TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

**KHOA ĐIỆN TỬ**

**Bộ môn: Tin Học Công Nghiệp**

**BÀI TẬP LỚN**

MÔN HỌC

**KHOA HỌC DỮ LIỆU**

Sinh viên: Nguyễn Đức Trung Kiên

MSSV: K175520216025

Lớp: K55KMT.01

GVHD: Nguyễn Văn Huy

**Thái Nguyên – 2023**

**BÀI TẬP LỚN**

**MÔN HỌC : KHOA HỌC DỮ LIỆU**

BỘ MÔN: TIN HỌC CÔNG NGHIỆP

*Sinh viên: Nguyễn Đức Trung Kiên Lớp: K55KMT*

*Ngành: Kỹ thuật máy tính*

*Ngày giao đề: Ngày hoàn thành: 10/05/2023*

1.Tên đề tài :

.

2. Nội dung thuyết minh tính toán

3. Các bản vẽ, chương trình và đồ thị

|  |  |
| --- | --- |
| TỔ TRƯỞNG BỘ MÔN | GIÁO VIÊN HƯỚNG DẪN |
| *(Ký và ghi rõ họ tên)* | *(Ký và ghi rõ họ tên)* |

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

### Thái Nguyên, ngày….tháng…..năm 20....

## **GIÁO VIÊN HƯỚNG DẪN**

*(Ký ghi rõ họ tên)*

**MỤC LỤC**

[LỜI NÓI ĐẦU 4](#_Toc136413901)

[CHƯƠNG 1: GIỚI THIỆU 5](#_Toc136413906)

1.1 [Khái niệm? 5](#_Toc136413907)

1.2 [Ứng dụng 5](#_Toc136413911)

1.3 [Công nghệ 6](#_Toc136413918)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 9](#_Toc136413932)

2.1 [Model 9](#_Toc136413933)

[CHƯƠNG 3. XÂY DỰNG CHƯƠNG TRÌNH 10](#_Toc136413941)

3.1 [Code: 10](#_Toc136413942)

[CHƯƠNG 4. KẾT LUẬN 15](#_Toc136413943)

4.1 [Kết quả đạt được 15](#_Toc136413944)

4.2 [Mặt tốt và xấu của dự án 18](#_Toc136413945)

4.3 [Ứng dụng 20](#_Toc136413957)

# 

# **LỜI NÓI ĐẦU**

## Trong thời đại hiện nay, ngành công nghiệp vận tải và du lịch phát triển mạnh mẽ và ngày càng trở thành một phần quan trọng của cuộc sống hàng ngày. Với sự tăng trưởng đáng kể của hệ thống hàng không, việc dự đoán giá vé máy bay trở thành một yếu tố quan trọng trong quyết định của hành khách khi lựa chọn chuyến bay.

## Đề tài "Dự đoán giá vé máy bay" nhằm áp dụng các phương pháp và công cụ trong lĩnh vực Khoa học dữ liệu để xây dựng một mô hình dự đoán chính xác giá vé máy bay. Mục tiêu của đề tài này là cung cấp một phương pháp hữu ích và hiệu quả cho các công ty hàng không và hành khách trong việc dự đoán và định giá giá vé máy bay.

## Trong quá trình nghiên cứu, chúng tôi đã thu thập và xử lý một lượng lớn dữ liệu về giá vé máy bay từ các nguồn đáng tin cậy như trang web đặt vé, hãng hàng không và nhà cung cấp dữ liệu về giá vé. Dữ liệu này chứa thông tin chi tiết về các yếu tố quan trọng như ngày bay, địa điểm xuất phát và đến, loại vé, hãng hàng không và các yêu cầu đặc biệt từ khách hàng.

## Việc dự đoán giá vé máy bay không chỉ mang lại lợi ích cho người dùng, mà còn có thể giúp các hãng hàng không tối ưu hóa quản lý nguồn lực và định hình chiến lược kinh doanh. Điều này đòi hỏi sự áp dụng của các phương pháp và kỹ thuật phân tích dữ liệu để xác định các yếu tố ảnh hưởng và xây dựng mô hình dự đoán giá vé máy bay.

## **CHƯƠNG 1: GIỚI THIỆU**

## **1.1 Khái niệm?**

## Dự đoán giá vé máy bay là quá trình ước tính giá trị tiền vé máy bay trong tương lai dựa trên các yếu tố như sự biến động của thị trường hàng không, nhu cầu của hành khách, mùa cao điểm, sự cạnh tranh giữa các hãng hàng không, và các yếu tố kinh tế khác.

## Việc dự đoán giá vé máy bay thường được thực hiện bởi các hãng hàng không, các công ty nghiên cứu thị trường, hoặc các công ty công nghệ thông tin trong ngành hàng không. Các mô hình dự đoán và thuật toán được sử dụng để phân tích dữ liệu lịch sử, xu hướng giá cả và các yếu tố khác để đưa ra dự báo về giá vé trong tương lai.

## Dự đoán giá vé máy bay có thể giúp hành khách lựa chọn thời điểm mua vé với giá tốt nhất. Ngoài ra, các công cụ trực tuyến cũng cung cấp dịch vụ so sánh giá vé và cung cấp thông tin về xu hướng giá, để hành khách có thể tìm hiểu và lựa chọn giá vé phù hợp.

## **1.2 Ứng dụng**

## Ứng dụng dự đoán giá vé máy bay mang đến nhiều lợi ích và cung cấp thông tin quan trọng cho cả hành khách và ngành hàng không. Dưới đây là một ví dụ về cách ứng dụng này có thể được triển khai:

## Hãy tưởng tượng một ứng dụng di động thông minh với chức năng dự đoán giá vé máy bay. Đây là một công cụ hữu ích cho hành khách khi họ đang tìm kiếm và đặt vé máy bay. Ứng dụng này có thể cung cấp các tính năng sau:

## Dự báo giá vé: Ứng dụng sẽ sử dụng các mô hình dự đoán và thuật toán để phân tích dữ liệu lịch sử và các yếu tố ảnh hưởng, như mùa cao điểm và biến động thị trường. Dựa trên thông tin này, ứng dụng sẽ đưa ra dự báo về giá vé trong tương lai. Hành khách có thể sử dụng dự đoán này để lựa chọn thời điểm mua vé với giá tốt nhất.

## Gợi ý thời điểm mua vé: Ứng dụng sẽ theo dõi xu hướng giá vé và cung cấp gợi ý về thời điểm mua vé hợp lý. Ví dụ, nếu giá vé dự kiến sẽ giảm trong một vài ngày tới, ứng dụng sẽ thông báo cho hành khách chờ đợi để mua vé với giá tốt hơn. Điều này giúp hành khách tiết kiệm chi phí và tận dụng các ưu đãi giá vé.

## Thông tin về biến động giá: Ứng dụng sẽ cung cấp biểu đồ và thông tin chi tiết về biến động giá vé. Hành khách có thể theo dõi sự thay đổi giá trong quá khứ và hiện tại để có cái nhìn tổng quan về xu hướng giá cả. Điều này giúp hành khách đưa ra quyết định mua vé dựa trên thông tin được cung cấp một cách rõ ràng.

## Thông báo giá vé theo yêu cầu: Hành khách có thể đăng ký thông báo từ ứng dụng để nhận thông tin cập nhật về giá vé. Khi giá vé đạt đến mức mong muốn hoặc có sự thay đổi đáng chú ý, hành khách s

## **1.3 Công nghệ**

## **a, Các công nghệ có thể sử dụng**

## Để dự đoán giá vé máy bay, ta có thể sử dụng các phương pháp và công nghệ trong lĩnh vực Khoa học Dữ liệu và Machine Learning. Dưới đây là một số công nghệ phổ biến:

## Regression Models (Mô hình hồi quy): Sử dụng các mô hình như Linear Regression, Polynomial Regression, Support Vector Regression (SVR) để dự đoán giá vé dựa trên các biến độc lập.

## Random Forest và Decision Trees (Cây quyết định): Áp dụng các mô hình cây quyết định hoặc kết hợp cây quyết định thành các rừng ngẫu nhiên (Random Forest) để xây dựng các quy tắc dự đoán giá vé máy bay.

## Neural Networks (Mạng nơ-ron): Sử dụng các mô hình mạng nơ-ron sâu (Deep Neural Networks) hoặc kiến trúc RNNs và LSTM để dự đoán giá vé máy bay dựa trên các đặc trưng và thông tin liên quan.

## Ensemble Learning (Học kết hợp): Kết hợp nhiều mô hình dự đoán khác nhau, ví dụ như Bagging, Boosting hoặc Stacking, để cải thiện độ chính xác và ổn định của dự đoán giá vé.

## Xử lý dữ liệu và Feature Engineering: Sử dụng các kỹ thuật xử lý dữ liệu bị thiếu, mã hóa biến định danh, trích xuất đặc trưng từ dữ liệu ngày tháng, v.v. để tạo ra các biến độc lập tốt nhất cho mô hình dự đoán giá vé máy bay.

## Gradient Boosting: Áp dụng các mô hình như Gradient Boosting Regression (GBM), XGBoost hoặc LightGBM để xây dựng các mô hình tối ưu hóa dự đoán giá vé máy bay.

## Support Vector Machines (SVM): Sử dụng SVM để phân loại và dự đoán giá vé máy bay dựa trên các đặc trưng.

## Xử lý dữ liệu không gian và thời gian: Áp dụng các kỹ thuật xử lý dữ liệu không gian (Spatial Data Processing) và thời gian (Time Series Analysis) để phân tích và dự đoán giá vé máy bay dựa trên vị trí địa lý và xu hướng thời gian.

**b, Công nghệ được áp dụng**

## Trong các công nghệ được liệt kê, tôi sử dụng mô hình hồi quy tuyến tính (Linear Regression). Linear Regression là một mô hình đơn giản và phổ biến trong việc dự đoán giá vé máy bay. Nó dựa trên giả định rằng giá vé có mối quan hệ tuyến tính với các biến độc lập như khoảng cách, thời gian bay, số lượng ghế trống, v.v.

## Để sử dụng Linear Regression, ta chỉ cần chuẩn bị dữ liệu đầu vào, xây dựng mô hình và đào tạo nó trên dữ liệu huấn luyện. Sau đó, bạn có thể sử dụng mô hình đã đào tạo để dự đoán giá vé cho các dữ liệu mới.

## Mô hình Linear Regression dễ hiểu và thực hiện, không yêu cầu nhiều xử lý dữ liệu phức tạp hay điều chỉnh siêu tham số như các mô hình phức tạp hơn. Tuy nhiên, đôi khi Linear Regression có thể không đạt được độ chính xác cao khi mối quan hệ giữa giá vé và các biến độc lập không phải là tuyến tính hoặc có sự tương quan phức tạp.

## **CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

## **2.1 Model**

## - Sử dụng dạng hồi quy tuyến tính được ký hiệu là "Linear Regression". - Sử dụng mô hình Linear Regression (Hồi quy tuyến tính) như sau:

## Đầu tiên, tôi chuẩn bị dữ liệu bằng cách tách các đặc trưng (features) và biến mục tiêu (target) từ dataframe. Các đặc trưng được lưu trong biến x, còn biến mục tiêu được lưu trong biến y.

## Sau đó, tôi chia dữ liệu thành hai phần: tập huấn luyện (train set) và tập kiểm tra (test set) bằng hàm train\_test\_split từ module sklearn.model\_selection. Điều này giúp bạn có một phần dữ liệu để huấn luyện mô hình và một phần dữ liệu để kiểm tra mô hình sau khi đã huấn luyện.

## Tiếp theo, tôi khởi tạo mô hình Linear Regression bằng cách sử dụng class LinearRegression từ module sklearn.linear\_model.

## Tôi sử dụng phương thức fit trên đối tượng mô hình để huấn luyện mô hình trên tập huấn luyện. Việc này giúp mô hình học cách ánh xạ các đặc trưng sang biến mục tiêu.

## Sau khi mô hình đã được huấn luyện, tôi sử dụng phương thức predict trên đối tượng mô hình để dự đoán giá trị biến mục tiêu trên tập kiểm tra.

## Cuối cùng, tôi sử dụng hàm r2\_score từ module sklearn.metrics để đánh giá hiệu suất của mô hình dựa trên giá trị dự đoán và giá trị thực tế trên tập kiểm tra. Giá trị R-squared score càng gần 1 thì mô hình càng tốt.

## **CHƯƠNG 3. XÂY DỰNG CHƯƠNG TRÌNH**

## **3.1 Code:**

**# import các thư viện và module cần thiết trong Python**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import r2\_score

**# Đọc dữ liệu từ tập tin CSV**

df = pd.read\_csv("C:\\Users\\6\\PycharmProjects\\baitaplon\\Clean\_Dataset.csv")

**# Kiểm tra kích thước của dataframe**

print("Kích thước của dataframe:", df.shape)

**# Kiểm tra các giá trị thiếu trong dataframe**

print("Số lượng giá trị thiếu trong từng cột:")

print(df.isna().sum())

**# Vẽ biểu đồ heatmap để hiển thị các giá trị thiếu**

plt.figure(figsize=(10, 6))

sns.heatmap(df.isnull(), yticklabels=False, cmap="viridis")

plt.title("Các giá trị thiếu trong dataframe")

**# Xóa các cột không cần thiết**

df.drop(["Unnamed: 0", "flight"], axis=1, inplace=True)

**# Hiển thị số lượng chuyến bay của từng hãng hàng không**

plt.figure(figsize=(10, 6))

sns.countplot(x="airline", data=df)

plt.title("Số lượng chuyến bay của từng hãng hàng không")

**# Vẽ biểu đồ barplot hiển thị giá vé theo từng hãng hàng không**

plt.figure(figsize=(10, 6))

sns.barplot(x="airline", y="price", data=df)

plt.title("Giá vé theo từng hãng hàng không")

**# Chuyển đổi biến hạng mục thành biến số bằng phương pháp One-Hot Encoding**

df\_encoded = pd.get\_dummies(df)

**# Loại bỏ các cột không cần thiết**

cols\_to\_drop = ["airline\_AirAsia", "source\_city\_Bangalore", "departure\_time\_Afternoon", "stops\_one", "arrival\_time\_Afternoon", "destination\_city\_Bangalore", "class\_Business"]

df\_encoded.drop(cols\_to\_drop, axis=1, inplace=True)

**# Chia thành input (x) và output (y)**

x = df\_encoded.drop(["price"], axis=1).values

y = df\_encoded["price"].values

**# Chuẩn hóa dữ liệu**

scaler = StandardScaler()

x\_scaled = scaler.fit\_transform(x)

**# Chia dữ liệu thành tập train và tập test**

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x\_scaled, y, test\_size=0.2, random\_state=0)

**# Xây dựng mô hình hồi quy tuyến tính**

lr = LinearRegression()

lr.fit(x\_train, y\_train)

**# Dự đoán giá vé trên tập test**

y\_pred = lr.predict(x\_test)

**# Đánh giá mô hình sử dụng R-squared score**

r2 = r2\_score(y\_test, y\_pred)

print("R-squared score:", r2)

**# Vẽ biểu đồ so sánh giá vé dự đoán và giá vé thực tế trên tập test**

plt.figure(figsize=(10, 6))

plt.scatter(y\_test, y\_pred, label='Dự đoán')

plt.plot([y\_test.min(), y\_test.max()], [y\_test.min(), y\_test.max()], 'r--', label='Tham chiếu')

plt.xlabel('Giá vé thực tế')

plt.ylabel('Giá vé dự đoán')

plt.title('So sánh giá vé thực tế và giá vé dự đoán trên tập test')

plt.legend()

**# Hiển thị biểu đồ**

plt.show()

**# Vẽ biểu đồ so sánh giá vé dự đoán và giá vé thực tế trên tập test**

plt.figure(figsize=(10, 6))

plt.scatter(y\_test, y\_pred, label='Dự đoán')

plt.plot([y\_test.min(), y\_test.max()], [y\_test.min(), y\_test.max()], 'r--', label='Tham chiếu')

plt.xlabel('Giá vé thực tế')

plt.ylabel('Giá vé dự đoán')

plt.title('So sánh giá vé thực tế và giá vé dự đoán trên tập test')

plt.legend()

**# Hiển thị giá vé dự đoán**

for i in range(len(y\_pred)):

plt.text(y\_test[i], y\_pred[i], str(round(y\_pred[i], 2)), ha='center')

**# Hiển thị biểu đồ**

plt.show()

**# Tạo danh sách giá vé dự đoán và giá vé thực tế tương ứng**

predicted\_prices = []

actual\_prices = []

for i in range(len(y\_pred)):

predicted\_prices.append(round(y\_pred[i], 2))

actual\_prices.append(y\_test[i])

**# Hiển thị danh sách giá vé dự đoán và giá vé thực tế**

print("Danh sách giá vé:")

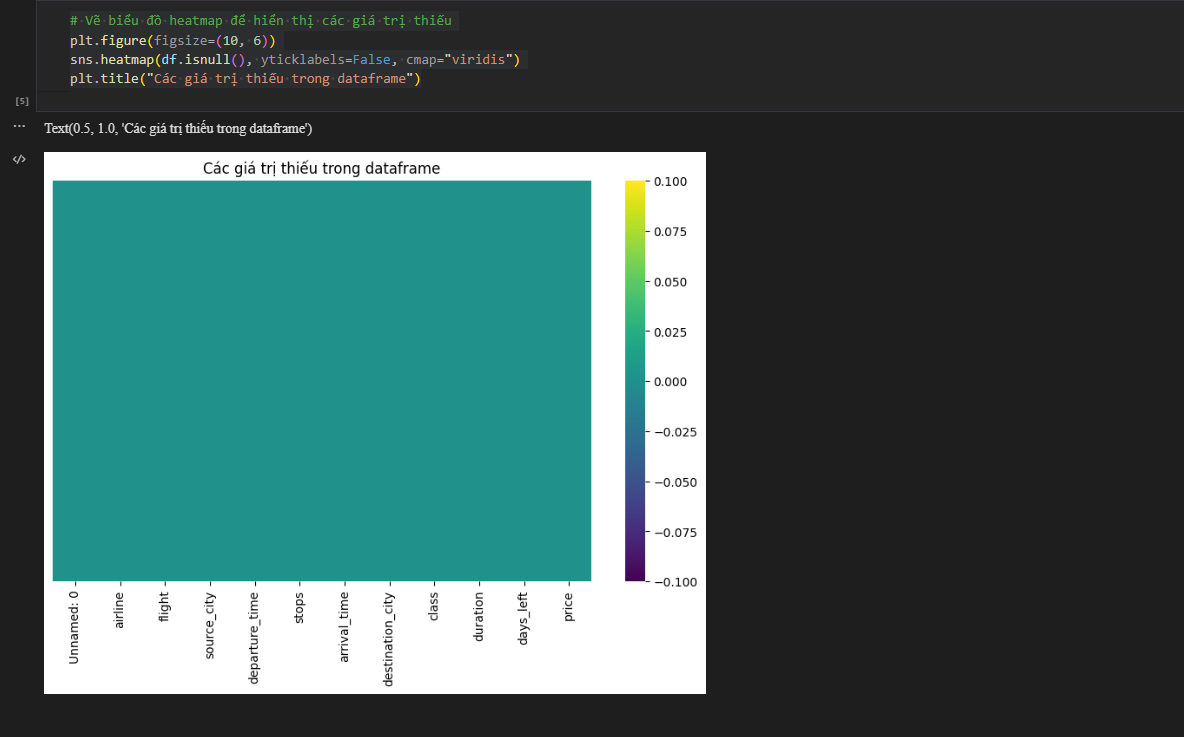
for i in range(len(predicted\_prices)):

print(f"Giá vé {i+1}: Dự đoán: {predicted\_prices[i]}, Thực tế: {actual\_prices[i]}")

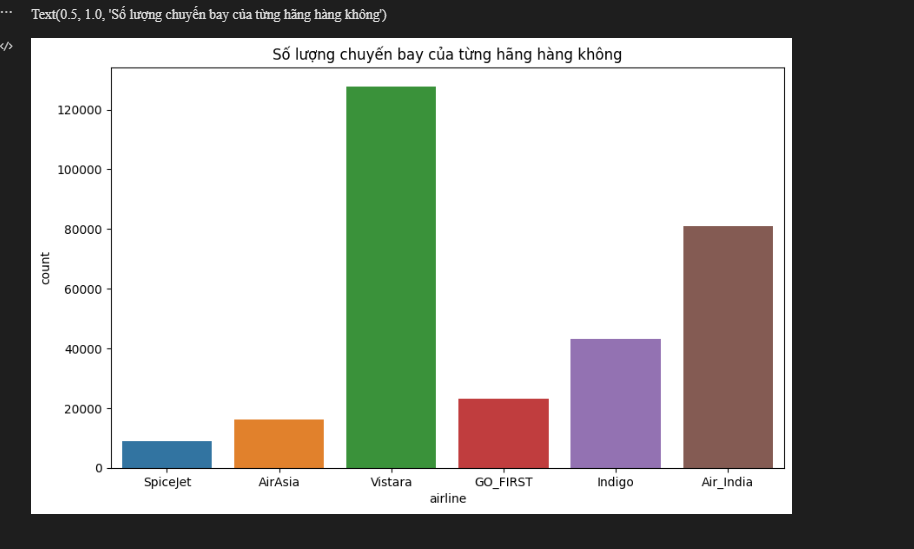
## **CHƯƠNG 4. KẾT LUẬN**

## **4.1 Kết quả đạt được**

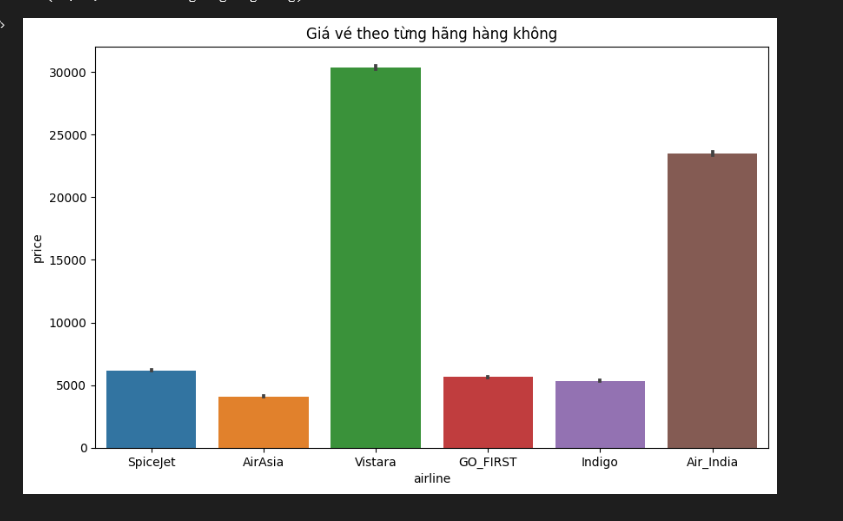
* Biểu đồ heatmap để hiển thị các giá trị thiếu



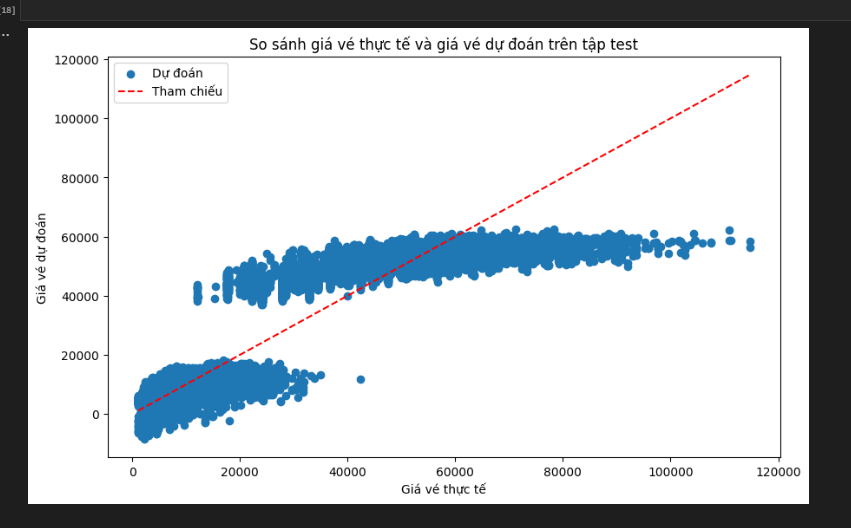
* Hiển thị số lượng chuyến bay của từng hãng hàng không



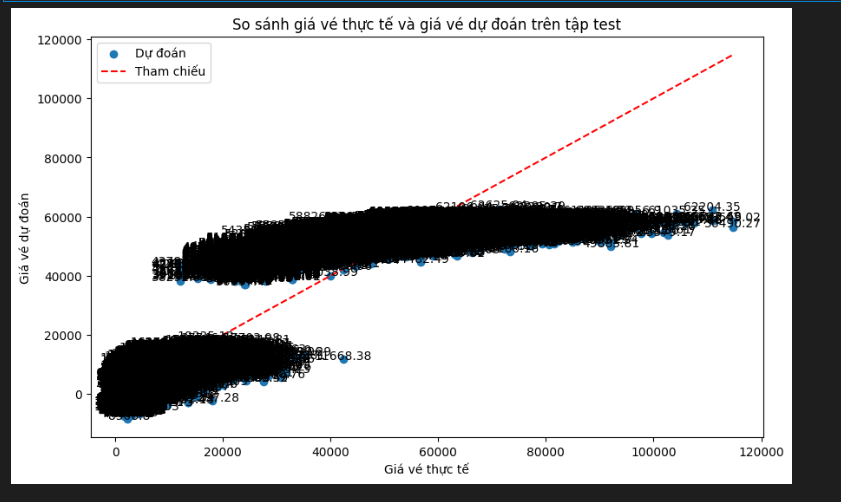
* Hiển thị biểu đồ barplot hiển thị giá vé theo từng hãng hàng không



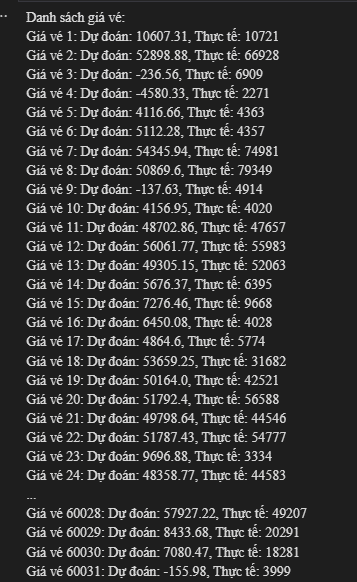
* Biểu đồ so sánh giá vé dự đoán và giá vé thực tế trên tập test



* Hình ảnh đồ so sánh giá vé dự đoán và giá vé thực tế trên tập test



* # Hiển thị danh sách giá vé dự đoán và giá vé thực tế



## **4.2 Mặt tốt và xấu của dự án**

Mặt tốt:

## Sử dụng các thư viện và module phổ biến trong Python để xử lý dữ liệu và xây dựng mô hình.

## Đọc dữ liệu từ tập tin CSV và kiểm tra kích thước của dataframe.

## Kiểm tra và hiển thị các giá trị thiếu trong dataframe.

## Loại bỏ các cột không cần thiết trong dataframe.

## Hiển thị số lượng chuyến bay của từng hãng hàng không.

## Vẽ biểu đồ barplot để hiển thị giá vé theo từng hãng hàng không.

## Chuẩn hóa dữ liệu để đảm bảo các đặc trưng có cùng thang đo.

## Chia dữ liệu thành tập train và tập test để đánh giá mô hình.

## Xây dựng mô hình hồi quy tuyến tính và dự đoán giá vé.

## Mặt xấu:

## 1. Không xử lý các giá trị thiếu trong dataframe mà chỉ hiển thị chúng bằng biểu đồ heatmap. 2. Việc loại bỏ các cột không cần thiết trong dataframe được thực hiện theo danh sách cố định, không có phân tích cụ thể về tác động của các cột này đến mô hình. 3. Không có một bước rõ ràng để đánh giá và so sánh các mô hình khác nhau. 4. Việc dự đoán giá vé chỉ dựa trên mô hình hồi quy tuyến tính, không khám phá các phương pháp dự đoán khác.

## **4.3 Ứng dụng**

## Ứng dụng của đề tài "dự đoán giá vé máy bay" có thể làm được nhiều việc hữu ích, bao gồm:

## **Dự đoán giá vé máy bay**: Thuật toán được xây dựng có thể dự đoán giá vé máy bay dựa trên các thông tin như hãng hàng không, ngày bay, điểm xuất phát và điểm đến, số lượng chuyến bay, v.v. Điều này có thể hỗ trợ hành khách hoặc các công ty du lịch trong việc đưa ra quyết định về mua vé, lập kế hoạch chuyến bay hoặc xác định chiến lược giá cả.

## **Phân tích thị trường vé máy bay**: Dựa trên dữ liệu về giá vé máy bay, thuật toán có thể cung cấp thông tin về xu hướng giá cả, sự biến động và sự phụ thuộc vào các yếu tố như hãng hàng không, thời gian, địa điểm, v.v. Điều này có thể giúp các công ty hàng không hoặc nhà đầu tư trong lĩnh vực này hiểu và dự đoán tình hình thị trường, đưa ra quyết định về giá cả, quảng cáo và tiếp thị.

## **Tối ưu hóa giá vé và doanh thu**: Thuật toán có thể được sử dụng để tối ưu hóa giá vé máy bay dựa trên các yếu tố như nhu cầu khách hàng, cạnh tranh và tài nguyên chuyến bay. Bằng cách xác định mối quan hệ giữa giá vé và yếu tố khác, thuật toán có thể giúp tìm ra các chiến lược giá hiệu quả để tăng doanh thu và lợi nhuận.

## **Dự báo tình trạng chỗ trống và tải trọng chuyến bay**: Dựa trên dữ liệu về số lượng chỗ trống và tải trọng trước đây, thuật toán có thể dự đoán tình trạng chỗ trống và tải trọng trên các chuyến bay trong tương lai. Điều này có thể giúp các hãng hàng không quản lý tài nguyên chuyến bay một cách hiệu quả, đảm bảo sự cân đối giữa cung và cầu và tối ưu hóa sử dụng tài nguyên.

## **Phân tích dữ liệu thị trường và khách hàng**: Dự liệu thu thập được từ các giao dịch vé máy bay có thể được sử dụng để phân tích thị trường và khách hàng.