Dự đoán Khả năng Sống sót Sau Thảm họa Titanic bằng Các Thuật toán Học Máy

Nguyễn Đăng Khoa, Lê Thị Trúc Ly, Lê Đoàn Kim Ngân, Lâm Tú Nhi

Nội dung trình bày

01 Introduction

02 Related Work

03 Proposed methods

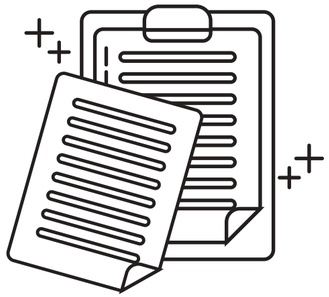
04 Experience and result

05 Conclusion

1. Introduction

Vấn đề:

- Input: Đặc điểm cá nhân và thông tin chuyến đi của khách hàng
- Output: Dự đoán khả năng sống sót của khách hàng (0: không sống sót, 1: sống sót)



MODELS

Not survived (0): 61.6%

Survived (1): 38.4%

1. Introduction

Dữ liệu:

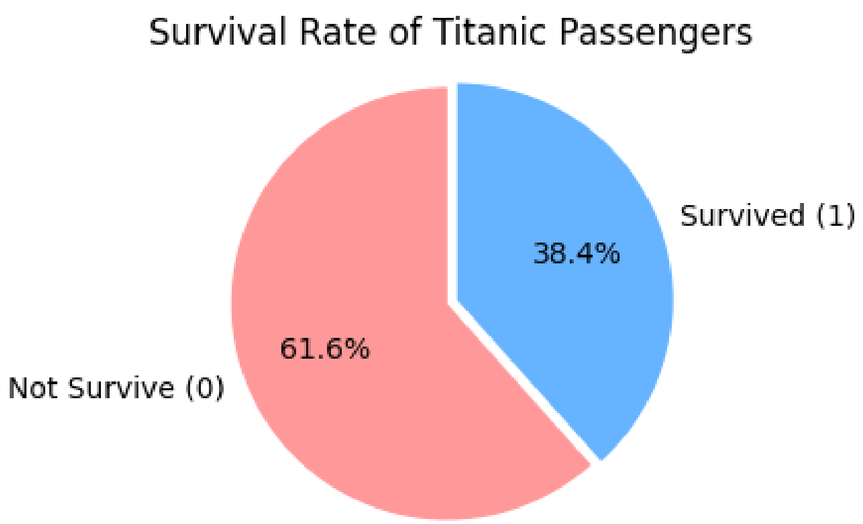
- **Bộ dữ liệu Titanic từ Kaggle (891 mẫu huấn luyện, 418 mẫu kiểm thử).**
- **Titanic - Machine Learning from Disaster**

PassengerId	Mã khách hàng	Age	Độ tuổi
Survived	Trạng thái sống	SibSp	Số lượng anh/ chị/ em hoặc vợ chồng đi cùng
Pclass	Hạng vé	Parch	Số lượng cha/ mẹ hoặc con đi cùng
Name	Họ tên khách hàng	Ticket	Số vé
Sex	Giới tính	Fare	Giá vé khách trả
Cabin	Số cabin	Embarked	Nơi lên tàu

1. Introduction

Thử thách:

- Dữ liệu chứa nhiều giá trị bị thiếu (missing value)
- Tính không cân bằng giữa số người sống sót và tử vong
- Nhiều biến định tính cần xử lý (Sex, PClass, Embarked, ...)
- Mô hình dữ liệu nhỏ dễ dẫn đến overfitting



	Missing Values	Percent (%)
Cabin	687	77.10
Age	177	19.87
Embarked	2	0.22

	Missing Values	Percent (%)
Cabin	327	78.23
Age	86	20.57
Fare	1	0.24

2. Related Work

Classical Models:

- Logistic Regression – mô hình tuyến tính cơ bản cho bài toán nhị phân

Ensemble Methods:

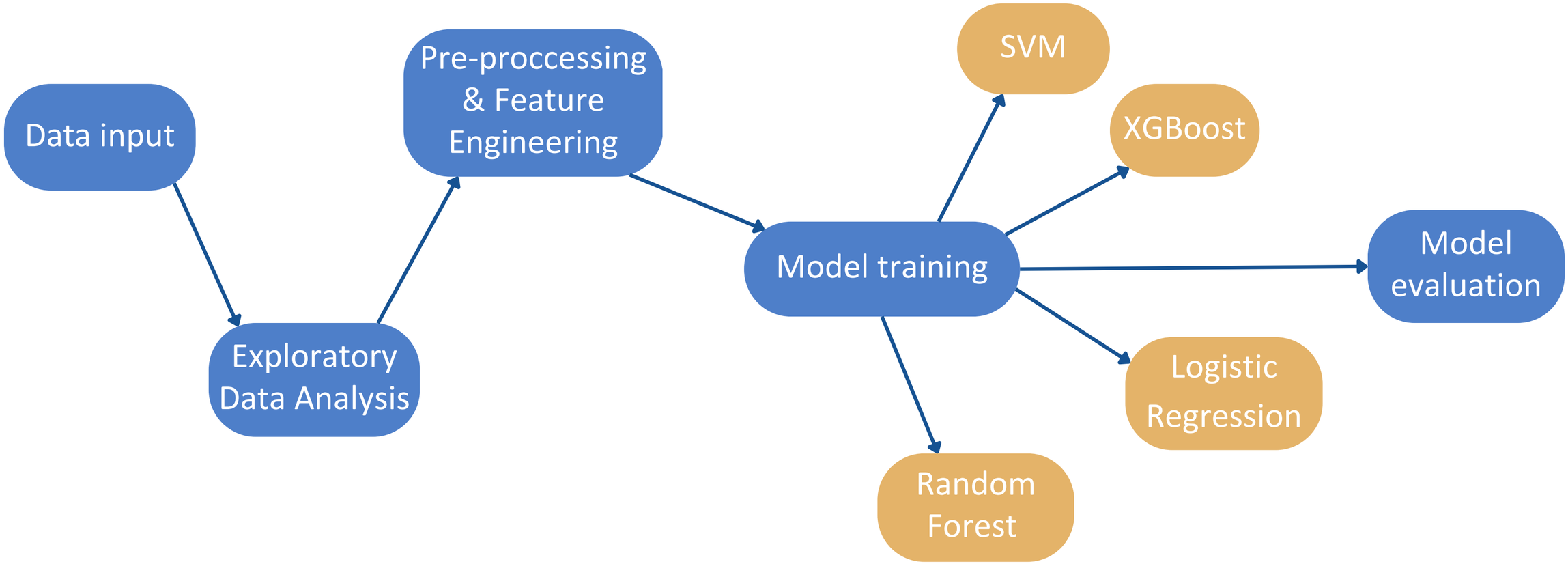
- Random Forest, Gradient Boosting (XGBoost) – cải thiện độ chính xác nhờ kết hợp nhiều cây quyết định
- Các mô hình này đã được chứng minh hiệu quả trong Titanic Dataset trên Kaggle

Feature Engineering Approaches:

- Tạo đặc trưng mới từ các cột như FamilySize, Title, Deck
- Xử lý dữ liệu bị thiếu bằng trung vị, mode, hoặc dự đoán bằng mô hình phụ trợ

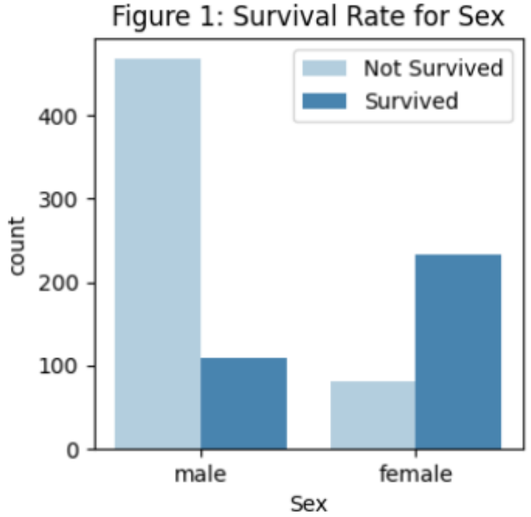
3. Proposed methods

Quá trình chính:

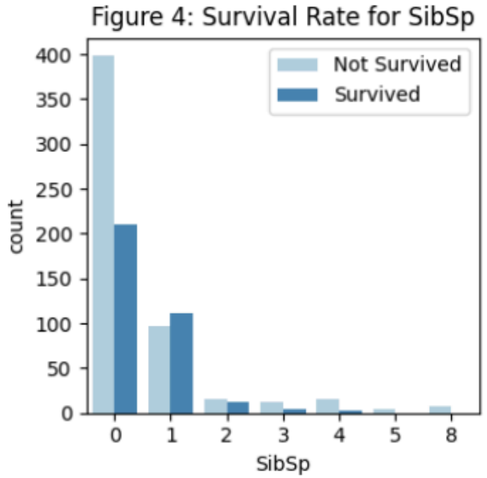


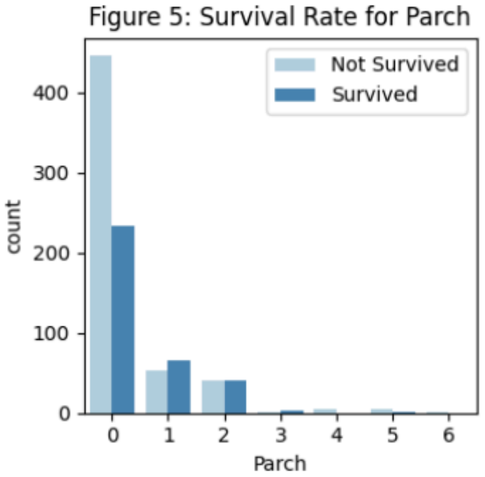
3. Proposed methods

Exploratory data analysis (EDA)



- Tỉ lệ sống sót ở nữ giới cao hơn nam giới do ưu tiên “phụ nữ và trẻ em trước”.
- Là yếu tố dự đoán mạnh mẽ về khả năng sống sót



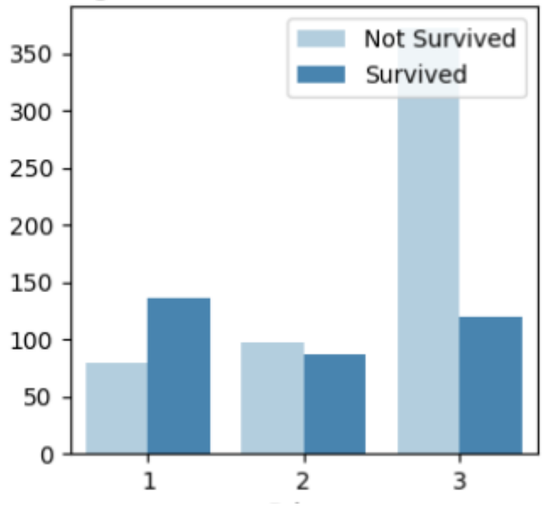


- Các gia định có quy mô vừa và nhỏ có nhiều cơ hội sống sót hơn những người đi một mình

3. Proposed methods

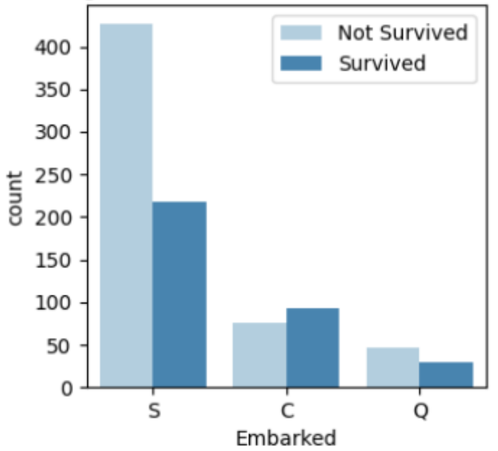
Exploratory data analysis (EDA)

Figure 3: Survival Rate for Pclass



- Hạng vé 1: có khả năng sống cao nhất trong 3 hạng vé
- Hạng vé 3: số người chết cao hơn rất nhiều so với số người sống
- Hạng vé đóng vai trò quan trọng trong khả năng sống của từng khách hàng

Figure 2: Survival Rate for Embarked



- Cảng S: số lượng khách đông nhất và có tỉ lệ không sống sót cao nhất
- Cảng C: số người sống sót và không sống sót gần bằng nhau nhưng tỉ lệ sống sót cao nhất trogn 3 cảng.

3. Proposed methods

Feature engineering

Cột mới	Ý nghĩa
Title	Phản ánh giới tính, địa vị, tầng lớp xã hội.
FamilySize	Quy mô gia đình ảnh hưởng tỉ lệ sống, đi cùng thành viên gia đình có tỉ lệ sống sót cao hơn (SibSp + Parch + 1).
IsChild	Đánh dấu trẻ em (dưới 12 tuổi). Trong Titanic, trẻ em thường được cứu trước theo nguyên tắc “women and children first”.
IsMother	Đánh dấu phụ nữ trưởng thành có con đi cùng. Mẹ thường được ưu tiên cứu cùng con. (Nữ, tiile: Mrs, độ tuổi: hơn 18 tuổi)
Deck	Boong (deck) tàu – ký tự đầu của Cabin (A, B, C, D...), gán Nan cho “U” - Unknow Deck thấp → gần đáy tàu → nguy hiểm hơn.

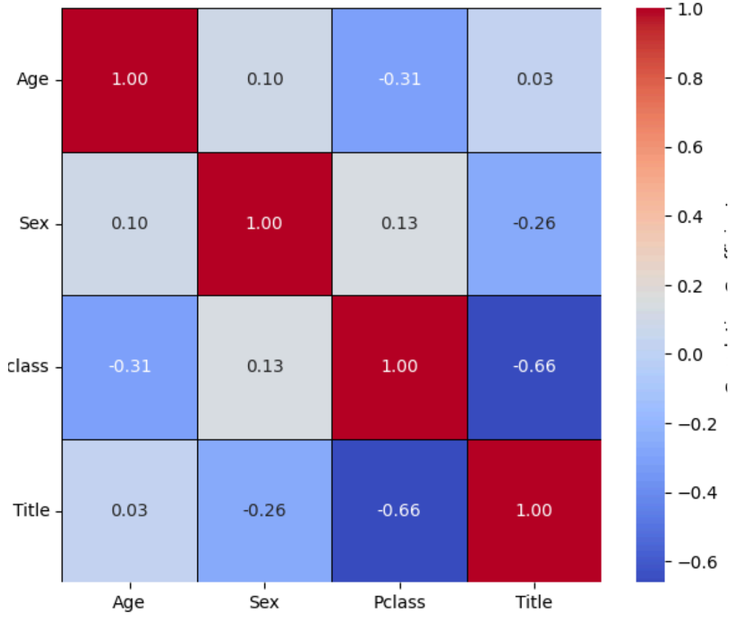
3. Proposed methods

Tiền xử lí dữ liệu:

- Xử lí các giá trị bị thiếu
  - Age (thiếu 20%): Điền bằng median dựa trên nhóm (Sex, Pclass, Title) để ước tính tuổi chính xác hơn.
  - Fare (thiếu 1): Điền bằng median của nhóm Pclass tương ứng.
  - Embarked (thiếu 2): Điền bằng giá trị phổ biến nhất.

	Missing Values	Percent (%)
Cabin	687	77.10
Age	177	19.87
Embarked	2	0.22

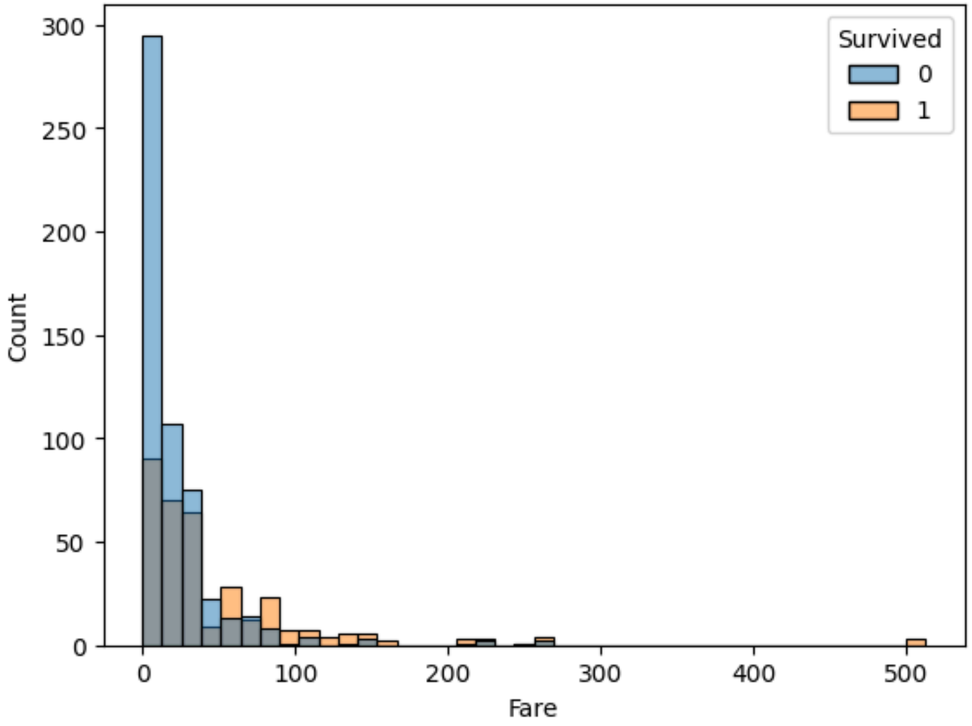
	Missing Values	Percent (%)
Cabin	327	78.23
Age	86	20.57
Fare	1	0.24

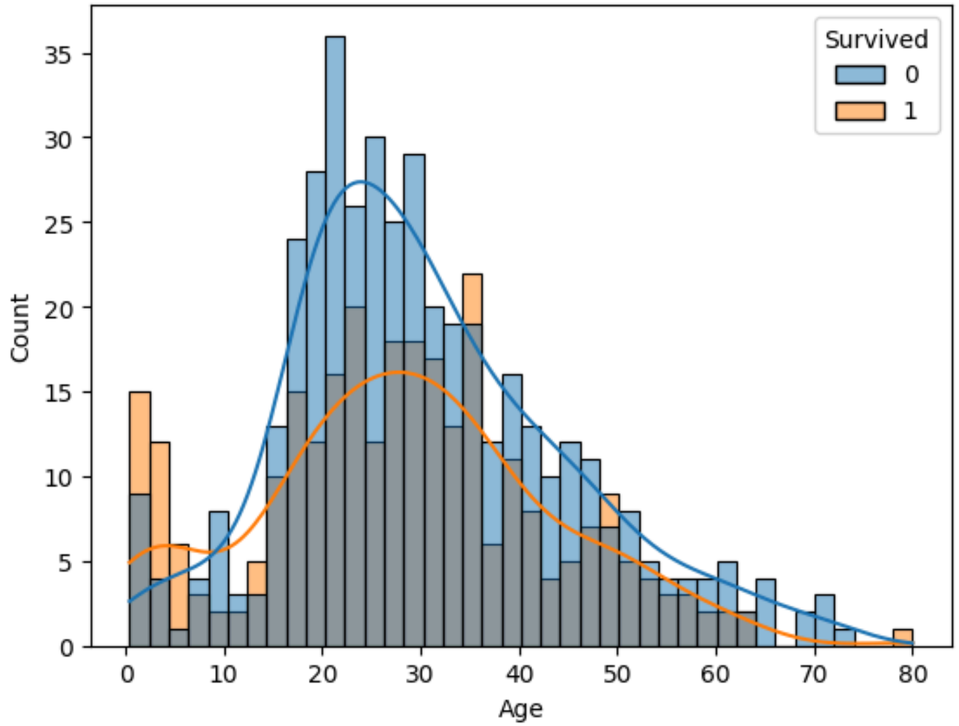


3. Proposed methods

Chuẩn hóa (Normalization):

- Áp dụng Log Transformation cho Fare và Age để giảm độ lệch (skewness) của dữ liệu.
- Các đặc trưng phân loại (Pclass, Sex, Embarked, ...) → OneHotEncder
- Các đặc trưng số (Age, Fare, FamilySize,...) → StandardScaler

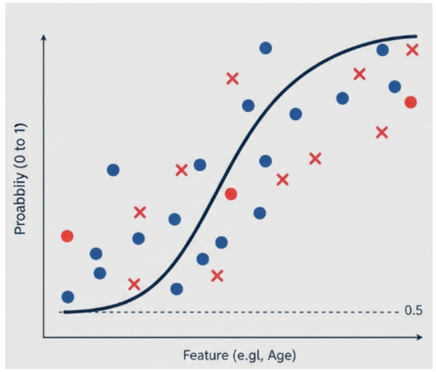




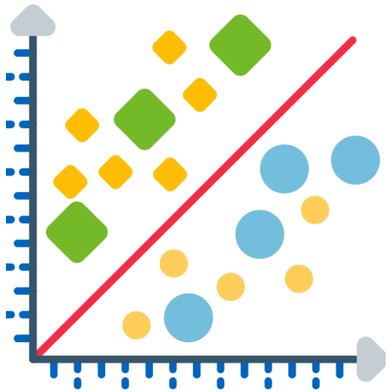
3. Proposed methods

Model training:

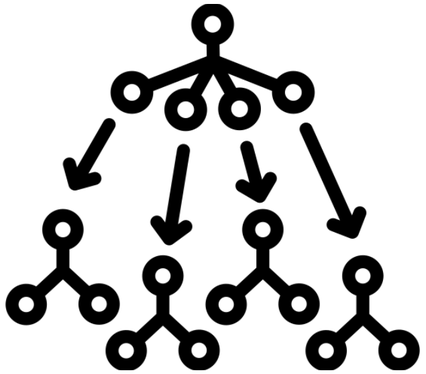
- Các Mô hình Được Thử nghiệm
  - Phân loại Cơ bản (Baseline): Logistic Regression, Support Vector Machine (SVM).
  - Mô hình Ensemble: Random Forest và XGBoost (Extreme Gradient Boosting).



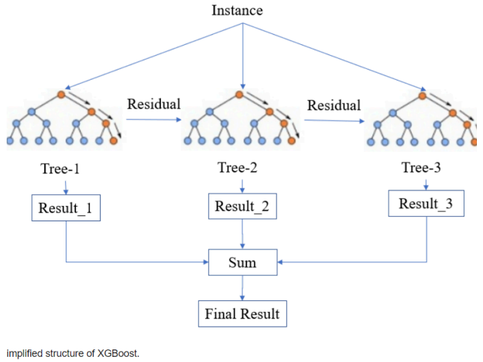
Logistic Regression



SVM



Random Frest



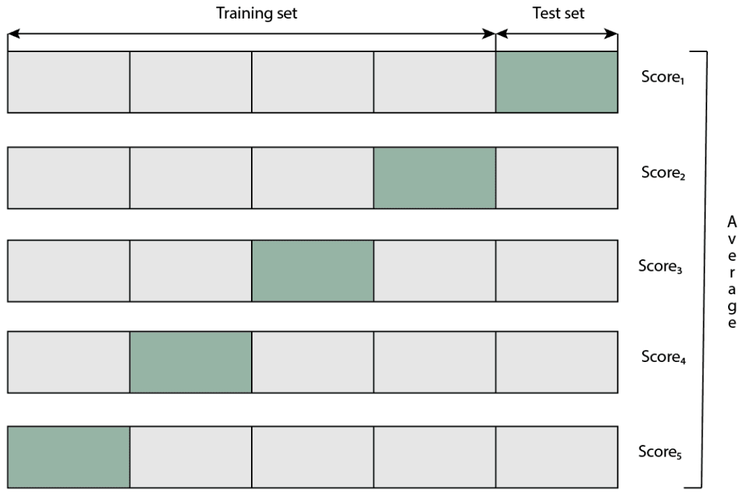
implified structure of XGBoost.

XGBoost

3. Proposed methods

Model evaluation:

- Đánh giá bằng Cross-Validation
  - Mục đích: Đánh giá độ ổn định và hiệu suất tổng quát của mô hìn.
  - Đảm bảo rằng tỉ lệ Survived được duy trì đồng đều trong mỗi 10 tập con, đặc biệt quan trọng khi lớp mục tiêu có sự mất cân bằng nhỏ (chỉ 38% sống sót).
  - Thực thi: Huấn luyện và đánh giá mô hình 10 lần, sau đó lấy giá trị trung bình của các chỉ số (Accuracy, F1 Score, ROC AUC).

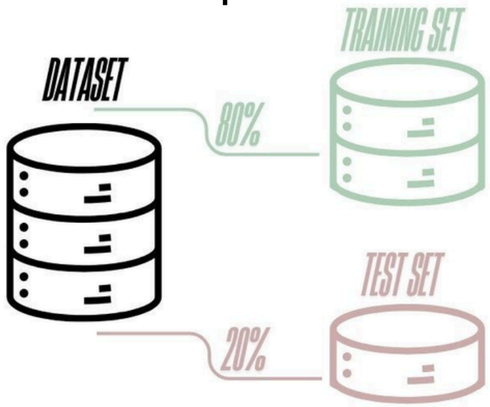


3. Proposed methods

Model evaluation:

- **Đánh giá bằng train_split_test:**
  - Mục đích: Đánh giá hiệu suất của mô hình trên một tập validation riêng biệt (hold-out set) sau khi mô hình đã được huấn luyện trên phần còn lại.
  - Phân chia: Tập dữ liệu được chia thành 80% cho huấn luyện (x_train, y_train) và 20% cho đánh giá (x_val, y_val).
  - Ứng dụng: Kết quả này thường được dùng để so sánh nhanh và hỗ trợ quyết định tinh chỉnh siêu tham số.

Train/Test Split Evaluation



4. Experience and results

Kết quả của cross-validation

Model	Accuracy	F1-Score	ROC AUC
Logistic Regression	0,8238	0,7625	0,8743
Random Forest	0,8304	0,7579	0,8695
XGBoost	0,8238	0,764	0,8745
SVM	**0,8328**	**0,7635**	**0,8749**

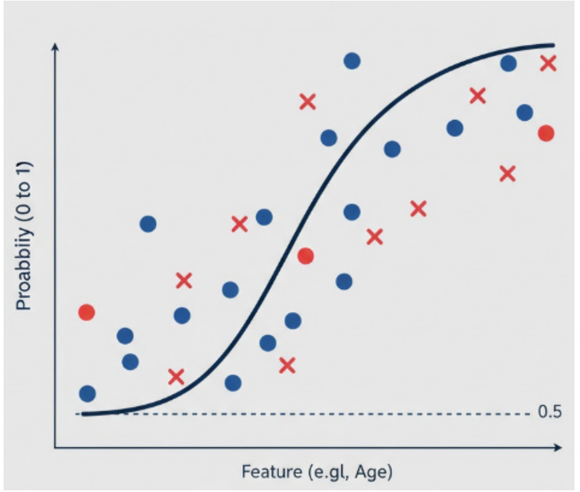
Kết quả của train-split-test

Model	Accuracy	F1-Score	ROC AUC
Logistic Regression	0,8156	0,7724	0,888
Random Forest	0,8101	0,7536	0,9001
XGBoost	0,8156	0,7755	0,8932
SVM	**0,8212**	**0,7746**	**0,8656**

4. Experience and results

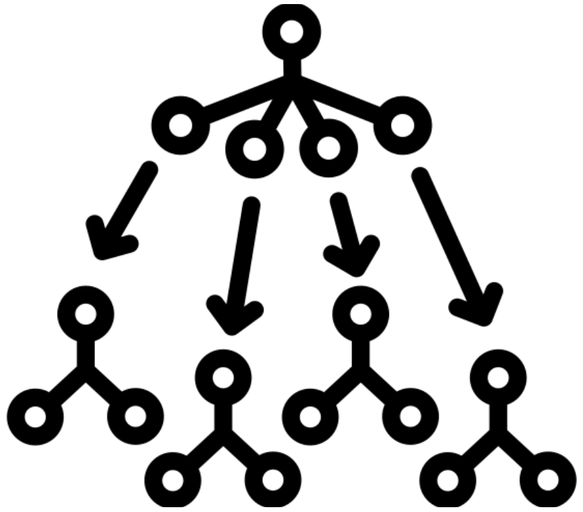
Logistic Regression setup:

- Regularization strength: 1
- Max_iter: 1000
- Random state: 42



Random Forest setup:

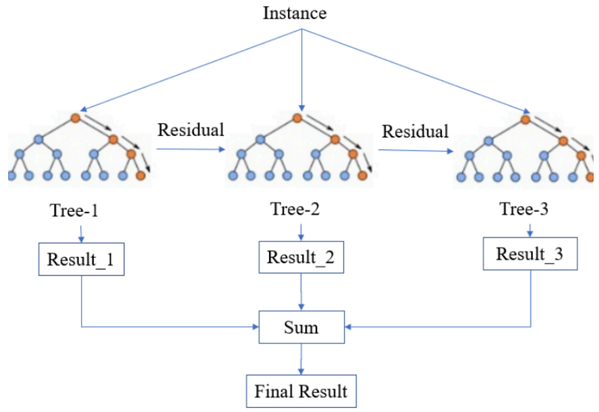
- Min simple split: 3
- Min simple leaf: 1
- n_estimators: 600
- n_jobs: -1



4. Experience and results

XGBoost setup:

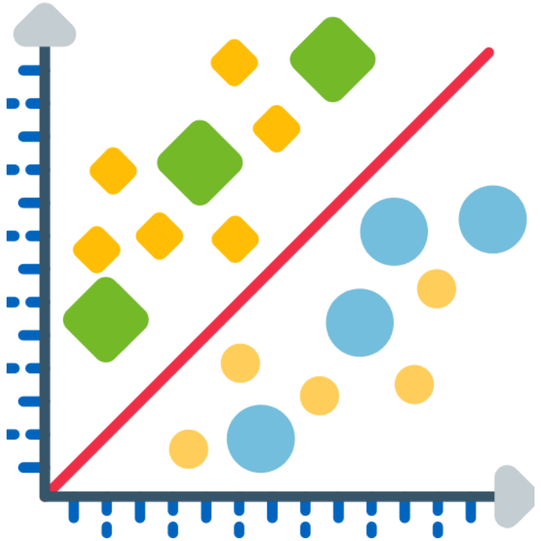
- n_estimators : 500
- learning rate: 0.03
- colsample_bytree : 0.8
- eval_metric: 'logloss'



implified structure of XGBoost.

SVM setup:

- Regularization strength: 1
- gamma: 'scale'
- probability: True



5. Conclusion

Mô hình SVM cho thấy hiệu suất mạnh mẽ và ổn định nhất, đạt điểm ROC AUC và F1 Score cao nhất trên tập Validation, xác nhận là lựa chọn tối ưu cho việc dự đoán.

Quá trình của xử lý và tạo các đặc trưng đóng vai trò quan trọng, giúp mã hóa các quy tắc sống còn của thảm họa thành các yếu tố dự đoán rõ ràng.

THANK YOU