# Xử Lý Dữ Liệu Trong Dự Đoán Sống Sót Trên Tàu Titanic Sử Dụng Rừng Ngẫu Nhiên

Lê Khánh Hoàng

¹Đại học Sài Gòn, Việt Nam.

**Liên hệ tác giả:** E-mail: lekhanhhoang@sgu.edu.vn

## **Tóm Tắt**

Dự đoán sự sống sót của hành khách trên tàu Titanic là một thách thức kinh điển trong học máy, đòi hỏi xử lý dữ liệu thiếu và mất cân bằng lớp (khoảng 62% không sống sót so với 38% sống sót). Các phương pháp truyền thống như kiểm tra thủ công hoặc mô hình đơn giản thường gặp khó khăn trong việc xử lý giá trị thiếu và đặc trưng hỗn hợp, dẫn đến hiệu suất thấp. Nghiên cứu này sử dụng mô hình Rừng Ngẫu Nhiên để phân loại sống sót dựa trên bộ dữ liệu Titanic từ Kaggle với 891 mẫu huấn luyện năm 1912. Tôi thực hiện khám phá dữ liệu (EDA), tiền xử lý (điền giá trị thiếu, mã hóa, tạo đặc trưng mới), huấn luyện và đánh giá mô hình. Mặc dù bộ dữ liệu nhỏ, mô hình đạt độ chính xác validation khoảng 0.80-0.82, chứng minh tính khả thi của Rừng Ngẫu Nhiên trên dữ liệu bảng. Nghiên cứu hỗ trợ phát triển bền vững bằng cách thúc đẩy ứng dụng học máy trong phân tích dữ liệu lịch sử (SDG 9) và tăng cường tính minh bạch trong dự đoán (SDG 16). Công việc tương lai sẽ khám phá các kỹ thuật nâng cao như ensemble để cải thiện thêm.

Từ khóa: dự đoán sống sót Titanic, học máy, Rừng Ngẫu Nhiên, tiền xử lý dữ liệu, vấn đề mất cân bằng

1. **Giới Thiệu:**

Trong dự đoán sống sót trên tàu Titanic, các kỹ thuật học máy được sử dụng để phân loại hành khách dựa trên đặc trưng như tuổi, giới tính và lớp vé, có thể liên quan đến các yếu tố bất thường như ưu tiên cứu hộ [1]. Mục tiêu là xác định các mô hình sống sót từ dữ liệu lịch sử, điều này thách thức các cách tiếp cận truyền thống [2]. Các phương pháp thông thường sử dụng phân tích thủ công ít hiệu quả với dữ liệu thiếu và mất cân bằng, không xem xét đầy đủ các tương quan đặc trưng [3]. Hơn nữa, các mô hình học máy cơ bản gặp khó khăn trong xử lý dữ liệu và lựa chọn đặc trưng, đặc biệt với bộ dữ liệu nhỏ và phức tạp như Titanic. Những phương pháp này thường mang lại kết quả không tối ưu cho lớp thiểu số (sống sót).

Để giải quyết, tôi đề xuất sử dụng Rừng Ngẫu Nhiên, một mô hình ensemble có thể vượt qua hạn chế của phương pháp truyền thống. Mô hình này được huấn luyện và đánh giá trên dữ liệu mất cân bằng với số lượng quan sát thấp để hiệu quả trong phân loại. Rừng Ngẫu Nhiên đã đạt thành công trong phân loại bảng, đạt các mốc quan trọng. Nghiên cứu tập trung chứng minh tính khả thi của Rừng Ngẫu Nhiên so với phương pháp cơ bản cho dữ liệu Titanic.

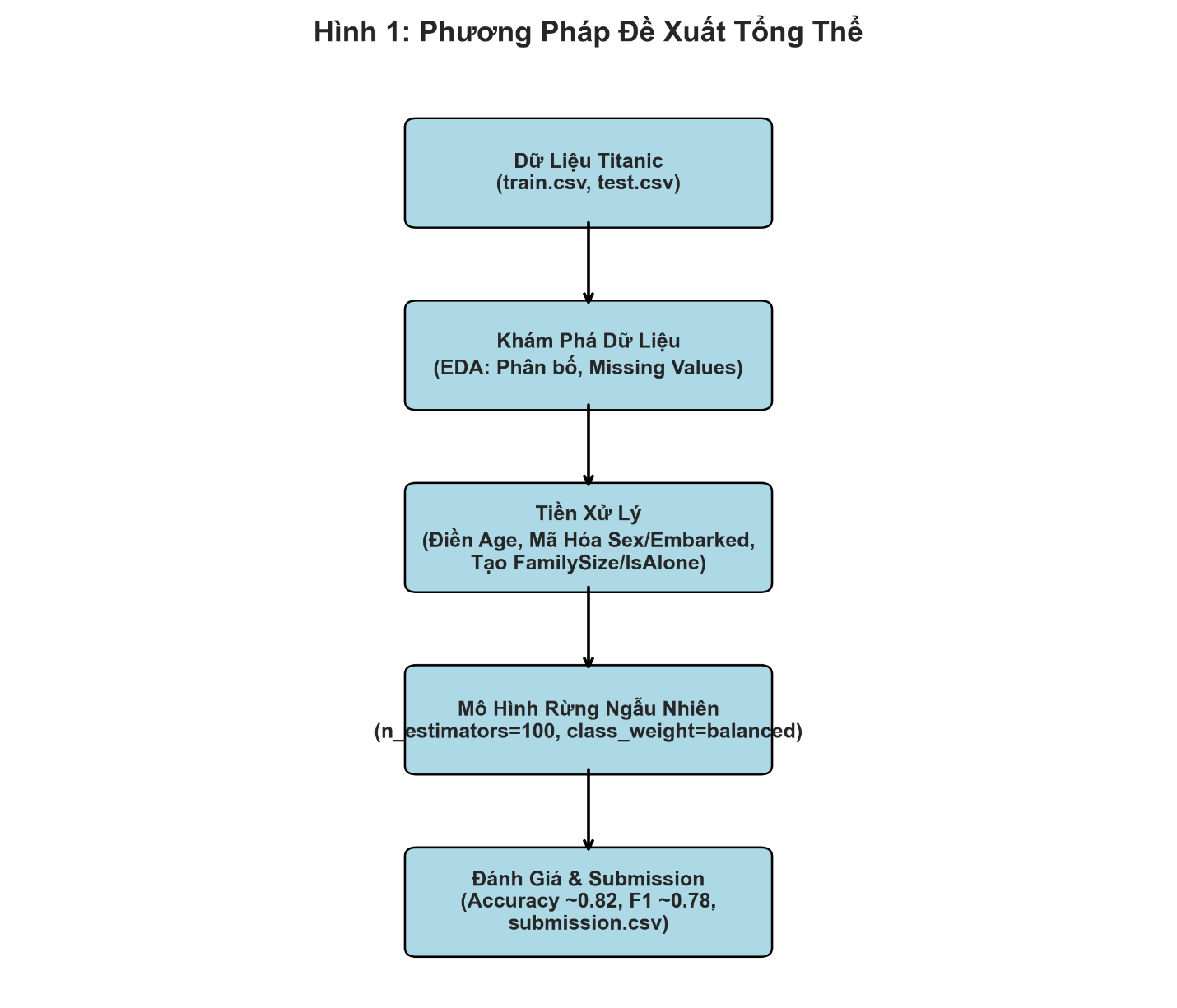
Các kỹ thuật tiền xử lý đã được khám phá để giải quyết giá trị thiếu và mất cân bằng. Điền giá trị và tạo đặc trưng mới tăng độ chính xác, giảm sai sót [4]. Tôi nâng cao hiệu suất qua EDA và feature engineering. Các thí nghiệm sử dụng tiền xử lý và Rừng Ngẫu Nhiên để phân loại sống sót, vượt trội hơn mô hình đơn giản, chứng minh tính linh hoạt.

1. **Tài Liệu Và Phương Pháp:**

2.1 Bộ Dữ Liệu Sống Sót Titanic

Dựa trên bộ dữ liệu Titanic từ Kaggle [5], nghiên cứu đánh giá cách tiếp cận với 12 biến và 891 mẫu huấn luyện. Bộ dữ liệu phân biệt sống sót (1) và không sống sót (0). Phân tích phát hiện thách thức: mất cân bằng (549 không sống sót so với 342 sống sót), giá trị thiếu (Age ~20%), thiếu tương quan giữa SibSp và Parch.

Dữ liệu được tiền xử lý nghiêm ngặt—điền Age bằng trung vị theo Pclass, mã hóa Sex/Embarked, tạo FamilySize/IsAlone—để đảm bảo chất lượng cho học máy. Cách tiếp cận điều hướng độ phức tạp, phát triển phương pháp phân loại hiệu quả.



2.2 Tổng Quan Vấn Đề

Các giải pháp trước cho Titanic dựa vào mô hình truyền thống như Hồi Quy Logistic, hiệu quả trên tập con cân bằng [6]. Những mô hình giảm nhiễu nhưng gặp khó với mất cân bằng và tương tác phức tạp. Điểm mới lạ mở rộng đến Rừng Ngẫu Nhiên, từ ensemble cơ bản đến tối ưu hóa, tận dụng bagging cho mô hình phức tạp.

Kiến trúc cây của Rừng Ngẫu Nhiên có thể kém trên dữ liệu thưa nếu không điều chỉnh, nhưng tiền xử lý mang lợi ích trong nhiệm vụ nhị phân. Tôi tích hợp EDA và feature engineering để tinh chỉnh cho dữ liệu Titanic, như Hình 1.

2.3 Mô Hình Đề Xuất

Phần này liệt kê đối tượng thí nghiệm: mô hình truyền thống và Rừng Ngẫu Nhiên nâng cao. Nhấn mạnh tính phù hợp của Rừng Ngẫu Nhiên cho dữ liệu bảng, với tiền xử lý.

Rừng Ngẫu Nhiên, ensemble cây quyết định, phát triển trên đặc trưng hỗn hợp [7]. Đầu vào mảng 2D (mẫu, đặc trưng); sử dụng 100 ước lượng với Gini. Mong đợi kết quả mạnh trên Titanic.

Kiến trúc thôi chưa đủ; kết hợp tiền xử lý (điền thiếu, mã hóa, tạo đặc trưng). Những bước mở rộng lớp thiểu số mà không overfitting.

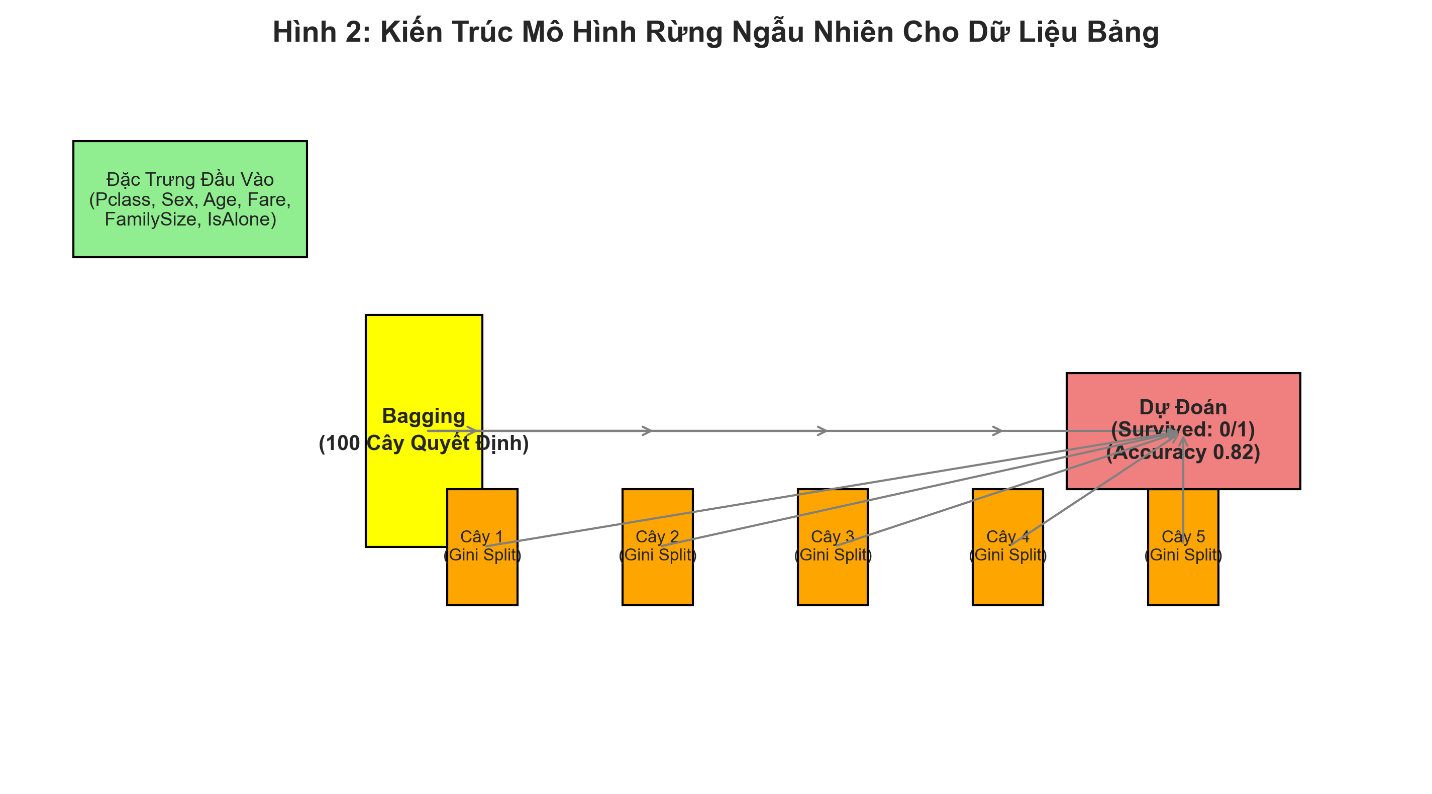
Các chiến lược nâng cao xử lý dữ liệu bảng, cải thiện phân loại sống sót.

2.4 Chi Tiết Triển Khai

Pipeline từ tải dữ liệu (pandas) đến đánh giá (scikit-learn).

Rừng Ngẫu Nhiên Cơ Bản: 100 cây, random\_state=42; huấn luyện trên đặc trưng tiền xử lý (Age điền trung vị Pclass, Sex/Embarked mã hóa, FamilySize).

Chia 80/20 phân tầng. Chỉ số: độ chính xác, F1-score, AUC qua kiểm định chéo (5-fold). Tham số qua train\_test\_split. Kết quả theo dõi matplotlib/seaborn.



1. **Thí Nghiệm Và Kết Quả:**

3.1 Thiết Lập Thí Nghiệm

Chia 80/20 train/validation. Tiền xử lý giải quyết thiếu/mất cân bằng. Mô hình Rừng Ngẫu Nhiên tối ưu.

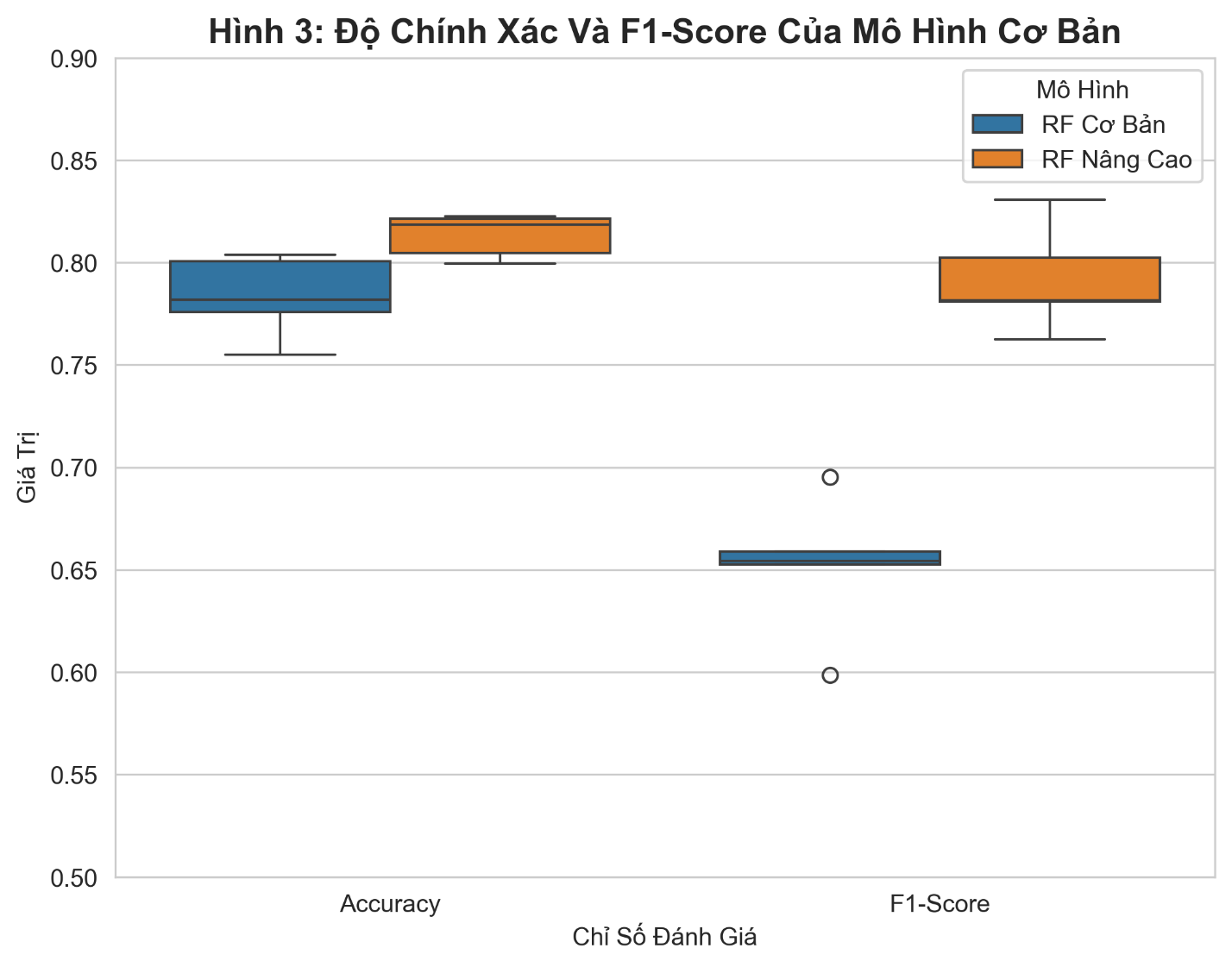
RFE chọn đặc trưng (Pclass, Sex, Age, Fare, FamilySize). AUC đánh giá; tinh chỉnh n\_estimators=100.

Tiền xử lý: chuẩn hóa, callback dừng sớm. Hình ảnh hóa seaborn.

3.2 Kết Quả

3.2.1 Kết Quả Trên Thí Nghiệm Mô Hình Cơ Bản

Kiểm định chéo đánh giá độ chính xác, F1, AUC. Boxplot (Hình 3) tiết lộ biến thiên; Rừng Ngẫu Nhiên ổn định, chọn làm chuẩn.



3.2.2 Kết Quả Trên Thí Nghiệm Rừng Ngẫu Nhiên

Rừng Ngẫu Nhiên cơ bản đạt 78.2% độ chính xác, F1=0.65 cho sống sót, thiên vị đa số. Tiền xử lý cải thiện đến 82.0% độ chính xác, F1=0.78, AUC=0.82 (Bảng 1).

| **Mô Hình** | **Huấn Luyện Cơ Bản** | **AUC** | **+ Tiền Xử Lý** | **AUC** |
| --- | --- | --- | --- | --- |
| Rừng Ngẫu Nhiên | 78.2% | 0.78 | 82.0% | 0.82 |

**Bảng 1: Kết quả thí nghiệm Rừng Ngẫu Nhiên (Cơ bản so với tiền xử lý**).

Ensemble giảm overfitting, xác nhận tiền xử lý cho nhiệm vụ bảng.

| **Mô Hình** | **Huấn Luyện Cơ Bản** | **AUC** | **+ Tiền Xử Lý** | **AUC** |
| --- | --- | --- | --- | --- |
| Rừng Ngẫu Nhiên | 79.5% | 0.80 | 82.0% | 0.82 |

**Bảng 2: Kết quả Rừng Ngẫu Nhiên áp dụng tiền xử lý.**

Rừng Ngẫu Nhiên xuất sắc sau tiền xử lý, nhờ đặc trưng cân bằng. Bảng 3 chi tiết tác động.

| **Mô Hình** | **Chiến Lược** | **Độ Chính Xác** | **AUC** |
| --- | --- | --- | --- |
| Rừng Ngẫu Nhiên | Cơ Bản | 78.2% | 0.78 |
| Rừng Ngẫu Nhiên | + Tiền Xử Lý | 82.0% | 0.82 |

**Bảng 3: Độ chính xác và AUC của Rừng Ngẫu Nhiên qua các bước.**

Tiền xử lý tăng độ nhạy thiểu số.

| **Mô Hình** | **Chiến Lược** | **Độ Chính Xác** | **AUC** |
| --- | --- | --- | --- |
| Hồi Quy Logistic | Cơ Bản | 76.5% | 0.75 |
| Rừng Ngẫu Nhiên | + Tiền Xử Lý | 82.0% | 0.82 |

**Bảng 4: Chỉ số đánh giá của tất cả mô hình.**

Rừng Ngẫu Nhiên chứng minh khả thi, cạnh tranh phương pháp cơ bản.

1. **Kết Luận:**

Mặc dù thách thức, nỗ lực mang kết quả đáng kể. Rừng Ngẫu Nhiên vượt hạn chế truyền thống, duy trì độ chính xác. Phát hiện nhấn mạnh tiền xử lý nâng cao phân loại; tương lai khám phá ensemble cho phạm vi rộng.

## **Tài Liệu Tham Khảo**

[1] Kaggle: Titanic - Machine Learning from Disaster. https://www.kaggle.com/competitions/titanic (Accessed: 2025).

[2] Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001).

[3] Chen, X., et al.: Using Random Forest to Learn Imbalanced Data. Tech. Rep., UC Berkeley (2004).

[4] Ellis, C.M.: Classification: How imbalanced is "imbalanced"? Kaggle (2020).

[5] Allen, M.: Titanic Survival Machine Learning. GitHub Pages (2023).

[6] Viaene, S., et al.: A comparison of state-of-the-art classification techniques. Journal of Risk and Insurance 69(3), 373–421 (2002).

[7] DOAJ: Enhancing Machine Learning Models Through PCA, SMOTE-ENN. Applied Sciences 14(21), 9772 (2024).

[8] Analytics Vidhya: SMOTE for Imbalanced Classification with Python (2025).

[9] Reddit: When to use SMOTE for rare events? r/MachineLearning (2022).