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

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

Mở đầu

Input

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

Output

Dự đoán trạng thái sống sót của từng hành khách. (0 hoặc 1)

Thách thức

Dữ liệu Titanic thiếu, mất cân bằng và nhiễu, cần tiền xử lý và trích xuất đặc trưng hiệu quả.

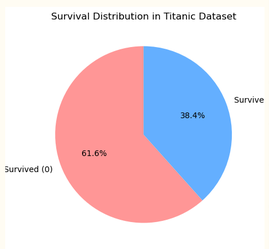
Tập dữ liệu

Nguồn dữ liệu: Kaggle – Titanic: Machine Learning from Disaster

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

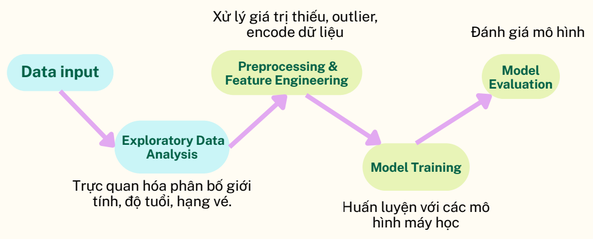
891 giá trị (training) - 418 giá trị (test)

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

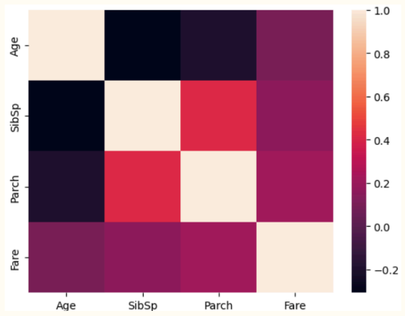


PassengerId	Id hành khách	Parch	ROC AUC
Pclass	Hạng vé	Ticket	Mã vé
Name	Tên hành khách	Fare	Giá vé
Sex	Giới tính	Cabin	Số cabin
Age	Tuổi	Embarked	Nơi lên tàu
SibSp	Số lượng anh/ chị/ em, vợ chồng trên tàu	Survived	Kết quả sống/chết

Phương pháp đề xuất



Phân tích khám phá dữ liệu



	Missing Values	Percent (%)
Cabin	687	77.1
Age	177	19.87
Embarked	2	0.22

Tiền xử lý và Kỹ thuật tạo đặc trưng

Top 5: Feature tương quan với Fare

Pclass	0.388032
Cabin	0.532923
FamilySiz	0.528907
SibSp	0.447113
Parch	0.410074

Missing Value Imputation:

- Embarked: Mode
- Fare: Median theo Pclass
- Age: Median theo (Sex, Pclass, Title)

Normalization:

- Log transform cho Fare & Age

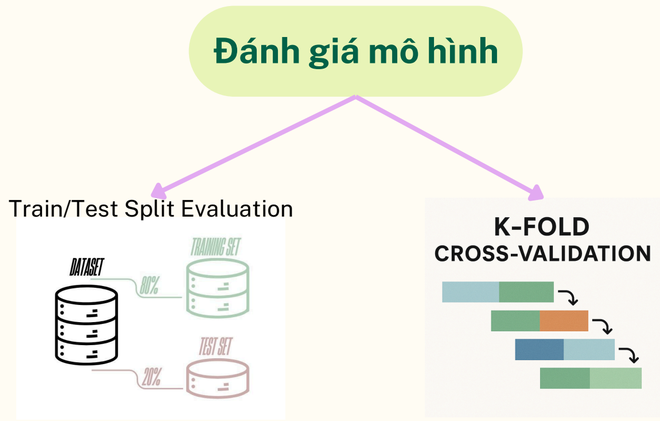
Kỹ thuật tạo đặc trưng

	Feature	Data Type	Missing Values	Missing (%)
Cabin	Cabin	object	687	77.104377
Age	Age	float64	177	19.865320
Embarked	Embarked	object	2	0.224467
PassengerId	PassengerId	int64	0	0.000000
Name	Name	object	0	0.000000
Pclass	Pclass	int64	0	0.000000
Survived	Survived	int64	0	0.000000
Sex	Sex	object	0	0.000000
Parch	Parch	int64	0	0.000000
SibSp	SibSp	int64	0	0.000000
Fare	Fare	float64	0	0.000000
Ticket	Ticket	object	0	0.000000

Feature	Ý nghĩa ngắn gọn	Feature	Ý nghĩa ngắn gọn
Title	Giới tính, địa vị, tầng lớp	**HasCabin**	Có cabin → hạng vé cao → giàu
IsFemale	Nữ → ưu tiên sống sót	**Deck**	Boong tàu (A,B,C...) → vị trí trên tàu
FamilySize	Quy mô gia đình → ảnh hưởng sống sót	**Age*Pclass**	Tuổi × hạng vé → tầng lớp xã hội
IsChild	Trẻ em (<12 tuổi) → ưu tiên	**Fare_Pclass**	Giá vé / hạng vé
IsMother	Phụ nữ trưởng thành có con	**TicketPrefix**	Loại vé → hãng tàu hoặc loại vé

Huấn luyện mô hình

Đánh giá mô hình



Thí nghiệm

Thí nghiệm 1 – Train/Test Split Evaluation:

Model	Accuracy	F1-Score	ROC AUC
Logistic	0.8156	0.7724	0.888
Random	0.8101	0.7536	0.9001
XGBoost	0.8156	0.7755	0.8932
SVM	**0.8212**	**0.7746**	**0.8656**

Thí nghiệm 2 – K-Fold Cross-Validation:

Model	Accuracy	F1-Score	ROC AUC
Logistic	0.8238	0.7625	0.8743
Random	0.8304	0.7579	0.8695
XGBoost	0.8238	0.764	nan
SVM	**0.8328**	**0.7635**	**0.8749**

Mô hình tốt nhất → SVM

Kết quả

Thành tích khi submit trên Kaggle:

→ 0.79425