**TRƯỜNG ĐẠI HỌC BÌNH DƯƠNG**



**PHÂN HIỆU ĐẠI HỌC BÌNH DƯƠNG TẠI CÀ MAU**

**BÁO CÁO**

**MÔN: KHOA HỌC DỮ LIỆU**

**(Đề tài: DỰ ĐOÁN LƯỢNG KHÁCH ĐẾN NHÀ HÀNG NHẬT BẢN)**

**Giảng viên hướng dẫn: Trần Thanh Nhã**

**Sinh viên thực hiện:**

**Trần Thành Phố 200501022**

**Trần Chí Hài 200501017**

**Vũ Nguyễn Thái Dương 200501013**

**Nguyễn Minh Thuận 200501019**

**Niên khóa: 2020-2024**

***Cà Mau, tháng 04 năm 2023***

**TRƯỜNG ĐẠI HỌC BÌNH DƯƠNG**



**PHÂN HIỆU ĐẠI HỌC BÌNH DƯƠNG TẠI CÀ MAU**

**BÁO CÁO**

**MÔN: KHOA HỌC DỮ LIỆU**

**(Đề tài:DỰ ĐOÁN LƯỢNG KHÁCH ĐẾN NHÀ HÀNG NHẬT BẢN)**

**Giảng viên hướng dẫn:**

**Sinh viên thực hiện: Trần Thanh Nhã**

**Trần Thành Phố 200501022**

**Trần Chí Hài 200501017**

**Vũ Nguyễn Thái Dương 200501013**

**Nguyễn Minh Thuận 200501019**

**Niên khóa: 2020-2024**

***Cà Mau, tháng 04 năm 2023***

**Bảng PHÂN CÔNG THÀNH VIÊN TRONG NHÓM**

|  |  |
| --- | --- |
| **Tên** | **Nhiệm Vụ** |
| **Trần Thành Phố** | **coder** |
| **Vũ Nguyễn Thái Dương** | **Chỉnh sửa Word, Thiết Kế PowerPoint** |
| **Trần Chí Hài** | **Tìm Tài Liệu , hỗ trợ Word** |
| **Nguyễn Minh Thuận** | **Tìm Tài Liệu , coder** |

**NHẬN XÉT CỦA GIẢNG VIÊN**

**….……………………………………………………………………………………**

**….……………………………………………………………………………………**

**….……………………………………………………………………………………**

**….……………………………………………………………………………………**

**….……………………………………………………………………………………**

**….……………………………………………………………………………………**

**….……………………………………………………………………………………**

**….……………………………………………………………………………………**

**….……………………………………………………………………………………**

**….……………………………………………………………………………………**

**….……………………………………………………………………………………**

**….……………………………………………………………………………………**

**….……………………………………………………………………………………**

|  |  |
| --- | --- |
|  | **Giáo viên hướng dẫn**  **Ký tên** |

***LỜI CẢM ƠN***

Đầu tiên, em xin gửi lời cảm ơn sâu sắc đến Trường Đại Học Bình Dương phân hiệu Cà Mau đã đưa bộ môn Khoa Học Dữ Liệu vào chương trình giảng dạy. Đặc biệt, em xin bày tỏ lòng biết ơn sâu sắc đến giảng viên bộ môn - thầy Trần Thanh Nhã chính thầy là người đã tận tình dạy dỗ và truyền đạt những kiến thức quý báu cho chúng em trong suốt học kỳ vừa qua. Trong thời gian tham dự lớp học của thầy ,chúng em đã được tiếp cận với nhiều kiến thức bổ ích và rất cần thiết cho quá trình học tập, làm việc sau này của em.

Bộ môn Khoa Học Dữ Liệu là một môn học thú vị và vô cùng bổ ích. Tuy nhiên, những kiến thức và kỹ năng về môn học này của em vẫn còn nhiều hạn chế. Do đó, bài tiểu luận của em khó tránh khỏi những sai sót. Kính mong thầy xem xét và góp ý giúp bài tiểu luận của em được hoàn thiện hơn**.**

**MỤC lỤC**

[**Chương 1. Giới Thiệu Về Đề Tài 9**](#_Toc133071656)

[**1.1 Lý do chọn đề tài 9**](#_Toc133071657)

[**1.2 Mục tiêu nghiên cứu 9**](#_Toc133071658)

[**1.3 Phạm vi nghiên cứu 10**](#_Toc133071659)

[**1.4 Ý nghĩa 11**](#_Toc133071660)

[**1.5 Mô tả nghiệp vụ 11**](#_Toc133071661)

[**1.6 Phương hướng nghiên cứu 12**](#_Toc133071662)

[**1.7 Phân tích bài toán 13**](#_Toc133071663)

[**Chương 2. Cơ sở lý thuyết 15**](#_Toc133071664)

[**2.1 Cách thức thu thập dữ liệu 15**](#_Toc133071665)

[**2.2 Thông tin dữ liệu 16**](#_Toc133071666)

[**2.3 Phân tích các thuộc tính dữ liệu 17**](#_Toc133071667)

[**2.3 Công cụ thực hiện 18**](#_Toc133071668)

[**2.3.1 Google Colab 18**](#_Toc133071669)

[**2.3.2 Thư viện numpy 19**](#_Toc133071670)

[**2.3.3 Thư viện pandas 19**](#_Toc133071671)

[**2.3.4 Thư viện Scikit-learn 20**](#_Toc133071672)

[**2.3.5 Data preprocessing 20**](#_Toc133071673)

[**2.3.6 Mô đun selection 20**](#_Toc133071674)

[**2.4 Mô hình máy học 20**](#_Toc133071675)

[**2.4.1 Mô hình Xgboost 21**](#_Toc133071676)

[**Ưu điểm 21**](#_Toc133071677)

[**Nhược điểm 22**](#_Toc133071678)

[**2.4.2 Hồi quy tuyến tính là gì? 22**](#_Toc133071679)

[**2.4.3 Tại sao hồi quy tuyến tính lại quan trọng? 23**](#_Toc133071680)

[**2.4.4 Hồi quy tuyến tính hoạt động như thế nào? 23**](#_Toc133071681)

[**2.4.5 Hồi quy tuyến tính được áp dụng như thế nào trước khi đưa vào mô hình XGBoost? 24**](#_Toc133071682)

[**CHƯƠNG 3.XÂY DỰNG DỮ LIỆU 25**](#_Toc133071683)

[**3.1 Biểu đồ trực quan 25**](#_Toc133071684)

[**3.1.1 Biểu đồ 1. Số lượng khách hành đến nhà hàng theo từng ngày trong tuần 25**](#_Toc133071685)

[**3.1.2Biểu đồ 2. biểu thị giá trị trung vị của số lượng khách du lịch (visitors) vào các ngày trong tháng. 26**](#_Toc133071686)

[**3.1.3 Biểu đồ 3. biểu thị tổng số lượng khách du lịch(visitors) 27**](#_Toc133071687)

[**3.1.4 Biểu đồ 4. Số lượng khách thăm quan vào các ngày trong khoảng thời gian từ ngày 15/4/2016 đến ngày 15/6/2016. 28**](#_Toc133071688)

[**3.1.5 Biểu đồ 5. Biểu thị tổng số lượng khách đặt chỗ (reserve visitors) vào các ngày khác nhau. 29**](#_Toc133071689)

**[3.1.6 Biểu đồ 6. biểu thị tổng số lượng khách đặt chỗ (reserve visitors) vào mỗi giờ của ngày. 30](#_Toc133071690)**

[**3.1.7 Biểu đồ 7. Biểu đồ đường để biểu thị tổng số lượng khách (all visitors) đặt bàn trong từng ngày. 31**](#_Toc133071691)

[**3.1.8 Biểu đồ 8. Tổng số lượng khách đặt chỗ (reserve visitors) trong từng giờ trong ngày. 32**](#_Toc133071692)

[**3.1.9Biểu đồ 9. Để hiển thị số lượng ngày là ngày lễ. 33**](#_Toc133071693)

[**3.1.10 Biểu đồ 10. Phân tán để biểu thị giá trị trung bình số lượng khách du lịch (visitors) vào các ngày trong tuần và trong các ngày lễ. 34**](#_Toc133071694)

[**CHƯƠNG 4. KẾT LUẬN 36**](#_Toc133071695)

[**4.1 Kết quả đạt được 36**](#_Toc133071696)

[**4.2 Định hướng tương lai 36**](#_Toc133071697)

[**Tài liệu tham thảo 37**](#_Toc133071698)

# 

# Chương 1. Giới Thiệu Về Đề Tài

* 1. **Lý do chọn đề tài**

Nhật Bản là một điểm đến du lịch phổ biến trên thế giới, với nhiều điểm đến du lịch hấp dẫn như Tokyo, Kyoto, Osaka, Hokkaido và Okinawa. Vì vậy, việc dự đoán lượng khách du lịch đến Nhật Bản sẽ mang lại nhiều giá trị cho ngành du lịch của đất nước này. Dự đoán lượng khách hàng đến Nhật Bản có thể giúp các tổ chức và doanh nghiệp trong lĩnh vực du lịch và dịch vụ liên quan lên kế hoạch cho các hoạt động kinh doanh, bao gồm quảng cáo, marketing, khuyến mại, và đặc biệt là lên kế hoạch về lực lượng lao động, đảm bảo các dịch vụ được cung cấp đúng lượng khách hàng.

Ví dụ, thông qua kết quả dự đoán lưu lượng khách hàng, nhà hàng có thể lên kế hoạch nhân sự sao cho đảm bảo có đủ nhân viên phục vụ khách hàng vào các thời điểm có lưu lượng khách hàng cao nhất. Điều này giúp tránh tình trạng chờ đợi của khách hàng và đảm bảo dịch vụ tốt nhất.

Bên cạnh đó, kết quả dự đoán cũng có thể được sử dụng để đặt hàng vật liệu, ví dụ như đồ dùng bếp, thực phẩm và nước uống, để đảm bảo đủ nguồn cung cấp vào các thời điểm có lưu lượng khách hàng dự đoán cao nhất. Điều này giúp tránh tình trạng thiếu hụt nguồn cung cấp và đảm bảo dịch vụ đúng thời điểm và chất lượng tốt nhất.

* 1. **Mục tiêu nghiên cứu**

Mục tiêu của dự đoán lượng khách đến nhà hàng ở Nhật Bản là giúp cho những người quan tâm đến việc tìm kiếm một nhà hàng phù hợp cho mình có thể đưa ra quyết định dựa trên số liệu và thông tin có sẵn. Dự đoán lượng khách đến nhà hàng sẽ giúp những chủ nhà hàng có thể lên kế hoạch về cung cấp dịch vụ, số lượng nhân viên phục vụ và tổ chức các hoạt động quảng cáo phù hợp. Đồng thời, dự đoán lượng khách đến cũng giúp cho khách hàng có thể biết được tình trạng đông đúc của nhà hàng và lựa chọn thời điểm đến ăn một cách thuận lợi hơn.

* 1. **Phạm vi nghiên cứu**

Phạm vi nghiên cứu của dự án này là sử dụng dữ liệu từ tập dữ liệu “Recruit Restaurant Visitor Forecasting”, có sẵn trên GitHub, để dự đoán lượng khách đến các nhà hàng ở Nhật Bản trong tương lai. Tập dữ liệu này chứa thông tin về số lượng khách đến các nhà hàng từ tháng 1 năm 2016 đến tháng 4 năm 2017, bao gồm thông tin về địa điểm, thời gian và các yếu tố khác như tiền tệ, ngày lễ và thời tiết.

Phương pháp nghiên cứu sẽ sử dụng các kỹ thuật phân tích dữ liệu và mô hình học máy để xây dựng một mô hình dự đoán đáng tin cậy cho lượng khách đến các nhà hàng. Cụ thể, phạm vi nghiên cứu bao gồm:

Thu thập và tiền xử lý dữ liệu: thu thập dữ liệu từ tập dữ liệu Recruit Restaurant Visitor Forecasting, tách và tiền xử lý dữ liệu để chuẩn bị cho việc xây dựng mô hình.

Phân tích và khai phá dữ liệu: sử dụng các kỹ thuật phân tích dữ liệu để khám phá sự tương quan giữa các biến đầu vào và lượng khách đến các nhà hàng. Điều này sẽ giúp xác định các yếu tố ảnh hưởng đến lượng khách và đưa ra các giả định về mô hình dự đoán.

Xây dựng mô hình dự đoán: sử dụng các kỹ thuật học máy để xây dựng mô hình dự đoán. Trong phạm vi này, chúng tôi sẽ tập trung vào các mô hình học máy phổ biến như Linear Regression, Random Forest Regression, Gradient Boosting Regression và các mô hình học sâu như LSTM.

Đánh giá và tinh chỉnh mô hình: đánh giá độ chính xác của các mô hình dự đoán bằng cách so sánh kết quả dự đoán với giá trị thực tế. Sau đó, tinh chỉnh các tham số của mô hình để cải thiện độ chính xác.

Dự đoán và đánh giá kết quả: sử dụng mô hình tốt nhất để dự đoán lượng khách đến các nhà hàng trong tương lai và đánh giá độ chính xác của dự đoán.

* 1. **Ý nghĩa**

Ý nghĩa của dự đoán lượng khách đến nhà hàng ở Nhật Bản là rất quan trọng trong việc quản lý hoạt động kinh doanh của nhà hàng. Nó giúp cho chủ nhà hàng có thể lên kế hoạch về việc cung cấp dịch vụ và sản phẩm một cách chính xác, giảm thiểu việc lãng phí tài nguyên như thức ăn và nhân viên, tối ưu hóa việc quản lý và phân bổ tài nguyên.

Ngoài ra, dự đoán lượng khách đến còn giúp cho chủ nhà hàng có thể xác định được lượng doanh thu dự kiến và có thể đưa ra các chiến lược kinh doanh phù hợp để tăng doanh thu và lợi nhuận. Đối với khách hàng, dự đoán lượng khách đến nhà hàng giúp cho họ có thể lựa chọn thời gian đến nhà hàng phù hợp với sở thích và nhu cầu của mình, tránh tình trạng đông đúc quá mức hoặc không có chỗ ngồi. Nó cũng giúp khách hàng có thể lựa chọn đượcnhà hàng phù hợp với mình dựa trên thông tin về lượng khách đến trong quá khứ, đánh giá chất lượng và dịch vụ của nhà hàng.

* 1. **Mô tả nghiệp vụ**

Thu thập dữ liệu: Cần thu thập thông tin từ hai trang web hpg và air, bao gồm thông tin về nhà hàng, địa điểm, thời gian, số lượng khách đến, giá cả và đánh giá của khách hàng.

Xử lý dữ liệu: Sau khi thu thập dữ liệu, cần tiền xử lý dữ liệu để làm sạch và chuẩn hóa dữ liệu để đưa vào mô hình dự đoán.

Xây dựng mô hình dự đoán: Sau khi tiền xử lý dữ liệu, cần xây dựng mô hình dự đoán để dự đoán lượng khách đến của một nhà hàng trong tương lai. Có thể sử dụng các phương pháp dự đoán chuỗi thời gian hoặc học máy để xây dựng mô hình.

Đánh giá mô hình: Sau khi xây dựng mô hình, cần đánh giá mô hình để kiểm tra tính chính xác và độ tin cậy của mô hình.

Có thể sử dụng các phương pháp đánh giá như chia tập dữ liệu thành tập huấn luyện và tập kiểm tra hoặc sử dụng các độ đo chính xác như RMSE, MAE, R-squared. Sử dụng mô hình để chọn nhà hàng phù hợp: Cuối cùng, có thể sử dụng mô hình đã xây dựng để dự đoán lượng khách đến của các nhà hàng và chọn ra nhà hàng phù hợp với yêu cầu của mình, bao gồm địa điểm, giá cả, đánh giá của khách hàng và dự đoán lượng khách đến trong tương lai.

* 1. **Phương hướng nghiên cứu**

Sử dụng các phương pháp học máy và mô hình dự đoán chuỗi thời gian để phân tích dữ liệu lượng khách đến nhà hàng từ các nguồn khác nhau, bao gồm hpg và air.

Các mô hình này có thể dự đoán lượng khách đến nhà hàng dựa trên các biến như ngày tháng, giờ đồng hồ, điều kiện thời tiết, khu vực địa lý, giá cả và các yếu tố khác.

Thu thập các dữ liệu thêm từ các nguồn khác như mạng xã hội, các trang đánh giá và đặt chỗ khác, và các hệ thống phản hồi khách hàng của nhà hàng. Các dữ liệu này có thể cung cấp thông tin về sở thích, nhu cầu và hành vi của khách hàng, giúp cải thiện độ chính xác của mô hình dự đoán.

Tối ưu hóa việc sử dụng các nguồn dữ liệu khác nhau để tăng độ chính xác của mô hình dự đoán. Điều này có thể đòi hỏi việc sử dụng các phương pháp tiền xử lý dữ liệu, kết hợp các mô hình khác nhau và tối ưu hóa các tham số của mô hình. Áp dụng các kỹ thuật như học sâu và mạng nơ-ron để cải thiện độ chính xác của mô hình dự đoán.

Phân tích độ ảnh hưởng của các yếu tố khác nhau đến lượng khách đến nhà hàng và đưa ra các đề xuất để tăng lượng khách đến nhà hàng, bao gồm các chiến lược quảng cáo, giảm giá, nâng cao chất lượng dịch vụ và menu.

* 1. **Phân tích bài toán**

Bài toán dự đoán lượng khách đến nhà hàng ở Nhật Bản có thể được giải quyết bằng cách sử dụng mô hình XGBoost (Extreme Gradient Boosting), một trong những mô hình học máy được sử dụng phổ biến hiện nay để giải quyết các bài toán dự đoán và phân loại.

Phân tích bài toán dự đoán lượng khách đến nhà hàng có thể được thực hiện như sau:

Đọc và xử lý dữ liệu: Tải dữ liệu từ đường dẫn <https://github.com/pai4451/Recruit-Restaurant-Visitor-Forecasting> và thực hiện các bước xử lý dữ liệu cần thiết, bao gồm xóa các dữ liệu thiếu, chuyển đổi định dạng ngày tháng và các bước tiền xử lý khác.

Tiền xử lý dữ liệu: Trích xuất các đặc trưng (features) từ dữ liệu, bao gồm thông tin về ngày, tháng, năm, giá trị đặt trước, giá trị trung bình và các thông tin khác liên quan đến nhà hàng.

Xây dựng mô hình XGBoost: Tiến hành chia dữ liệu thành các tập huấn luyện và kiểm tra và xây dựng mô hình XGBoost. Áp dụng kỹ thuật tăng cường độ dốc (Gradient Boosting) để cải thiện độ chính xác của mô hình. Sử dụng các tham số phù hợp để tinh chỉnh mô hình XGBoost.

Đánh giá mô hình: Tiến hành kiểm tra độ chính xác của mô hình trên tập kiểm tra và đánh giá độ chính xác của mô hình bằng các chỉ số đánh giá như độ chính xác, F1 score, độ phân loại chính xác và các chỉ số khác.

Dự đoán lượng khách đến nhà hàng: Áp dụng mô hình đã xây dựng để dự đoán lượng khách đến nhà hàng trong tương lai và đưa ra kết quả dự đoán.

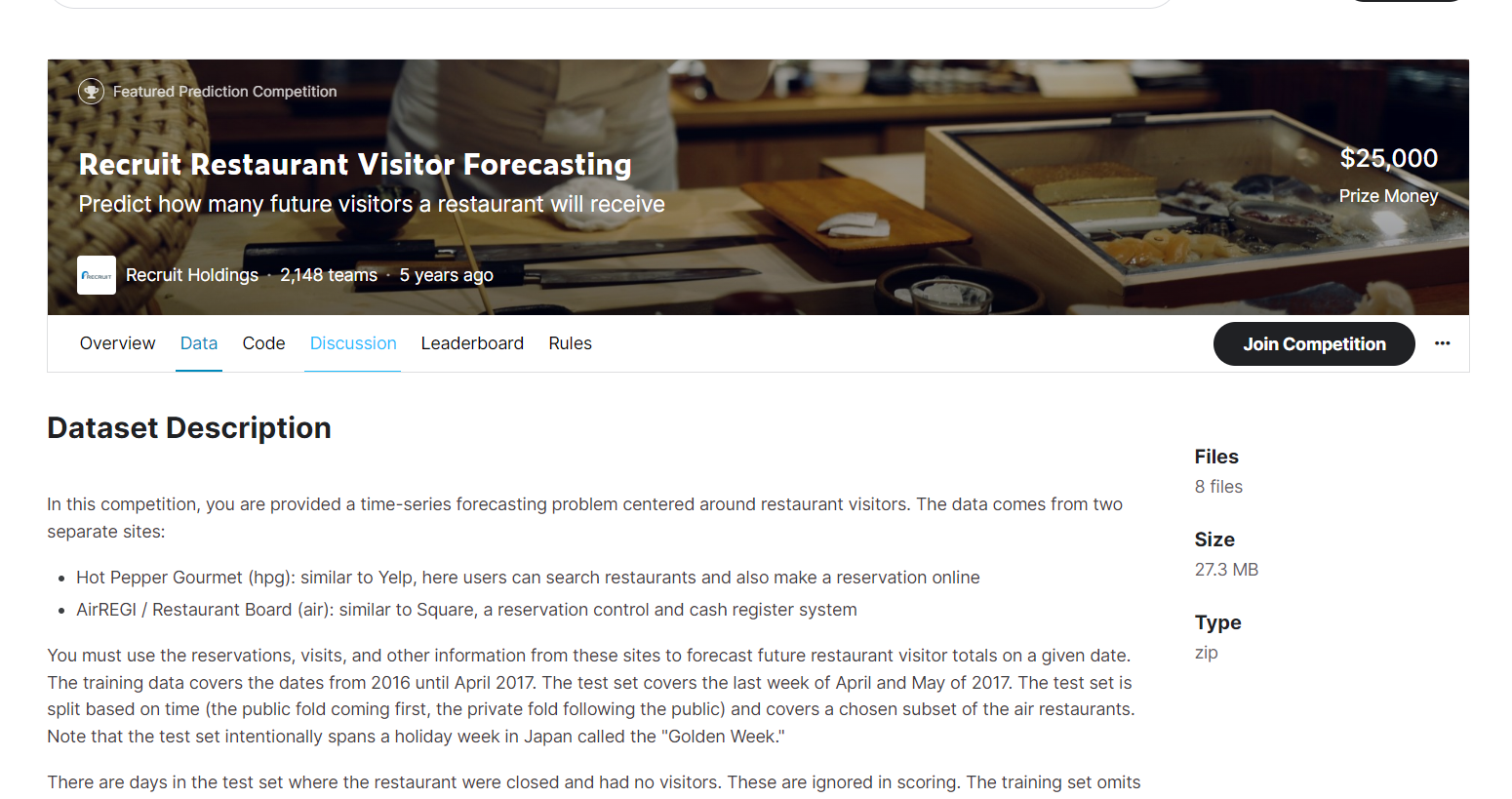
Đánh giá kết quả và cải thiện mô hình: Kiểm tra kết quả dự đoán và đánh giá hiệu suất của mô hình để tìm cách cải thiện mô hình. Các cách để cải thiện mô hình có thể bao gồm thay đổi các tham số mô hình, tăng cường dữ liệu huấn luyện hoặc sử dụng các phương pháp học máy khác.

Điểm mạnh của các phương pháp này là chúng có thể giúp dự đoán số lượng khách đến các nhà hàng với độ chính xác khá cao, tùy thuộc vào các biến đầu vào và cách tiếp cận.

**Chương 2. Cơ sở lý thuyết**

**2.1 Cách thức thu thập dữ liệu**

Truy cập vào trang mạng :<http://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/data> để lấy dữ liệu.

****

Dữ liệu dự đoán lượng khách đến nhà hàng ở Nhật Bản có thể được thu thập từ nhiều nguồn khác nhau, bao gồm các hệ thống quản lý đặt bàn trực tuyến, các ứng dụng đặt chỗ nhà hàng và các hệ thống thanh toán điện tử. Tuy nhiên, trong trường hợp này, dữ liệu đã được thu thập sẵn và được cung cấp trên đường dẫn <https://github.com/pai4451/Recruit-Restaurant-Visitor-Forecasting>.

Dữ liệu này bao gồm thông tin về lượng khách đến nhà hàng của chuỗi nhà hàng Recruit Holdings ở Nhật Bản trong khoảng thời gian từ tháng 1 năm 2016 đến tháng 4 năm 2017. Dữ liệu bao gồm các cột như:

air\_store\_id: mã số nhà hàng

visit\_date: ngày khách đến nhà hàng

visitors: số lượng khách đến nhà hàng trong ngày đó

day\_of\_week: thứ trong tuần của ngày đó

holiday\_flg: ngày đó có phải là ngày lễ hay không

air\_genre\_name: thể loại nhà hàng

air\_area\_name: vị trí nhà hàng

latitude: vĩ độ

longitude: kinh độ

Dữ liệu này được cung cấp dưới dạng tập tin CSV và có thể được đọc vào các công cụ phân tích dữ liệu như Jupyter Notebook sử dụng các thư viện Python như Pandas.

Khi đọc dữ liệu vào Jupyter Notebook, bạn có thể sử dụng các thư viện Pandas và Numpy để xử lý và phân tích dữ liệu, đồng thời sử dụng mô hình XGBoost để dự đoán lượng khách đến nhà hàng. Bạn cũng có thể tìm kiếm và tham khảo các bộ dữ liệu khác để bổ sung thông tin và cải thiện độ chính xác của mô hình.

**2.2 Thông tin dữ liệu**

Trong cuộc thi này, bạn được cung cấp một bài toán dự báo chuỗi thời gian xoay quanh những vị khách đến nhà hàng. Dữ liệu đến từ hai trang web riêng biệt:

* Hot Pepper Gourmet (hpg): tương tự như Yelp, tại đây người dùng có thể tìm kiếm nhà hàng và đặt chỗ trực tuyến .
* AirREGI / Restaurant Board (air): tương tự như Square, hệ thống kiểm soát đặt chỗ và tính tiền.

Bạn phải sử dụng thông tin đặt trước, lượt truy cập và các thông tin khác từ các trang web này để dự đoán tổng số khách truy cập nhà hàng trong tương lai vào một ngày nhất định.

Dữ liệu đào tạo bao gồm các ngày từ năm 2016 đến tháng 4 năm 2017. Tập kiểm tra bao gồm tuần cuối cùng của tháng 4 và tháng 5 năm 2017. Tập kiểm tra được phân chia dựa trên thời gian (phần mở rộng công khai đến trước, phần mở rộng riêng tư theo sau công khai) và bao gồm một tập hợp con được chọn của các nhà hàng trên không. Lưu ý rằng bộ thử nghiệm cố tình kéo dài một tuần nghỉ lễ ở Nhật Bản được gọi là "Tuần lễ vàng". Có những ngày trong tập thử nghiệm, nhà hàng đóng cửa và không có khách. Những điều này được bỏ qua trong tính điểm. Tập huấn luyện bỏ qua những ngày mà các nhà hàng đóng cửa.

**2.3 Phân tích các thuộc tính dữ liệu**

Bảng 1. Phân tích các thuộc tính dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| Tên thuộc tính | Mô tả thuộc tính | Kiểu dữ liệu | Mô tả minh hoạ |
| ID | ID Khách hàng | String | hpg\_6622b62385aec8bf air\_877f79706adbfb06 |
| Visitors | Số lượng khách đến | int | 0,1,2,3 |
| Visit\_datetime | Ngày và giờ đặt bàn | string | 1/1/2016 7:00:00 PM |
| reserve\_datetime | Ngày và giờ đặt bàn trước | string | 1/1/2016 4:00:00 PM |
| genre\_name(hpg\_genre\_name)(air\_genre\_name) | Tên nhà hàng | string | Italian,French |
| area\_name (air\_area\_name)(hpg\_area\_name) | Vị trí nhà hàng | String | Hyōgo-kenKōbe-shi,Kumoidōri,Tōkyō-to Minato-ku,Shibakōen |
| Latitude,Longitude | Tọa độ vĩ độ của nhà hàng(Kinh độ, Vĩ độ) | Float(số thực) | 34.695124, 135.197853 |
| Calendar\_date | Ngày tháng năm | datetime | 2016-01-01 |
| Date\_of\_week | Thứ trong tuần | String | Friday, Saturday, Sunday |
| Holiday\_flg | Cho biết ngày đó có phải là ngày nghỉ không | boolean | 1,0,1,0 |

## 2.3 Công cụ thực hiện

### **2.3.1 Google Colab**

Google Colab chỉ đơn giản là một đại diện trực tuyến của Jupyter Notebook. Mặc dù Jupyter Notebook cần cài đặt trên máy tính và chỉ có thể sử dụng tài nguyên máy cục bộ, nhưng Colab là một ứng dụng đám mây chính thức để mã hóa Python. Bạn có thể viết mã Python bằng Colab trên trình duyệt web Google Chrome hoặc Mozilla Firefox của mình. Bạn cũng có thể thực thi các mã đó trên trình duyệt mà không cần bất kỳ môi trường thời gian chạy hoặc giao diện dòng lệnh nào.

Hơn nữa, bạn có thể cung cấp cho sổ ghi chép dự án Python của mình một giao diện chuyên nghiệp bằng cách thêm các phương trình toán học, đồ thị, bảng, hình ảnh và đồ họa khác. Ngoài ra, bạn có thể mã hóa hình ảnh dữ liệu bằng Python và Colab sẽ hiển thị mã trong một tài sản trực quan. Colab cho phép bạn sử dụng lại các tệp Jupyter Notebook từ GitHub. Ngoài ra, bạn cũng có thể nhập các dự án khoa học dữ liệu và máy học tương thích từ các nguồn khác. Colab xử lý nội dung đã nhập một cách hiệu quả để hiển thị mã Python rõ ràng và không có lỗi.

### **2.3.2 Thư viện numpy**

NumPy chứa các tính năng khác nhau bao gồm những tính năng quan trọng sau:

* Đối tượng mảng N-chiều mạnh mẽ
* Các chức năng broadcasting
* Phép biến đổi Fourier, khả năng số ngẫu nhiên
* Các công cụ để tích hợp mã C / C ++ và Fortran.

Bên cạnh những công dụng khoa học rõ ràng, NumPy cũng có thể được sử dụng như một nơi chứa dữ liệu chung đa chiều hiệu quả. Các kiểu dữ liệu tùy ý có thể được xác định bằng cách sử dụng NumPy, cho phép NumPy tích hợp liền mạch và nhanh chóng với nhiều loại cơ sở dữ liệu.

### **2.3.3 Thư viện pandas**

Pandas là một thư viện Python toàn diện; một nguồn lực để thực hiện phân tích và thao tác dữ liệu; bất kỳ loại xử lý, phân tích, lọc và tổng hợp dữ liệu nào. Thư viện này được xây dựng dựa trên ngôn ngữ lập trình Python và có thể được sử dụng cho bất kỳ quy trình thu thập thông tin chi tiết từ dữ liệu nào.

Trong nghiên cứu khoa học dữ liệu, Pandas là một trong những công cụ quan trọng trong việc hỗ trợ, xử lý và phân tích dữ liệu với mã nguồn mở nhanh, mạnh, linh hoạt và dễ sử dụng .

### **2.3.4 Thư viện Scikit-learn**

Scikit-learn (trước đây là scikits.learn và còn được gọi là sklearn) là một thư viện máy học phần mềm miễn phí dành cho ngôn ngữ lập trình Python . Nó có nhiều thuật toán phân loại, hồi quy và phân cụm bao gồm máy vectơ hỗ trợ , rừng ngẫu nhiên , tăng cường độ dốc , k -means và DBSCAN và được thiết kế để tương tác với các thư viện số và khoa học Python NumPy và khoa học . Scikit-learning là một dự án được tài trợ bởi NumFOCUS.

### **2.3.5 Data preprocessing**

Data Preprocessing là Dữ liệu tiền xử lý. Đây là nghĩa tiếng Việt của thuật ngữ Data Preprocessing - một thuật ngữ thuộc nhóm Technology Terms - Công nghệ thông tin.

Tiền xử lý dữ liệu là một kỹ thuật khai thác dữ liệu có liên quan đến chuyển dữ liệu thô thành một định dạng dễ hiểu. dữ liệu thực tế thường không đầy đủ, không nhất quán, và / hoặc thiếu trong hành vi hay xu hướng nhất định là, và có khả năng chứa nhiều lỗi. tiền xử lý dữ liệu là một phương pháp đã được chứng minh giải quyết vấn đề như vậy. tiền xử lý dữ liệu chuẩn bị dữ liệu thô để chế biến tiếp. tiền xử lý dữ liệu được sử dụng các ứng dụng cơ sở dữ liệu-driven như quản lý quan hệ khách hàng và các ứng dụng dựa trên luật lệ (như mạng thần kinh).

### **2.3.6 Mô đun selection**

Cung cấp các công cụ để chuẩn hóa và xử lý dữ liệu.

## 2.4 Mô hình máy học

### **2.4.1 Mô hình Xgboost**

XGBoost là viết tắt của Extreme Gradient Boosting. Đây là thuật toán state-of-the-art nhằm giải quyết bài toán supervised learning cho độ chính xác khá cao bên cạnh mô hình Deep learning như chúng ta từng tìm hiểu. Nếu Deep learning chỉ nhận đầu vào là raw data dạng numerical (ta thường phải chuyển đổi sang n-vector trong không gian số thực) thì XGBoost nhận đầu vào là tabular datasets với mọi kích thước và dạng dữ liệu bao gồm cả categorical mà dạng dữ liệu này thường được tìm thấy nhiều hơn trong business model, đây là lý do đầu tiên tại sao các cá nhân tham gia Kaggle thường sử dụng.

Bên cạnh đó, XGboost có tốc độ huấn luyện nhanh, có khả năng scale để tính toán song song trên nhiều server, có thể tăng tốc bằng cách sử dụng GPU, nhờ vậy mà Big Data không phải là vấn đề của mô hình này. Vì thế, XGBoost thường được sử dụng và đã giành được nhiều chiến thắng trong các cuộc thi tại Kaggle.

Mô hình XGBoost Regression sử dụng trong bài toán này có tác dụng học các mẫu từ dữ liệu huấn luyện, sau đó dự đoán kết quả cho các mẫu mới. Mô hình này có khả năng xử lý các biến đầu vào phức tạp và xử lý các tương tác giữa chúng để đưa ra dự đoán chính xác. Ngoài ra, mô hình XGBoost cũng có tính ổn định và khả năng điều chỉnh các siêu tham số để tối ưu hóa độ chính xác của mô hình.

Mô hình XGBoost được huấn luyện trên dữ liệu từ ngày 1/1/2016 đến ngày 31/3/2017 và sau đó được sử dụng để dự đoán lượng khách đến nhà hàng trong khoảng thời gian còn lại. Kết quả dự đoán được so sánh với dữ liệu thực tế từ ngày 1/4/2017 đến ngày 31/5/2017 để đánh giá độ chính xác của mô hình.

Vì vậy, dữ liệu dự đoán thành công sẽ là dữ liệu lượng khách đến nhà hàng trong khoảng thời gian từ ngày 1/4/2017 đến ngày 31/5/2017 được dự đoán bởi mô hình XGBoost và có độ chính xác đáng tin cậy.

## Ưu điểm

Ưu điểm của việc sử dụng thuật toán XGBoost để dự đoán lượng khách đến nhà hàng là:

* Hiệu quả cao: XGBoost là một trong những thuật toán dự đoán tốt nhất hiện nay, đặc biệt là với các bài toán có dữ liệu lớn và phức tạp.
* Khả năng xử lý dữ liệu thiếu: XGBoost có thể tự động xử lý các giá trị thiếu trong dữ liệu mà không cần phải thực hiện bước tiền xử lý riêng biệt.
* Độ ổn định cao: XGBoost có khả năng xử lý tốt các nhiễu trong dữ liệu và có khả năng chống overfitting, giúp mô hình đạt được độ ổn định cao.
* Đơn giản để sử dụng: XGBoost có API đơn giản và dễ hiểu, giúp người dùng có thể sử dụng và tinh chỉnh mô hình một cách dễ dàng.

## Nhược điểm

Việc sử dụng XGBoost cũng có một số nhược điểm như:

* Thời gian huấn luyện lâu: XGBoost có thể yêu cầu thời gian huấn luyện lâu hơn so với các thuật toán khác khi xử lý các tập dữ liệu lớn hoặc phức tạp.
* Khả năng ứng dụng có hạn: XGBoost không phải là một thuật toán đa nhiệm và chỉ có thể áp dụng cho các bài toán dự đoán với đầu vào là các đặc trưng dạng số.
* Độ chính xác phụ thuộc vào các tham số: Độ chính xác của mô hình XGBoost phụ thuộc vào các tham số được chọn cho mô hình, việc chọn sai tham số có thể dẫn đến việc mô hình không hoạt động hiệu quả.

Tuy nhiên, với tập dữ liệu về lượng khách đến nhà hàng cụ thể, thuật toán XGBoost đã cho kết quả dự đoán chính xác và độ chính xác cao. Tuy thời gian huấn luyện không nhanh lắm nhưng có thể chấp nhận được với tập dữ liệu này. Do đó, XGBoost có thể là một lựa chọn hiệu quả để dự đoán lượng khách đến nhà hàng trong trường hợp này.

### **2.4.2 Hồi quy tuyến tính là gì?**

Hồi quy tuyến tính là một kỹ thuật phân tích dữ liệu dự đoán giá trị của dữ liệu không xác định bằng cách sử dụng một giá trị dữ liệu liên quan và đã biết khác. Nó mô hình toán học biến không xác định hoặc phụ thuộc và biến đã biết hoặc độc lập như một phương trình tuyến tính. Ví dụ, giả sử rằng bạn có dữ liệu về chi phí và thu nhập của bạn trong năm ngoái. Kỹ thuật hồi quy tuyến tính phân tích dữ liệu này và xác định rằng chi phí của bạn là một nửa thu nhập của bạn. Sau đó, họ tính toán một chi phí trong tương lai không rõ bằng cách giảm một nửa thu nhập được biết đến trong tương lai.

### **2.4.3 Tại sao hồi quy tuyến tính lại quan trọng?**

Các mô hình hồi quy tuyến tính tương đối đơn giản và cung cấp một công thức toán học dễ giải thích để đưa ra các dự đoán. Hồi quy tuyến tính là một kỹ thuật thống kê được sử dụng từ lâu và áp dụng dễ dàng cho phần mềm và tính toán. Các doanh nghiệp sử dụng nó để chuyển đổi dữ liệu thô một cách đáng tin cậy và có thể dự đoán được thành nghiệp vụ thông minh và thông tin chuyên sâu hữu ích. Các nhà khoa học trong nhiều lĩnh vực, bao gồm sinh học và các ngành khoa học hành vi, môi trường, và xã hội, sử dụng hồi quy tuyến tính để tiến hành phân tích dữ liệu sơ bộ và dự đoán các xu hướng tương lai. Nhiều phương pháp khoa học dữ liệu, chẳng hạn như máy học và trí tuệ nhân tạo, sử dụng hồi quy tuyến tính để giải quyết các bài toán phức tạp.

### **2.4.4 Hồi quy tuyến tính hoạt động như thế nào?**

Về bản chất, một kỹ thuật hồi quy tuyến tính đơn giản cố gắng vẽ một đồ thị đường giữa hai biến dữ liệu, x và y. Là biến độc lập, x được vẽ dọc theo trục hoành. Các biến độc lập còn được gọi là biến giải thích hoặc biến dự báo. Biến phụ thuộc, y, được vẽ trên trục tung. Bạn cũng có thể tham chiếu các giá trị y như các biến phản hồi hoặc các biến dự báo.

### **2.4.5 Hồi quy tuyến tính được áp dụng như thế nào trước khi đưa vào mô hình XGBoost?**

Mô hình hồi quy tuyến tính có tác dụng tìm mối quan hệ tuyến tính giữa biến đầu vào và biến mục tiêu, trong trường hợp này là lượng khách đến nhà hàng. Từ đó, mô hình sẽ tìm ra các trọng số phù hợp cho các biến đầu vào để đưa ra dự đoán chính xác cho biến mục tiêu.

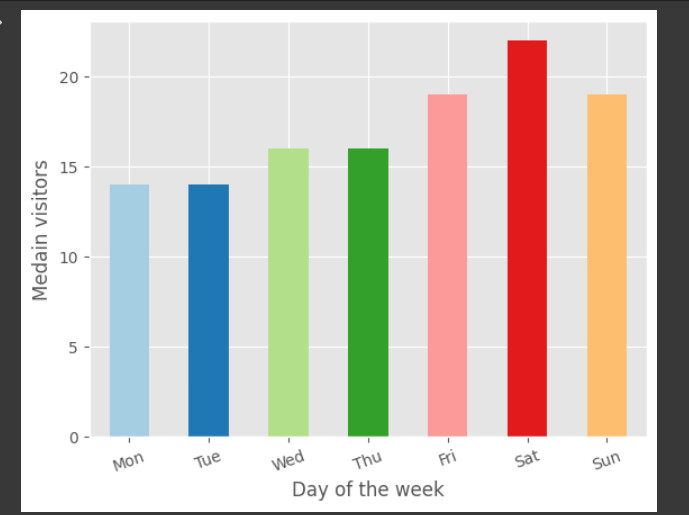
Việc sử dụng mô hình hồi quy tuyến tính để xử lý dữ liệu trước khi đưa vào mô hình XGBoost giúp cải thiện độ chính xác của mô hình dự đoán, bởi vì việc tìm mối quan hệ tuyến tính giữa các biến đầu vào có thể giúp loại bỏ các yếu tố không quan trọng và giảm thiểu sự phân tán của dữ liệu. Từ đó, mô hình XGBoost sẽ dễ dàng học được mối quan hệ giữa các biến và đưa ra dự đoán chính xác cho lượng khách đến nhà hàng ở Nhật Bản.

# CHƯƠNG 3.XÂY DỰNG DỮ LIỆU

## 3.1 Biểu đồ trực quan

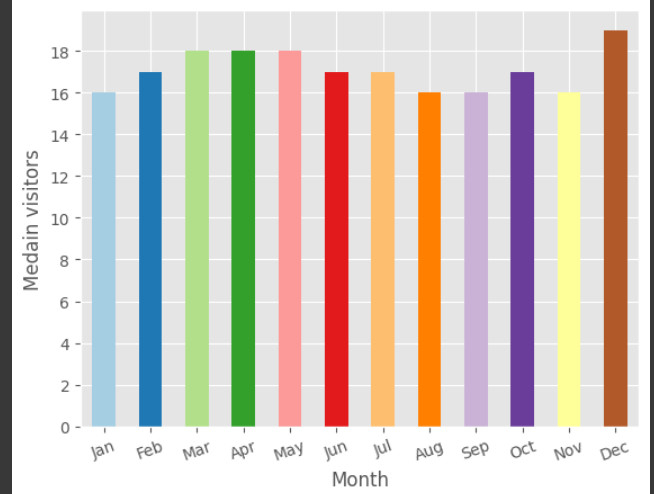
### **3.1.1 Biểu đồ 1. Số lượng khách hành đến nhà hàng theo từng ngày trong tuần**

* Sử dụng để tạo biểu đồ cột (bar chart) thể hiện số lượng khách hàng đến một nhà hàng theo từng ngày trong tuần. Biểu đồ này giúp cho người dùng có thể xem được các ngày trong tuần mà số lượng khách hàng đến nhà hàng nhiều nhất hoặc ít nhất.

****

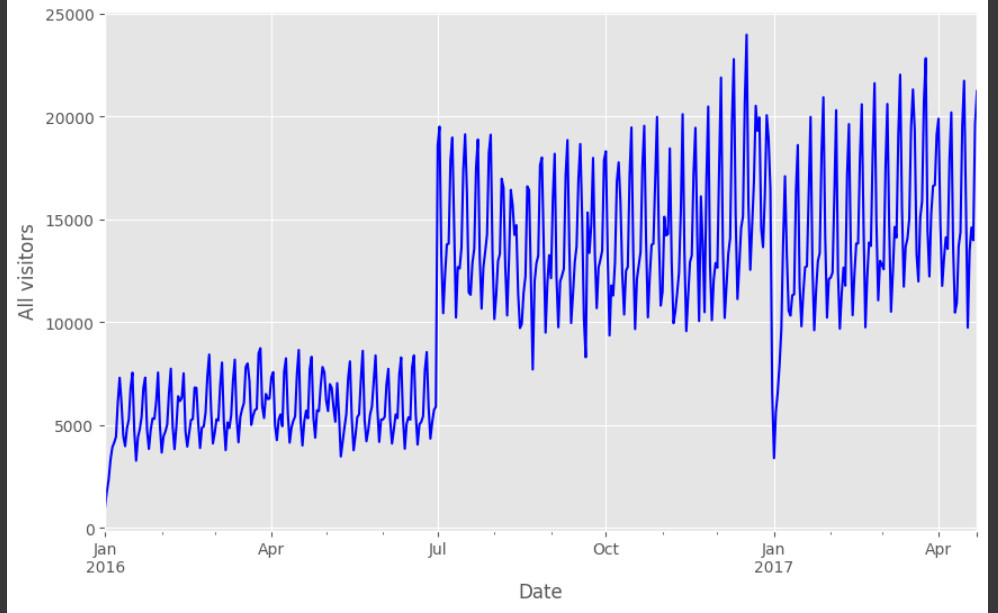
### **3.1.2Biểu đồ 2. biểu thị giá trị trung vị của số lượng khách du lịch (visitors) vào các ngày trong tháng.**

* Trực quan hóa dữ liệu trong tập tin air\_visit\_data bằng cách tạo một biểu đồ cột dọc để biểu thị giá trị trung vị của số lượng khách du lịch (visitors) vào các ngày trong tháng.

****

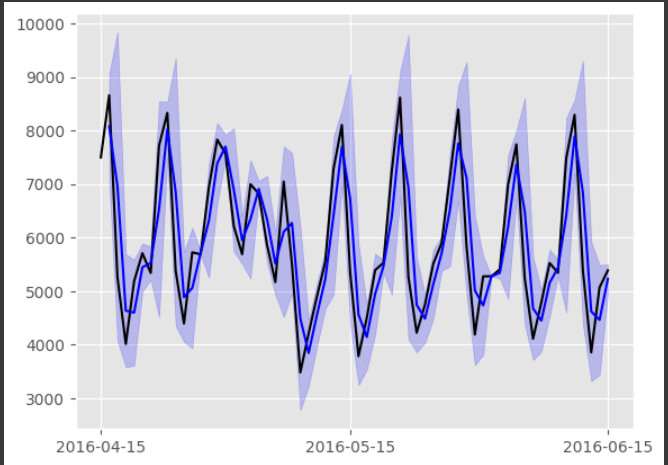
### **3.1.3 Biểu đồ 3. biểu thị tổng số lượng khách du lịch(visitors)**

* Trực quan hóa dữ liệu trong tập tin air\_visit\_data bằng cách tạo một biểu đồ đường để biểu thị tổng số lượng khách du lịch (visitors)

****

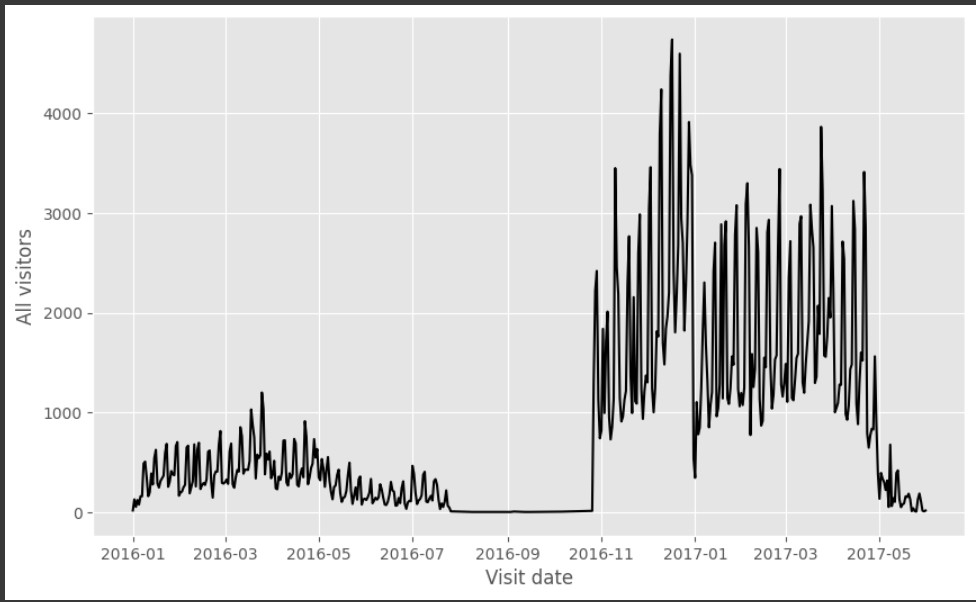
### **3.1.4 Biểu đồ 4. Số lượng khách thăm quan vào các ngày trong khoảng thời gian từ ngày 15/4/2016 đến ngày 15/6/2016.**

* Trực quan hóa dữ liệu số lượng khách thăm quan vào các ngày trong khoảng thời gian từ ngày 15/4/2016 đến ngày 15/6/2016. Xữ lý nhóm dữ liệu trong tập tin air\_visit\_data theo ngày thăm quan (visit\_date) và tính tổng số lượng khách du lịch trong mỗi ngày

****

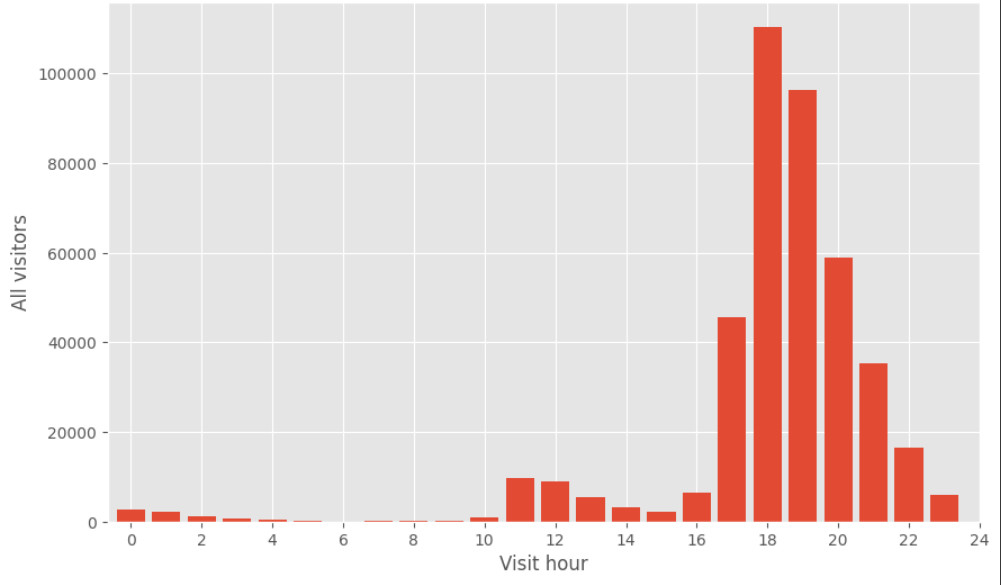
### **3.1.5 Biểu đồ 5. Biểu thị tổng số lượng khách đặt chỗ (reserve visitors) vào các ngày khác nhau.**

* Trực quan hóa dữ liệu trong tập tin air\_reserve.csv bằng cách tạo một biểu đồ đường để biểu thị tổng số lượng khách đặt chỗ (reserve visitors) vào các ngày khác nhau.

****

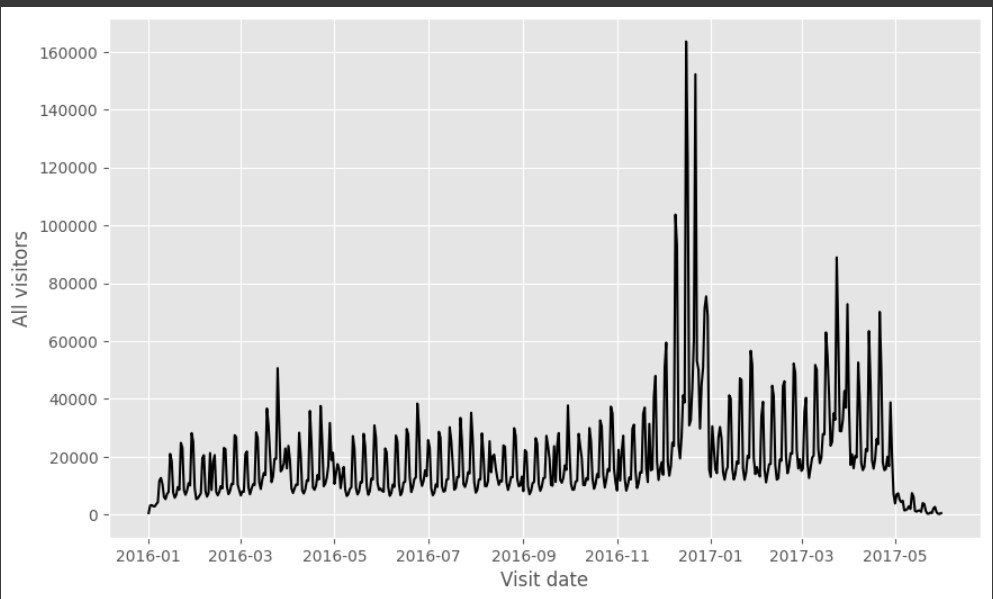
### **3.1.6 Biểu đồ 6. biểu thị tổng số lượng khách đặt chỗ (reserve visitors) vào mỗi giờ của ngày.**

* Trực quan hóa dữ liệu trong tập tin air\_reserve bằng cách tạo một biểu đồ cột dọc để biểu thị tổng số lượng khách đặt chỗ (reserve visitors) vào mỗi giờ của ngày.

****

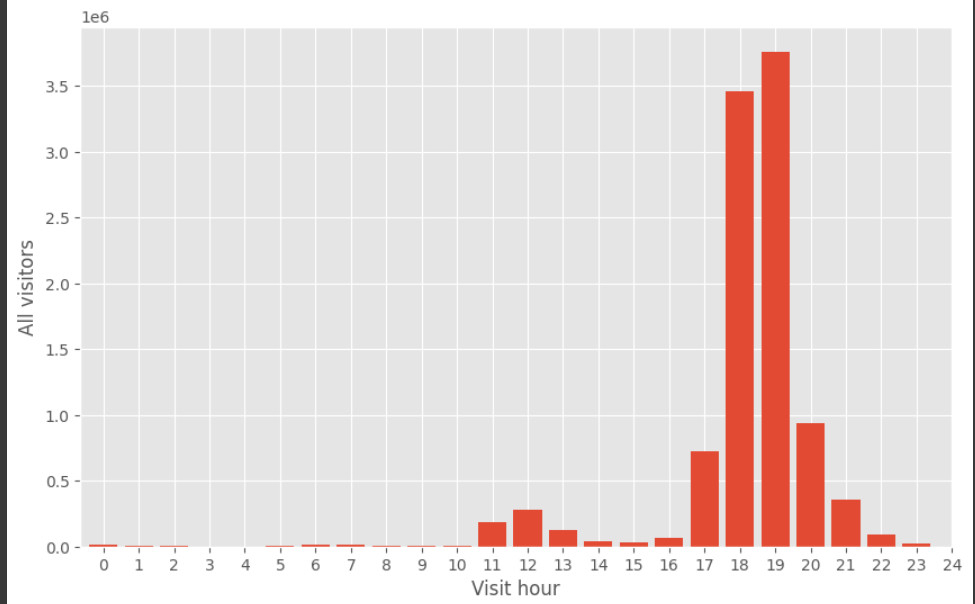
### **3.1.7 Biểu đồ 7. Biểu đồ đường để biểu thị tổng số lượng khách (all visitors) đặt bàn trong từng ngày.**

* Trực quan hóa dữ liệu trong tập tin hpg\_reserve bằng cách tạo một biểu đồ đường để biểu thị tổng số lượng khách (all visitors) đặt bàn trong từng ngày.

****

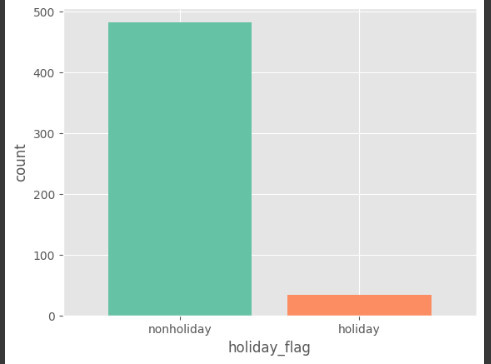
### **3.1.8 Biểu đồ 8. Tổng số lượng khách đặt chỗ (reserve visitors) trong từng giờ trong ngày.**

* Biểu thị tổng số lượng khách đặt chỗ (reserve visitors) trong từng giờ trong ngày (visit hour), dựa trên dữ liệu trong tập tin hpg\_reserve.csv

****

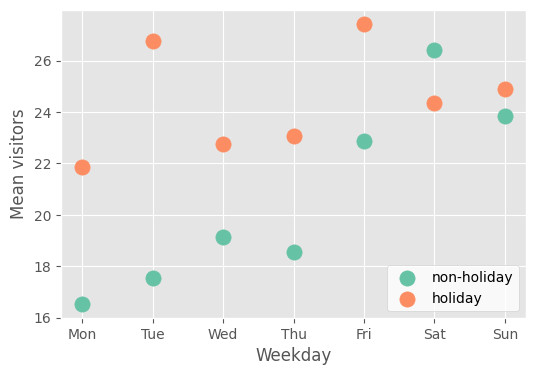
### **3.1.9Biểu đồ 9. Để hiển thị số lượng ngày là ngày lễ.**

* Trực quan hóa dữ liệu trong tập tin date\_info bằng cách tạo một biểu đồ cột dọc, để hiển thị số lượng ngày là ngày lễ.

****

### **3.1.10 Biểu đồ 10. Phân tán để biểu thị giá trị trung bình số lượng khách du lịch (visitors) vào các ngày trong tuần và trong các ngày lễ.**

* Trực quan hóa dữ liệu trong tập tin air\_visit bằng cách tạo một biểu đồ phân tán để biểu thị giá trị trung bình số lượng khách du lịch (visitors) vào các ngày trong tuần và trong các ngày lễ (holiday\_flg).

****

# CHƯƠNG 4. KẾT LUẬN

## 4.1 Kết quả đạt được

Trong nghiên cứu dự đoán lượng khách hàng đến nhà hàng Nhật Bản rất khả thi , giúp chủ nhà hàng biết được số lượng khách hàng đến tháng có lượng khách nhiều nhất . Tuy nhiên còn nhiều điều hạn chế .

## 4.2 Định hướng tương lai

Để định hướng tương lai dự đoán lượng khách đến nhà hàng Nhật Bản thông qua khoa học dữ liệu, có thể áp dụng các phương pháp và công cụ như sau:

* Thu thập dữ liệu: Thu thập dữ liệu về số lượng khách đến nhà hàng từ các nguồn khác nhau như các hệ thống đặt chỗ trực tuyến, thẻ thành viên, hệ thống thanh toán, đo lường lưu lượng khách qua cổng,v.v...
* Xử lý dữ liệu: Xử lý dữ liệu thu thập được bằng các công cụ và phương pháp như xử lý dữ liệu bị thiếu, chuẩn hóa dữ liệu, rút trích đặc trưng, phân tích thống kê, Machine Learning, AI,v.v…
* Xây dựng mô hình dự đoán: Dựa trên dữ liệu đã xử lý, có thể áp dụng các phương pháp như hồi quy tuyến tính, Random Forest, Neural Network, Gradient Boosting,... để xây dựng mô hình dự đoán lượng khách đến nhà hàng.
* Đánh giá và cải thiện mô hình: Đánh giá mô hình dự đoán bằng các phương pháp như Cross-validation, Confusion matrix, ROC curve,.. để cải thiện chất lượng của mô hình và đưa ra các dự đoán chính xác hơn.
* Triển khai và giám sát mô hình: Sau khi xây dựng và đánh giá mô hình, có thể triển khai mô hình vào sản phẩm thực tế và giám sát kết quả dự đoán để đưa ra các cải tiến và tối ưu hóa mô hình.

# Tài liệu tham thảo

1. <https://github.com/pai4451/Recruit-Restaurant-Visitor-Forecasting>

2. <https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/data>

3. <https://www.toponseek.com/blogs/google-colab>

4. <https://vn.got-it.ai/blog/tong-quan-thu-vien-numpy-trong-python>

5.https://t3h.edu.vn/tin-tuc/pandas-trong-python-la-gi-gioi-thieu-chi-tiet-ve-pandas-cho-nguoi-moi-bat-dau.

6. <https://en.wikipedia.org/wiki/Scikit-learn>

7. <https://filegi.com/tech-term/data-preprocessing-4357/>

8.<https://ongxuanhong.wordpress.com/2017/12/21/xgboost-thuat-toan-gianh-chien-thang-tai-nhieu-cuoc-thi-kaggle/>