng thay đổi theo tháng/mùa. Các kết quả này hỗ trợ trả lời các câu hỏi nghiên cứu về xu hướng đặt phòng theo giai đoạn và là nền tảng cho bài toán **chuỗi thời gian** cũng như các phân tích liên quan đến nhu cầu dịch vụ.

*Cuối cùng*, các kết quả EDA cung cấp định hướng rõ ràng cho giai đoạn mô hình hóa: (i) ưu tiên các biến có tín hiệu liên quan đến hủy đặt phòng như *lead\_time* và các biến phản ánh mức độ cam kết (*deposit\_type*), (ii) chuẩn bị dữ liệu phù hợp cho phân cụm và luật kết hợp thông qua chuẩn hóa/rồi rạc hóa, và (iii) xây dựng chuỗi dữ liệu theo tháng để phục vụ dự báo nhu cầu. Nhìn chung, các insight thu được từ EDA đóng vai trò nền tảng để thiết kế thí nghiệm, lựa chọn mô hình và diễn giải kết quả trong Chương 4.

## 3.4. Lựa chọn mô hình

Dựa trên mục tiêu nghiên cứu và các insight thu được từ quá trình phân tích khám phá dữ liệu (EDA), nghiên cứu này đề xuất triển khai nhiều nhóm kỹ thuật khai phá dữ liệu nhằm khai thác tri thức tiềm ẩn trong bộ dữ liệu *Hotel Booking Demand*. Cụ thể, các thuật toán được lựa chọn bao gồm: (i) **phân lớp** để dự đoán khả năng hủy đặt phòng (*is\_canceled*), (ii) **phân cụm** nhằm nhận diện các phân khúc khách hàng dựa trên hành vi đặt phòng, (iii) **khai phá luật kết hợp** để phát hiện các mẫu hành vi đặt phòng phổ biến, và (iv) **phân tích chuỗi thời gian** để đánh giá xu hướng và tính mùa vụ của nhu cầu đặt phòng theo thời gian.

Việc lựa chọn mô hình được thực hiện dựa trên mức độ phù hợp của thuật toán với đặc điểm dữ liệu (bao gồm phân bố lệch, ngoại lai và sự đa dạng của biến phân loại), khả năng diễn giải kết quả và hiệu quả thực nghiệm. Đồng thời, các mô hình sẽ được đánh giá theo các tiêu chí/độ đo phù hợp với từng bài toán nhằm đảm bảo tính khách quan và khả năng so sánh.

### 3.4.1. Mô hình phân lớp (Classification)

**Mục tiêu:** Bài toán phân lớp được xây dựng nhằm dự đoán khả năng hủy đặt phòng của khách hàng thông qua biến mục tiêu *is\_canceled*.

Các mô hình được lựa chọn. Nghiên cứu triển khai và so sánh ba mô hình học máy phổ biến:

- *Logistic Regression (LR)*: được sử dụng như mô hình cơ sở (baseline) nhờ cấu trúc đơn giản, khả năng diễn giải và giúp thiết lập mức hiệu năng tham chiếu cho các mô hình phức tạp hơn.
- *Decision Tree (DT)*: có khả năng mô hình hóa các quan hệ phi tuyến và tương tác giữa các biến, đồng thời dễ diễn giải thông qua cấu trúc phân tách của cây.
- *Random Forest (RF)*: mô hình tập hợp (ensemble) từ nhiều cây quyết định, giúp cải thiện khả năng tổng quát hóa, giảm phương sai và hạn chế hiện tượng quá khớp (overfitting) so với một cây đơn lẻ.

Sau khi hoàn tất tiền xử lý, tập dữ liệu được chia thành **tập huấn luyện và tập kiểm tra theo tỷ lệ 80/20**. Việc chia dữ liệu được thực hiện theo phương pháp **stratified split** dựa trên biến mục tiêu *is\_canceled* nhằm **giữ nguyên (xấp xỉ) tỷ lệ các lớp** ở cả hai tập. Cách tiếp cận này giúp giảm sai lệch do mất cân bằng lớp và đảm bảo tính khách quan khi đánh giá hiệu năng mô hình trên dữ liệu chưa quan sát.

**Chỉ số đánh giá:** Hiệu quả mô hình được đánh giá thông qua các chỉ số gồm Accuracy, Precision, Recall, F1-score và Confusion Matrix. Trong đó, Precision/Recall/F1-score được ưu tiên để phản ánh hiệu năng dự đoán trong trường hợp phân bố lớp không cân bằng, còn Confusion Matrix giúp phân tích chi tiết các dạng sai số (FP/FN) và hỗ trợ lựa chọn mô hình phù hợp với mục tiêu ứng dụng.

Việc sử dụng nhiều mô hình và nhiều thước đo đánh giá cho phép so sánh hiệu quả một cách toàn diện, từ đó lựa chọn mô hình tối ưu cho bài toán dự đoán hủy đặt phòng.

Ngoài ra, chỉ số *ROC-AUC* có thể được sử dụng để đánh giá khả năng phân biệt hai lớp của mô hình một cách tổng quát.

### 3.4.2. Mô hình phân cụm (Clustering)

**Mục tiêu:** Bài toán phân cụm được thực hiện nhằm phân nhóm khách hàng dựa trên hành vi đặt phòng, từ đó phát hiện các nhóm khách hàng có đặc điểm tương đồng và hỗ trợ phân tích phân khúc.

**Các mô hình được lựa chọn.** Nghiên cứu sử dụng hai phương pháp phân cụm:

- **K-Means:** phân cụm dựa trên khoảng cách, phù hợp với dữ liệu số sau khi được chuẩn hóa. Số cụm tối ưu được xác định thông qua phương pháp **Elbow (Inertia)** kết hợp với **Silhouette Score** để cân bằng giữa độ chật chẽ trong cụm và mức độ tách biệt giữa các cụm.

**Thuộc tính sử dụng:** Các biến số phản ánh hành vi đặt phòng được lựa chọn gồm: *lead\_time, adr, total\_stay (tổng số đêm lưu trú), total\_guests*. Nhóm thuộc tính này đại diện cho thời điểm đặt phòng, mức chi tiêu và đặc điểm lưu trú, phù hợp cho mục tiêu phân khúc khách hàng.

**Chỉ số đánh giá:** Hiệu quả phân cụm được đánh giá bằng **Silhouette Score** (mức độ gắn kết trong cụm và tách biệt giữa các cụm) và **Inertia** đối với **K-Means**. Sau khi xác định số cụm, nghiên cứu thực hiện **mô tả đặc trưng cụm (cluster profiling)** dựa trên thống kê trung tâm (mean/median) của các biến để diễn giải ý nghĩa từng phân khúc khách hàng.

Kết quả phân cụm cung cấp góc nhìn tổng quan về hành vi khách hàng, làm cơ sở hỗ trợ đề xuất định hướng kinh doanh và chiến lược marketing theo từng phân khúc.

### 3.4.3. Khai phá luật kết hợp (Association Rules Mining)

**Mục tiêu.** Khai phá luật kết hợp được thực hiện nhằm khám phá các mẫu hành vi đặt phòng phổ biến và các mối quan hệ đồng xuất hiện giữa các thuộc tính trong bộ dữ liệu *Hotel Booking Demand*, từ đó hỗ trợ diễn giải hành vi khách hàng và ra quyết định quản lý.

**Thuật toán được lựa chọn.** Nghiên cứu sử dụng thuật toán **Apriori** để tìm **tập mục phổ biến (frequent itemsets)** và sinh **luật kết hợp (association rules)**. Trước khi áp dụng Apriori, các thuộc tính được **rời rạc hóa và chuyển đổi về dạng giao dịch (transaction/binary)** nhằm phù hợp với yêu cầu đầu vào của thuật toán.

**Tiêu chí đánh giá và sàng lọc luật.** Các luật được lọc dựa trên ba thước đo:

- *Support:* mức độ phổ biến của tập mục trong toàn bộ dữ liệu;
- *Confidence:* xác suất về phải xảy ra khi về trái xảy ra;
- *Lift:* mức độ phụ thuộc giữa hai về của luật (*lift > 1* cho thấy mối liên hệ dương đáng chú ý).

Các ngưỡng *support/confidence* được thiết lập sau quá trình thử nghiệm nhằm cân bằng giữa số lượng luật và mức độ ý nghĩa.”

#### Ví dụ luật kết hợp.

$$\{deposit\_type = Non\ Refund\} \Rightarrow \{is\_canceled = 0\}$$

Luật này gợi ý rằng các booking thuộc nhóm **Non Refund** có xu hướng **không hủy**, phản ánh mức độ cam kết cao hơn của khách hàng đối với loại đặt phòng này.

Nhìn chung, các luật kết hợp giúp làm rõ các yếu tố gắn liền với hành vi đặt phòng và cung cấp ý nghĩa phục vụ tối ưu hóa chính sách đặt phòng, quản lý rủi ro hủy và thiết kế chiến lược vận hành/marketing.

#### 3.4.4. Phân tích chuỗi thời gian (Time Series Analysis)

**Mục tiêu:** Bài toán chuỗi thời gian được thực hiện nhằm phân tích và dự báo xu hướng đặt phòng theo thời gian, hỗ trợ lập kế hoạch vận hành và quản lý nguồn lực.

**Xây dựng chuỗi thời gian:** Chuỗi thời gian được tổng hợp theo **tháng nhận phòng**, với biến quan sát là **số lượng booking theo tháng**. Dữ liệu theo tháng cho phép nhận diện xu hướng và tính mùa vụ, đồng thời phù hợp cho mục tiêu dự báo trung hạn.

**Mô hình được đề xuất.**

- **SARIMAX:** được lựa chọn để mô hình hóa đồng thời **xu hướng và mùa vụ** của chuỗi booking, đồng thời tích hợp **các biến ngoại sinh** nhằm cải thiện năng lực dự báo. Trong nghiên cứu này, các biến ngoại sinh được sử dụng bao gồm: **[liệt kê biến ngoại sinh theo tháng, ví dụ: adr trung bình theo tháng / tỷ lệ hủy theo tháng / số lượng special requests theo tháng / ...]**. Việc bổ sung các biến này giúp mô hình nắm bắt tác động của các yếu tố bên ngoài chuỗi chính đến nhu cầu đặt phòng.
- **XGBoost:** được áp dụng như phương pháp học máy cho dự báo chuỗi thời gian bằng cách xây dựng các đặc trưng **độ trễ (lag features)** và đặc trưng lịch (tháng/mùa), qua đó mô hình hóa các quan hệ phi tuyến và tương tác phức tạp trong dữ liệu.

Các mô hình được sử dụng nhằm đánh giá khả năng dự báo nhu cầu đặt phòng theo tháng và so sánh hiệu quả giữa hướng tiếp cận thống kê (SARIMAX) và học máy (XGBoost).

Hiệu năng dự báo được đánh giá bằng các thước đo sai số như *MAE/RMSE/MAPE* trên tập kiểm tra theo thứ tự thời gian.

#### 3.4.5. Tổng kết lựa chọn mô hình

Việc kết hợp nhiều nhóm mô hình và kỹ thuật khai phá dữ liệu cho phép tiếp cận bộ dữ liệu dưới nhiều góc độ khác nhau, bao gồm dự đoán, phân khúc, khai phá mẫu và phân tích xu hướng theo thời gian. Cách tiếp cận này giúp **đối chiếu hiệu quả giữa các phương pháp**, đồng thời nâng cao tính tin cậy của kết luận thông qua việc kiểm chứng chéo các insight thu được từ các bài toán khác nhau.

Các mô hình được lựa chọn sẽ được triển khai và đánh giá chi tiết trong *Chương 4 – Thực nghiệm, Kết quả và Thảo luận*, nhằm làm rõ mức độ phù hợp của từng phương pháp đối với các câu hỏi nghiên cứu đã đặt ra.

### 3.5. Quy trình mô hình hóa đề xuất

Dựa trên mục tiêu nghiên cứu và các phương pháp đã được lựa chọn, nghiên cứu này đề xuất một **quy trình mô hình hóa khai phá dữ liệu tổng thể** nhằm đảm bảo tính khoa học, khả

năng tái lập và phù hợp với yêu cầu của học phần *Khai phá dữ liệu*. Quy trình bao phủ các bước chính gồm: **tiền xử lý và làm sạch dữ liệu, phân tích khám phá (EDA), xây dựng/biến đổi đặc trưng, lựa chọn mô hình theo từng bài toán (phân lớp, phân cụm, luật kết hợp, chuỗi thời gian), thiết kế thí nghiệm và đánh giá bằng các thước đo phù hợp**, và cuối cùng là **diễn giải kết quả** để rút ra tri thức phục vụ trả lời câu hỏi nghiên cứu.

Cách tiếp cận theo pipeline giúp đảm bảo các thí nghiệm được thực hiện nhất quán, hạn chế sai lệch trong đánh giá và hỗ trợ kiểm chứng lại kết quả khi cần thiết.

Trong toàn bộ quy trình, các bước xử lý và tham số được ghi nhận nhất quán nhằm đảm bảo khả năng tái lập thí nghiệm và tính minh bạch của kết luận.

### 3.5.1. Xác định bài toán khai phá dữ liệu

Từ bộ dữ liệu *Hotel Booking Demand*, các bài toán khai phá dữ liệu được xác định nhằm đáp ứng các mục tiêu nghiên cứu và khai thác tri thức theo nhiều góc độ khác nhau, bao gồm:

- *Phân lớp*: dự đoán khả năng hủy đặt phòng của khách hàng thông qua biến mục tiêu *is\_canceled*.
- *Phân cụm*: phân nhóm khách hàng dựa trên các đặc trưng hành vi đặt phòng để nhận diện các phân khúc có đặc điểm tương đồng.
- *Khai phá luật kết hợp*: phát hiện các mẫu hành vi đặt phòng phổ biến và các mối quan hệ đồng xuất hiện giữa các thuộc tính.
- *Phân tích chuỗi thời gian*: phân tích xu hướng, tính mùa vụ và dự báo nhu cầu đặt phòng theo thời gian.

Việc xác định rõ các bài toán ngay từ đầu giúp định hướng lựa chọn tập thuộc tính phù hợp, lựa chọn thuật toán tương ứng và xây dựng hệ tiêu chí đánh giá nhất quán, qua đó đảm bảo tính khoa học và khả năng diễn giải của kết quả thực nghiệm.

### 3.5.2. Chuẩn bị dữ liệu cho mô hình hóa

Sau khi hoàn tất các bước tiền xử lý (mục 3.2), tập dữ liệu được sử dụng cho giai đoạn mô hình hóa theo quy trình sau:

- *Lựa chọn thuộc tính theo từng bài toán*: xác định tập đặc trưng phù hợp cho phân lớp, phân cụm, luật kết hợp và chuỗi thời gian nhằm phản ánh đúng mục tiêu phân tích.
- *Chuẩn hóa dữ liệu khi cần thiết*: áp dụng chuẩn hóa đối với các mô hình nhạy cảm với thang đo và khoảng cách (ví dụ: K-Means, các mô hình dựa trên khoảng cách), đảm bảo các biến số đóng góp công bằng vào quá trình huấn luyện.
- *Chia dữ liệu để đánh giá khách quan*: dữ liệu được tách thành tập huấn luyện và tập kiểm tra để đánh giá hiệu quả mô hình trên dữ liệu chưa quan sát, hạn chế hiện tượng quá khớp và đảm bảo tính khách quan của kết quả thực nghiệm.

Quy trình này giúp đảm bảo dữ liệu đầu vào được chuẩn bị nhất quán và phù hợp cho các thí nghiệm mô hình hóa ở các phần tiếp theo.

### 3.5.3. Huấn luyện mô hình

Các mô hình được triển khai và huấn luyện **độc lập theo từng bài toán** nhằm đảm bảo tính phù hợp của thuật toán với mục tiêu phân tích:

- *Phân lớp*: huấn luyện các mô hình *Logistic Regression, Decision Tree, Random Forest* trên tập huấn luyện để dự đoán biến mục tiêu *is\_canceled*, sau đó đánh giá trên tập kiểm tra bằng các thước đo phù hợp.
- *Phân cụm*: áp dụng *K-Means* trong đó thử nghiệm với nhiều cấu hình (đặc biệt là **số cụm k** và/hoặc liên kết phân cấp) để xác định phương án phân cụm tối ưu dựa trên các chỉ số chất lượng cụm.
- *Luật kết hợp*: sử dụng thuật toán **Apriori** để tìm tập mục phổ biến và sinh các luật kết hợp, sau đó sàng lọc luật theo các ngưỡng *support, confidence* và *lift* nhằm đảm bảo tính ý nghĩa và khả năng diễn giải.
- *Chuỗi thời gian*: xây dựng mô hình dự báo dựa trên chuỗi **số lượng booking theo tháng**, từ đó phân tích xu hướng/mùa vụ và đánh giá khả năng dự báo nhu cầu trong tương lai.

Việc triển khai nhiều mô hình theo từng bài toán cho phép **so sánh hiệu quả giữa các phương pháp**, đồng thời tăng độ tin cậy của kết luận thông qua đối chiếu kết quả từ các góc nhìn khai phá dữ liệu khác nhau.

Chi tiết cấu hình tham số và kết quả đánh giá của từng mô hình sẽ được trình bày trong Chương 4.

### 3.5.4. Đánh giá và so sánh mô hình

Hiệu quả của các mô hình được đánh giá bằng các thước đo phù hợp với đặc thù của từng bài toán:

- *Phân lớp*: *Accuracy, Precision, Recall, F1-score* và **Confusion Matrix** nhằm đánh giá toàn diện khả năng dự đoán, đồng thời phân tích chi tiết các dạng sai sót (FP/FN).
- *Phân cụm*: **Silhouette Score** (đánh giá mức độ gắn kết và tách biệt giữa các cụm) và **Inertia** (đối với K-Means) để hỗ trợ lựa chọn số cụm tối ưu.
- *Luật kết hợp*: *Support, Confidence* và *Lift* nhằm sàng lọc các luật có tính phổ biến và mức độ liên hệ đáng chú ý giữa các thuộc tính.
- *Chuỗi thời gian*: sai số dự báo như **MAE** và **RMSE** trên tập kiểm tra theo thứ tự thời gian để đánh giá mức độ chính xác của mô hình dự báo.

Các kết quả đánh giá được sử dụng để so sánh hiệu quả giữa các mô hình và lựa chọn phương án phù hợp nhất cho từng bài toán, đồng thời làm cơ sở thảo luận cho các câu hỏi nghiên cứu trong chương kết quả.

Việc sử dụng nhiều thước đo giúp hạn chế thiên lệch khi dữ liệu có phân bố lớp không cân bằng hoặc khi mục tiêu ưu tiên giảm FN/FP.

### 3.5.5. Trực quan hóa và diễn giải kết quả

Nhằm nâng cao khả năng diễn giải và trình bày kết quả một cách trực quan, nghiên cứu sử dụng các phương pháp trực quan hóa phù hợp với từng nhóm bài toán. Việc trực quan hóa không chỉ hỗ trợ so sánh hiệu năng giữa các mô hình mà còn giúp rút ra các insight có ý nghĩa thực tiễn từ dữ liệu.

*Đối với phân lớp*, kết quả được trình bày thông qua **Confusion Matrix** để quan sát chi tiết các dạng sai sót (FP/FN), đồng thời so sánh các chỉ số **Precision**, **Recall**, **F1-score** giữa các mô hình. Ngoài ra, mức độ ảnh hưởng của các biến đầu vào được diễn giải thông qua **feature importance** (đối với Random Forest) và/hoặc các hệ số của mô hình (đối với Logistic Regression), từ đó làm rõ các yếu tố liên quan đến rủi ro hủy đặt phòng.

*Đối với phân cụm*, kết quả phân cụm được trực quan hóa bằng biểu đồ thể hiện chất lượng cụm như **Elbow (Inertia)** và **Silhouette Score** để lựa chọn số cụm phù hợp. Sau khi xác định cụm, nghiên cứu thực hiện **mô tả đặc trưng cụm (cluster profiling)** thông qua thống kê trung tâm (mean/median) và biểu đồ so sánh, qua đó diễn giải ý nghĩa của từng nhóm khách hàng.

*Đối với luật kết hợp*, các luật được trình bày theo thứ tự mức độ ý nghĩa dựa trên **support**, **confidence** và **lift**; đồng thời có thể trực quan hóa bằng biểu đồ (ví dụ: scatter giữa support-confidence, hoặc sắp xếp theo lift) nhằm làm nổi bật các mẫu hành vi đặt phòng phổ biến và các mối quan hệ đồng xuất hiện đáng chú ý.

*Đối với chuỗi thời gian*, kết quả được trực quan hóa bằng biểu đồ **xu hướng số lượng booking theo tháng**, đồng thời so sánh **giá trị thực tế và giá trị dự báo** trên tập kiểm tra. Sai số dự báo (MAE, RMSE) được báo cáo để đánh giá định lượng hiệu năng mô hình và hỗ trợ lựa chọn phương án dự báo phù hợp.

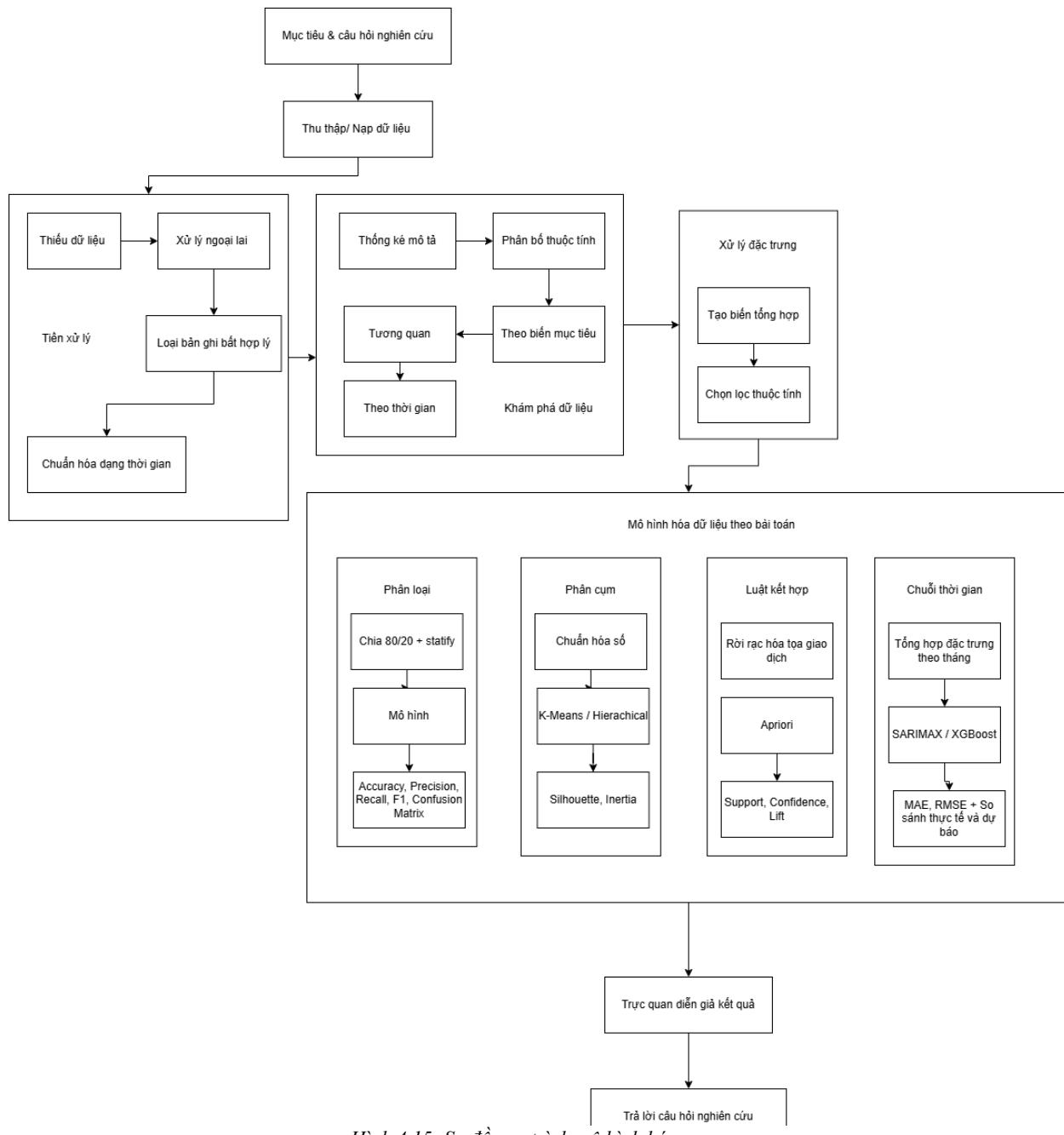
Các kết quả trực quan hóa và diễn giải sẽ được trình bày chi tiết trong Chương 4, nhằm trả lời các câu hỏi nghiên cứu và rút ra các kết luận có ý nghĩa ứng dụng.

### 3.5.6. Tổng kết quy trình mô hình hóa

Quy trình mô hình hóa được đề xuất đảm bảo **tính hệ thống và logic** trong toàn bộ tiến trình khai phá dữ liệu, từ chuẩn bị dữ liệu đến xây dựng mô hình, đánh giá và diễn giải kết quả. Bên cạnh đó, quy trình cho phép **áp dụng và đổi chiều nhiều kỹ thuật khác nhau trên cùng một bộ dữ liệu**, qua đó hỗ trợ kiểm chứng chéo các phát hiện và nâng cao độ tin cậy của kết luận. Các kết quả thu được được tổng hợp và diễn giải nhằm cung cấp **insight có ý nghĩa** phục vụ trả lời các câu hỏi nghiên cứu và hỗ trợ ra quyết định.

Quy trình này là cơ sở để triển khai thực nghiệm và thảo luận kết quả trong **Chương 4 – Thực nghiệm, Kết quả và Thảo luận**.

## SƠ ĐỒ QUY TRÌNH MÔ HÌNH HÓA



Hình 4.15: Sơ đồ quy trình mô hình hóa

## CHƯƠNG IV: THỰC NGHIỆM

### 4.1. Thiết lập thực nghiệm

#### 4.1.1. Môi trường và công cụ thực nghiệm

Các thí nghiệm trong nghiên cứu này được thực hiện bằng ngôn ngữ lập trình Python, sử dụng các thư viện phổ biến trong khai phá dữ liệu và học máy như Pandas, NumPy, Scikit-learn, Matplotlib và Seaborn.

Toàn bộ quá trình xử lý dữ liệu, xây dựng mô hình và đánh giá kết quả được triển khai trên môi trường Jupyter Notebook nhằm đảm bảo tính minh bạch và khả năng tái lập của nghiên cứu.

#### 4.1.2. Dữ liệu và phương pháp chia tập

Tập dữ liệu cuối cùng sau khi loại bỏ nhiễu và các giá trị khuyết thiêу đạt quy mô **87.229 bản ghi trên không gian 40 chiều thuộc tính**. Việc triển khai các mô hình khai phá dữ liệu được thực hiện như sau:

*Phân loại nhị phân:* Thực hiện phân chia tập dữ liệu theo tỷ lệ 8:2. Nhằm giảm thiểu sai số do sự lệch pha trong phân bổ lớp mục tiêu, phương pháp hiệu chỉnh trọng số đơn vị (class\_weight) được áp dụng, giúp mô hình tăng cường độ nhạy (sensitivity) đối với các trường hợp hủy phòng.

*Khai phá tri thức không giám sát:* Đối với phân cụm và tìm kiếm luật kết hợp, chúng tôi duy trì tính toàn vẹn của tập dữ liệu gốc để đảm bảo các mẫu hình (patterns) được trích xuất mang tính đại diện cao nhất cho tổng thể.

*Dự báo chuỗi thời gian:* Các quan sát được làm mịn và tổng hợp theo đơn vị tháng. Để đảm bảo tính khách quan và tránh hiện tượng "nhìn trước tương lai" (look-ahead bias), chúng tôi thiết lập điểm cắt thời gian (cutoff): dữ liệu lịch sử (2015–2016) làm cơ sở huấn luyện và dữ liệu kế tiếp (2017) dùng để kiểm chứng hiệu năng dự báo.

### 4.2. Kết quả mô hình phân loại

#### 4.2.1. So sánh hiệu quả các mô hình

Ba mô hình phân loại gồm *Logistic Regression*, *Decision Tree* và *Random Forest* được huấn luyện và đánh giá trên cùng tập dữ liệu.

#### KẾT QUẢ SO SÁNH HIỆU QUẢ CÁC MÔ HÌNH

Mô hình	Accuracy	Precision	Recall	F1-Score	AUC - ROC
Logistic Regression	0.77	0.77	0.79	0.78	0.86

Random Forest	0.77	0.83	0.77	0.75	0.88
Decision Tree	0.67	0.8	0.67	0.68	0.8

Bảng 2: So sánh hiệu quả các mô hình

Dựa trên dữ liệu thực nghiệm, chúng ta có thể rút ra các kết luận quan trọng sau:

*Về khả năng phân loại tổng quát:* Cả Logistic Regression và Random Forest đều đạt độ chính xác (Accuracy) tương đương nhau là **0.77**, vượt trội so với Decision Tree (**0.67**). Điều này cho thấy các mô hình phức tạp hơn đã học được những quy luật tiềm ẩn hiệu quả hơn mô hình cây quyết định đơn lẻ.

*Ưu thế của Random Forest về độ tin cậy (Precision):* Random Forest đạt chỉ số Precision cao nhất (**0.83**). Trong bài toán dự báo hủy phòng, điều này có nghĩa là khi mô hình dự báo một khách hàng sẽ hủy, khả năng dự báo đó chính xác là rất cao. Điều này giúp khách sạn tránh được việc đưa ra các biện pháp xử lý sai lầm đối với những khách hàng thực sự có ý định ở lại.

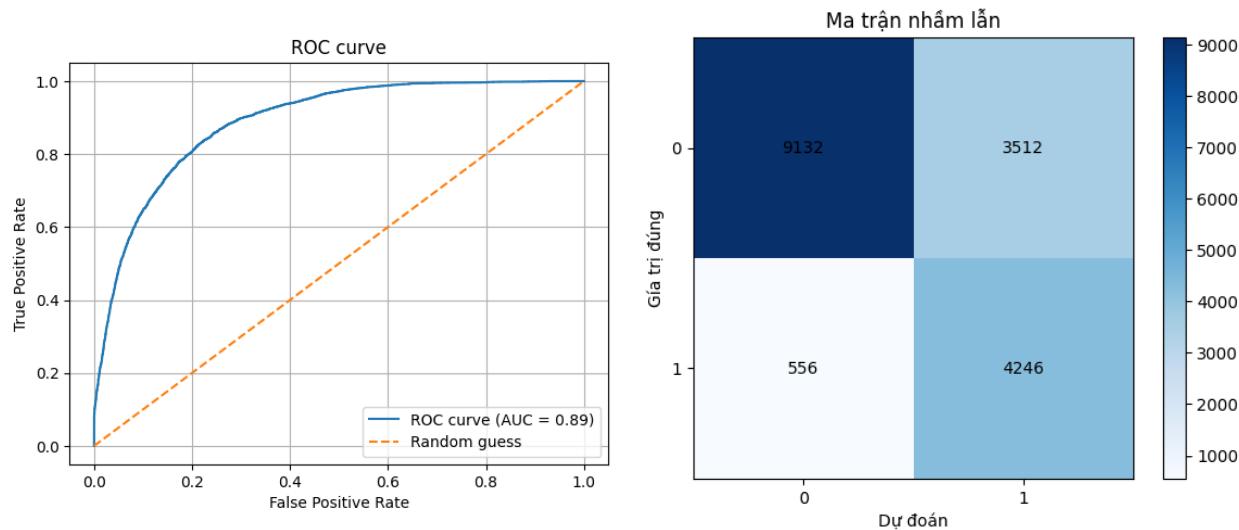
*Khả năng phân biệt và sự ổn định (AUC-ROC):* Random Forest dẫn đầu với AUC = 0.88. Chỉ số này khẳng định rằng về tổng thể, Random Forest có khả năng phân tách giữa hai lớp (**Hủy và Không hủy**) tốt hơn các mô hình còn lại, ít bị ảnh hưởng bởi việc thay đổi ngưỡng phân loại (threshold).

*Sự cân bằng của Logistic Regression:* Một điểm đáng chú ý là Logistic Regression lại có *F1-Score* và *Recall* nhỉnh hơn. Điều này cho thấy thuật toán này bao quát được nhiều trường hợp hủy phòng hơn (**Recall = 0.79**), dù độ chính xác trên từng dự báo đơn lẻ không cao bằng Random Forest.

*Đánh giá chung:* Mặc dù Logistic Regression thể hiện sự cân bằng tốt, nhưng Random Forest được đánh giá là mô hình tối ưu nhất cho bài toán này nhờ chỉ số **AUC (0.88)** và **Precision (0.83)** vượt trội. Với một bài toán thực tế như kinh doanh khách sạn, việc tối ưu hóa

độ chính xác của dự báo (*Precision*) thường được ưu tiên để tối ưu hóa chi phí vận hành và giữ chân khách hàng.

Đồ thị đường cong học tập (Learning Curve) mang lại thông tin quan trọng về hiện tượng

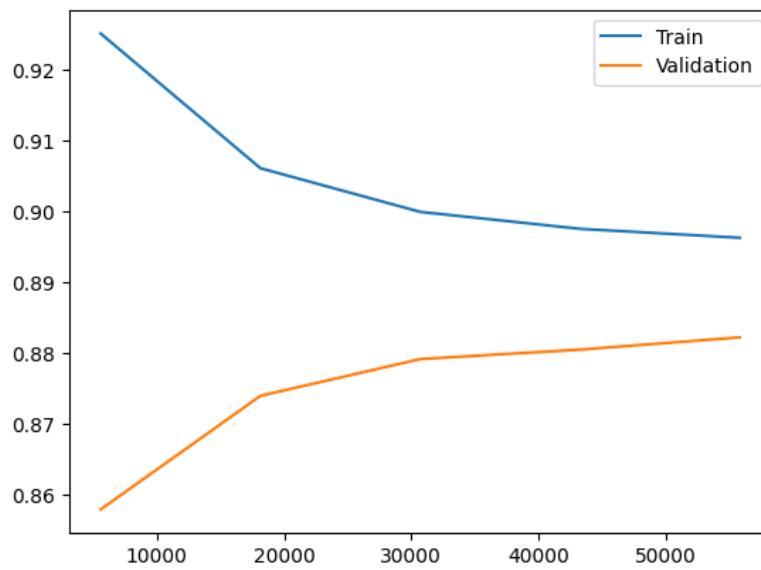


Hình 4.16: Biểu đồ đường ROC và Ma trận nhầm lẫn

Overfitting/Underfitting:

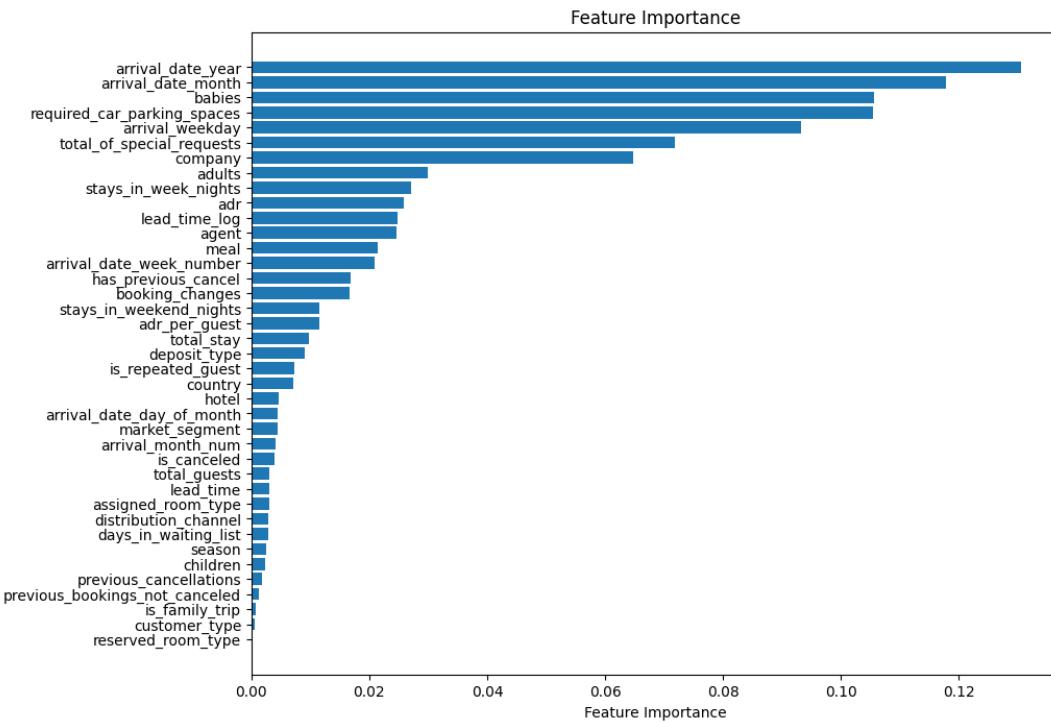
**Sự hội tụ (Convergence):** Khi kích thước mẫu tăng lên, khoảng cách giữa đường Train và Validation thu hẹp dần và hội tụ ở mức  $\sim 0.88$ . Điều này chứng minh mô hình không bị hiện tượng quá khớp (Overfitting) nghiêm trọng.

**Độ ổn định:** Đường Validation có xu hướng đi lên và ổn định dần, cho thấy mô hình đã học được các mẫu hình (patterns) cốt lõi từ dữ liệu và có khả năng tổng quát hóa tốt trên các tập dữ liệu chưa từng thấy.



Hình 4.17: Biểu đồ thể hiện mức độ hội tụ của tập train và tập đánh giá

#### 4.2.2. Phân tích tầm quan trọng của đặc trưng



Hình 4.18: Biểu đồ thể hiện các đặc trưng quan trọng

Phân tích mức độ quan trọng của thuộc tính (*Feature Importance*) đóng vai trò then chốt trong việc giải thích hành vi của mô hình học máy, góp phần nâng cao tính minh bạch và khả năng diễn giải của hệ thống dự báo. Kết quả cho thấy các thuộc tính liên quan đến yếu tố thời gian, đặc biệt là *arrival\_date\_year* và *arrival\_date\_month*, có mức đóng góp cao nhất vào quyết định phân loại hủy phòng. Điều này phản ánh rõ rệt ảnh hưởng của xu hướng theo thời gian và tính mùa vụ trong hành vi đặt phòng của khách hàng, vốn là đặc trưng phổ biến trong lĩnh vực du lịch – khách sạn.

Bên cạnh đó, thuộc tính *babies* thể hiện vai trò phân biệt đáng kể giữa các đơn đặt phòng, gợi ý sự tồn tại của các nhóm khách hàng có hành vi đặt phòng khác biệt. Kết quả này cho thấy tiềm năng áp dụng các phương pháp phân cụm nhằm xác định các phân khúc khách hàng đặc thù, từ đó hỗ trợ xây dựng các chính sách quản trị phù hợp cho từng nhóm đối tượng.

Các thuộc tính phản ánh mức độ cam kết của khách hàng, bao gồm *required\_car\_parking\_spaces* và *total\_of\_special\_requests*, cũng cho thấy mức độ quan trọng cao trong mô hình. Cụ thể, số lượng yêu cầu đặc biệt và nhu cầu về chỗ đỗ xe có xu hướng tỷ lệ nghịch với khả năng hủy phòng, cho thấy những khách hàng đầu tư nhiều công sức và thời gian vào quá trình đặt phòng thường có mức độ ổn định cao hơn. Phát hiện này mở ra hướng khai thác các phương pháp luật kết hợp nhằm phát hiện các mẫu hành vi đặt phòng có rủi ro hủy thấp, góp phần hỗ trợ ra quyết định trong quản trị vận hành khách sạn.

Về hiệu năng dự báo, mô hình Random Forest đạt được kết quả khả quan với chỉ số AUC bằng 0,8878, cho thấy khả năng phân biệt tốt giữa các đơn đặt phòng bị hủy và không bị hủy. Mặc dù mô hình chấp nhận một tỷ lệ dương tính giả nhất định nhằm tối ưu hóa độ bao phủ (Recall = 0,88), đây là sự đánh đổi hợp lý trong bối cảnh bài toán quản trị rủi ro hủy phòng, nơi chi phí của việc bỏ sót các trường hợp hủy phòng tiềm ẩn thường lớn hơn chi phí dự báo dư thừa.

Ngoài ra, đồ thị Learning Curve cho thấy quá trình huấn luyện của mô hình hội tụ ổn định, phản ánh sự cân bằng hợp lý giữa độ phức tạp mô hình và khả năng tổng quát hóa trên dữ liệu chưa quan sát. Kết quả này khẳng định tính ổn định của mô hình và khả năng ứng dụng trong các hệ thống hỗ trợ ra quyết định thực tế trong ngành dịch vụ lưu trú.

### **4.3. Kết quả mô hình phân cụm**

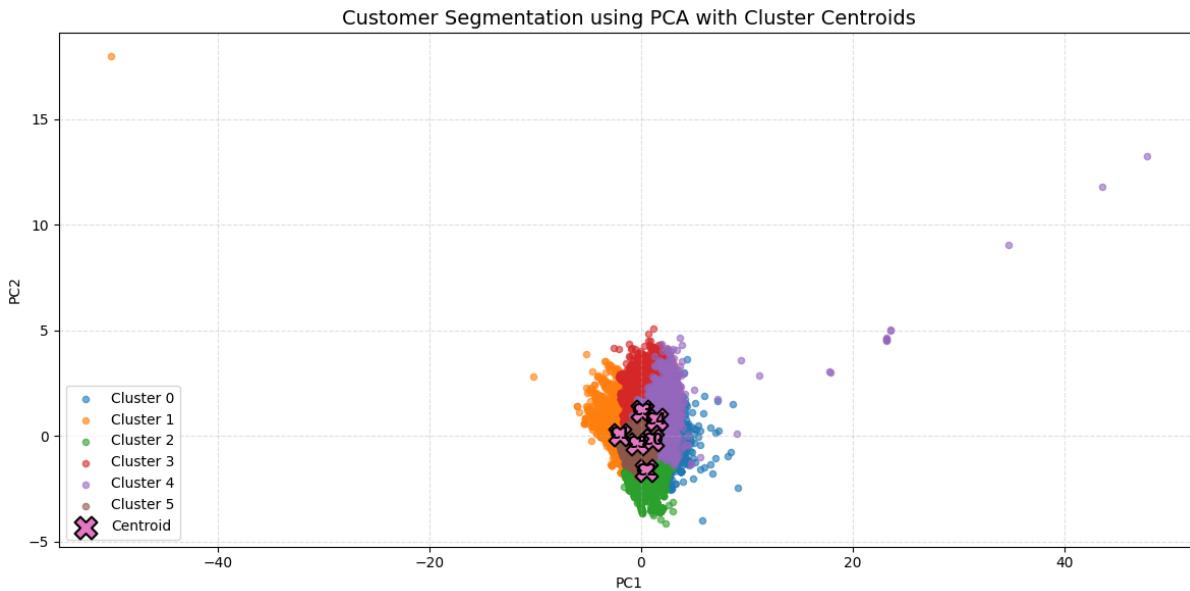
#### **4.3.1. Xác định số cụm tối ưu**

Phương pháp Elbow và Silhouette Score được sử dụng nhằm xác định số cụm tối ưu cho thuật toán K-Means. Kết quả thực nghiệm cho thấy giá trị K = 6 đạt Silhouette Score cao nhất (0,22), cho thấy mức độ phân tách giữa các cụm ở mức trung bình.

Giá trị Silhouette tương đối thấp có thể được lý giải bởi đặc thù của bộ dữ liệu Hotel Booking Demand, vốn phản ánh hành vi đặt phòng phức tạp và không đồng nhất của con người. Dữ liệu bao gồm nhiều biến định tính và định lượng, chịu ảnh hưởng mạnh bởi yếu tố mùa vụ, hành vi hủy phòng và sự đa dạng trong mục đích đặt phòng, dẫn đến ranh giới giữa các cụm không rõ ràng. Do đó, việc đạt được các cụm có độ tách biệt cao bằng K-Means là một thách thức.

#### **4.3.2. Diễn giải các cụm khách hàng**

Hình minh họa kết quả phân cụm sau khi giảm chiều dữ liệu bằng phương pháp PCA cho thấy các cụm khách hàng có xu hướng chồng lấn đáng kể trong không gian hai chiều (PC1, PC2). Các tâm cụm (centroids) nằm tương đối gần nhau, phản ánh sự khác biệt giữa các nhóm khách hàng ở mức trung bình. Kết quả này phù hợp với giá trị Silhouette Score đạt 0.22, cho thấy mức độ tách biệt cụm không cao nhưng vẫn tồn tại cấu trúc phân nhánh tiềm ẩn trong dữ liệu.



Hình 4.18: Minh họa các cụm và tâm của cụm đó

Hiện tượng chồng lấn giữa các cụm xuất phát từ bản chất của bộ dữ liệu Hotel Booking Demand, vốn phản ánh hành vi đặt phòng đa dạng và liên tục của con người, chịu ảnh hưởng bởi nhiều yếu tố như mục đích chuyến đi, thời gian đặt phòng, mùa vụ và khả năng chi tiêu.

#### **Cụm 0 – Khách lên kế hoạch sớm, lưu trú dài ngày**

Nhóm khách này có **thời gian đặt phòng trước cao** và **thời gian lưu trú dài nhất** trong các cụm. Mức chi tiêu trung bình và số yêu cầu đặc biệt ở mức vừa phải, phản ánh nhóm khách có kế hoạch rõ ràng, thường là khách nghỉ dưỡng dài ngày hoặc khách gia đình.

#### **Cụm 1 – Khách lưu trú ngắn hạn, chi tiêu cao**

Cụm này có **ADR trên mỗi khách cao nhất**, thời gian lưu trú ngắn và số lượng khách thấp. Đây là nhóm khách có giá trị doanh thu cao trên mỗi đêm, thường liên quan đến khách công tác hoặc khách du lịch ngắn ngày.

#### **Cụm 2 – Khách đặt rất sớm, nhạy cảm về giá**

Đây là cụm có **lead time cao nhất** nhưng **mức chi tiêu thấp**, chủ yếu thuộc phân khúc **Groups**. Nhóm này thường đặt sớm để hưởng ưu đãi và có xu hướng nhạy cảm về giá.

#### **Cụm 3 – Khách có nhiều yêu cầu đặc biệt**

Cụm này nổi bật với **số lượng yêu cầu đặc biệt cao nhất**, cho thấy nhóm khách chủ trọng đến trải nghiệm và dịch vụ. Mặc dù thời gian lưu trú không dài, nhóm này đòi hỏi sự chú ý nhiều hơn trong khâu vận hành.

#### **Cụm 4 – Khách đi theo nhóm/gia đình**

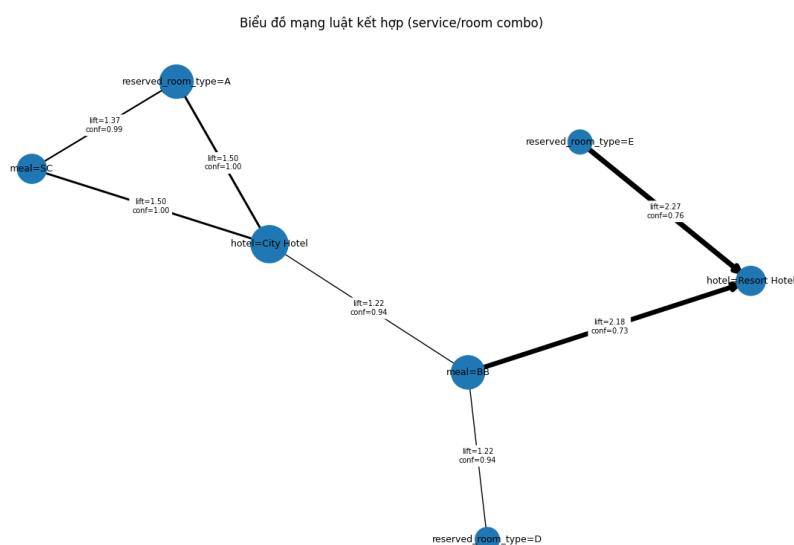
Nhóm khách này có **số lượng khách trung bình cao nhất**, thời gian lưu trú ở mức trung bình và số yêu cầu đặc biệt tương đối cao. Đây là nhóm khách gia đình hoặc nhóm bạn, phù hợp với các gói dịch vụ trọn gói.

### **Cụm 5 – Khách đơn giản, ít tương tác**

Cụm này có mức chi tiêu thấp và gần như **không có yêu cầu đặc biệt**, phản ánh nhóm khách đơn giản, dễ phục vụ nhưng giá trị kinh tế không cao.

Mặc dù giá trị *Silhouette Score* không cao, kết quả phân cụm vẫn mang ý nghĩa thực tiễn khi giúp xác định các nhóm khách hàng có **đặc điểm hành vi và giá trị kinh doanh khác nhau**. Điều này cho thấy trong các bài toán dữ liệu hành vi thực tế, **tính dien giải và giá trị ứng dụng** quan trọng hơn việc tối ưu các chỉ số đánh giá thuần túy.

#### **4.4. Kết quả luật kết hợp**



Hình 4.19: Minh họa biểu đồ luật kết hợp

Biểu đồ mạng luật kết hợp thể hiện các mối quan hệ đồng xuất hiện giữa **loại khách sạn (City/Resort)**, **loại phòng** và **dịch vụ ăn uống**. Kết quả cho thấy dữ liệu hình thành **hai cụm rõ rệt**, tương ứng với hai mô hình kinh doanh khách sạn.

*Cụm City Hotel* tập trung quanh các phòng tiêu chuẩn (đặc biệt là room type A) và dịch vụ **Self Catering (SC)** hoặc không kèm ăn. Các cạnh trong cụm này có giá trị lift > 1 và confidence cao, phản ánh xu hướng khách tại City Hotel ưu tiên sự linh hoạt và tối giản dịch vụ.

*Cụm Resort Hotel* thể hiện mối liên hệ mạnh với các phòng cao cấp (room type E, D) và dịch vụ **Bed & Breakfast (BB)**. Các cạnh nối giữa Resort Hotel và các dịch vụ này có độ dày lớn hơn, cho thấy lift và confidence cao hơn so với City Hotel.

Đặc biệt, luật liên quan đến *reserved\_room\_type = E → Resort Hotel* có giá trị lift cao nhất trong mạng, cho thấy loại phòng này gần như chỉ xuất hiện tại Resort Hotel. Điều này phản ánh rõ đặc trưng của mô hình khách sạn nghỉ dưỡng, nơi khách hàng có xu hướng lựa chọn phòng cao cấp đi kèm các dịch vụ trọn gói.

Kết quả khai phá luật kết hợp bằng thuật toán **Apriori** cho thấy các dịch vụ và loại phòng **không được lựa chọn ngẫu nhiên**, mà có xu hướng được đặt cùng nhau theo từng loại hình khách sạn:

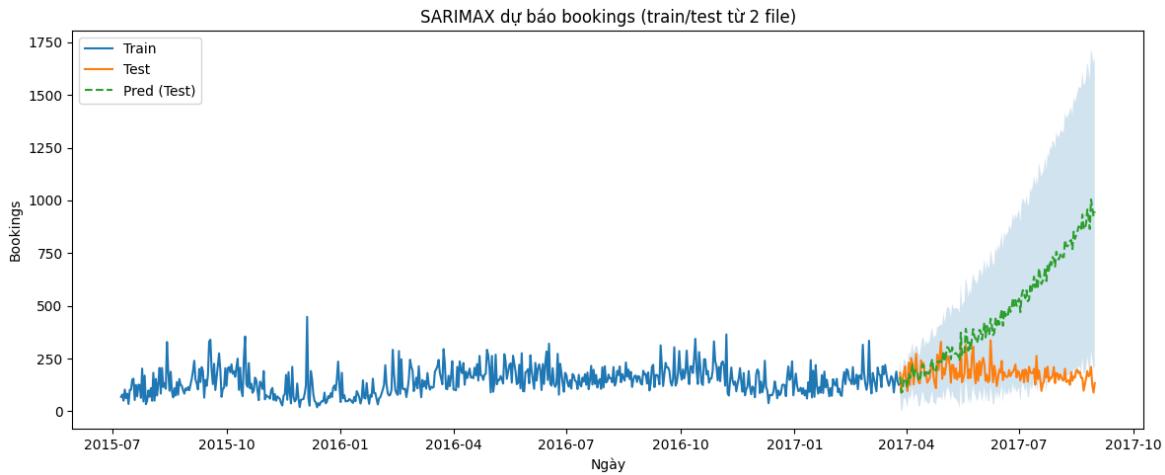
- **Resort Hotel** gắn liền với:
  - phòng cao cấp (room type E, D)
  - dịch vụ ăn sáng (BB)
- **City Hotel** chủ yếu đi kèm với:
  - phòng tiêu chuẩn (room type A)
  - hình thức tự phục vụ hoặc không kèm ăn (SC)

Các luật thu được đều có **lift > 1** và **confidence cao**, cho thấy mối liên hệ có ý nghĩa thống kê và mang giá trị thực tiễn. Những kết quả này có thể được ứng dụng trong:

- thiết kế **gói dịch vụ phù hợp từng loại khách sạn**
- **gợi ý bán chéo (cross-selling)** phòng và dịch vụ
- tối ưu chiến lược giá và phân khúc khách hàng

Mặc dù Apriori chỉ khai thác mối quan hệ đồng xuất hiện và không phản ánh quan hệ nhân quả, kết quả thu được vẫn cung cấp cái nhìn trực quan và dễ diễn giải về hành vi đặt phòng của khách hàng. Điều này đặc biệt hữu ích trong bối cảnh dữ liệu hành vi, nơi mục tiêu chính là **hiểu môi trường tiêu dùng** hơn là dự đoán chính xác.

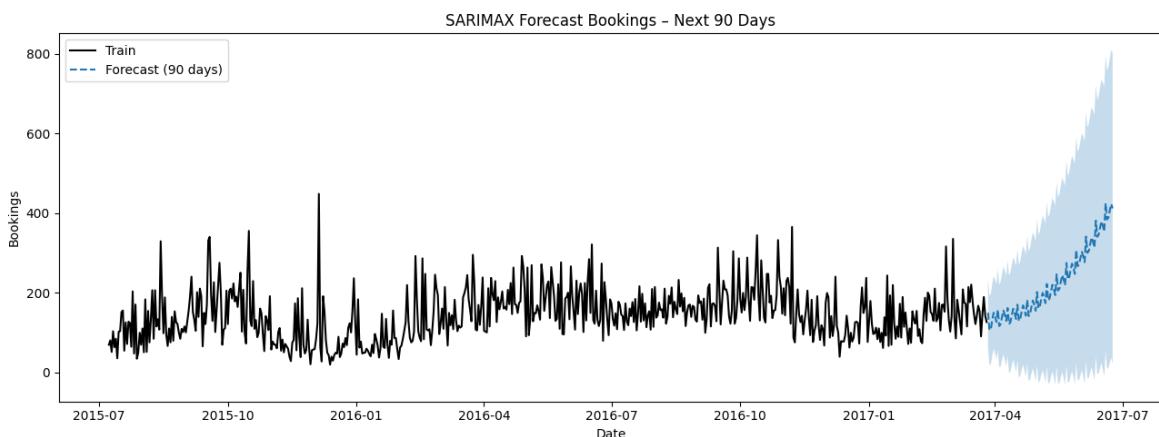
#### 4.5. Kết quả chuỗi thời gian



Biểu đồ dự báo SARIMAX trên tập huấn luyện và kiểm tra cho thấy mô hình đã **nắm bắt được xu hướng tổng thể và tính mùa vụ** của số lượng đặt phòng theo thời gian. Trong giai đoạn huấn luyện, chuỗi dữ liệu thể hiện sự biến động mạnh theo ngày, phản ánh đặc trưng thực tế của nhu cầu đặt phòng khách sạn.

Tuy nhiên, khi so sánh kết quả dự báo với dữ liệu kiểm tra, có thể nhận thấy mô hình **có xu hướng dự báo cao hơn giá trị thực tế** trong giai đoạn test. Khoảng tin cậy của dự báo (confidence interval) mở rộng đáng kể về cuối chuỗi, cho thấy **độ không chắc chắn tăng dần theo thời gian dự báo**.

Biểu đồ dự báo 90 ngày tiếp theo cho thấy số lượng đặt phòng **có xu hướng tăng**, phù hợp với giả định về tính mùa vụ và xu hướng tăng trưởng nhu cầu trong giai đoạn tiếp theo. Đường dự báo trung tâm thể hiện mức tăng ổn định, trong khi khoảng tin cậy mở rộng theo thời gian, phản ánh mức độ bất định của dự báo dài hạn.



Hình 4.20: Kết quả minh họa dự đoán chuỗi thời gian trong 90 ngày

Kết quả này cho thấy mô hình SARIMAX **phù hợp cho dự báo ngắn hạn và trung hạn**,

Tuy nhiên, với khoảng dự báo dài hơn, việc chỉ dựa vào dữ liệu lịch sử là chưa đủ, cần bổ sung thêm các biến ngoại sinh để cải thiện độ tin cậy.

Mô hình SARIMAX đã khai thác hiệu quả tính mùa vụ và xu hướng trong dữ liệu đặt phòng.

Sai lệch giữa dự báo và dữ liệu thực tế trong tập test cho thấy dữ liệu chịu ảnh hưởng mạnh từ các yếu tố bên ngoài như hành vi hủy phòng, sự kiện, hoặc thay đổi nhu cầu theo thời điểm.

Khoảng tin cậy rộng phản ánh tính bất định cao của dữ liệu hành vi, điều này là đặc trưng phổ biến trong các bài toán dự báo nhu cầu thực tế.

Do đó, kết quả dự báo cần được sử dụng như công cụ hỗ trợ ra quyết định, thay vì coi là giá trị dự đoán tuyệt đối.

#### **4.6. Tổng hợp & thảo luận đa mô hình**

Kết quả thực nghiệm cho thấy việc kết hợp nhiều kỹ thuật khai phá dữ liệu mang lại cái nhìn toàn diện và đa chiều về hành vi đặt phòng khách sạn. Mỗi phương pháp đóng một vai trò khác nhau nhưng hỗ trợ lẫn nhau trong việc phân tích và hỗ trợ ra quyết định.

Cụ thể, mô hình phân loại giúp nhận diện và kiểm soát rủi ro hủy phòng, hỗ trợ khách sạn trong việc quản lý công suất và doanh thu. Phân cụm khách hàng cho phép phân khúc khách hàng dựa trên hành vi đặt phòng và mức độ chi tiêu, từ đó giúp xây dựng các chiến lược tiếp thị và dịch vụ phù hợp cho từng nhóm khách. Luật kết hợp làm rõ các mối quan hệ đồng xuất hiện giữa loại phòng và dịch vụ, cung cấp cơ sở cho việc thiết kế các gói dịch vụ và chiến lược bán chéo hiệu quả. Cuối cùng, mô hình chuỗi thời gian hỗ trợ dự báo nhu cầu đặt phòng theo thời gian, góp phần tối ưu hóa kế hoạch vận hành và phân bổ nguồn lực.

Nhìn chung, cách tiếp cận kết hợp này không chỉ nâng cao khả năng hiểu biết về dữ liệu mà còn tăng tính ứng dụng thực tiễn của kết quả phân tích, đặc biệt trong bối cảnh dữ liệu hành vi phức tạp và biến động mạnh như lĩnh vực kinh doanh khách sạn.

## CHƯƠNG V: KẾT LUẬN

### 5.1. Tóm tắt kết quả

Nghiên cứu này đã xây dựng và đánh giá một khung phân tích khai phá dữ liệu tích hợp nhằm phân tích hành vi đặt phòng khách sạn, dự đoán khả năng hủy phòng và dự báo nhu cầu lưu trú trong ngắn hạn. Trên cơ sở bộ dữ liệu đặt phòng khách sạn giai đoạn 2015–2017, các kỹ thuật phân tích mô tả, phân loại, phân cụm, khai phá luật kết hợp và phân tích chuỗi thời gian đã được triển khai một cách hệ thống theo quy trình CRISP-DM.

Kết quả phân tích mô tả cho thấy nhu cầu đặt phòng có tính mùa vụ rõ rệt và hành vi của khách hàng chịu ảnh hưởng mạnh bởi các yếu tố như thời gian đặt trước, loại khách sạn và kênh đặt phòng. Các mô hình phân loại được xây dựng cho thấy khả năng hủy phòng không phải là hiện tượng ngẫu nhiên mà có thể được dự đoán hiệu quả ngay tại thời điểm đặt phòng, trong đó mô hình *Random Forest* đạt hiệu suất tốt nhất.

Bên cạnh đó, phân tích phân cụm đã xác định được các nhóm khách hàng có hành vi đặt phòng khác biệt rõ rệt về thời gian đặt, mức chi tiêu và tỷ lệ hủy phòng. Khai phá luật kết hợp làm rõ các mối quan hệ đồng xuất hiện giữa loại phòng, dịch vụ đi kèm và đặc điểm khách hàng. Cuối cùng, mô hình chuỗi thời gian cho phép dự báo nhu cầu đặt phòng trong khoảng 30–90 ngày tới với sai số chấp nhận được, phản ánh đúng xu hướng và tính mùa vụ của ngành khách sạn.

### 5.2. Trả lời câu hỏi nghiên cứu

Đối với nhóm câu hỏi nghiên cứu mô tả, nghiên cứu khẳng định rằng nhu cầu đặt phòng và hành vi khách hàng có sự biến động đáng kể theo thời gian và mùa vụ, đồng thời tồn tại sự khác biệt rõ rệt giữa các phân khúc khách hàng. Các yếu tố như thời gian đặt trước, loại phòng và kênh đặt phòng đóng vai trò quan trọng trong việc hình thành hành vi đặt phòng và hủy phòng.

Đối với nhóm câu hỏi nghiên cứu dự đoán, nghiên cứu cho thấy khả năng hủy phòng có thể được dự đoán với độ chính xác cao dựa trên thông tin đặt phòng ban đầu. Các mô hình phân loại được xây dựng đều cho kết quả khả quan, trong đó mô hình Random Forest thể hiện ưu thế rõ rệt nhờ khả năng khai thác các mối quan hệ phi tuyến trong dữ liệu.

Đối với nhóm câu hỏi về phân cụm, kết quả cho thấy khách hàng có thể được phân nhóm thành các cụm hành vi khác biệt dựa trên các đặc trưng như thời gian đặt, thời gian lưu trú và mức chi tiêu. Các cụm này thể hiện sự khác biệt đáng kể về giá trị mang lại cho khách sạn cũng như mức độ rủi ro hủy phòng.

Ngoài ra, kết quả khai phá luật kết hợp đã phát hiện các mối quan hệ có ý nghĩa giữa loại phòng, dịch vụ và đặc điểm khách hàng, trong khi phân tích chuỗi thời gian đã cung cấp khả năng dự báo nhu cầu đặt phòng trong ngắn hạn, hỗ trợ hiệu quả cho công tác lập kế hoạch vận hành và quản trị doanh thu.

### **5.3. Hạn chế của nghiên cứu**

Mặc dù đạt được các kết quả tích cực, nghiên cứu vẫn tồn tại một số hạn chế. Thứ nhất, dữ liệu được sử dụng chỉ giới hạn trong giai đoạn 2015–2017, do đó có thể chưa phản ánh đầy đủ sự thay đổi hành vi khách hàng trong bối cảnh hiện nay của ngành du lịch – khách sạn.

*Thứ hai*, nghiên cứu chưa xem xét các yếu tố ngoại sinh như thời tiết, sự kiện đặc biệt, biến động kinh tế hoặc chính sách du lịch, trong khi đây là những yếu tố có thể ảnh hưởng đáng kể đến nhu cầu đặt phòng và tỷ lệ hủy phòng.

*Thứ ba*, các mô hình được sử dụng chủ yếu là các phương pháp học máy truyền thống, chưa khai thác các mô hình học sâu hoặc mô hình lai có khả năng nắm bắt tốt hơn các mối quan hệ phức tạp trong dữ liệu.

### **5.4. Hướng mở rộng**

Trong các nghiên cứu tiếp theo, đề tài có thể được mở rộng theo một số hướng. Trước hết, việc sử dụng các bộ dữ liệu mới hơn, đa dạng hơn hoặc dữ liệu thời gian thực sẽ giúp nâng cao tính cập nhật và khả năng ứng dụng của mô hình. Bên cạnh đó, các mô hình học sâu như LSTM hoặc Transformer có thể được áp dụng cho bài toán dự báo chuỗi thời gian nhằm cải thiện độ chính xác dự báo trong bối cảnh dữ liệu có tính phi tuyến cao.

Ngoài ra, việc tích hợp thêm các yếu tố ngoại sinh như thời tiết, sự kiện và giá cạnh tranh sẽ giúp mô hình phản ánh sát hơn bối cảnh vận hành thực tế của khách sạn. Cuối cùng, kết quả nghiên cứu có thể được triển khai dưới dạng hệ thống hỗ trợ ra quyết định hoặc dashboard trực quan, góp phần đưa các kết quả khai phá dữ liệu vào ứng dụng thực tiễn trong quản trị khách sạn.

Tóm lại, nghiên cứu cho thấy việc kết hợp nhiều kỹ thuật khai phá dữ liệu trên cùng một bộ dữ liệu mang lại cái nhìn toàn diện hơn so với việc sử dụng từng kỹ thuật đơn lẻ, đồng thời chứng minh tiềm năng ứng dụng thực tế của học máy trong quản trị khách sạn hiện đại.

### Tài liệu tham khảo

**Arreeras, T., Arimura, M., Asada, T., & Arreeras, S. (2019).** *Association rule mining of tourist-attractive destinations.* Sustainability.

**Chiang, W. K., Chen, Y. C., & Chen, M. Y. (2017).** The impact of inaccurate demand forecasts on hotel revenue management: A simulation study. *Journal of Hospitality & Tourism Research*, 41(2), 175-200.

**Chen, S., Ngai, E. W. T., Ku, Y., Xu, Z., & Gou, X. (2023).** *Prediction of hotel booking cancellations using machine learning.* Decision Support Systems

**Dolnicar, S., & Le, K. (2017).** A review of data-driven market segmentation in tourism. *Journal of Travel Research*, 56(3), 346-359.

**Deldadehasl, M. (2025).** *Customer clustering and marketing optimization in hospitality industry.* MDPI Hospitality.

**Eibl, S. (2024).** *Clustering user characteristics in hotel booking situations.* ACM Proceedings.

**UNWTO (2025).** World Tourism Barometer. *Tạp chí du lịch thế giới (Vol. 23, Iss. 1).*

**Kim, H., Kim, K., & Lee, S. (2020).** Digital Transformation in Small and Medium-Sized Hotels: Barriers and Driving Forces. *Journal of Hospitality & Tourism Technology*, 11(3), 443-461.

**Moro, S., Rita, P., & Coelho, J. (2019).** *Predicting hotel booking cancellations.* Expert Systems with Applications.

**Nguyen Van Chuc & Dao Thi Giang (2015).** *Ứng dụng kỹ thuật phân cụm và luật kết hợp khai phá dữ liệu khách hàng sử dụng dịch vụ khách sạn.* Tạp chí KH&CN – ĐH Đà Nẵng.

**Law, R., Lin, H., & Chen, W. (2018).** A comparison of traditional time series models and advanced machine learning models for hotel booking forecast. *International Journal of Contemporary Hospitality Management*, 30(6), 2465-2485.

**Hotel booking demand datasets** — Nuno António, Ana de Almeida & Luís Nunes (2019). *Data in Brief*, 22, 41–49.

**Song, H. & Li, G. (2008).** *Tourism demand modelling and forecasting.* Tourism Management.

