**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

---------o0o---------



**Bài tập lớn**

Dự đoán sự hài lòng của khách hàng với dịch vụ hàng không

**Nhóm 04**

**Giảng viên hướng dẫn :**

Giảng viên : *Nguyễn Quốc Tuấn*

**Sinh viên thực hiện :**

|  |  |  |
| --- | --- | --- |
| 1 | Đỗ Tuấn Khải | 211210438 |
| 2 | Trịnh Thành Nam | 211201058 |
| 3 | Nguyễn Xuân Bách | 211243054 |
| 4 | Trần Công Chiến | 211201878 |
| 5 | Nguyễn Đức Mạnh | 211241820 |

*Hà Nội, 2024*

---------------o0o---------------

Mục Lục

[Mục Lục 2](#_Toc182128672)

[I. Tổng quan về Phân lớp dữ liệu 4](#_Toc182128673)

[1. Khái niệm 4](#_Toc182128674)

[2. Các bước của quá trình xây dựng mô hình phân lớp dữ liệu 4](#_Toc182128675)

[3. Ví dụ 5](#_Toc182128676)

[4. Các kỹ thuật phân lớp dữ liệu 5](#_Toc182128677)

[II. Tìm hiểu về cây và thuật toán C4.5 6](#_Toc182128678)

[1. Giới thiệu về Cây Quyết Định 6](#_Toc182128679)

[1.1. Cây Quyết Định là gì ? 6](#_Toc182128680)

[1.2. Cấu trúc của Cây Quyết Định 6](#_Toc182128681)

[1.3. Quy trình hoạt động của Cây Quyết Định 7](#_Toc182128682)

[1.4. Ưu điểm và nhược điểm của Cây Quyết Định 7](#_Toc182128683)

[2. Thuật Toán C4.5 7](#_Toc182128684)

[2.1. Cách thức hoạt động của C4.5 8](#_Toc182128685)

[2.2. Công thức tính tỷ lệ thông tin (Gain Ratio) 8](#_Toc182128686)

[2.3. Ưu điểm và nhược điểm của thuật toán C4.5 9](#_Toc182128687)

[2.4. Cắt tỉa cây (Pruning) 10](#_Toc182128688)

[2.5. Ví dụ 10](#_Toc182128689)

[III. Ứng dụng 14](#_Toc182128690)

[1. Tên ứng dụng 14](#_Toc182128691)

[2. Nguồn dữ liệu 14](#_Toc182128692)

[3. Cách thức thực hiện 22](#_Toc182128693)

[3.1. Khám phá dữ liệu 22](#_Toc182128694)

[3.2. Tiền xử lí dữ liệu 22](#_Toc182128695)

[3.3. Xây dựng cây quyết định C4.5 24](#_Toc182128696)

[4. Kết quả 24](#_Toc182128697)

[4.1. Kết quả đầu ra 24](#_Toc182128698)

[4.2. Đánh giá mô hình 24](#_Toc182128699)

[5. Các lần điều chỉnh mô hình 26](#_Toc182128700)

[IV. Kết Luận 27](#_Toc182128709)

# Tổng quan về Phân lớp dữ liệu

## Khái niệm

Phân lớp là quá trình gán nhãn cho các đối tượng của 1 dataset dựa trên các đặc tính của đối tượng đó. Quá trình có thể được dùng để hỗ trợ việc xây dựng các mô hình để tự động hóa việc phân lớp các đối tượng.

Có 2 loại phân lớp là phân loại nhị phân và phân loại đa lớp (binary classification và multi-class classification) :

* Phân loại nhị phân là việc chia các đối tượng vào 2 lớp duy nhất, như ‘spam’ và ‘not spam’.
* Phân loại đa lớp là việc chia các đối tượng vào nhiều hơn 2 lớp.

## Các bước của quá trình xây dựng mô hình phân lớp dữ liệu

Bước 1 : Thu thập dữ liệu

* Thu thập dữ liệu cần thiết cho mô hình từ những nguồn có thể :
* Làm khảo sát
* Tìm từ các database có sẵn trên Internet

Bước 2 : Xử lý dữ liệu

* Thực hiện tiền xử lý cho những dữ liệu thu thập được. Giải quyết các vấn đề như : bị thiếu giá trị, xử lý ngoại lai, chuyển đổi định dạng dữ liệu sao cho phù hợp với việc phân tích

Bước 3 : Lựa chọn thuộc tính

* Lựa chọn ra những thuộc tính chính sẽ dùng cho việc phân loại
* Có thể dùng các phương pháp như : phân tích tương quan, đo lường tính hữu dụng của thông tin, phân tích thành phần chính

Bước 4 : Lựa chọn mô hình

* Ở bước này, ta cần lựa chọn ra thuật toán phân lớp phù hợp với mong muốn của mình
* Một số thuật toán có thể sử dụng : decision trees, support vector machines, và neural network

Bước 5 : Đào tạo mô hình và đánh giá mô hình

* Những dữ liệu đã xử lý sẽ chia làm 2 loại : tập dữ liệu đào tạo và tập dữ liệu xác thực.
* Mô hình sẽ được đào tạo bằng tập dữ liệu đánh giá và được đánh giá/ kiểm thử 1 lần nữa bằng tập dữ liệu xác thực.

## Ví dụ

Đề tài : Giám đốc tín dụng ngân hàng cần phân tích dữ liệu để xác định người vay nào được coi là "an toàn" và người nào "rủi ro" đối với ngân hàng.

Bước 1 : Thu thập dữ liệu :

* Thu thập những thông tin liên quan của người vay như : tuổi, mức thu nhập, nghề nghiệp, lịch sử tín dụng, số tiền vay, thời hạn vay

Bước 2 : Xử lý dữ liệu

* Xử lý các vấn đề như thiếu thông tin, thông tin nhiễu/ngoại lai
* Chuẩn hóa thông tin

Bước 3 : Lựa chọn thuộc tính

* Xác định các thuộc tính có liên quan nhất góp phần vào việc dự đoán vỡ nợ

Bước 4 : Lựa chọn mô hình

* Chọn thuật toán phù hợp dựa trên dữ liệu thu được

Bước 5 : Đào tạo và đánh giá mô hình

* Chia lượng thông tin thu được thành 2 phần : tập huấn luyện và tập xác thực
* Huấn luyện mô hình trên tập huấn luyện để tìm hiểu mối quan hệ giữa các thuộc tính và khả năng vỡ nợ của đối tượng
* Đánh giá hiệu suất của mô hình bằng cách tập xác thực

## Các kỹ thuật phân lớp dữ liệu

* Cây quyết định (Decision Trees)
* Máy vector hỗ trợ (Support Vector Machines - SVM)
* Naïve Bayes
* K-Nearest Neighbors (KNN)
* Mạng nơron nhân tạo (Artificial Neural Networks - ANN)
* Hồi quy logistic (Logistic Regression)
* Cây tăng cường độ chính xác (Gradient Boosted Trees)
* Rừng ngẫu nhiên (Random Forest)
* Phân lớp Bayes (Bayesian Classification)
* Máy học sâu (Deep Learning)

# Tìm hiểu về cây và thuật toán C4.5

## Giới thiệu về Cây Quyết Định

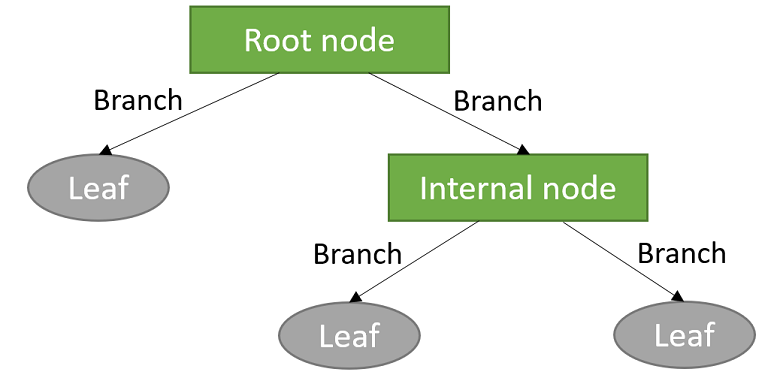
### Cây Quyết Định là gì ?

Cây quyết định là một mô hình dự đoán dưới dạng cấu trúc phân nhánh giống như cây. Mỗi nút trong cây đại diện cho một thuộc tính trong tập dữ liệu mà cần phải đánh giá, trong khi mỗi nhánh từ nút đại diện cho một kết quả khả dĩ của thuộc tính đó. Quá trình phân nhánh này tiếp tục cho đến khi đạt đến một lá của cây, là điểm mà tất cả các mẫu thuộc về một lớp cụ thể (trong bài toán phân loại) hoặc có một giá trị dự đoán cụ thể (trong bài toán hồi quy).

### Cấu trúc của Cây Quyết Định

Cấu trúc của cây quyết định bao gồm ba loại thành phần chính :

* Nút gốc (Root Node) : Là điểm bắt đầu của cây quyết định và chứa toàn bộ tập dữ liệu. Nút gốc đại diện cho thuộc tính đầu tiên mà mô hình sử dụng để phân chia dữ liệu.
* Nút nội bộ (Internal Node) : Mỗi nút nội bộ tương ứng với một thuộc tính trong tập dữ liệu. Mỗi nhánh từ nút nội bộ đại diện cho một trong những giá trị của thuộc tính đó. Quyết định tại mỗi nút nội bộ dựa trên giá trị của thuộc tính tại nút đó.
* Branches (Nhánh) : Các kết nối giữa các nút, biểu thị sự phân chia dữ liệu dựa trên giá trị của thuộc tính. Chỉ ra các hướng đi từ một nút đến các nút con, thể hiện các kết quả của điều kiện phân chia.
* Nút lá (Leaf Node) : Các nút lá là những nút cuối cùng trong cây, không còn phân nhánh. Mỗi nút lá đại diện cho một lớp cụ thể trong bài toán phân loại hoặc một giá trị trong bài toán hồi quy. Nút lá đưa ra kết quả dự đoán cho các mẫu mà nó đại diện.



### Quy trình hoạt động của Cây Quyết Định

Cây quyết định hoạt động bằng cách sử dụng các thuộc tính của dữ liệu để phân chia tập dữ liệu thành các tập con, sao cho mỗi tập con thuộc về một lớp nhất định. Quá trình phân chia tiếp tục cho đến khi không còn thuộc tính nào để chia hoặc dữ liệu đã hoàn toàn được phân loại. Khi một mẫu mới cần phân loại, cây sẽ dẫn dắt mẫu đó từ nút gốc qua các nút nội bộ, đến nút lá đại diện cho lớp dự đoán.

### Ưu điểm và nhược điểm của Cây Quyết Định

* **Ưu điểm :**
  + Dễ hiểu và giải thích : Cấu trúc cây quyết định gần giống với cách con người suy nghĩ để ra quyết định. Nó có thể được trực quan hóa và giải thích một cách dễ dàng, giúp việc truyền đạt kết quả mô hình trở nên đơn giản.
  + Ít yêu cầu tiền xử lý dữ liệu : Cây quyết định có thể làm việc với cả dữ liệu số và dữ liệu phân loại mà không cần nhiều xử lý trước. Không yêu cầu chuẩn hóa hay chuẩn hóa dữ liệu.
  + Khả năng xử lý dữ liệu phi tuyến : Cây quyết định hiệu quả với các dữ liệu phi tuyến, nhờ vào quá trình phân chia cây ở mỗi cấp độ nút.
  + Lựa chọn tự động thuộc tính quan trọng : Các thuộc tính quan trọng sẽ tự động được chọn làm các nút phân chia đầu tiên, giúp xác định những thuộc tính có ảnh hưởng nhất đến kết quả.
* **Nhược điểm :**
  + Dễ bị quá khớp (overfitting) : Cây quyết định có thể phân chia dữ liệu quá chi tiết, dẫn đến hiện tượng quá khớp. Điều này đặc biệt nghiêm trọng với các cây có độ sâu lớn, làm giảm khả năng tổng quát hóa mô hình.
  + Hiệu suất thấp trên dữ liệu phức tạp : Trên các bài toán lớn và phức tạp, cây quyết định có thể không đủ mạnh và chính xác so với các mô hình học máy khác như rừng ngẫu nhiên (Random Forest) hoặc mô hình hồi quy tuyến tính.
  + Nhạy cảm với dữ liệu nhiễu : Dữ liệu chứa nhiễu (outlier) có thể ảnh hưởng lớn đến mô hình cây quyết định, làm cây phân nhánh nhiều và phức tạp hơn, dẫn đến mất tính khái quát.
  + Không ổn định : Cây quyết định có thể dễ dàng thay đổi nếu có sự thay đổi nhỏ trong dữ liệu đầu vào, đặc biệt là khi không có chiến lược cắt tỉa (pruning) hợp lý.

## Thuật Toán C4.5

Thuật toán C4.5, được phát triển bởi Ross Quinlan, là một trong những thuật toán xây dựng cây quyết định phổ biến nhất nhờ khả năng xử lý dữ liệu liên tục và giá trị bị thiếu. Đây là phiên bản cải tiến của ID3, sử dụng một số kỹ thuật giúp tối ưu hóa độ chính xác và khả năng tổng quát của cây quyết định.

### Cách thức hoạt động của C4.5

Quy trình cơ bản của thuật toán C4.5 gồm các bước :

1. Chọn thuộc tính tốt nhất để chia dữ liệu : Thuật toán sử dụng một chỉ số gọi là "tỷ lệ thông tin" (Gain Ratio) để chọn thuộc tính có lợi nhất cho việc phân chia.
2. Phân chia dữ liệu dựa trên thuộc tính : Dữ liệu được chia thành các tập con dựa trên các giá trị của thuộc tính đã chọn.
3. Lặp lại : Quy trình này tiếp tục cho đến khi đạt điều kiện dừng (tất cả mẫu thuộc cùng một lớp hoặc không còn thuộc tính để chia).
4. Cắt tỉa cây : Sau khi cây hoàn thành, thuật toán cắt tỉa cây để loại bỏ các nhánh không cần thiết nhằm tránh hiện tượng quá khớp.

### Công thức tính tỷ lệ thông tin (Gain Ratio)

#### Entropy

Entropy đo lường mức độ hỗn loạn hoặc không chắc chắn của một tập dữ liệu. Công thức Entropy cho một tập dữ liệu với các lớp là:

Trong đó:

* là xác suất của lớp trong tập dữ liệu .

#### Information Gain (Lợi ích thông tin)

Lợi ích thông tin cho biết mức độ giảm Entropy khi chia dữ liệu theo một thuộc tính. Công thức tính Gain của một thuộc tính là:

Trong đó:

* là các giá trị có thể có của thuộc tính .
* là tập con của chứa các mẫu có giá trị cho thuộc tính .

#### Split Information (Thông tin phân chia)

C4.5 không chỉ dựa vào Gain mà còn tính toán thông tin phân chia để tránh sự thiên vị đối với các thuộc tính có nhiều giá trị. Công thức tính Split Information của một thuộc tính là:

#### Gain Ratio (Tỷ lệ lợi ích thông tin)

Gain Ratio là chỉ số được C4.5 sử dụng để quyết định thuộc tính phân chia, nhằm cân bằng giữa Gain và Split Information. Công thức tính Gain Ratio là:

Thuộc tính có Gain Ratio cao nhất sẽ được chọn làm thuộc tính phân chia.

### Ưu điểm và nhược điểm của thuật toán C4.5

* **Ưu điểm :**
* Xử lý dữ liệu liên tục : C4.5 có khả năng xử lý dữ liệu liên tục bằng cách chia nhỏ các thuộc tính liên tục thành các khoảng giá trị. Điều này giúp mở rộng ứng dụng của cây quyết định trên các tập dữ liệu phức tạp hơn.
* Xử lý dữ liệu có giá trị thiếu : Khác với ID3, C4.5 có thể xử lý các mẫu có giá trị thuộc tính bị thiếu mà không cần phải loại bỏ chúng. Thuật toán đưa ra các quy tắc để phân chia dựa trên những phần dữ liệu có sẵn, giúp giảm thiểu mất mát thông tin.
* Sử dụng Gain Ratio để chọn thuộc tính : Gain Ratio trong C4.5 giúp tránh thiên vị các thuộc tính có nhiều giá trị, cải thiện khả năng tổng quát hóa của cây so với ID3.
* Cắt tỉa cây (Pruning) : C4.5 có tích hợp kỹ thuật cắt tỉa để loại bỏ các nhánh không cần thiết sau khi xây dựng cây, giúp giảm nguy cơ quá khớp và cải thiện hiệu suất trên dữ liệu mới.
* **Nhược điểm :**
* Tính toán phức tạp : So với ID3, thuật toán C4.5 phức tạp hơn và đòi hỏi nhiều tài nguyên tính toán hơn, đặc biệt trên các tập dữ liệu lớn với nhiều thuộc tính liên tục.
* Không hiệu quả trên dữ liệu có nhiều giá trị : Mặc dù Gain Ratio giúp giảm thiểu sự thiên vị, C4.5 vẫn gặp khó khăn với các thuộc tính có số lượng giá trị quá lớn, dẫn đến cây phức tạp và khó hiểu.
* Nhạy cảm với dữ liệu nhiễu : Dữ liệu chứa nhiễu có thể dẫn đến cây quyết định sâu và phức tạp hơn, mặc dù kỹ thuật cắt tỉa có thể giúp cải thiện, nhưng vẫn không hoàn toàn loại bỏ được tác động của nhiễu.
* Bị hạn chế bởi khả năng mở rộng : Trên các tập dữ liệu rất lớn hoặc các bài toán với yêu cầu thời gian thực, C4.5 có thể không phải là lựa chọn tối ưu vì thời gian xử lý lâu và yêu cầu bộ nhớ cao.

### Cắt tỉa cây (Pruning)

2.4.1. Khái niệm về Cắt Tỉa Cây (Pruning)

Cắt tỉa cây là quá trình loại bỏ các nhánh không cần thiết hoặc ít thông tin trong cây quyết định sau khi cây đã được xây dựng đầy đủ. Mục tiêu của việc cắt tỉa là đơn giản hóa cây để tránh tình trạng quá khớp (overfitting), cải thiện hiệu suất khi áp dụng trên dữ liệu mới, và giảm độ phức tạp của mô hình.

2.4.2. Phương Pháp Cắt Tỉa trong C4.5

Thuật toán C4.5 áp dụng kỹ thuật cắt tỉa hậu xử lý, nghĩa là sau khi cây quyết định được xây dựng hoàn chỉnh, thuật toán sẽ cắt bớt những nhánh không mang lại giá trị phân loại cao. Cụ thể :

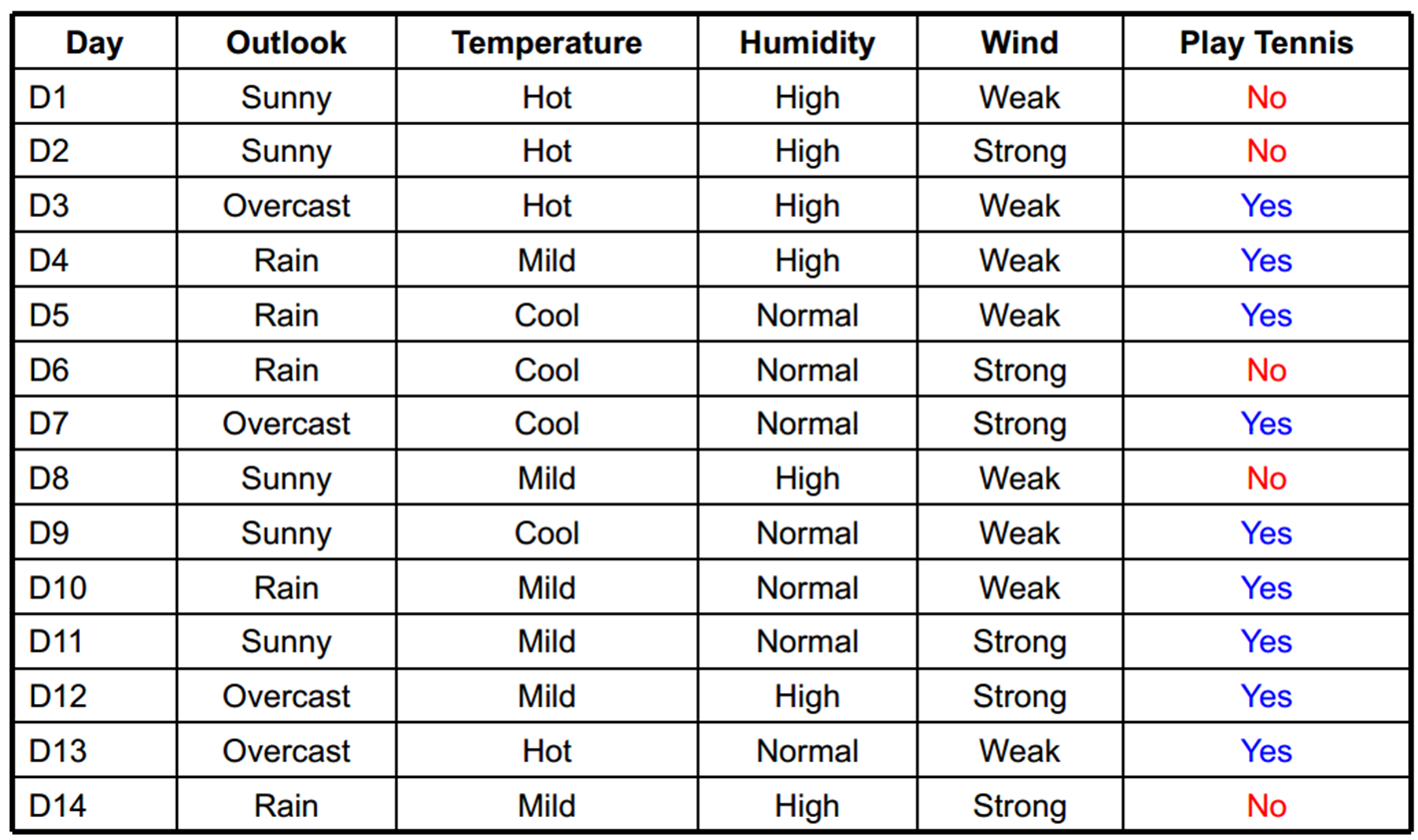
* Cắt tỉa bằng thống kê : C4.5 sử dụng phương pháp thống kê để đánh giá từng nhánh, xem liệu việc loại bỏ nhánh đó có làm giảm hiệu suất dự đoán hay không. Nếu không ảnh hưởng đáng kể đến kết quả, nhánh sẽ được loại bỏ.
* Cắt tỉa dựa trên độ tin cậy : Một số nhánh có thể không đại diện tốt cho tập dữ liệu do số lượng mẫu nhỏ. Thuật toán sẽ xem xét độ tin cậy của nhánh, và loại bỏ nhánh nào không đạt yêu cầu tin cậy nhất định.

2.4.3. Lợi Ích của Cắt Tỉa Cây

* Giảm quá khớp : Việc cắt tỉa loại bỏ các nhánh phức tạp, giúp cây không quá chi tiết đối với dữ liệu huấn luyện và cải thiện hiệu suất trên dữ liệu mới.
* Tăng hiệu suất dự đoán : Cây quyết định cắt tỉa thường đạt độ chính xác tốt hơn vì loại bỏ được các mẫu nhiễu hoặc các mẫu ít đại diện.
* Giảm độ phức tạp của mô hình : Cây quyết định trở nên dễ hiểu hơn, và việc triển khai hoặc giải thích kết quả mô hình cũng đơn giản hơn.

### Ví dụ

Yêu cầu : xây dựng cây quyết định c4.5 để có thể xác định xem ngày hôm đó có đi chơi tennis không với bảng dữ liệu sau

**Tính entropy (E) của play tennis:**

**Tính entropy(E) của outlook**

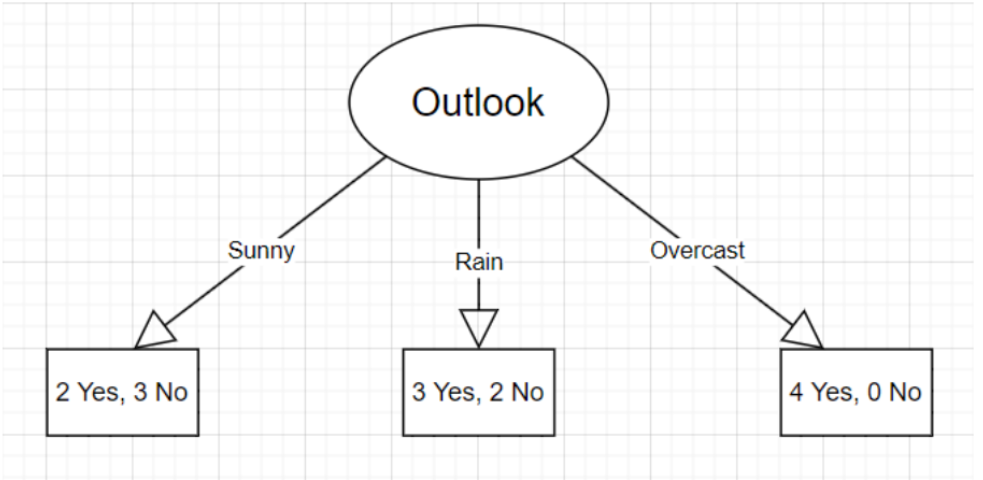
**Tính Information Gain (IG) của outlook**

**Tính Split Info (SI)**

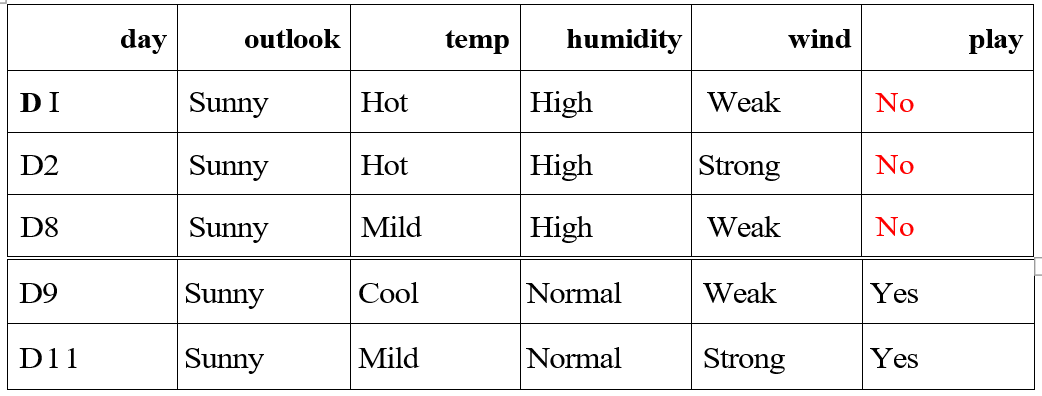
**Tính Gain Ratio (GR)**

**Tương tự ta tìm được:**

=> Dựa vào hệ số gain ratio của từng thuộc tính thì khi chọn outlook là gốc ta sẽ lấy được thông tin với chính xác cao nhất. Như vậy ta sẽ dựng cây với thuộc tính Outlook làm nút gốc. Đưa ra các cạnh là các giá trị của thuộc tính Outlook. Từ đó ta có cây như sau:



Theo như cây được minh họa như trên cùng với tập dữ liệu, ta có thể cho rằng khi Outlook có giá trị là Overcast thì khả năng quyết định chơi Tennis sẽ là Yes. Vì vậy giá trị Overcast sẽ không cần được xét đến ở những lần tính tiếp theo. Tiếp tục tính các hệ số khi xét đến điều kiện của Outlook là Sunny. Từ đó ta rút ngắn được bảng giá trị để xét



**Tính entropy(E)** của **play tennis** nhưng không phải của bảng đầu mà là khi **Outlook** = **Sunny**

E (Play tennis) = 0.971

**Các bước lại lặp lại như phần đầu**

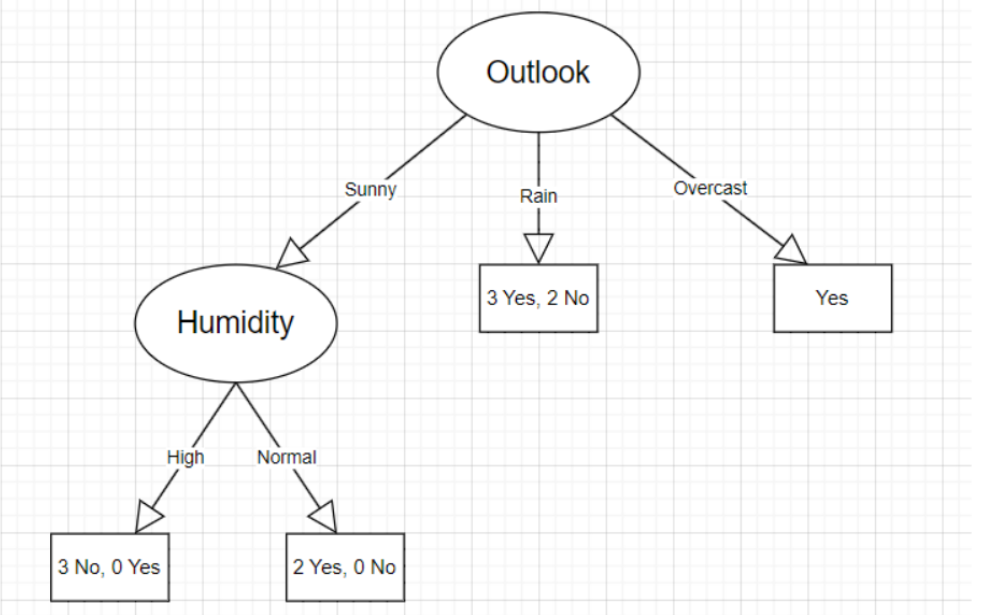
**Gain Ratio (GR)**

GR (Play tennis, temp) = 0.3752

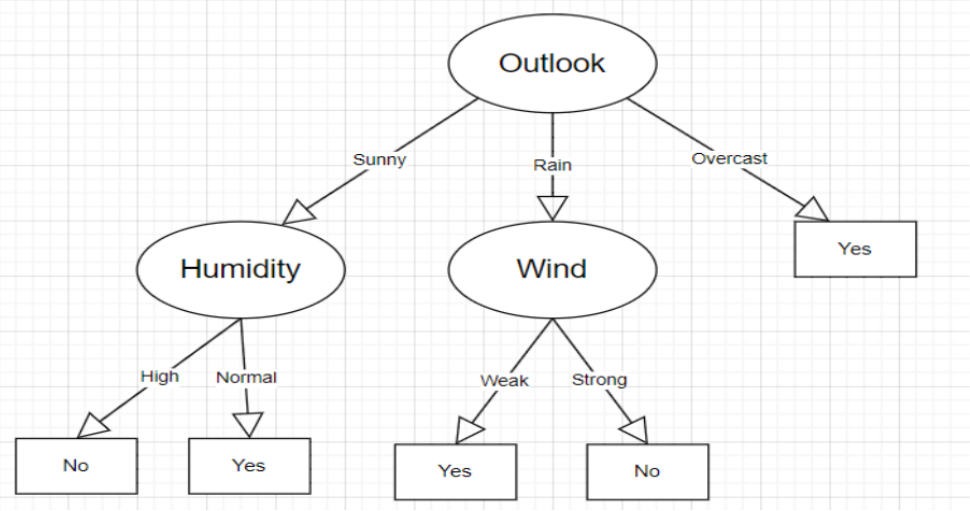
GR (Play tennis, humidity) = 1

GR (Play tennis, wind) = 0.021

Như vậy, ta sẽ chọn thuộc tính Humidity để làm nút nối từ Outlook qua giá trị sunny. Từ đó cây sẽ được tiếp tục phát triển như sau:



Cứ lặp lại 4 bước cuối cùng chúng ta sẽ có



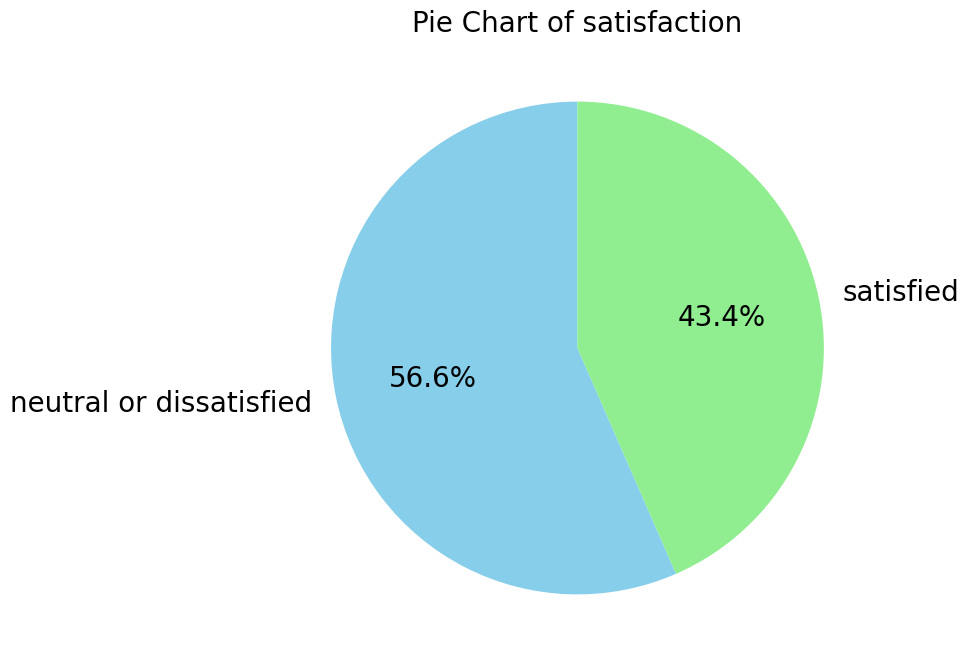
# Ứng dụng

## Tên ứng dụng

Dự đoán sự hài lòng của khách hàng với dịch vụ hàng không

## Nguồn dữ liệu

* Dataset : "Airline Passenger Satisfaction"
* Link dataset : [https ://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction](https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction)
* Đặc điểm của dataset :
* Số lượng mẫu : 129,880
* Số lượng thuộc tính : 24
* Biến mục tiêu : Satisfaction



* Các thuộc tính chính trong dataset :
* Gender : Giới tính (male, female)

A pie chart of gender

Description automatically generated

* Customer Type : Loại khách hàng (Loyal Customer, disloyal Customer)

A pie chart of customer type

Description automatically generated

* Age : Tuổi (dao động từ 7-80 tuổi)

A blue rectangular object with black lines

Description automatically generated

* Type of Travel : Loại chuyến bay (Business, Personal)

A pie chart of type of travel

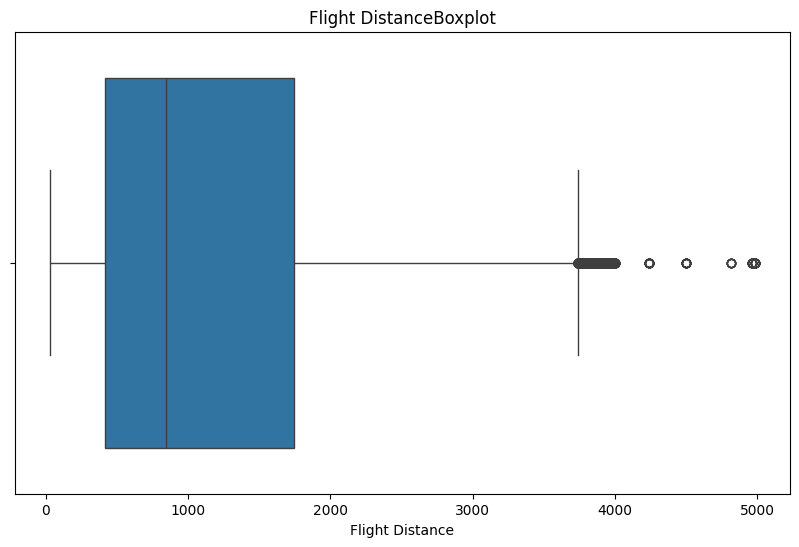
Description automatically generated

* Class : Hạng vé (Business, Eco, Eco Plus)

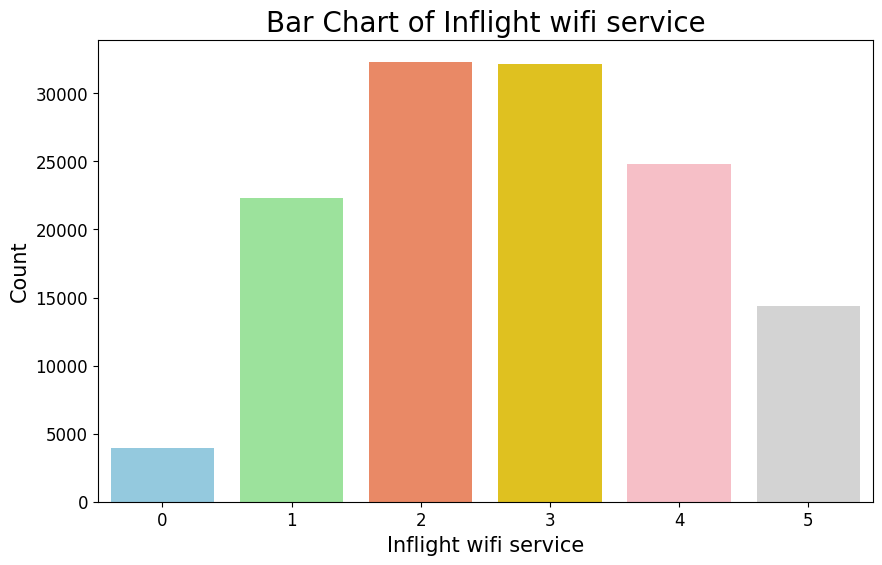
A pie chart of class

Description automatically generated

* Flight Distance : Khoảng cách bay



* Các đánh giá về dịch vụ (thang điểm 0-5) :
* Inflight wifi service - Dịch vụ wifi trên máy bay



* Departure/Arrival time convenience - Sự thuận tiện của giờ khởi hành/đến

A bar chart of departure

Description automatically generated

* Ease of Online booking - Sự dễ dàng đặt vé trực tuyến

A bar chart of online booking

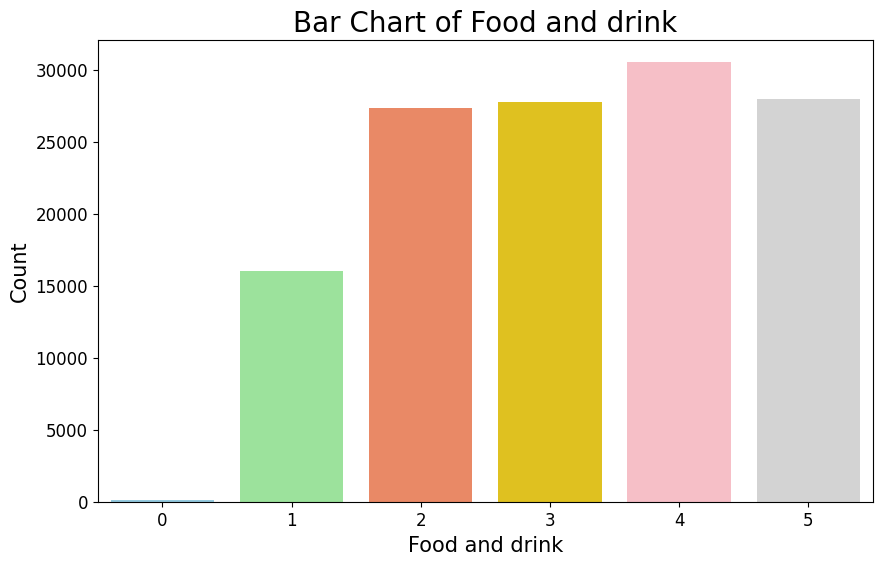
Description automatically generated

* Gate location - Vị trí cổng lên máy bay

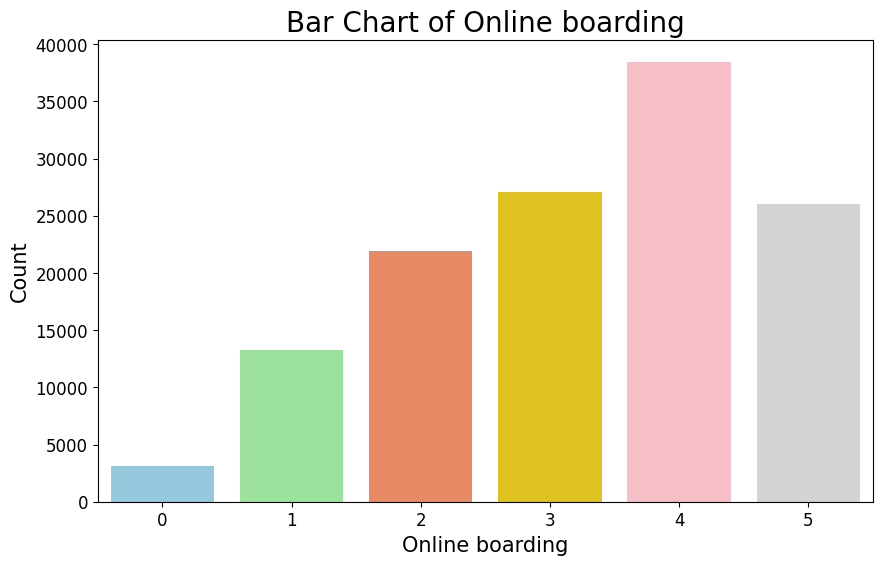
A bar chart of gate location

Description automatically generated

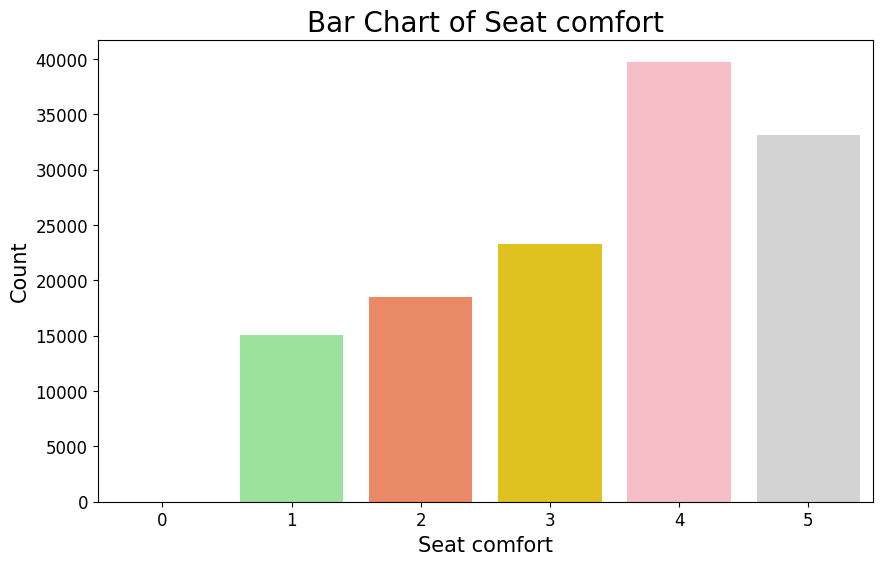
* Food and drink - Đồ ăn và thức uống



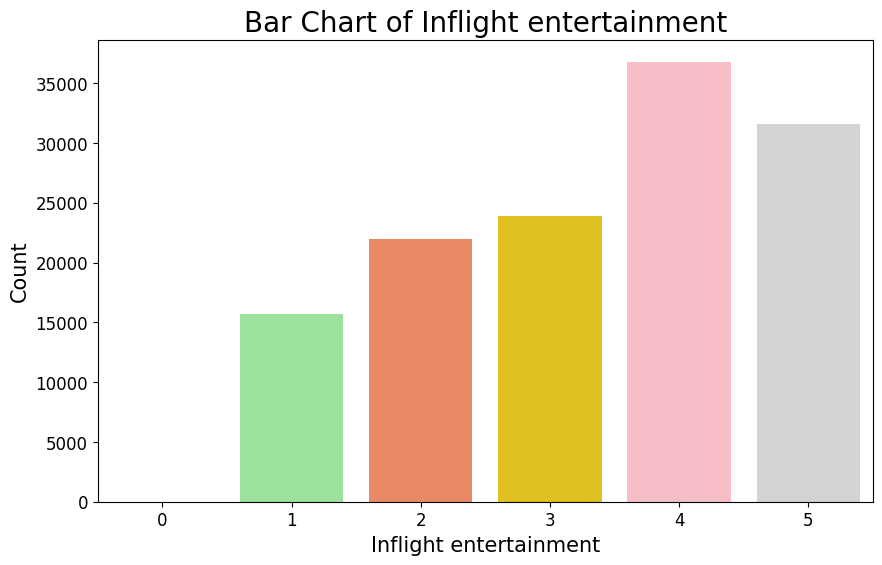
* Online boarding - Làm thủ tục lên máy bay trực tuyến



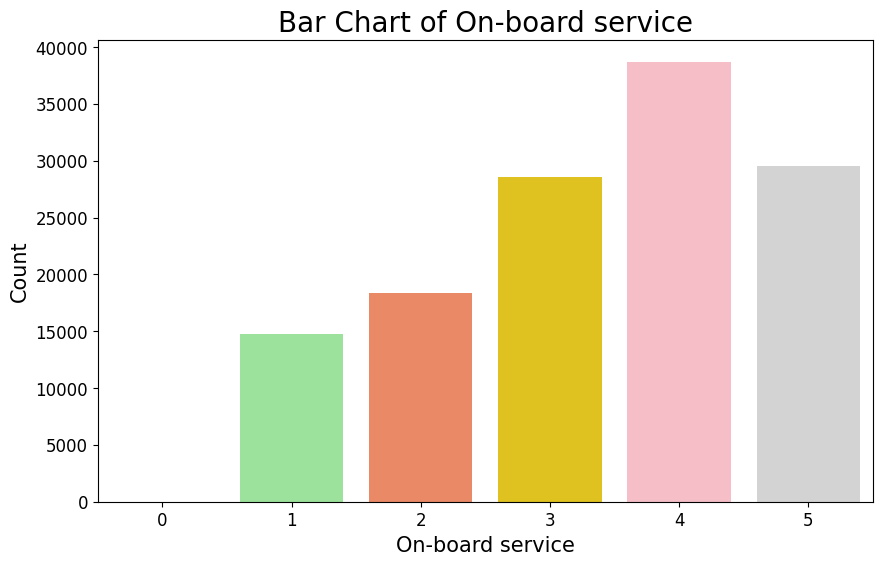
* Seat comfort - Độ thoải mái của ghế ngồi



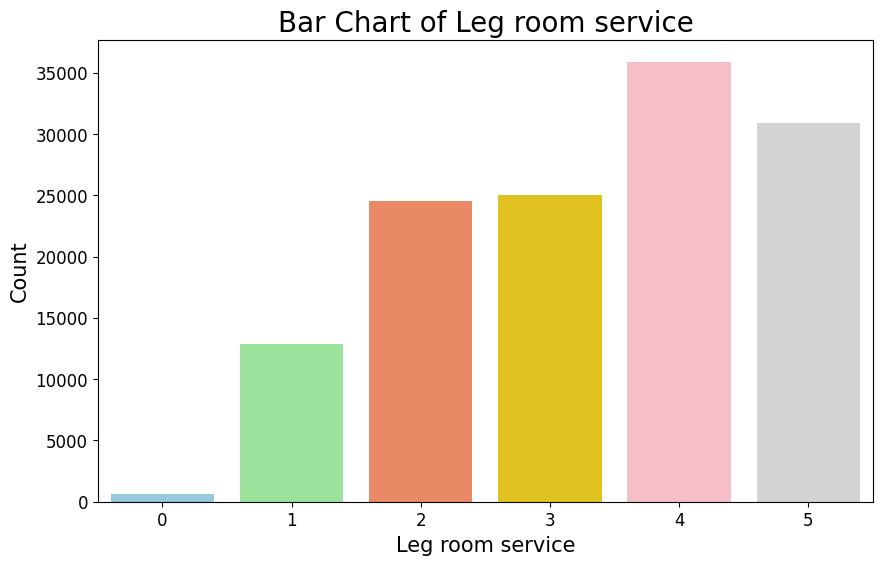
* Inflight entertainment - Giải trí trên máy bay



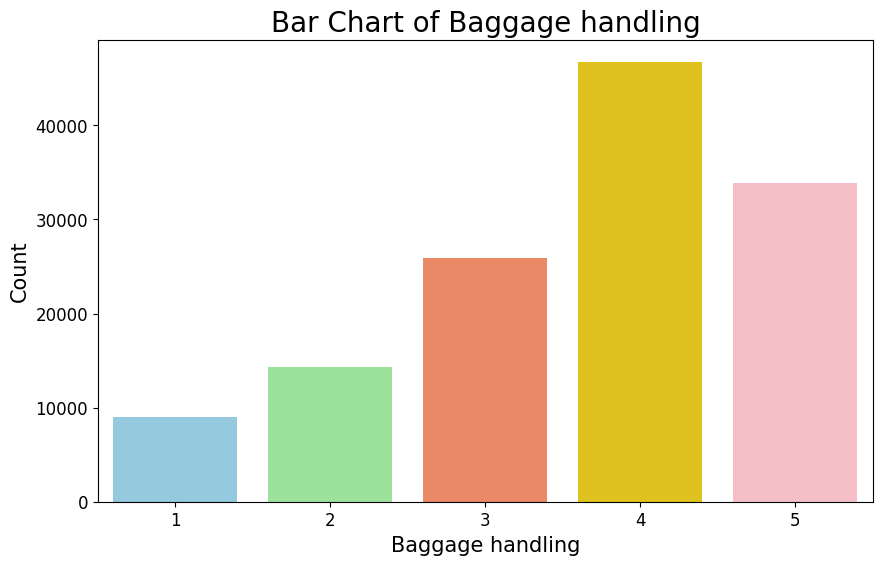
* On-board service - Dịch vụ trên máy bay



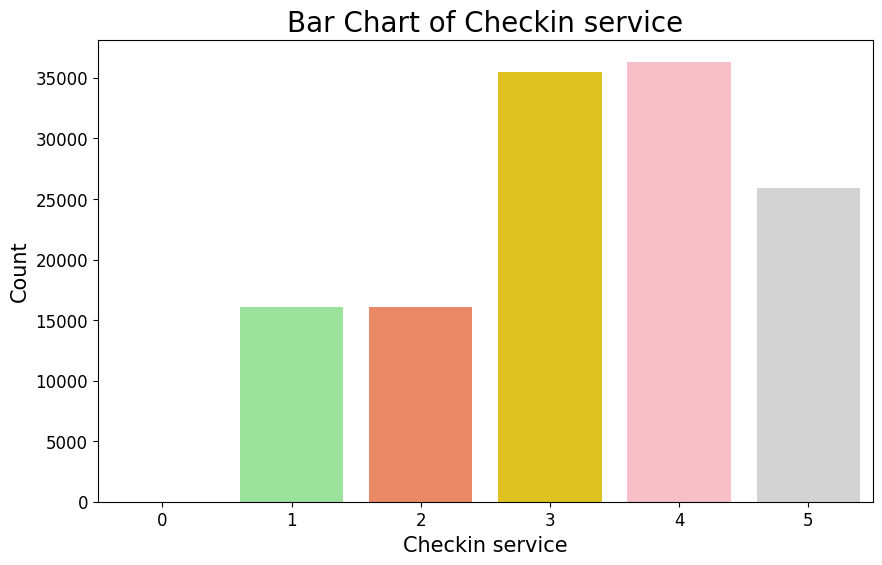
* Leg room service - Không gian để chân



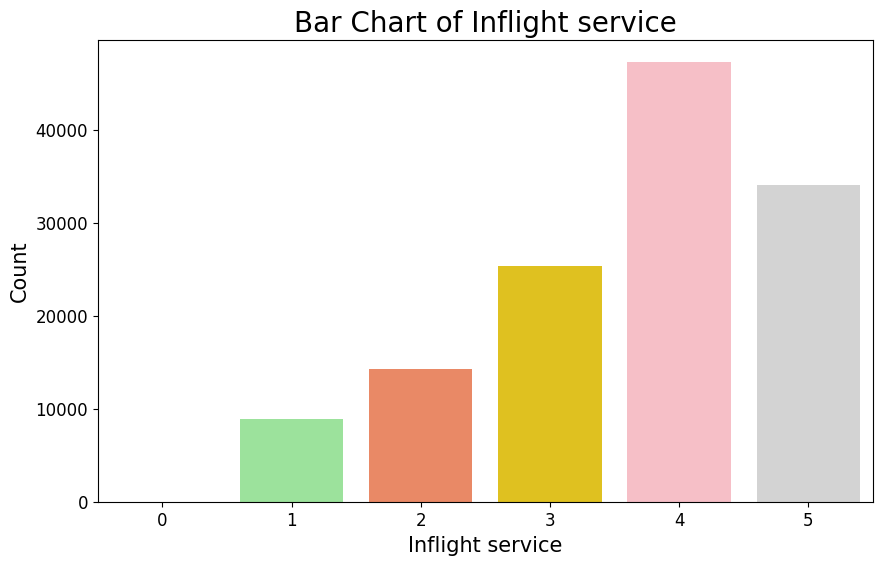
* Baggage handling - Xử lý hành lý



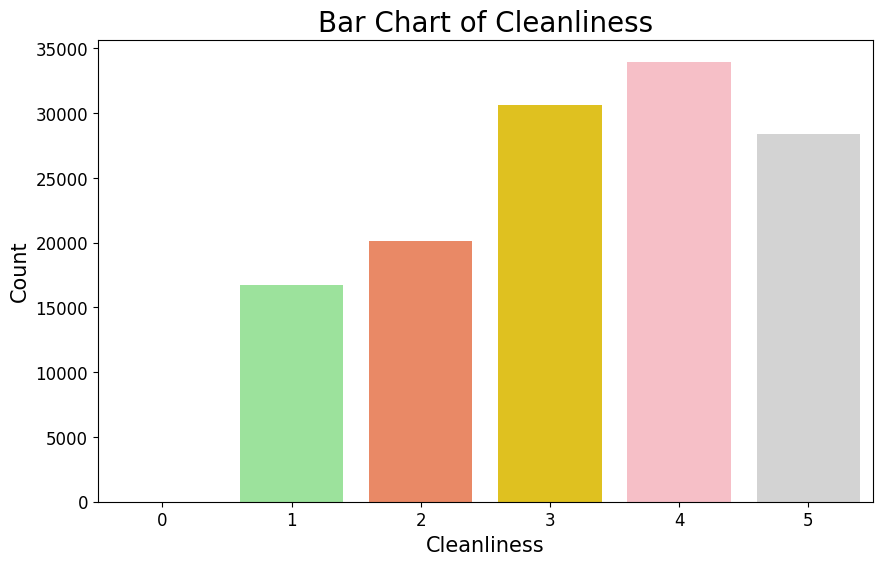
* Check-in service - Dịch vụ làm thủ tục



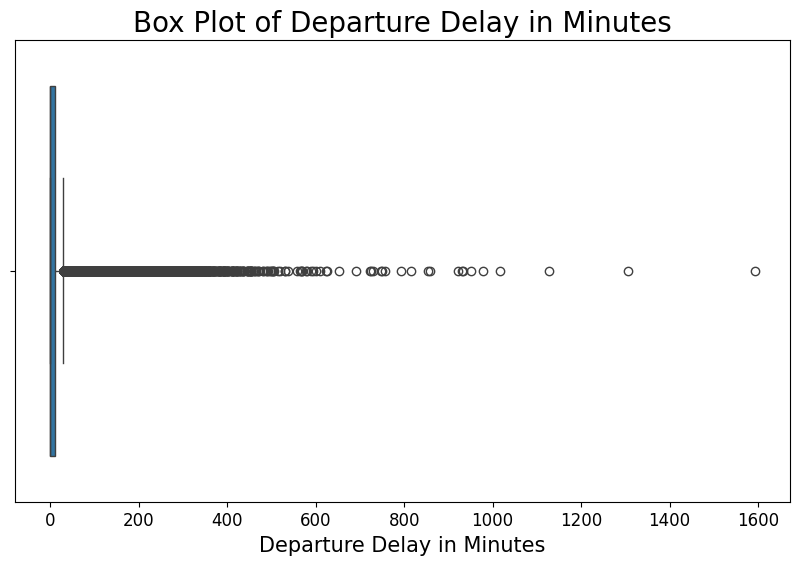
* Inflight service - Dịch vụ trong chuyến bay



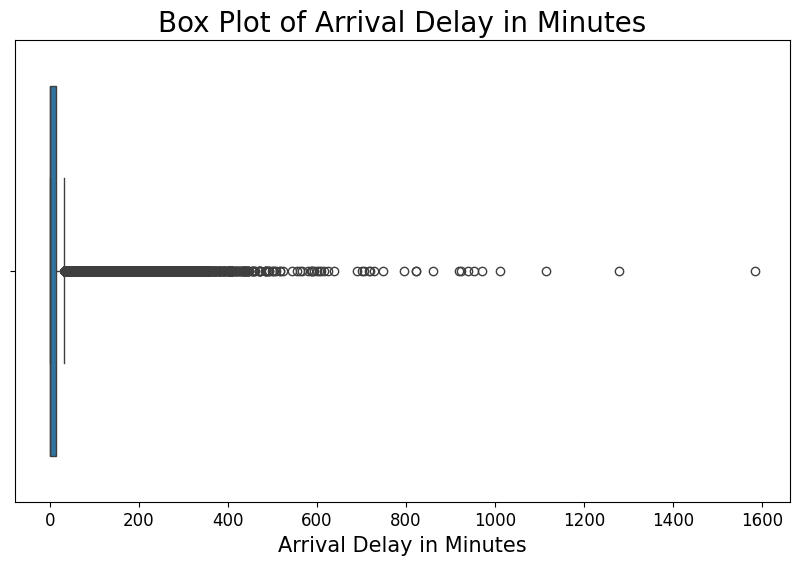
* Cleanliness - Sự sạch sẽ



* Departure Delay in Minutes



* Arrival Delay in Minutes



* Lý do dataset này phù hợp với C4.5 :
* Bài toán phân loại nhị phân : Dự đoán khách hàng hài lòng hay không hài lòng.
* Kết hợp cả loại giá tri định danh (nominal) và số (numerical)

## Cách thức thực hiện

Ngôn ngữ sử dụng : Python

### Khám phá dữ liệu

* Data set gồm 25 thuộc tính (attribute) và 129880
* Các loại thuộc tính gồm
  + Định danh (nominal attribute) :id
  + Nhị phân (binary attribute) : gender, customer type, type of travel, satisfaction
  + Ordinary : class, Inflight wifi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Checkin service, Inflight service, Cleanliness
  + Số (numeric): age, Flight Distance, Departure Delay in Minutes, Arrival Delay in Minutes
* Giá trị thiếu (missing value): Arrival Delay in Minutes là 393 giá trị

### Tiền xử lí dữ liệu

* Xóa các đặc trưng không cần thiết : xóa id
* Kiểm tra và xử lí các giá trị trùng lặp
* Kiểm tra và xử lí các giá trị thiếu : thay thế các đặc trưng bằng median
* Xử lí giá trị ngoại lai với IQR
* Rời rạc hóa dữ liệu : rời rạc hóa các đặt trưng số
* Mã hóa dữ liệu
* Tính hệ số tương quan của các thuộc tính với thuộc tính **satisfaction** :

|  |  |
| --- | --- |
| **Feature** | **Correlation Coefficient** |
| Gender | 0.011111 |
| Customer Type | -0.184593 |
| Age | 0.147241 |
| Type of Travel | -0.487225 |
| Class | -0.471642 |
| Flight Distance | 0.300979 |
| Inflight wifi service | 0.277998 |
| Departure/Arrival time convenient | -0.074133 |
| Ease of Online booking | 0.181683 |
| Gate location | 0.007245 |
| Food and drink | 0.195326 |
| Online boarding | 0.508265 |
| Seat comfort | 0.345562 |
| Inflight entertainment | 0.390191 |
| On-board service | 0.306936 |
| Leg room service | 0.313230 |
| Baggage handling | 0.227046 |
| Checkin service | 0.182724 |
| Inflight service | 0.220526 |
| Cleanliness | 0.296636 |
| Departure Delay in Minutes | -0.039918 |
| Arrival Delay in Minutes | -0.074396 |
| satisfaction | 1.000000 |

* Chia tập dữ liệu : chia thành tập train (70%) và tập test (30%) để huấnluyện model
* Chuẩn hóa dữ liệu : sử dụng chuẩn hóa Z-score

Sau khi xử lý thu được :

* Tập huấn luyền gồm 22 đăc trưng và 65553 giá trị
* Tập kiểm tra gồm 22 đặc trưng và 28095 giá trị

### Xây dựng cây quyết định C4.5

Sử dụng thư viện c45-decision-tree để xây dựng model

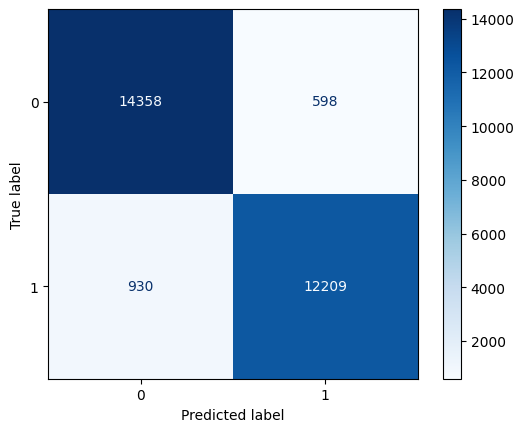
## Kết quả

### Kết quả đầu ra

Kết quả gồm 2 giá trị 0 và 1 đại diện cho có satisfaction (hài lòng) và neutral or dissatisfied (không hài lòng)

### Đánh giá mô hình

- Ma trận nhầm lẫn



* Thông số đánh giá với tập test

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| 0 (neutral or dissatisfied) | 0.94 | 0.96 | 0.95 | 14956 |
| 1 (satisfaction) | 0.95 | 0.93 | 0.94 | 13139 |

* **Hiệu Suất Tổng Thể**
  + Độ Chính Xác (Accuracy): Mô hình đạt độ chính xác 95%, cho thấy mô hình hoạt động tốt trên tập dữ liệu kiểm tra và dự đoán đúng phần lớn các mẫu.
  + Macro avg và Weighted avg: Các chỉ số precision, recall, và f1-score đều đạt 95%, cho thấy mô hình hoạt động đồng đều và ổn định trên cả hai lớp.
* **Hiệu Suất Của Từng Lớp**
  + Lớp 0 (neutral or dissatisfied):
    - Precision: 0.94, cho thấy mô hình ít dự đoán sai các mẫu thuộc lớp neutral or dissatisfied.
    - Recall: 0.96, cho thấy mô hình có khả năng phát hiện các mẫu neutral or dissatisfied rất tốt.
    - F1-score: 0.95, là sự cân bằng giữa precision và recall, cho thấy mô hình hoạt động tốt trên lớp này.
  + Lớp 1 (satisfied):
    - Precision: 0.95, cho thấy mô hình ít dự đoán sai các mẫu thuộc lớp satisfied.
    - Recall: 0.93, cho thấy mô hình có khả năng phát hiện các mẫu satisfied tốt, nhưng không tốt bằng lớp neutral or dissatisfied.
    - F1-score: 0.94, là sự cân bằng giữa precision và recall, cho thấy mô hình hoạt động tốt trên lớp này.
* **Các Điểm Cần Cải Thiện**
  + Lớp 1 (satisfied): Có recall thấp hơn một chút so với lớp 0, điều này có nghĩa là mô hình có thể bỏ sót một số mẫu thuộc lớp satisfied. Bạn có thể xem xét lại các thuộc tính và quy tắc quyết định liên quan đến lớp này để cải thiện hiệu suất.

## Các lần điều chỉnh mô hình

* satisfied: 1
* neutral or dissatisfied: 0

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nội dung điều chỉnh** | | **Chưa rời rạc hóa và thực hiện cắt tỉa** | **Đã rời rạc hóa các giá trị numeric** | **Đã thực hiện cắt tỉa tiền nghiệm** |
| **Confuse matrix** | TP | 12284 | 12325 | 12209 |
| FP | 4061 | 1145 | 598 |
| TN | 10895 | 13811 | 14358 |
| FN | 855 | 814 | 930 |
| **Precision** | 1 | 0.75 | 0.91 | 0.95 |
| 0 | 0.93 | 0.94 | 0.94 |
| **Recall** | 1 | 0.93 | 0.94 | 0.93 |
| 0 | 0.73 | 0.92 | 0.96 |
| **F1-score** | 1 | 0.83 | 0.93 | 0.94 |
| 0 | 0.82 | 0.93 | 0.95 |

# Kết Luận

Phân lớp dữ liệu là một công cụ quan trọng trong việc dự đoán sự hài lòng của khách hàng với dịch vụ hàng không. Nó giúp phân loại khách hàng thành các nhóm khác nhau dựa trên mức độ hài lòng, từ đó hỗ trợ các hãng hàng không cải thiện dịch vụ và ra quyết định chiến lược.

Mô hình C4.5 của hoạt động tốt với độ chính xác cao và các chỉ số precision, recall, và f1-score đều đạt 95%. Điều này cho thấy mô hình có khả năng phân loại chính xác sự hài lòng của khách hàng. Mô hình hoạt động tốt trên cả hai lớp neutral or dissatisfied và satisfied, nhưng có thể cần cải thiện một chút về recall cho lớp satisfied. Các yếu tố quan trọng nhất ảnh hưởng đến sự hài lòng của khách hàng :

* Online boarding
* Inflight entertainment
* Seat comfort
* Leg room service
* On-board service
* Flight Distance
* Cleanliness
* Inflight wifi service

Các insight để cải thiện dịch vụ khách hàng :

* Cải thiện quy trình lên máy bay trực tuyến : Đảm bảo quy trình lên máy bay trực tuyến dễ dàng và thuận tiện cho khách hàng.
* Nâng cao chất lượng giải trí trên chuyến bay : Cung cấp nhiều lựa chọn giải trí phong phú và chất lượng cao để khách hàng có trải nghiệm tốt hơn.
* Tăng cường sự thoải mái của ghế ngồi : Đảm bảo ghế ngồi thoải mái và có đủ không gian cho chân để khách hàng cảm thấy dễ chịu trong suốt chuyến bay.
* Cải thiện dịch vụ trên máy bay : Đào tạo nhân viên phục vụ trên máy bay để cung cấp dịch vụ chuyên nghiệp và thân thiện.
* Duy trì vệ sinh sạch sẽ : Đảm bảo máy bay luôn sạch sẽ và vệ sinh để tạo cảm giác thoải mái cho khách hàng.
* Cải thiện dịch vụ wifi trên chuyến bay : Cung cấp dịch vụ wifi ổn định và nhanh chóng để khách hàng có thể kết nối internet dễ dàng.