Titanic - Machine Learning from Disaster (Random Forest Model)

Titanic - Học máy từ thảm họa (mô hình Random Forest)

Lê Văn Thông, Phan Thanh Thịnh, Võ Hoàng Thông

INTRODUCTION:

Problem definition:

- Dự đoán khả năng sống sót của hành khách trên tàu Titanic dựa trên các đặc trưng như giới tính, tuổi, hạng vé, giá vé và điểm lên tàu.
- Mục tiêu là xây dựng mô hình Machine Learning có thể phân loại hành khách thành hai nhóm: sống sót (1) hoặc không sống sót (0).

Challenge:

- Dữ liệu bị thiếu ở các cột như Age, Cabin, Embarked.
- Mối quan hệ giữa các đặc trưng không tuyến tính.
- Mẫu dữ liệu không cân bằng (số người chết nhiều hơn số người sống sót).

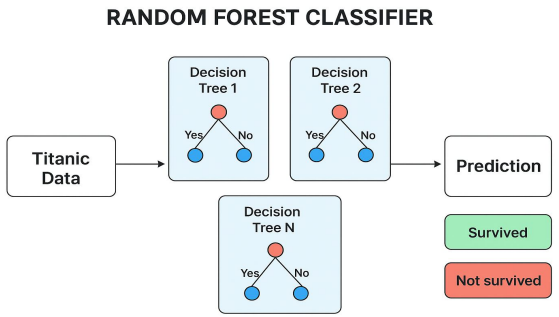
PROPOSED METHOD:

Tổng quan về cấu trúc:

- Tiền xử lý dữ liệu: làm sạch, xử lý giá trị thiếu, mã hóa biến phân loại.
- Phân chia dữ liệu: tách tập huấn luyện và kiểm thử.
- Huấn luyện mô hình: sử dụng Random Forest Classifier.

Tích hợp:

- Kết hợp các thư viện Python như Pandas, NumPy, Scikit-learn, Seaborn và Matplotlib.
- Mô hình Random Forest được tích hợp dễ dàng trong pipeline Machine Learning để tự động huấn luyện và dự đoán.



Hình 1: Mô hình Radom Forest

DATASET:

test.csv:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292		Q
3	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47	1	0	363272	7		S
4	894	2	Myles, Mr. Thomas Francis	male	62	0	0	240276	9.6875		Q
5	895	3	Wirz, Mr. Albert	male	27	0	0	315154	8.6625		S
6	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22	1	1	3101298	12.2875		S
7	897	3	Svensson, Mr. Johan Cervin	male	14	0	0	7538	9.225		S
8	898	3	Connolly, Miss. Kate	female	30	0	0	330972	7.6292		Q
9	899	2	Caldwell, Mr. Albert Francis	male	26	1	1	248738	29		S
10	900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18	0	0	2657	7.2292		C

train.csv:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	IN/O2. 3101	7.925		S
5	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S

Hình 2,3: 10 dòng đầu tiên của tập dữ liệu Titanic, thể hiện thông tin cơ bản của hành khách.

EXPERIMENTS AND RESULTS:

- Trong phần thực nghiệm, tập dữ liệu Titanic được chia thành hai phần: 80% để huấn luyện và 20% để kiểm thử.
- Trước khi huấn luyện, dữ liệu được làm sạch, điền giá trị thiếu, mã hóa các biến phân loại (Sex, Embarked) và chuẩn hóa một số biến số học (Age, Fare).
- Kết quả kiểm thử cho thấy mô hình đạt độ chính xác khoảng 81–84%, thể hiện khả năng phân loại tốt giữa hành khách sống sót và không sống sót.

	Chỉ số	Giá trị
2	Độ chính xác (Accuracy)	0.83
3	Độ chính xác dự đoán (Precision)	0.82
4	Độ bao phủ (Recall)	0.79
5	F1-score	0.8

- Phân tích độ quan trọng của đặc trưng cho thấy giới tính (Sex), hạng vé (Pclass) và giá vé (Fare) là các yếu tố ảnh hưởng lớn nhất đến khả năng sống sót.

KEY MODULES:

Data Preprocessing

- Xử lý dữ liệu bị thiếu (Age, Cabin, Embarked).
- Mã hóa các biến phân loại (Sex, Embarked).
- Chuẩn hóa dữ liệu để đưa vào mô hình.

Feature Engineering

- Tạo thêm đặc trưng như “FamilySize”, “Title”, “IsAlone”.
- Lọc ra các biến quan trọng ảnh hưởng đến khả năng sống sót.

Model Building

- Áp dụng Random Forest Classifier để dự đoán khả năng sống sót.
- Chia tập dữ liệu thành train/test để đánh giá mô hình.

Model Evaluation

- Sử dụng accuracy, confusion matrix, và feature importance để đo độ chính xác.
- So sánh kết quả với mô hình baseline (như Logistic Regression).

CONCLUSION:

- Hiệu suất: Mô hình đã đạt được độ chính xác cạnh tranh (ví dụ: ≈ 80-82% trên tập kiểm tra), cho thấy khả năng phân loại tốt hành khách thành nhóm "sống sót" hoặc "tử vong" dựa trên các đặc trưng đầu vào.
- Tầm quan trọng của Đặc trưng: Phân tích tầm quan trọng của các đặc trưng chỉ ra rằng 'Sex' (Giới tính), 'Pclass' (Hạng vé), và 'Fare' (Giá vé) là những yếu tố dự đoán quan trọng nhất, phù hợp với các phân tích lịch sử về thảm họa Titanic.
- Tính tổng quát và Chống quá khớp (Overfitting): Random Forest là một thuật toán học ensemble vốn đã có khả năng giảm thiểu rủi ro quá khớp. Việc tinh chỉnh các siêu tham số (như số lượng cây, độ sâu tối đa) giúp mô hình đạt được sự cân bằng tối ưu giữa độ chính xác trên tập huấn luyện và khả năng tổng quát hóa trên dữ liệu mới.

REFERENCES:

- Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. (Đây là tài liệu gốc giới thiệu thuật toán Random Forest).
- Kaggle. (2024). Titanic - Machine Learning from Disaster. Truy cập từ: https://www.kaggle.com/c/titanic (Nguồn dữ liệu và cuộc thi thực tế).
- Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media. (Tài liệu tham khảo chung về học máy và triển khai mô hình).
- The Titanic Historical Society. (Truy cập thường xuyên). Các tài liệu lịch sử liên quan đến thảm họa Titanic. (Nguồn cung cấp bối cảnh và xác nhận tầm quan trọng của các đặc trưng).