

**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**  
**VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC**

—o0o—



**HỆ HỖ TRỢ QUYẾT ĐỊNH**  
**TỐI ƯU HÓA PHÂN TÍCH GIỎ HÀNG BẰNG**  
**THUẬT TOÁN APRIORI**

**Chuyên ngành: Toán ứng dụng**

**Người hướng dẫn:**  
**Sinh viên thực hiện:**  
**MSSV:**  
**Lớp:**

**TS. Lê Hải Hà**  
**Ngô Quang Tùng**  
**20196006**  
**HTTTQL - K64**

**HÀ NỘI, 07/2023**

# Mục lục

<b>1</b>	<b>Giới thiệu bộ dữ liệu và bài toán</b>	<b>2</b>
1.1	Mục tiêu bài toán . . . . .	2
1.2	Mô tả bộ dữ liệu . . . . .	3
<b>2</b>	<b>Tối ưu hóa phân tích giỏ hàng bằng thuật toán Apriori</b>	<b>4</b>
2.1	Tiền xử lý dữ liệu . . . . .	4
2.1.1	Xem xét tổng quan dữ liệu . . . . .	4
2.2	Xây dựng mô hình . . . . .	5
2.3	Đánh giá mô hình . . . . .	6
	<b>Tài liệu tham khảo</b>	<b>8</b>

# LỜI CẢM ƠN

Đầu tiên, em xin gửi lời cảm ơn tới thầy Lê Hải Hà người đã giúp đỡ em trong báo cáo này và trong môn học Hệ hỗ trợ quyết định. Qua môn này, em đã được trang bị kiến thức về các hệ hỗ trợ ra quyết định quản lý, các kỹ thuật phân tích và khai phá dữ liệu, máy học và trí tuệ nhân tạo. Từ đó có thể áp dụng kiến thức vào phát triển một số hệ hỗ trợ quyết định trong thực tế.

Dù đã cố gắng xong vẫn không thể tránh khỏi những hạn chế cần khắc phục. Vì vậy, em rất mong quý thầy cô đưa ra những ý kiến góp ý để đồ án có thể phát triển và có những kết quả tốt hơn.

## 1 Giới thiệu bộ dữ liệu và bài toán

### 1.1 Mục tiêu bài toán

Market-basket analysis (Phân tích giỏ hàng) là gì? Phân tích giỏ hàng là một kỹ thuật phân tích hành vi mua dựa trên lịch sử giao dịch của họ. Kỹ thuật này thường được sử dụng để nắm được thị hiếu, thói quen tiêu dùng của khách hàng. Từ đó đưa ra những chiến lược Marketing một cách hợp lý. Ví dụ: Một khách hàng đi siêu thị thường có xu hướng mua một vài sản phẩm cùng lúc. Giỏ hàng khách hàng A bao gồm: sữa, bánh mì, bia, thuốc lá. Giỏ hàng của khách hàng B gồm: sữa, bánh mì, bàn chải đánh răng, kem đánh răng) ...

Câu hỏi đặt ra là khách hàng thường mua những sản phẩm gì? Sau khi mua một sản phẩm thì họ mua thêm những sản phẩm nào? Như vậy, việc phân tích giỏ hàng sẽ trả lời những câu hỏi trên.

Do việc đầu ra tương ứng cho mỗi đầu vào là không thể biết trước nên đây thuộc nhóm học không giám sát (unsupervised learning).

Phân tích giỏ hàng là một loại khai thác dữ liệu (data mining) xác định các mẫu hành vi của người tiêu dùng trong bất kỳ môi trường bán lẻ nào. Nói một cách đơn giản, phân tích giỏ thị trường trong khai thác dữ liệu kiểm tra các loại mặt hàng đã được mua cùng nhau

## 1.2 Mô tả bộ dữ liệu

Bộ dữ liệu được lấy từ bài viết Market Basket Optimization trên trang kaggle <https://www.kaggle.com/datasets/devchauhan1/market-basket-optimisationcsv>  
Bộ dữ liệu có 20 thuộc tính với những thông tin:

- Shrimp : Tôm
- almonds : Hạnh nhân
- avocado : Bơ
- vegetables mix : Rau củ hỗn hợp
- green grapes : Nho xanh
- whole weat flour : Bột mì nguyên chất
- yams : Khoai
- cottage cheese : Phô mai tươi
- energy drink : Nước tăng lực
- tomato juice : Nước cà chua
- low fat yogurt : sữa chua ít béo
- green tea : Trà xanh
- honey : Mật ong
- salad : Rau
- mineral water: Nước khoáng
- salmon : Cá hồi
- antioxydant juice : Nước chống oxy hóa
- frozen smoothie
- spinach
- olive oil

## 2 Tối ưu hóa phân tích giỏ hàng bằng thuật toán Apriori

### 2.1 Tiền xử lý dữ liệu

#### 2.1.1 Xem xét tổng quan dữ liệu

- Xem xét tổng quan bộ dữ liệu:

	shrimp	almonds	avocado	vegetables mix	green grapes	whole wheat flour	yams	cottage cheese	energy drink	tomato juice	low fat yogurt	green tea	honey	salad	mineral water	salmon	antioxydant juice	frozen smooth
0	burgers	meatballs	eggs	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	chutney	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	turkey	avocado	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	mineral water	milk	energy bar	whole wheat rice	green tea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	low fat yogurt	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Hình 2.1 Tổng quan bộ dữ liệu

- Số lượng dữ liệu:

```
len(data)
```

7500

Hình 2.2 Số lượng dữ liệu

- Kiểm tra các giá trị null trong bộ dữ liệu:

```

Data columns (total 20 columns):
#   Column                Non-Null Count  Dtype
---  -
0   shrimp                 7500 non-null   object
1   almonds                5746 non-null   object
2   avocado                4388 non-null   object
3   vegetables mix         3344 non-null   object
4   green grapes           2528 non-null   object
5   whole wheat flour      1863 non-null   object
6   yams                   1368 non-null   object
7   cottage cheese         980 non-null    object
8   energy drink            653 non-null    object
9   tomato juice           394 non-null    object
10  low fat yogurt         255 non-null    object
11  green tea              153 non-null    object
12  honey                   86 non-null     object
13  salad                   46 non-null     object
14  mineral water          24 non-null     object
15  salmon                  7 non-null      object
16  antioxidant juice      3 non-null      object
17  frozen smoothie        3 non-null      object
18  spinach                 2 non-null      object
19  olive oil              0 non-null      float64
dtypes: float64(1), object(19)
memory usage: 1.1+ MB

```

Hình 2.3 Kiểm tra giá trị Null

Kết quả trên cho thấy rằng ngoại trừ thuộc tính Shrimp là đầy đủ, các dữ liệu khác ít nhiều đều thiếu giá trị.

Ở đây ta thấy mọi giá trị đều có kiểu dữ liệu object trừ thuộc tính olive oil. Tuy nhiên, vì đây là thuộc tính rỗng nên ta không quan tâm.

## 2.2 Xây dựng mô hình

Trước tiên, chúng ta cần transaction dataset:

```

## We need a list of transactions
transactions=[]
for i in range(0,7500): ## rows
    transactions.append([str(data.values[i,j]) for j in range (0,20) ])# Columns
## must be strings

```

Hình 2.4 Transaction dataset

Tiếp theo ta sẽ sử dụng thuật toán Apriori để xây dựng mô hình:

```

from apyori import apriori

rules=apriori(transactions=transactions,
               min_support=0.003,## (min 3 times per day * 7 days)/7501
               min_confidence=0.2,## you can try different values depending on the business requirements
               min_lift = 3, ## relevance of the rule... good lift is at least 3. Rules below 3 are not relevant--based on the experience
               min_length=2, ## 2 products to have at the end, A and B
               max_length=2)

```

Hình 2.5 Xây dựng mô hình bằng thuật toán Apriori

Ở đây ta có `min_support` là một điều kiện giúp loại bỏ các tập không phổ biến trong bất kỳ cơ sở dữ liệu.

`min_confidence` là độ tin cậy tối thiểu. Có thể thử các giá trị khác nhau tùy thuộc vào yêu cầu.

### Trực quan hóa kết quả:

Ta sẽ đưa kết quả vừa có được vào khung:

```

def inspect(results):
    lhs = [tuple(result[2][0][0])[0] for result in results]
    rhs = [tuple(result[2][0][1])[0] for result in results]
    supports = [result[1] for result in results]
    confidences = [result[2][0][2] for result in results]
    lifts = [result[2][0][3] for result in results]
    return list(zip(lhs, rhs, supports, confidences, lifts))

resultsinDataFrame = pd.DataFrame(inspect(results), columns = ['Left Hand Side', 'Right Hand Side', 'Support', 'Confidence', 'Lift'])

```

Hình 2.6 Kết quả trực quan

## 2.3 Đánh giá mô hình

Bước tiếp theo là đánh giá mô hình dựa trên kết quả:

```
resultsinDataFrame.nlargest(n=10, columns="Lift") ## 10 rows
```

	Left Hand Side	Right Hand Side	Support	Confidence	Lift
3	fromage blanc	honey	0.003333	0.245098	5.178128
0	light cream	chicken	0.004533	0.290598	4.843305
2	pasta	escalope	0.005867	0.372881	4.700185
8	pasta	shrimp	0.005067	0.322034	4.514494
7	whole wheat pasta	olive oil	0.008000	0.271493	4.130221
5	tomato sauce	ground beef	0.005333	0.377358	3.840147
1	mushroom cream sauce	escalope	0.005733	0.300699	3.790327
4	herb & pepper	ground beef	0.016000	0.323450	3.291555
6	light cream	olive oil	0.003200	0.205128	3.120612

Hình 2.7 Đánh giá mô hình

Ở đây ta có Left Hand Side là những mặt hàng mà khách hàng sẽ mua đầu tiên và Right Hand Side là những mặt hàng mà khách hàng có thể mua kế tiếp.

Ta có thể thấy các giá trị mức độ tăng đều lớn hơn 1, tức là phản hồi mục tiêu có kết quả cao hơn mức trung bình. Do đó, các quy tắc kết hợp sẽ cải thiện cơ hội của kết quả.



## Tài liệu tham khảo

[1] Lê Hải Hà, Hệ hỗ trợ quyết định, Bài giảng, Bộ môn Toán tin – Viện Toán ứng dụng và Tin học – Đại học Bách khoa Hà Nội (lưu hành nội bộ).