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

Lê Đoàn Kim Ngân, Lê Thị Trúc Ly, Lâm Tú Nhi,

Nguyễn Đăng Khoa

Liên hệ tác giả: E-mail(s) : khoanguyen12062005@gmail.com;

Đồng tác giả: nhilam150@gmail.com;

imtrucly@gmail.com;

ledoankimngan2005@gmail.com;

Các tác giả đóng góp công bằng trong bài báo này.

Tóm tắt

Vào năm 1912, đã có một sự kiện hàng hải rung động cả thế giới được xem là thảm khốc nhất thế giới đó chính là thảm họa chìm tàu *RMS Titanic*, nó cũng đã cướp đi sinh mạng của hơn *1500 hành khách và thủy thủ* có trên con tàu này. Trong *nghiên cứu* này, chúng ta sẽ xây dựng một mô hình học máy nhằm có thể dự đoán được khả năng sống xót của các hành khách dựa trên các đặc trưng nhân khẩu và thông tin hành khách như giới tính, tuổi, giá vé, hạng vé và điểm khởi hành. Dữ liệu được sử dụng lấy từ bộ dữ liệu công khai *Titanic - Machine Learning from Disaster* trên *Kaggle*. Bằng cách sử dụng các mô hình học máy như: *Logistic Regression, Random Forest, Support Vector Machine (SVM), XGBosst*,.. thông qua các quy trình như : *tiền xử lý dữ liệu*, *phân tích tương quan* giữa các đặc trưng, trích xuất và tạo thêm các đặc trưng (*feature engineering*). Huấn luyện và đánh giá các mô hình thông qua các thông số như: *Accuracy, F1-Score, ROC AUC*. Kết quả thực nghiệm cho thấy mô hình **SVM** có khả năng dự đoán chính xác nhất. Nghiên cứu này không chỉ góp phần minh họa ứng dụng của học máy trong khai thác dữ liệu lịch sử mà còn khẳng định tầm quan trọng của việc chọn lọc và xử lý đặc trưng phù hợp trong bài toán dự đoán nhị phân.

Từ khóa: Titanic, học máy, khai thác dữ liệu, dự đoán sống sót, dự đoán nhị phân.

1. Giới thiệu

Với hậu quả do thảm họa *Titanic* mang lại, dù đã trôi qua hơn một thế kỉ nhưng sự kiện này vẫn rất thu hút sự quan tâm từ giới nghiên cứu bởi dữ liệu của hành khách được lưu trữ đầy đủ và phản ánh rõ các mối quan hệ giữa các yếu tố về nhân khẩu học và khả năng sống sót. *Trong thời đại 4.0*, nơi mà dữ liệu và trí tuệ nhân tạo đang *phát triển* một cách vô cùng mạnh mẽ, việc áp dụng các mô hình học máy để *khai thác, phân tích* và *dự đoán* từ những bộ dữ liệu lịch sử như Titanic không chỉ mang lại ý nghĩa về *học thuật* mà còn giúp chúng ta minh họa được *khả năng ứng dụng* của các công nghệ trong việc nhận diện được các *yếu tố rủi ro* và *đưa ra quyết định*.

Trước đây, trong các cuộc nghiên cứu người ta chỉ sử dụng các phương pháp *thống kê truyền thống* hoặc chỉ là các mô hình học máy *cơ bản* như *Logistic Regression* để dự đoán *xác xuất sống sót*. Tuy nhiên, với sự phát triển mạnh mẽ của máy học hiện nay, đã có rất nhiều thuật toán *tiên tiến* ra đời như: *Random Forest, Support Vector Machine (SVM), XGBoost*, ... đã chứng minh được khả năng xử lý được các loại *dữ liệu phi tuyến* và khai thác được các *đặc trưng phức tạp* tốt hơn. Dù là như vậy nhưng việc so sánh và đánh giá hiệu năng giữa các mô hình này vẫn còn *hạn chế* do các tiêu chí đánh giá *khác nhau* nên kết quả nghiên cứu *không thể so sánh trực tiếp* được.

Từ các *thực tế* đó, nghiên cứu này được hình thành và thực hiện nhằm mục đích xây dựng và *so sánh hiệu năng* của các mô hình học máy trong việc dự đoán được khả năng sống sót của các hành khách trên con tàu *Titanic.* Nghiên cứu này cần tập trung vào các quy trình *tiền xử lý dữ liệu*, *kỹ thuật tạo đặc trưng* (feature engineering) và *đánh giá* các mô hình bằng phương pháp *K-Fold Cross Validation*. Kết quả kỳ vọng sẽ xác định được mô hình nào *tối ưu* và làm rõ được *vai trò* của từng yếu tố trong việc ảnh hưởng đến khả năng sống sót, qua đó góp phần củng cố hiểu biết về *ứng dụng học máy* trong phân tích dữ liệu.

2. Đối tượng và phương pháp

2.1. Tập dữ liệu Titanic

Dựa trên tập dữ liệu về các thông tin của các hành khách đã có mặt trên con tàu Titanic. Với *10 biến số* và gần *1000 mẫu* quan sát, bộ dữ liệu được thiết kế nhằm phân biệt các hành khách còn *sống* và đã *chết*. Tuy nhiên, theo như phân tích của chúng tôi đã phát hiện ra mộ số khó khăn nhấn mạnh về *mức độ phức tạp* của nhiệm vụ này. Trước hết, dữ liệu bị *thiếu* ở các dòng nổi lên như một vấn đề đáng phải chú ý. Chúng tôi nhận thấy sự thiếu sót *nghiêm trọng* này, với số lượng giá trị bị thiếu ở cột *Cabin* lên tới 687 dòng (gần *80% tổng số mẫu*) và cột *Age* cũng không hề ít khi bị thiếu 177 dòng (*gần 20%*). Bên cạnh

đó cột *Age* và *Fare* có các giá trị *outlier rất lớn*, những điều này đã tạo ra một trở ngại đáng kể trong việc đào tạo mô hình của chúng tôi sao cho hiệu quả, vì sự chênh lệch này sẽ làm cho mô hình *thiên lệch* về các đặc trưng có giá trị lớn hơn. Việc này sẽ cản trở khả năng *học tập* của các mô hình. Hơn nữa, qua quá trình kiểm tra của chúng tôi đã nhận ra sự thiếu hụt về các mối tương quan giữa một số biến trong bộ dữ liệu. Việc thiếu các mối tương quan này đã làm phức tạp sự nỗ lực của chúng tôi trong việc xác định các mối quan hệ và cản trở *khả năng dự đoán* của mô hình.

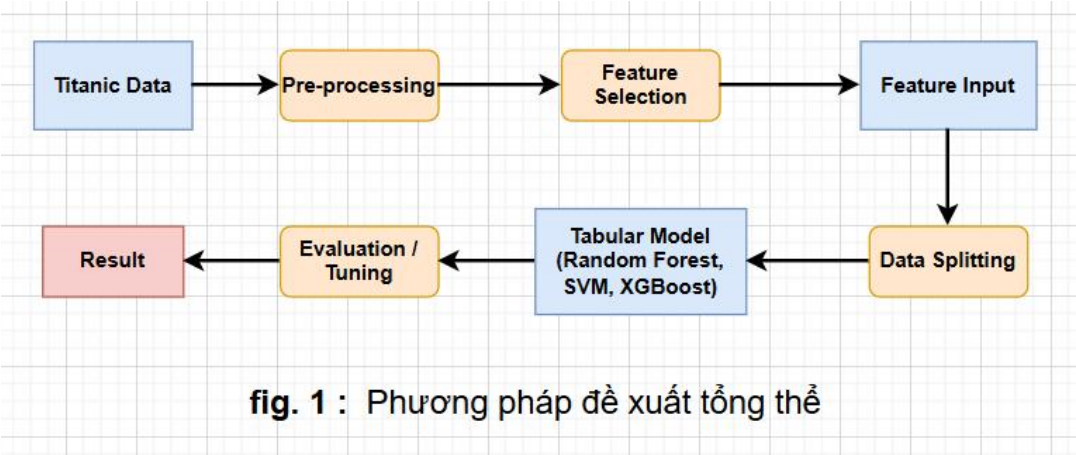


fig. 1 : Phương pháp đề xuất tổng thể

Nghiên cứu này gặp phải những thách thức do các *missing value, outlier*, các mối tương quan khá ít cũng như tập dữ liệu nhỏ. Nhưng dữ liệu đã đã được chúng tôi *phân tích*, *tiền xử lý* nghiệm ngặt để *đảm bảo chất lượng* và *tính phù hợp* của dữ liệu với các mô hình học máy.Phương pháp này đã giúp *xử lý sự phức tạp* của tập dữ liệu và *phát triển* các phương pháp mạnh mẽ để có thể *phân loại* khả năng sống/chết một cách hiệu quả.

2.2. Tổng quan về vấn đề

Nghiên cứu về bài toán dự đoán khả năng sống sót trên tàu Titanic là một trong những hướng đi kinh điển trong lĩnh vực học máy, được sử dụng rộng rãi để đánh giá và so sánh hiệu quả của các thuật toán *phân loại nhị phân*. Trong các phương pháp học máy *truyền thống*, các mô hình như *Logistic Regression*, Cây quyết định (*Decision Tree*), và Rừng ngẫu nhiên (*Random Forest*) thường được sử dụng nhờ khả năng diễn giải cao và tính *đơn giản* trong triển khai. Tuy nhiên, các mô hình này đôi khi không thể mô tả đầy đủ mối quan hệ *phi tuyến* giữa các đặc trưng, đặc biệt khi các yếu tố nhân khẩu học, địa vị xã hội và mối quan hệ gia đình tương tác lẫn nhau một cách *phức tạp*.

Trong nghiên cứu này, mô hình *Support Vector Machine – SVM* được xác định là mô hình đạt hiệu suất cao nhất trong việc dự đoán khả năng sống sót của hành khách. *SVM* thể hiện ưu thế nhờ khả năng xây dựng *siêu phẳng* (*hyperplane*) tối ưu để phân tách hai lớp

dữ liệu, đồng thời hoạt động hiệu quả ngay cả khi kích thước tập dữ liệu nhỏ. Bằng việc lựa chọn hàm *kernel* thích hợp (chẳng hạn ***rbf*** hoặc ***polynomial***), SVM có thể học được *ranh giới phi tuyến*, giúp mô hình nắm bắt các tương tác phức tạp giữa các đặc trưng như *Age, Sex, Fare, Pclass, và Embarked*. Kết quả huấn luyện cho thấy SVM không chỉ đạt độ chính xác (*accuracy*) cao mà còn giữ được sự cân bằng tốt giữa *precision* và *recall*, vượt trội so với các mô hình khác như *Logistic Regression, Random Forest và KNN.*

Bên cạnh đó, tập dữ liệu Titanic vẫn tồn tại một số *thách thức* cố hữu, bao gồm quy mô nhỏ (891 mẫu huấn luyện) và sự thiếu hụt ở các *đặc trưng quan trọng* như Age, Cabin và Embarked. Mặc dù tỷ lệ *sống sót* (38%) và không sống sót (62%) không tạo ra mất cân bằng nghiêm trọng, sự chênh lệch vừa phải này vẫn có thể gây *sai lệch* trong dự đoán nếu không được *xử lý cẩn thận*. Do đó, việc tiền xử lý dữ liệu, bao gồm *chuẩn hóa* (scaling), *mã hóa* (encoding), và *xử lý giá trị thiếu*, đóng vai trò quan trọng trong việc giúp mô hình SVM phát huy *hiệu quả* tối đa.

Từ những cơ sở trên, nghiên cứu tập trung xây dựng một quy trình huấn luyện toàn diện, kết hợp giữa tiền xử lý dữ liệu, chọn và tạo đặc trưng, và tối ưu tham số cho mô hình SVM nhằm đạt hiệu năng cao nhất. Cách tiếp cận này cho thấy rằng, với việc lựa chọn đặc trưng phù hợp và điều chỉnh tham số hợp lý, SVM là một trong những mô hình có độ ổn định và khả năng tổng quát hóa tốt nhất cho bài toán dự đoán khả năng sống sót trên tàu Titanic.

2.3. Mô hình đề xuất

Trong phần này, chúng ta sẽ liệt kê các đối tượng thử nghiệm của nghiên cứu bao gồm các mô hình học máy truyền thống như Logistics Regression và các mô hình tiên tiến hơn như Random Forest, Support Vector Machine - SVM. Sau đó chúng tôi sẽ tập trung vào phân tích chi tiết của từng nhóm mô hình học máy, làm rõ lý do tại sao chúng phù hợp với dữ liệu dạng bảng (tabular data), đồng thời thảo luận chuyên sâu về các kĩ thuật lấy mẫu và hàm mất mát được sử dụng.

Như đã đề cập trước đó, mô hình *Support Vector Machine – SVM* là một trong những phương pháp học máy nổi bật, đã đạt được nhiều thành tựu quan trọng trong các bài toán phân loại nhị phân và nhận dạng mẫu. SVM hoạt động dựa trên nguyên lý xác định siêu phẳng tối ưu để phân tách các lớp dữ liệu, đồng thời có khả năng xử lý tốt các mối quan hệ phi tuyến thông qua việc sử dụng các hàm nhân (kernel).

Trong nghiên cứu này, mô hình SVM được tinh chỉnh và tối ưu dựa trên hàm nhân Radial Basis Function (RBF), vốn được đánh giá là phù hợp cho dữ liệu có cấu trúc phi tuyến như tập dữ liệu Titanic. Các tham số điều chỉnh, bao gồm hệ số phạt (C) và tham số

gamma (γ), được lựa chọn thông qua quá trình thử nghiệm và xác thực chéo để đạt hiệu suất cao nhất. Chúng tôi kỳ vọng rằng mô hình SVM được xây dựng và tối ưu theo hướng này sẽ mang lại kết quả tích cực, thể hiện khả năng phân loại ổn định và tổng quát hóa tốt trên tập dữ liệu Titanic.

Phân tích chuyên sâu cho thấy việc lựa chọn mô hình phù hợp là chưa đủ; cần kết hợp với các kỹ thuật tiền xử lý và chọn đặc trưng để đạt hiệu quả tối ưu. Trong nghiên cứu này, dữ liệu được chuẩn hóa bằng StandardScaler, mã hóa bằng One-Hot Encoding và mở rộng với các đặc trưng như FamilySize, IsAlone và Title có thể giúp các mô hình đạt hiệu suất ổn định và chính xác hơn.

Hơn nữa, chúng tôi cho rằng việc lựa chọn tiêu chí đánh giá và chiến lược tối ưu hóa phù hợp có thể cải thiện đáng kể khả năng dự đoán của mô các hình. Các nghiên cứu trước đây đã chỉ ra rằng việc sử dụng các chỉ số đánh giá cân bằng như F1-score và ROC-AUC giúp phản ánh chính xác hơn hiệu năng của mô hình trên dữ liệu có sự chênh lệch giữa các lớp.

Bằng cách triển khai các chiến lược này, chúng tôi hướng tới việc tăng cường khả năng xử lý dữ liệu bảng bị mất cân bằng và cải thiện hiệu năng tổng thể của mô hình trong bài toán dự đoán khả năng sống sót.

2.4. Chi tiết triển khai

Nghiên cứu này tập trung khảo sát và so sánh ba mô hình học máy phổ biến, bao gồm Hồi quy Logistic (Logistic Regression), Rừng ngẫu nhiên (Random Forest) và Máy véc-tơ hỗ trợ (Support Vector Machine – SVM), nhằm đánh giá hiệu quả của chúng trong việc dự đoán khả năng sống sót của hành khách trên tàu Titanic.

Mô hình Logistic Regression: Là một trong những mô hình cơ bản và dễ diễn giải nhất trong học máy, hồi quy logistic được sử dụng rộng rãi cho các bài toán phân loại nhị phân. Mô hình này hoạt động dựa trên việc ước lượng xác suất sống sót của hành khách thông qua hàm logistic (sigmoid). Trong nghiên cứu này, hồi quy logistic được sử dụng như mô hình cơ sở (baseline) để đánh giá mức độ ảnh hưởng của các đặc trưng đã được tiền xử lý, đồng thời so sánh hiệu năng với các mô hình phi tuyến tính phức tạp hơn.

Mô hình Random Forest: Random Forest là một thuật toán học máy mạnh mẽ dựa trên tập hợp nhiều cây quyết định (decision trees). Mỗi cây được huấn luyện trên một mẫu ngẫu nhiên của dữ liệu, và kết quả cuối cùng được xác định bằng biểu quyết đa số (majority voting). Mô hình này có khả năng xử lý tốt dữ liệu chứa nhiễu, hạn chế hiện tượng quá khớp (overfitting) và cung cấp khả năng đánh giá tầm quan trọng của từng đặc

trưng. Trong bài toán Titanic, Random Forest được sử dụng nhằm kiểm tra khả năng nắm bắt mối quan hệ phi tuyến giữa các yếu tố như độ tuổi, giới tính, hạng ghế và mức giá vé đối với xác suất sống sót.

Mô hình Support Vector Machine (SVM): SVM là một trong những thuật toán học máy mạnh mẽ nhất cho các bài toán *phân loại*. Mô hình này xác định *siêu phẳng tối ưu* (*optimal hyperplane*) để phân tách hai lớp dữ liệu với khoảng cách biên lớn nhất. Để xử lý các mối quan hệ phi tuyến trong dữ liệu Titanic, nghiên cứu này sử dụng hàm nhân *Radial Basis Function* (**rbf**) cùng với quá trình *tối ưu tham số* (**C và γ**) thông qua *Grid Search* và *K-Fold Cross Validation*. Nhờ khả năng tạo ranh giới phân tách *linh hoạt* và *ổn định*, SVM được kỳ vọng đạt hiệu năng *cao nhất* trong việc dự đoán khả năng sống sót của hành khách.

Bên cạnh việc xử lý dữ liệu thông qua các kỹ thuật tiền xử lý và chọn đặc trưng, chúng tôi nhận thấy rằng việc lựa chọn các tiêu chí đánh giá và chiến lược tối ưu hóa phù hợp có thể giúp cải thiện đáng kể hiệu năng của mô hình. Do đó, nghiên cứu này tiến hành thử nghiệm và so sánh nhiều phương pháp đánh giá khác nhau nhằm xác định cách tiếp cận chính xác và khách quan nhất trong việc dự đoán khả năng sống sót của hành khách trên tàu Titanic.

F1-Score: Chỉ số này được sử dụng để đánh giá độ cân bằng giữa precision (độ chính xác) và recall (độ bao phủ). Trong các bài toán có chênh lệch giữa hai lớp, như tỷ lệ sống sót và không sống sót trong tập dữ liệu Titanic, việc tối ưu hóa F1-score giúp mô hình tránh thiên lệch về một lớp, đặc biệt khi một chỉ số có xu hướng bị bỏ qua nếu chỉ dựa vào độ chính xác thông thường.

$$\mathbf{F1} = \frac{\mathbf{2} \times (\mathbf{Precision} \times \mathbf{Recall})}{\mathbf{Precision} + \mathbf{Recall}} \quad (\mathbf{1})$$

ROC-AUC: Là thước đo khả năng phân biệt giữa hai lớp, chỉ số ROC-AUC phản ánh mức độ mô hình có thể xếp hạng các mẫu dương và âm chính xác. Trong nghiên cứu này, ROC-AUC được xem như một tiêu chí quan trọng để đánh giá tổng thể hiệu năng của các mô hình học máy, đặc biệt trong bối cảnh dữ liệu có sự chênh lệch lớp vừa phải.

$$\mathbf{AUC} = \int_{\mathbf{0}}^{\mathbf{1}} \mathbf{TPR}(\mathbf{FPR})\mathbf{d}(\mathbf{FPR}) \quad (\mathbf{2})$$

Grid Search & Cross Validation: Để đảm bảo khả năng tổng quát hóa của mô hình, chúng tôi sử dụng kỹ thuật Grid Search kết hợp với K-Fold Cross Validation nhằm tìm ra bộ siêu tham số tối ưu cho từng mô hình. Phương pháp này cho phép mô hình được đánh giá trên nhiều tập con khác nhau, giảm thiểu rủi ro quá khớp và cải thiện độ tin cậy của kết quả.

$$Score_{final} = \frac{1}{K}\sum_{i=1}^{K} Score_i \quad (3)$$

Model Optimization: Quá trình tối ưu được thực hiện cho ba mô hình chính, bao gồm điều chỉnh hệ số phạt (C) và tham số γ trong SVM, số lượng cây (n_estimators) và độ sâu (max_depth) trong Random Forest, cũng như phương pháp điều chuẩn (regularization) trong Logistic Regression. Việc tối ưu đồng bộ này cho phép đánh giá khách quan mức độ hiệu quả của từng thuật toán và xác định mô hình có khả năng dự đoán cao nhất.

3. Thí nghiệm và kết quả

3.1. Thiết lập thí nghiệm

Quy trình huấn luyện: Tập dữ liệu train được huấn luyện bằng kỹ thuật *K-Fold Cross Validation* (với K = 10) nhằm đảm bảo tính *khách quan* và giảm thiểu hiện tượng *quá khớp*. Phương pháp này cho phép mô hình được huấn luyện và kiểm thử luân phiên trên nhiều tập con khác nhau, từ đó cung cấp đánh giá *ổn định* và toàn diện hơn về *hiệu năng* mô hình.

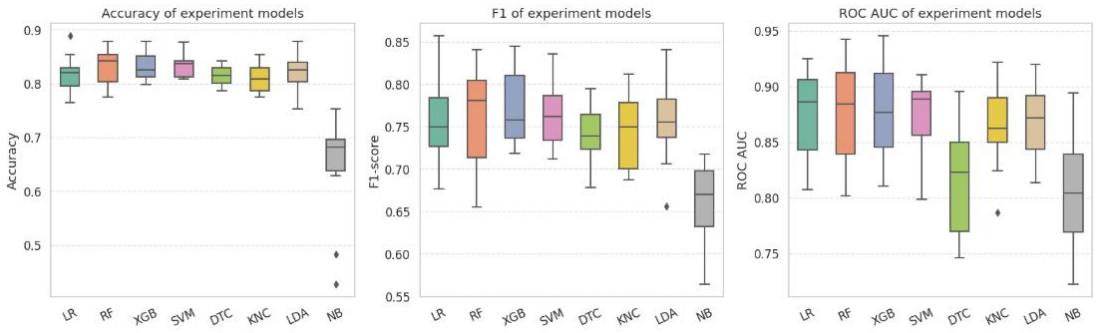
Trước khi huấn luyện, toàn bộ dữ liệu được xử lý theo các bước đã trình bày, bao gồm *xử lý giá trị khuyết*, chuẩn hóa bằng *StandardScaler* và mã hóa các biến phân loại bằng *One-Hot Encoding* hoặc *Label Encoding*. Ngoài ra, các đặc trưng mở rộng như *FamilySize, IsAlone* và *Title* cũng được bổ sung nhằm tăng khả năng mô hình hóa mối *quan hệ phi tuyến* giữa các yếu tố.

Trong quá trình huấn luyện, các mô hình *Logistic Regression, Random Forest* và *Support Vector Machine (SVM)* được tối ưu hóa thông qua *Grid Search* kết hợp với *K-Fold Cross Validation* để lựa chọn bộ siêu *tham số tối ưu*. Các chỉ số đánh giá cân bằng như *F1-score* và *ROC-AUC* được sử dụng thay cho độ chính xác (accuracy) truyền thống, nhằm phản ánh tốt hơn khả năng *phân loại* giữa hai lớp (sống sót và không sống sót).

3.2. Kết quả

Chúng tôi đã thực hiện quá trình *cross-validation* bằng phương pháp K-fold, sau đó tiến hành phân tích hiệu suất của mô hình thông qua các chỉ số đã đề cập gồm: *độ chính xác (accuracy), điểm F1 (F1 score) và đường cong AUC (AUC curve).*

Các kết quả thể hiện trong Hình 2 về độ chính xác (*Accuracy*) và *F1_Score* và *ROC AUC* trên tập kiểm thử được trực quan hóa bằng biểu đồ *Boxplot*, giúp chúng tôi hiểu rõ hơn về các chỉ số và mức độ biến thiên của từng mô hình.



Hình 2: Accuracy, F1_Score, ROC_AUC của các mô hình

Bảng 1: Các thông số ban đầu của từng mô hình

Model	Accuracy	F1-Score	ROC_AUC
Logistic Regression	0.8101	0.7385	0.8387
Random Forest	0.8212	0.7538	0.8345
XGBoost	0.7989	0.7353	0.8080
Support Vector Machine	0.8268	0.7559	0.8381
Decision Tree	0.7821	0.6667	0.7937
K-Nearest Neighbors	0.7877	0.7206	0.8298
Linear Discriminant Analysis	0.8156	0.7442	0.8324
Naive Bayes	0.7709	0.7172	0.8233

Trước tiên sau khi chạy demo thử các mô hình khi *chưa tạo thêm các đặc trưng* có mối tương quan từ dữ liệu có sẵn ta được các thông số như Bảng 1. Độ chính xác của các mô hình dao động từ *0.7709 đến 0.8268*, cho thấy sự khác biệt đáng kể về khả năng học và khái quát hóa giữa các thuật toán. Trong đó, *Support Vector Machine* (SVM) đạt kết quả cao nhất với độ chính xác 0.8268 và điểm F1-Score là 0.7559, phản ánh khả năng cân bằng tốt giữa độ nhạy (*recall*) và độ chính xác (*precision*). Mô hình này cũng duy trì giá trị ROC_AUC *tương đối cao* (0.8381), chứng tỏ hiệu quả trong việc phân biệt giữa các lớp dữ liệu.

Các mô hình *Random Forest* và *Linear Discriminant Analysis* (LDA) cũng thể hiện hiệu năng ổn định với độ chính xác lần lượt là *0.8212 và 0.8156*. Đặc biệt, *Random Forest* có F1-Score và ROC_AUC *khá cân bằng* (0.7538 và 0.8345), cho thấy mô hình có khả năng *tổng quát hóa tốt* nhờ cơ chế lấy mẫu và kết hợp nhiều cây quyết định.

Ngược lại, các mô hình như *Decision Tree* và *Naive Bayes* cho kết quả thấp hơn, lần lượt đạt 0.7821 và 0.7709 về độ chính xác. Điều này có thể xuất phát từ việc các mô hình này hoạt động *kém hiệu quả* trên dữ liệu có mối quan hệ *phi tuyến* phức tạp hoặc phân phối đặc trưng *không đồng nhất*.

Nhìn chung, các kết quả bước đầu cho thấy nhóm mô hình tuyến tính (*Logistic Regression, LDA*) và phi tuyến (*SVM, Random Forest*) đều có tiềm năng tốt cho bài toán. Tuy nhiên, để tối ưu hiệu suất, bước tiếp theo cần tập trung vào việc tạo thêm *đặc trưng mới*, *chuẩn hóa dữ liệu*, và *tinh chỉnh siêu tham số* (hyperparameter tuning) nhằm cải thiện khả năng *phân loại* và *độ ổn định* của mô hình.

Bảng 2: Các thông số sau khi cải tiến tương ứng với từng mô hình

Model	Accuracy	F1-Score	ROC_AUC
Logistic Regression	0.8192	0.7568	0.8749
Random Forest	0.8316	0.7602	0.8766
XGBoost	0.8316	0.7713	0.8790
Support Vector Machine	0.8350	0.7652	0.8727
Decision Tree	0.8148	0.7423	0.8154
K-Nearest Neighbors	0.8092	0.7434	0.8649
Linear Discriminant Analysis	0.8192	0.7561	0.8693
Naive Bayes	0.6419	0.6594	0.8048

Sau khi thực hiện các bước cải tiến và tạo đặc trưng tương ứng cho từng mô hình, kết quả thể hiện trong **Bảng 2** cho thấy hiệu năng của hầu hết các mô hình đều được cải thiện rõ rệt. Độ chính xác *tăng* trung bình từ 1–2%, đặc biệt ở các *mô hình phi tuyến* như XGBoost, Random Forest và Support Vector Machine.

Cụ thể, mô hình SVM tiếp tục thể hiện hiệu quả *vượt trội* với *độ chính xác* 0.8350 và ROC_AUC đạt 0.8727, chứng tỏ khả năng duy trì ổn định khi dữ liệu được mở rộng và tái biểu diễn qua các đặc trưng mới. Trong khi đó, XGBoost và Random Forest đều đạt cùng độ chính xác 0.8316 nhưng có F1-Score và ROC_AUC cao hơn so với giai đoạn ban đầu, phản ánh khả năng tổng quát hóa và xử lý quan hệ phi tuyến giữa các biến đầu vào.

Các mô hình tuyến tính như Logistic Regression và Linear Discriminant Analysis cũng ghi nhận mức tăng nhẹ về cả ba chỉ số, cho thấy việc tái cấu trúc dữ liệu và lựa chọn đặc trưng hợp lý giúp mô hình tuyến tính khai thác được tốt hơn mối tương quan tiềm ẩn giữa các biến độc lập.

Đáng chú ý, hiệu suất của mô hình Naive Bayes *giảm mạnh* (Accuracy chỉ còn 0.6419), cho thấy giả định độc lập giữa các biến không còn phù hợp khi dữ liệu đã được mở rộng và bổ sung các đặc trưng có tính tương quan cao. Điều này khẳng định tầm *quan trọng* của việc lựa chọn mô hình phù hợp với *bản chất phân phối* của dữ liệu.

Tổng thể, kết quả sau cải tiến cho thấy việc *thiết kế* và *bổ sung đặc trưng* đóng vai trò quan trọng trong việc *nâng cao hiệu quả* học máy. Mô hình *SVM* tỏ ra phù hợp nhất cho bài toán này nhờ khả năng thích ứng linh hoạt với các *đặc trưng phức tạp* và *phi tuyến* trong dữ liệu thực tế. Kết quả này củng cố nhận định rằng, đối với các bài toán có cấu trúc dữ liệu tương tự, SVM là lựa chọn đáng tin cậy khi được kết hợp với chiến lược trích chọn và xây dựng đặc trưng hợp lý.

4. Kết luận

Mặc dù nghiên cứu phải đối mặt với nhiều thách thức trong quá trình xử lý dữ liệu và lựa chọn đặc trưng, những nỗ lực không ngừng của nhóm nghiên cứu đã mang lại những kết quả khả quan. Các mô hình học máy được áp dụng đã chứng minh được hiệu quả trong việc dự đoán khả năng sống sót của hành khách, đặc biệt là mô hình Support Vector Machine (SVM), vốn đạt được độ chính xác và tính ổn định cao hơn so với các phương pháp khác.

Kết quả chi tiết trong nghiên cứu cho thấy việc kết hợp các bước tiền xử lý dữ liệu, chuẩn hóa đặc trưng và tối ưu tham số đã giúp nâng cao đáng kể hiệu suất của mô hình SVM, khẳng định tính phù hợp của thuật toán này trong các bài toán phân loại dữ liệu dạng

bảng. Trong tương lai, nhóm nghiên cứu dự kiến sẽ mở rộng quy mô dữ liệu và thử nghiệm thêm các kỹ thuật học máy tiên tiến khác như Gradient Boosting hoặc XGBoost, nhằm tiếp tục cải thiện độ chính xác và khả năng tổng quát hóa của mô hình trong dự đoán khả năng sống sót trên tàu Titanic.

References

1) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). "*Scikit-learn: Machine learning in Python." Journal of Machine Learning Research, 12, 2825-2830.*

2) Altman, N. S. (1992). "*An introduction to kernel and nearest-neighbor nonparametric regression." The American Statistician, 46(3), 175-185*.

3) Heaton, J. (2017). "*An empirical analysis of feature engineering for predictive modeling." Proceedings of the 14th International Conference on Machine Learning and Data Mining in Pattern Recognition, 1-13.*

4) Nguyen_Quoc_Huy. "*ICIT24_Fraud_Detect_Paper": Addressing data imbalance in insurance fraud prediction using sampling techniques and robust losses".*

5) Ahmed, M., Mahmood, A., & Afzal, S. (2019). "*Comparative Analysis of Machine Learning Algorithms for Titanic Survival Prediction." International Journal of Advanced Computer Science and Applications, 10(1), 70-75.*

6) Breiman, L. (2001). "*Random Forests." Machine Learning, 45(1), 5-32*