Ứng dụng các mô hình máy học phổ biến để dự đoán hành khách sống sót sau vụ chìm tàu RMS Titanic năm 1912

Đỗ Nguyễn Thanh Phong, Trịnh Minh Toàn, Dương Chí Khang, Cao Thanh Tài

Giới thiệu

Tổng quan vấn đề:

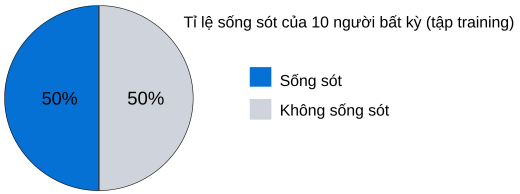
- Đây là bài toán kinh điển giúp người học làm quen với Machine Learning cơ bản.
- Mục tiêu nghiên cứu: áp dụng các mô hình học máy để xác định yếu tố ảnh hưởng đến khả năng sống sót.

Xác định vấn đề:

- Input: Danh sách các hành khách đã được thông kê trạng thái sống sót
- Output: Mô hình có độ chính xác cao nhất và danh sách dự đoán khả năng sống sót của các hành khách còn lại

Dữ liệu

- Số lượng: 891 bản ghi (training set), 418 bản ghi (testing set)
- Nguồn dữ liệu: Kaggle – *Titanic: Machine Learning from Disaster*



Tỉ lệ sống sót của 10 người bất kỳ (tập training)

- Các thuộc tính chính: Pclass (Hạng vé) (1, 2, 3), Sex (Giới tính), Age (Tuổi), Fare (Giá vé), Embarked (Cảng lên tàu), Survived (Trạng thái sống sót) (1 = sống sót, 0 = không)

Quy trình nghiên cứu

Làm sạch và xử lý dữ liệu

- Điền thiếu: Age bằng trung vị theo nhóm (Sex × Pclass), Embarked bằng mode, Fare bằng median.
- Tạo đặc trưng: FamilySize = SibSp + Parch + 1; IsAlone = (FamilySize==1); tách Title từ Name.
- Mã hóa: Sex (0/1), One-Hot Embarked, One-Hot/Label Title.
- Chuẩn hóa: StandardScaler cho các biến số.
- Xử lý mất cân bằng: Thử class_weight='balanced' và SMOTE (nếu cần).

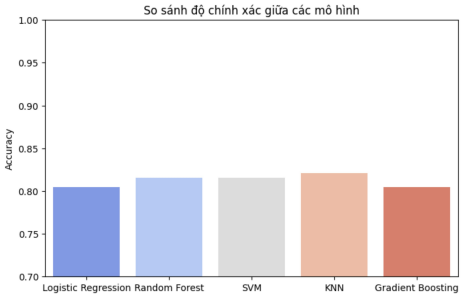
Chia tập Train / Validation

Chia 80% train – 20% validation, dùng StratifiedKFold (k=5) để đánh giá chéo, tránh thiên lệch do phân phối nhãn.

Thực nghiệm

- Thực nghiệm 1: Huấn luyện các mô hình phổ biến và só sánh
- Thực nghiệm 2: Chọn ra mô hình tốt nhất và đánh giá chi tiết

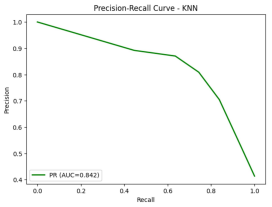
Thực nghiệm 1:

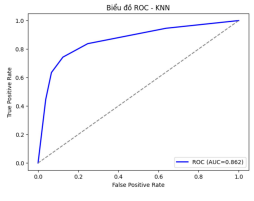


Thực nghiệm 2:

Sau khi so sánh, ta nhận thấy mô hình KNN cho ra độ chính xác cao và có kết quả đầu ra tốt nhất, nên chúng ta sẽ đánh giá mô hình

	precision	recall	f1-score	support
0	0.83	0.88	0.85	105
1	0.81	0.74	0.77	74
accurary	-	-	0.82	179
macro avg	0.82	0.81	0.81	179
weighted avg	0.82	0.82	0.82	179





Kết quả

Mô hình	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.80	0.78	0.76	0.77	0.84
KNN	0.82	0.80	0.78	0.79	0.85
SVM	0.81	0.79	0.77	0.78	0.84
Random Forest	0.85	0.83	0.81	0.82	0.89
Gradient Boosting	0.84	0.82	0.80	0.81	0.88

Kết luận và hướng phát triển

Kết luận:

- RandomForest là mô hình có kết quả tốt nhất với độ chính xác ~0.85.
- Các yếu tố quan trọng nhất ảnh hưởng đến khả năng sống sót: Giới tính, Hạng vé, Tuổi.
- Pipeline học máy đã giúp mô hình đạt hiệu quả cao, dễ tái sử dụng.

Hướng phát triển:

- Thử các kỹ thuật nâng cao như XGBoost, CatBoost.
- Tối ưu tham số (GridSearchCV).
- Tạo thêm đặc trưng mới như Title hoặc CabinType.
- Ứng dụng mạng nơ-ron sâu (Deep Learning).