Dự Đoán Khả Năng Sống Sót Thảm Họa Titanic Bằng Machine Learning

- Nguyễn Tuấn Đạt, Châu Hải Đăng, Trần Đại Thắng -

Problem Review

Thuộc dạng bài toán phân lớp (Classification)

Input:

- Dữ liệu hành khách gồm thông tin cá nhân và vé tàu từ **Kaggle Titanic Dataset**.

Output:

- Khả năng sống sót của của hành khách (0 hay 1).

Notice:

- Tập dữ liệu bị thiếu, mất mát và nhiễu. Cần tiền xử lí và phân tích đặc trưng.

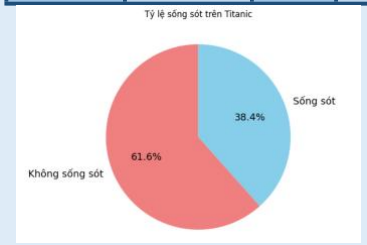
Dataset

Source:

- Kaggle – Titanic: Machine Learning from Disaster.

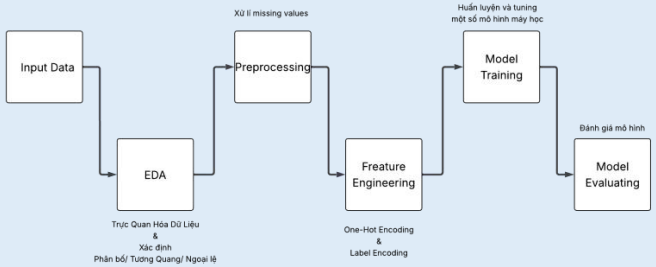
https://www.kaggle.com/competitions/titanic/data

Embarked	Bến khởi hành	Parch	Cha mẹ, Con cái	SibSp	Anh, chị, em và họ hàng
PClass	Hạng khoa	Fare	Giá vé	PassengerId	Mã hành khách
Age	Tuổi	Sex	Giới tính	Survived	Sống sót
Name	Tên	Cabin	Số cabin	Ticket	Mã vé

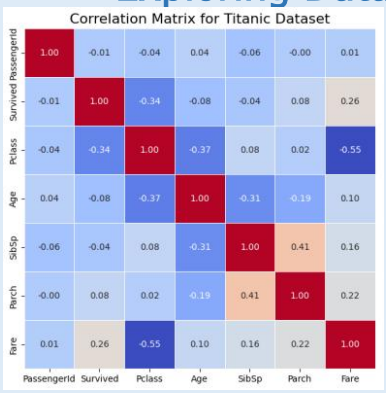


- Gồm 891 bộ train và 418 bộ test.
- Tỉ lệ phân lớp của tập train xấp xỉ 1:2

Propose Method



Exploring Data Analysis



Đáng chú ý	
Fare	Mật độ outlier dày đặc
Age	
Sibps	Mật thiết với và có thể xử lí chung
Parch	

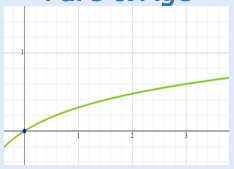
Thuộc tính	Thiếu	Tỉ lệ
Cabin	687	77.1%
Age	177	~20%
Embark	2	0.22%

Preprocessing

Missing Datas:

Thuộc tính	Phương Thức xử lí
Cabin	Chuyển tất cả dữ liệu thiếu về unknown.
Age	Median.
Embark	Mode.

Skrewness Reduction: Fare & Age



Log(1 + x)

Standardization

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

σ = Standard Deviation

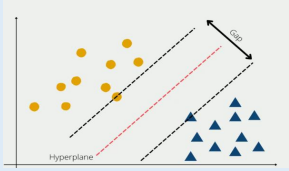
Áp dụng phân phối chuẩn cho việc chuẩn hóa dữ liệu.

Feature Engineering:

Đặc trưng	Phương Thức xử lí
PassengerID	Đặc trưng không hữu dụng drop.
Name	Trích xuất đặc trưng thành - Title. Loại bỏ name và One-hot Title
Age	Trích xuất ra đặc trưng IsChild và IsMother.
SibSp, Parch	Trích xuất đặc trưng thành FamilySize và loại bỏ đặc trưng cũ.
Ticket	Trích xuất đặc trưng thành Prefix Ticket phân loại vé. Và drop Ticket.
Embarked, PClass, Cabin, Sex, Ticket Prefix	One-Hot.

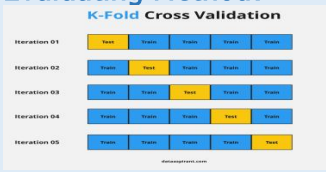
Model Training

Proposed model:



- Support Vector Machine (SVM) Classifier.

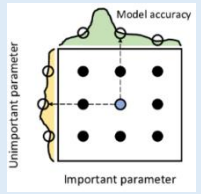
Evaluating Method:



- K-Fold Cross validation và kaggle submition score.

Tuning Method:

- Kết hợp giữa K-Fold và GridSearch Cho các hyperparameters của các model.



Model Evaluating

```
Logistic Regression CV Results:
  Accuracy: 0.8238
  F1 Score: 0.7620
  ROC AUC:  0.8796
--------------------------------------
Random Forest Classifier CV Results:
  Accuracy: 0.8305
  F1 Score: 0.7760
  ROC AUC:  0.8906
--------------------------------------
Support Vector CV Results:
  Accuracy: 0.8373
  F1 Score: 0.7738
  ROC AUC:  0.8708
--------------------------------------
XGBoost CV Results:
  Accuracy: 0.8260
  F1 Score: 0.7652
  ROC AUC:  0.8708
--------------------------------------
```

Baseline models

Best Of Interest → Selected SVM for fine tuned

```
SVM CV Results:
  Accuracy: 0.8384
  F1 Score: 0.7717
  ROC AUC:  0.8732
```

Fine tuned SVM

Có tăng tiến trong việc phân biệt người sống sót.

Result

Best Model:

- Support Vector Machine (SVM) of exp4.

Kaggle result: