# Titanic: Machine Learning from Disaster

## Predicting Survival Using Classification Models

Nguyễn Thành Đạt--Nguyễn Công Trí--Nguyễn Tấn Duy

## Introduction

### Problem Definition

- **Input: Dữ liệu hành khách gồm các thuộc tính như Age, Sex, Pclass, Fare, Cabin, Embarked, SibSp, Parch.**
- **Output: Nhãn phân loại Survived = 1 (sống sót) hoặc Survived = 0 (thiệt mạng).**

### Objective

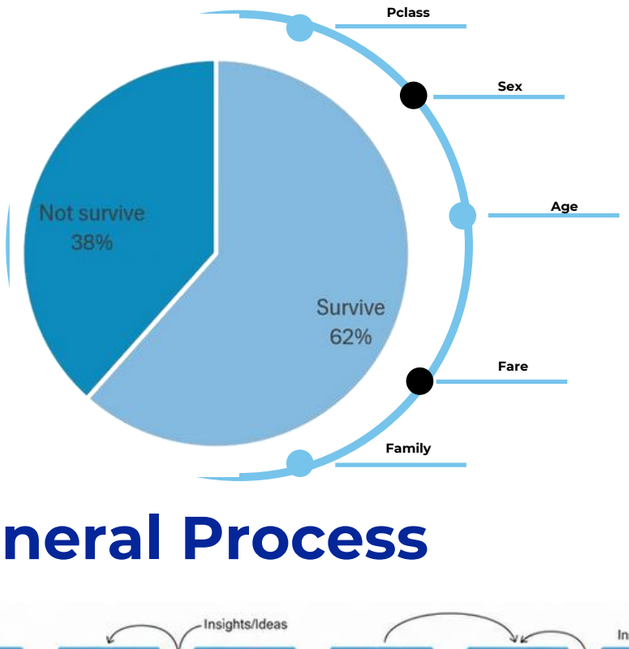
- **Xây dựng mô hình phân loại khả năng sống sót dựa trên thông tin hành khách (tuổi, giới tính, hạng vé...).**

### Challenges

- **Dữ liệu bị thiếu (null/nan)**
- **Dữ liệu ngoại lai nhiều**

### Dataset

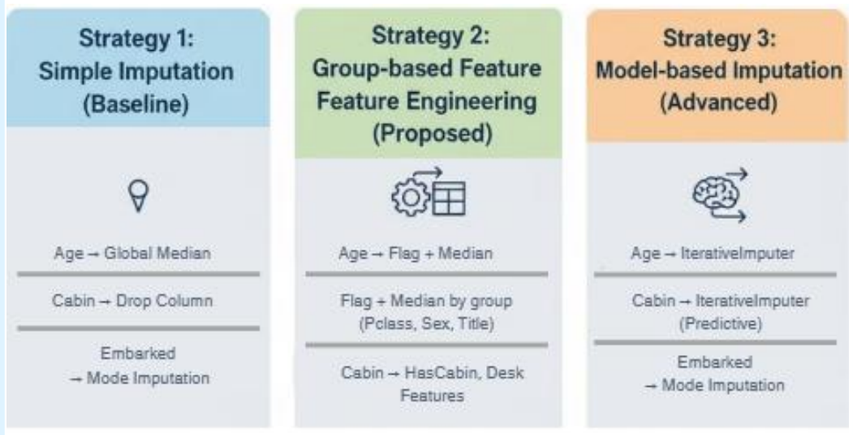
- **Titanic - Machine Learning from Disaster**
- **Data source https://www.kaggle.com/competitions/titanic/overview**



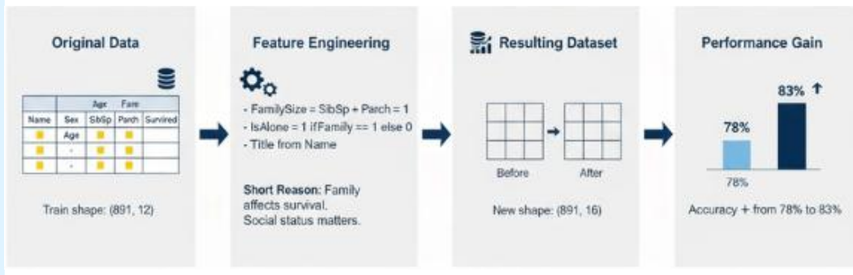

### General Process


## Methods

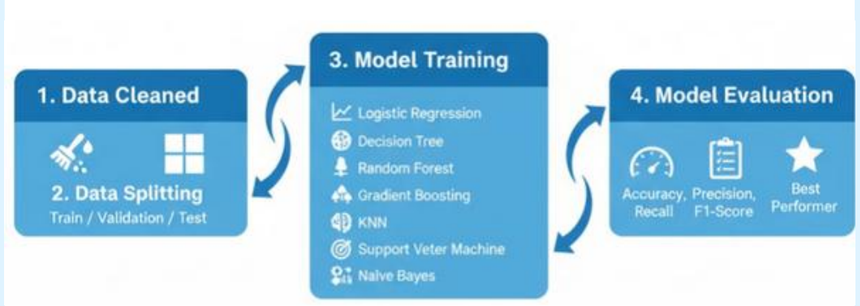
### Data Processing




### Feature creation

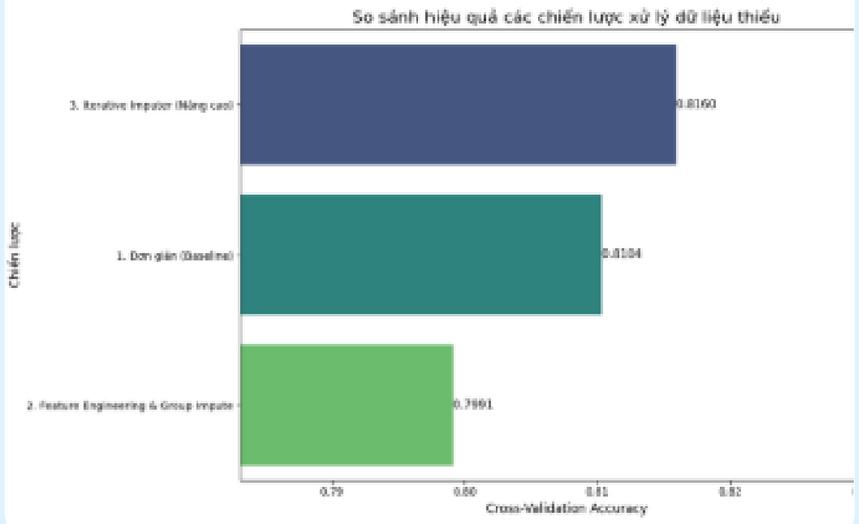



### Model Selection




### Feature Selection


### Process null/nan variable



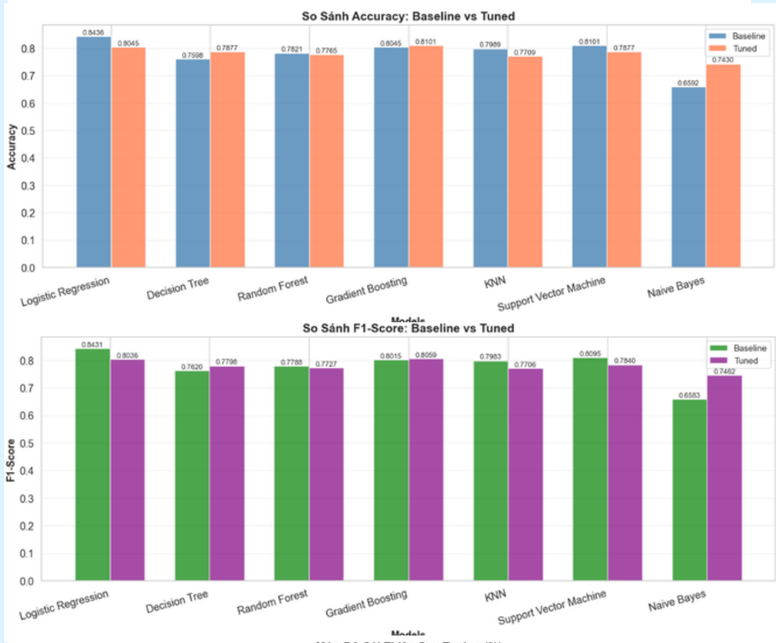

## Results and conclusion

### Experiment Results

| Strategy | Models | Accuracy |
|---|---|---|
| Baseline | Logistic Regression | 0.8104 |
| Feature enginering and Group Impute | SVM | 0.7991 |
| Interactive imputer Optimization | Gradient Boosting | 0.8160 |

### Top 3 Models:

- **LogisticRegression**
- **Gradient Boosting**
- **SVM**




### Key Insights:

- **Tiền xử lý dữ liệu quyết định lớn đến hiệu suất.**
- **Mô hình đơn giản (Logistic Regression) vẫn ổn định và hiệu quả.**

### Conclusion:

- **Mô hình Logistic Regression cho hiệu năng ổn định, dễ huấn luyện và tránh overfitting.**
- **Gradient Boosting cải thiện độ chính xác khi được tối ưu siêu tham số.**