



Trường Đại Học Sài Gòn

Báo cáo tổng kết đề tài nghiên cứu khoa học

# DỰ ĐOÁN SỰ HÀI LÒNG CỦA KHÁCH HÀNG BẰNG CÁC PHƯƠNG PHÁP HỌC MÁY/HỌC SÂU

Giáo viên hướng dẫn: Đỗ Như Tài

Thành viên: Nguyễn Hoàng Thanh Phương, Văn Hoàng Như Ý, Lê Hồng Sơn, Đỗ Hữu Lộc

Thành phố Hồ Chí Minh, ngày 18 tháng 05 năm 2025

# AGENDA

- ▶ 1. Introduction
- ▶ 2. Motivation
- ▶ 3. Related works
- ▶ 4. Proposed method
- ▶ 5. Experiment and Results
- ▶ 6. Conclusion

# 1. INTRODUCTION

*The situation:*



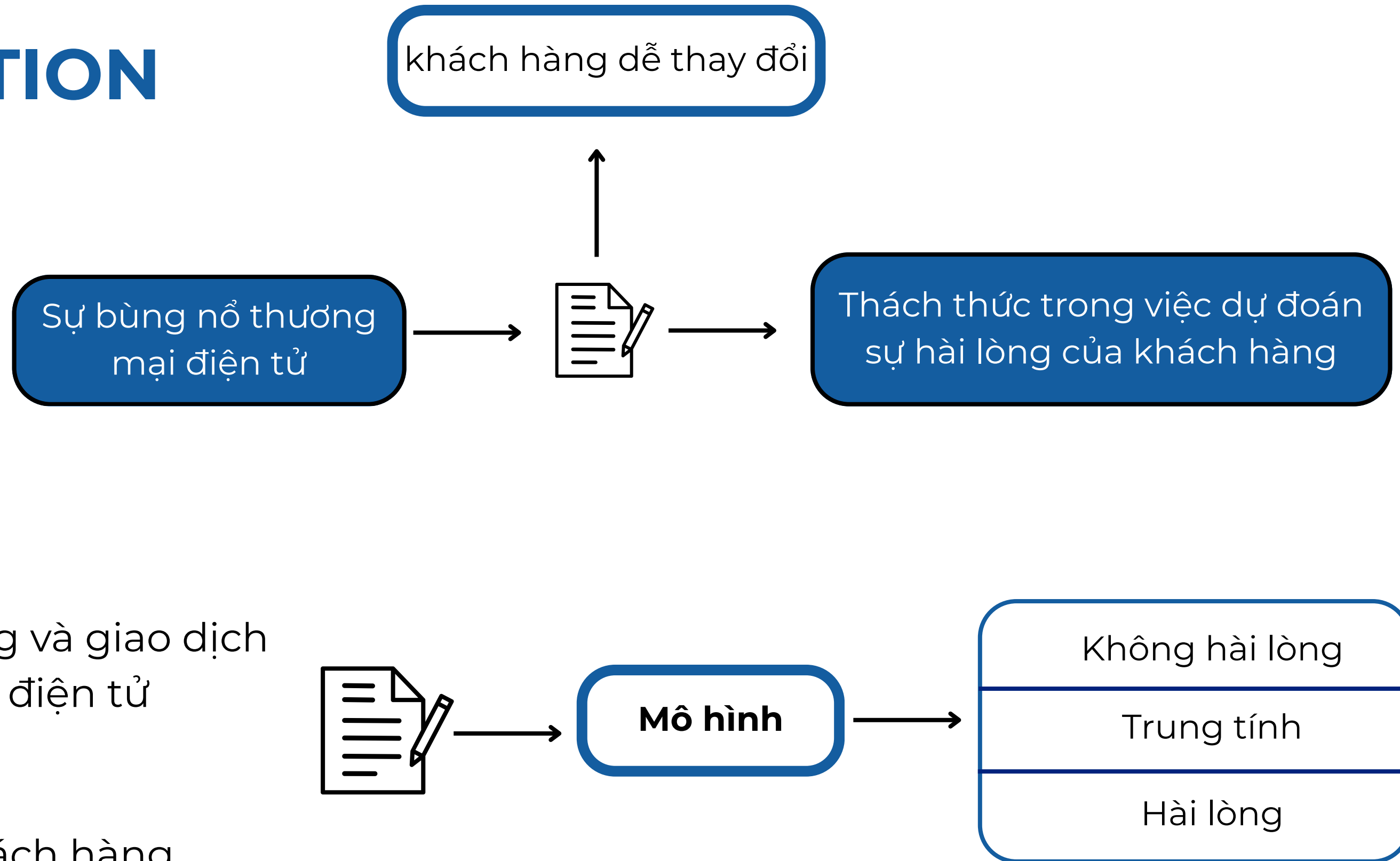
*The problem:*

**Input:**

Các đặc trưng người dùng và giao dịch trong dữ liệu thương mại điện tử

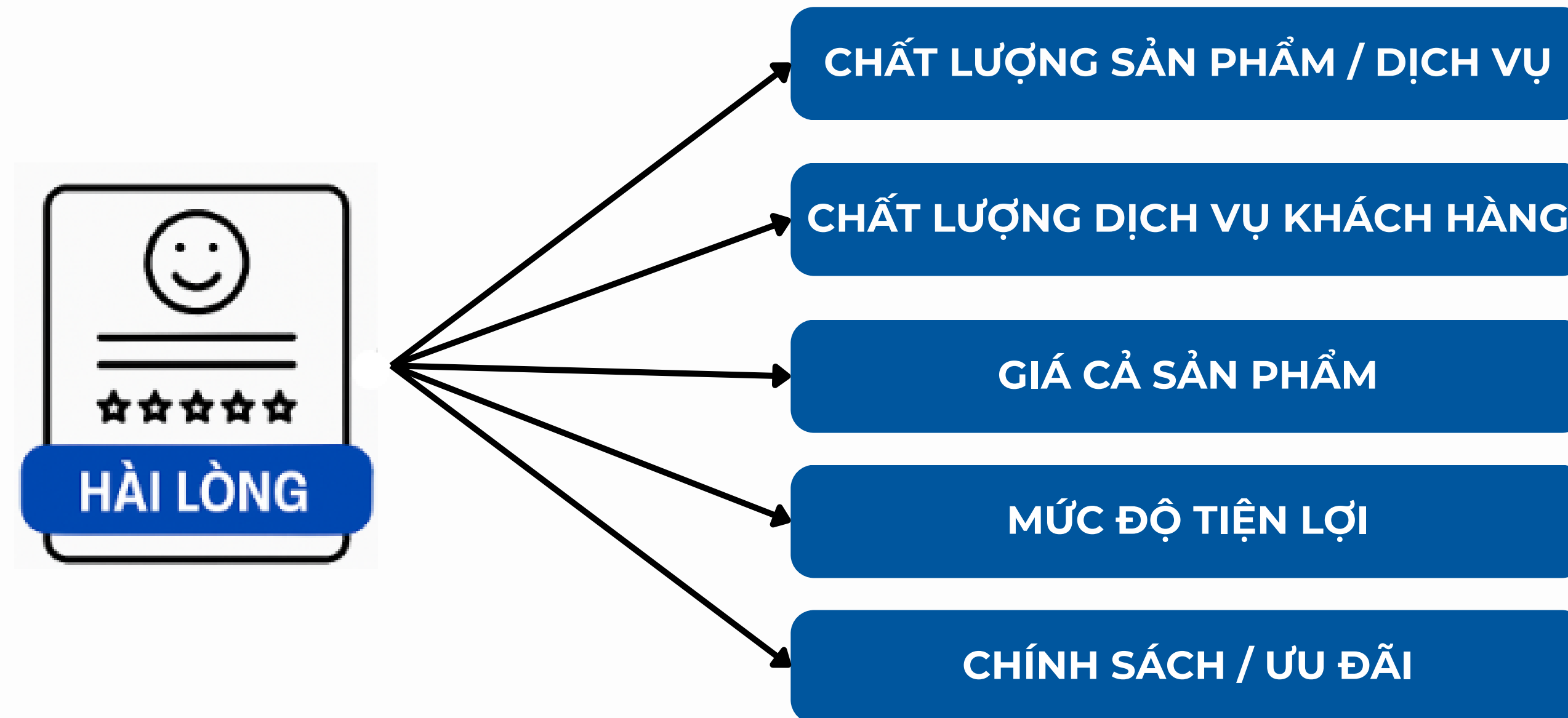
**Output:**

Cảm xúc, thái độ của khách hàng



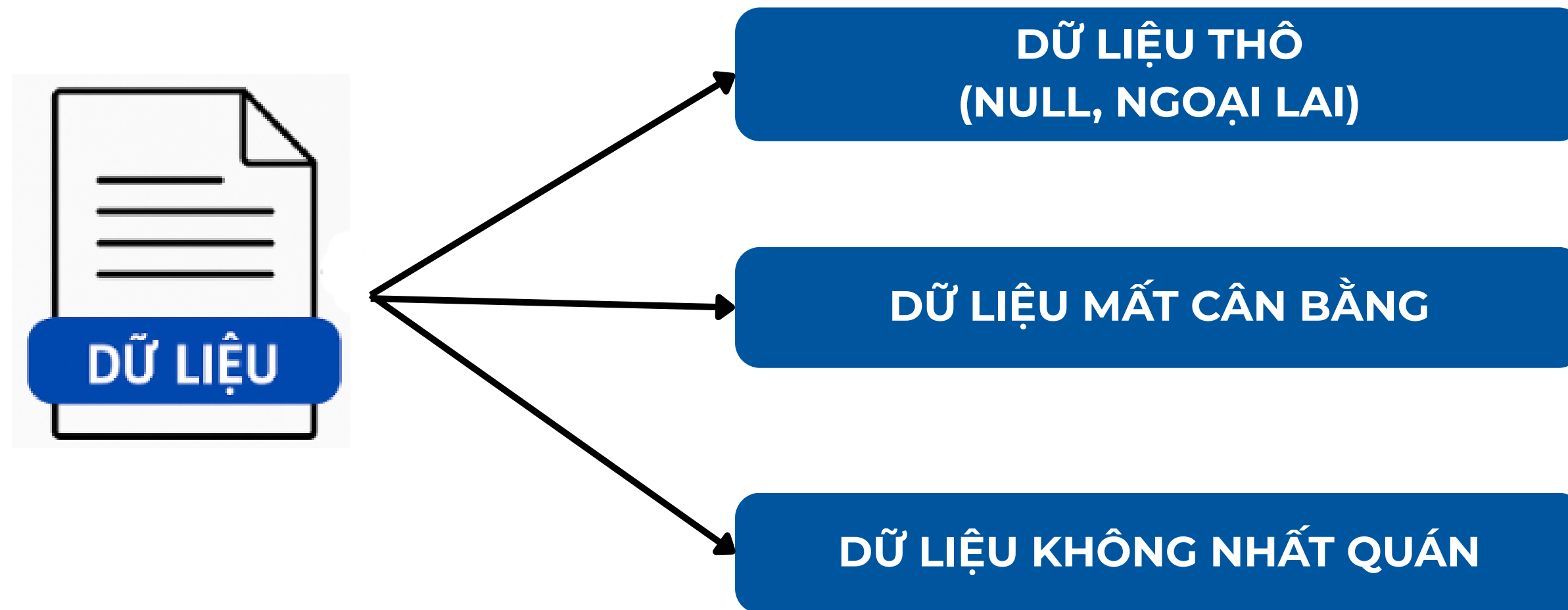
# 1. INTRODUCTION

- Khách hàng thường cung cấp phản hồi chủ quan, mang tính cảm xúc, sử dụng ẩn dụ, từ lóng hoặc không đầy đủ về mức độ hài lòng

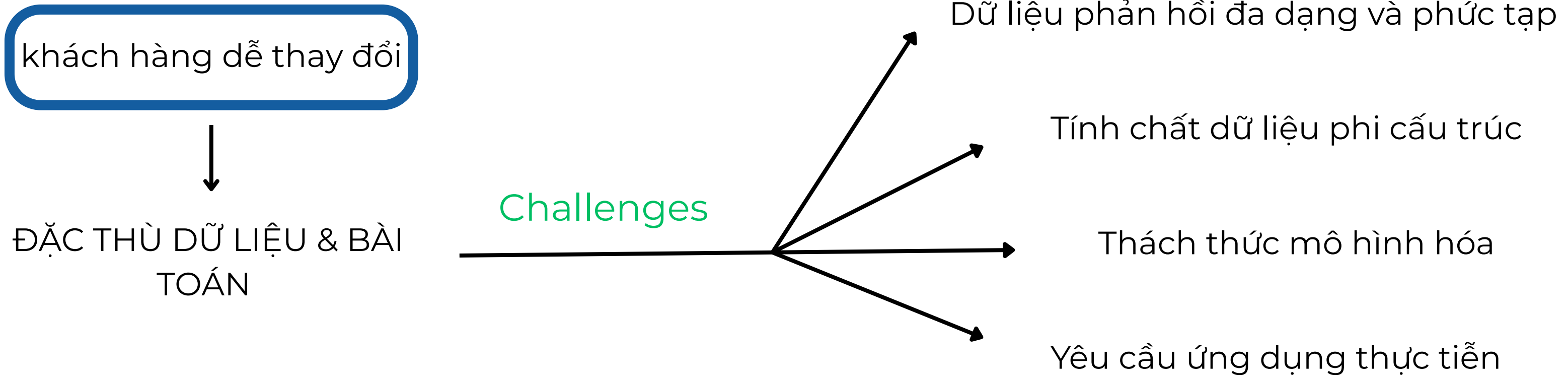


# 1. INTRODUCTION

- Dữ liệu thu thập được thường tồn tại nhiều vấn đề tiềm ẩn ảnh hưởng đến độ chính xác của mô hình dự đoán.



# 1. INTRODUCTION



**Làm thế nào để tiền xử lý và chọn mô hình phù hợp để dự đoán chính xác sự hài lòng từ dữ liệu thương mại điện tử?**

# 1. INTRODUCTION

## Dataset

Source: [Kaggle](#)

26 thuộc tính &  
206.969 mẫu



good 91388  
bad 68538  
neutral 47043



BALANCE

user_id
age
Gender
Date_Registered
Is_current_loyalty_program_member
loyalty_points_redeemed
loyalty_tier
Received_tier_discount_percentage
Received_card_discount_percentage
Received_coupon_discount_percentage
product_category
Product_value
transaction_id

order_id
payment_method
payment_datetime
purchased_datetime
purchase_medium
final_payment
released_date
estimated_delivery_date
received_date
shipping_method
tracking_number
customer_experience



Dữ liệu có nhiều biến chưa tương quan

Phân phối biến đa dạng, không đồng đều

Dữ liệu bị thiếu và có một số ngoại lai cần xử lý

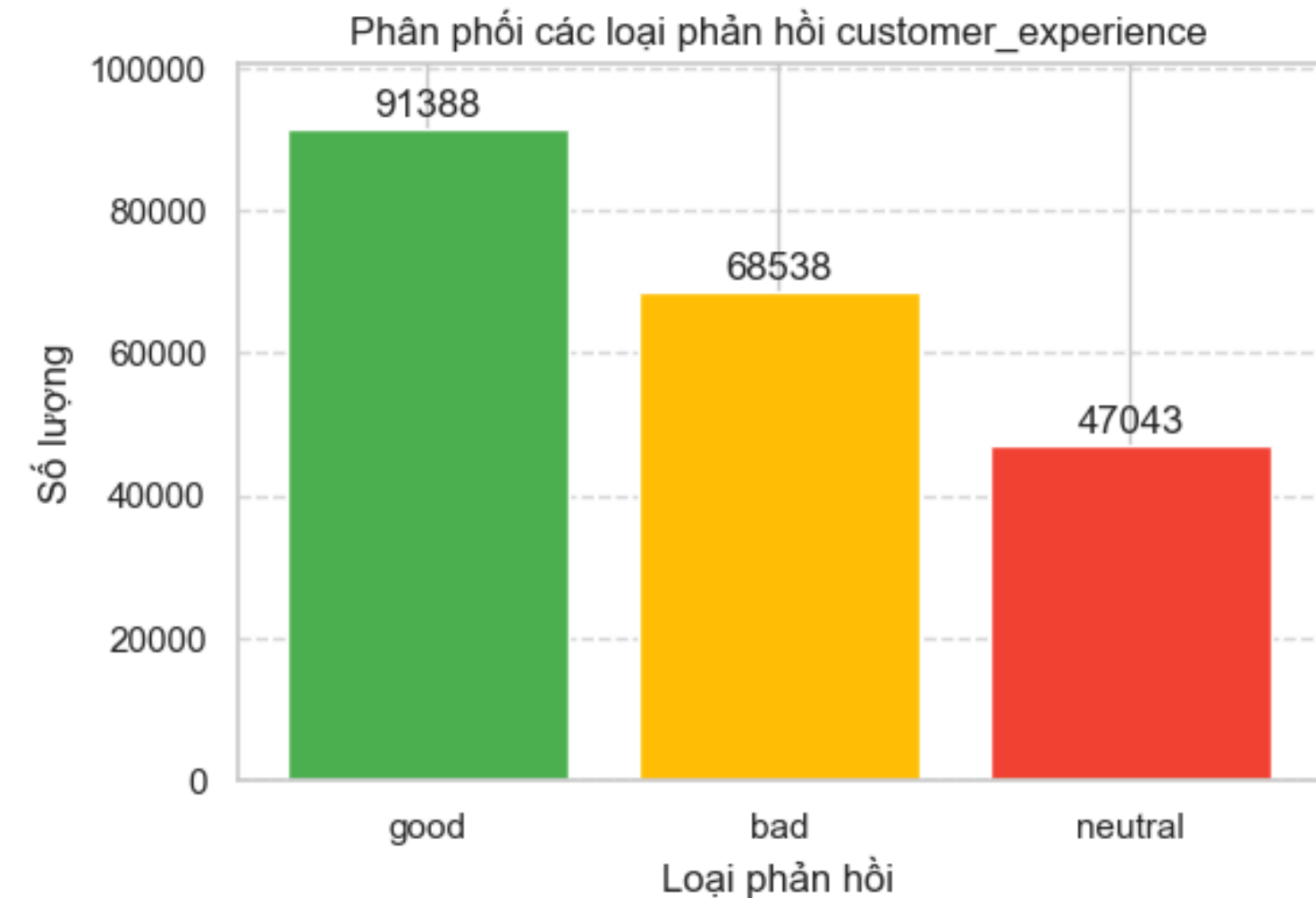
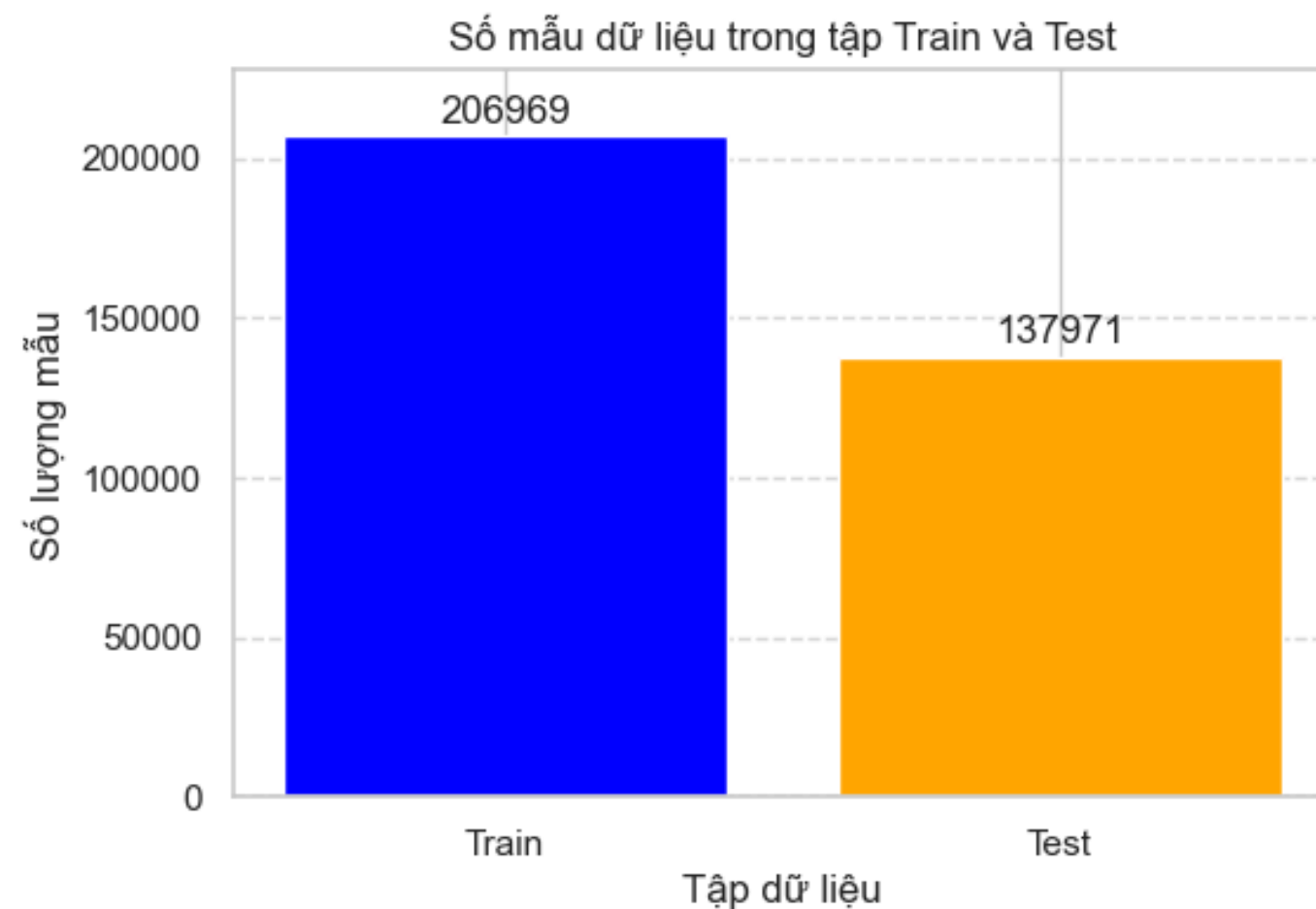
Dữ liệu chưa cân bằng giữa các lớp



**Cần xử lý kỹ càng trước khi xây dựng mô hình**

← Biến mục tiêu (good, neutral, bad)

## 2. MOTIVATION



### Nghiên cứu hướng đến các mục tiêu cụ thể:

- Phân tích các yếu tố ảnh hưởng đến sự hài lòng của khách hàng trên sàn thương mại điện tử.
- Xây dựng và huấn luyện các mô hình học máy/học sâu để dự đoán mức độ hài lòng khách hàng dựa trên dữ liệu phản hồi.
- Đánh giá và so sánh hiệu quả các mô hình dựa trên các chỉ số phổ biến như Accuracy, F1-Score và RMSE.
- Đề xuất mô hình tối ưu có khả năng ứng dụng thực tế, hỗ trợ doanh nghiệp cải tiến dịch vụ và ra quyết định chính xác hơn.



### 3. RELATED WORKS

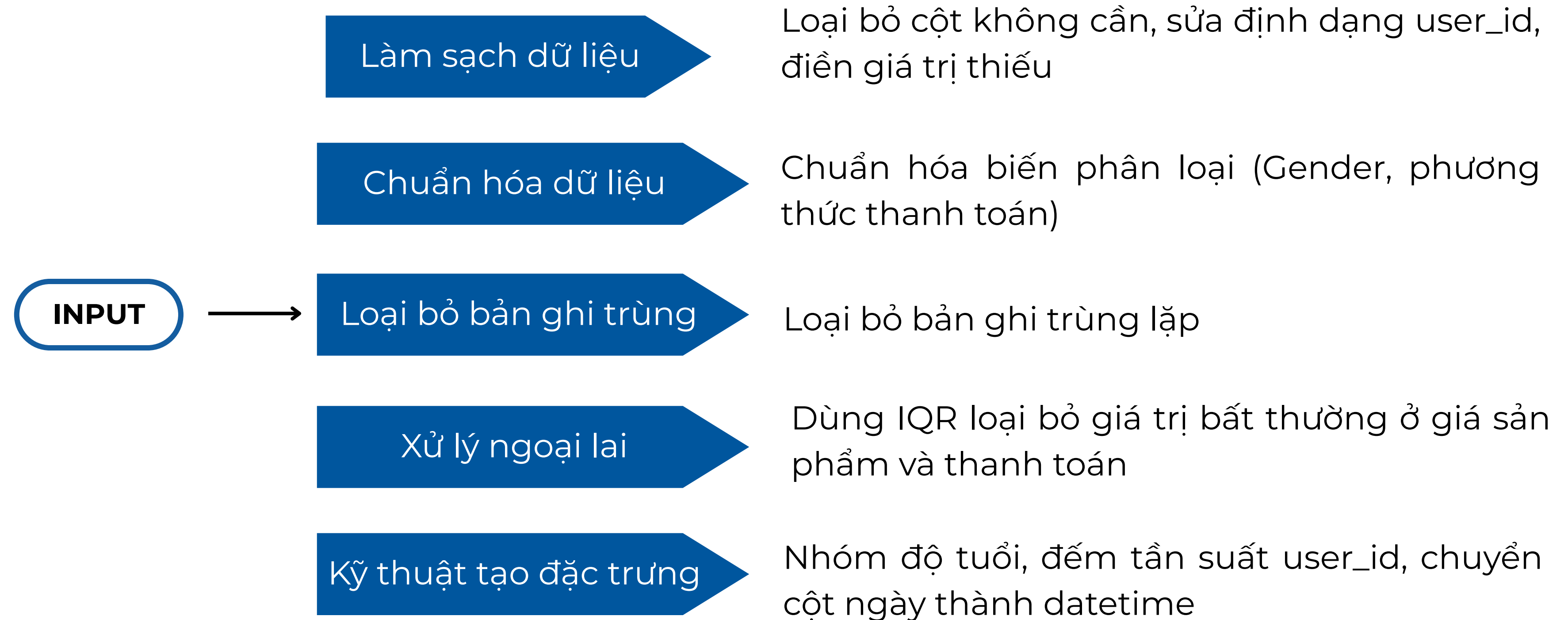
- Graph Neural Networks to Predict Customer Satisfaction Following Interactions with a Corporate Call Center
- A Machine Learning Approach to Detect Customer Satisfaction From Multiple Tweet Parameters

## 4. PROPOSED METHOD

- Xây dựng mô hình sử dụng mô hình bảng cho vấn đề này (Đánh giá tính ổn định của mô hình trên các tập dữ liệu nhỏ).
- Sử dụng phương pháp Lấy mẫu dữ liệu và mất cân bằng trên mô hình đề xuất.

# 4. PROPOSED METHOD

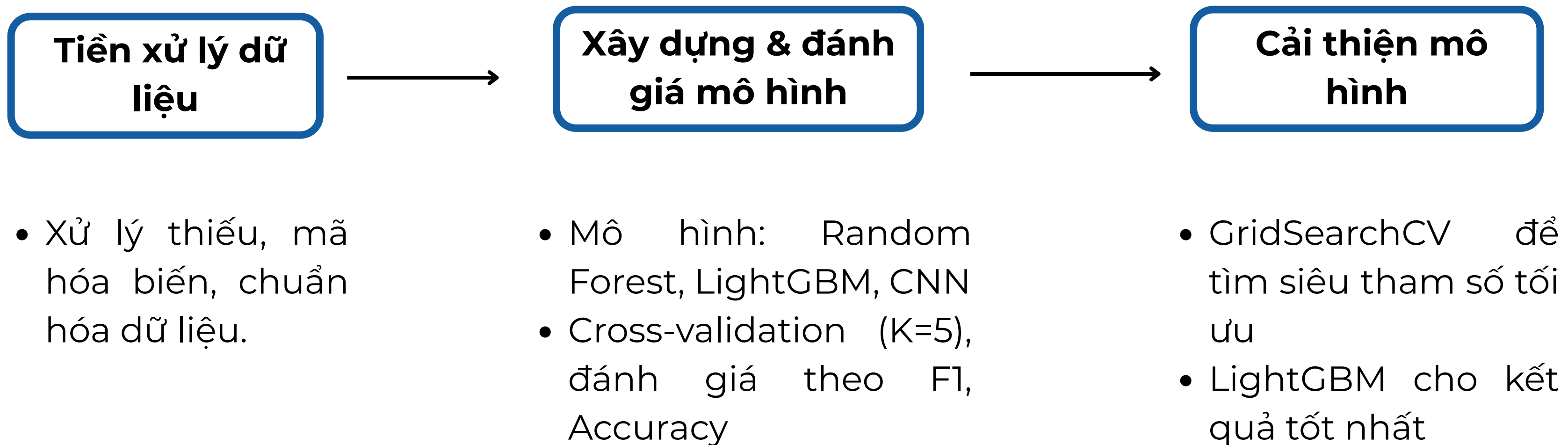
## Giải thích chi tiết: Tiền xử lý tập dữ liệu



## 4. PROPOSED METHOD

### Giải thích chi tiết: Xây dựng và đánh giá mô hình dự đoán

- Áp dụng các kỹ thuật tiền xử lý và lựa chọn đặc trưng phù hợp với từng mô hình.
- Đánh giá hiệu quả các mô hình học máy/học sâu và chọn mô hình tối ưu nhất.




# 4. PROPOSED METHOD

## Các mô hình:



Xây dựng LightGBM, mô hình boosting dựa trên cây quyết định, tối ưu tốc độ và hiệu quả trên dữ liệu lớn, nhiều đặc trưng.



Triển khai Random Forest, mô hình ensemble đơn giản, dễ huấn luyện, dùng để so sánh cơ sở với các mô hình tiên tiến hơn.



Ứng dụng CNN, sử dụng convolution 1D để khám phá khả năng học sâu với các đặc trưng tuần tự hoặc đã được mã hóa từ dữ liệu bảng.

# 5. EXPERIMENTS AND RESULTS

## 5.1. Mô hình truyền thống (Random Forest)

### Kết quả đánh giá:

- Là một mô hình ensemble đơn giản và phổ biến.
- Dễ triển khai, tốc độ nhanh nhưng không đạt hiệu suất cao như LightGBM trong bài toán này.

### Chi tiết theo từng lớp:

Lớp	Precision	Recall	F1-Score	Support
0	0.57	0.54	0.55	13,707
1	0.52	0.63	0.57	9,408
2	0.69	0.64	0.67	17,416

# 5. EXPERIMENTS AND RESULTS

## 5.2. Các mô hình học sâu (CNN)

**Kết quả đánh giá:**

**Chi tiết theo từng lớp:**

# 5. EXPERIMENTS AND RESULTS

## 5.3. Mô hình LightGBM

### Kết quả đánh giá:



- Accuracy: 0.6731
- F1-Score: 0.6741
- Macro Avg Precision: 0.68
- Macro Avg Recall: 0.68
- Macro Avg F1: 0.68



### Chi tiết theo từng lớp:

Lớp	Precision	Recall	F1-Score	Support
0	0.60	0.67	0.63	13,707
1	0.73	0.73	0.73	9,408
2	0.71	0.65	0.68	17,416



# 6. CONCLUSION

- LightGBM cho hiệu suất tốt nhất với Accuracy và F1-Score vượt trội.
- Có thể cải thiện bằng kỹ thuật xử lý mất cân bằng, chọn đặc trưng và tuning hyperparameter.
- Random Forest phù hợp làm baseline, không tối ưu cho bài toán.

*Bảng kết quả đánh giá mô hình (Model Evaluation Metrics Table).*

Chỉ số	LightGBM	Random Forest
Accuracy	6.731	6.043
F1-Score	6.741	6.054
F1 lớp 0	63	55
F1 lớp 1	73	57
F1 lớp 2	68	67
Ổn định	Tốt	Trung bình
Tổng quát	Tốt	Cần cải thiện

# 6. CONCLUSION

## Ý nghĩa thực tiễn

- Giúp doanh nghiệp dự đoán sớm mức độ hài lòng khách hàng.
- Tối ưu chiến lược tiếp thị, nâng cao hiệu quả và giữ chân khách hàng.
- Mô hình có thể ứng dụng rộng rãi trong nhiều lĩnh vực khác.
- Hỗ trợ chuyển đổi số và nâng cao trải nghiệm khách hàng.

## Ý nghĩa thực tiễn

- Accuracy và F1-Score còn thấp, dữ liệu có thể chứa nhiều, outlier.
- Giải pháp:
  - Tuning tham số nâng cao, ensemble học máy.
  - Loại bỏ outlier, feature engineering, chuẩn hóa dữ liệu.



Trường đại  
học Sài Gòn

# THANK YOU!

