Dự đoán tỉ lệ sống sót áp dụng mô hình học máy và khai phá dữ liệu

Nguyễn Đăng Tiến

Giới thiệu

Định nghĩa vấn đề:

- **Đầu vào :** Bộ dữ liệu Titanic Disaster được lấy và xuất từ cuộc thi Kaggle
- **Đầu ra:** Ứng dụng các mô hình học máy để phân loại, dự đoán tỷ lệ sống sót và áp dụng các

kỹ thuật khai phá dữ liệu

Thách thức:

- Chỉ nằm ở mức cơ bản, không tận dụng tốt các kỹ thuật xử lý nhằm tăng hiệu suất mô hình
- Chỉ áp dụng các mô hình học máy cơ bản, các kỹ thuật tiền sử lý chỉ nằm ở mức bắt đầu

Mục tiêu chính: Phân tích mối tương quan giữa các đặc trưng xung quanh tỷ lệ sống sót, mối quan

hệ, phân bố của các đặc trưng, sử dụng mô hình học máy để dự đoán khả năng sống sót

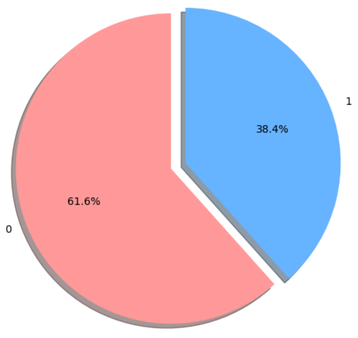
Bộ dữ liệu

- Mẫu: 891 hành khách xuất hiện trong mẫu cùng với 11 đặc trưng. 891 hành khách được sử dụng để

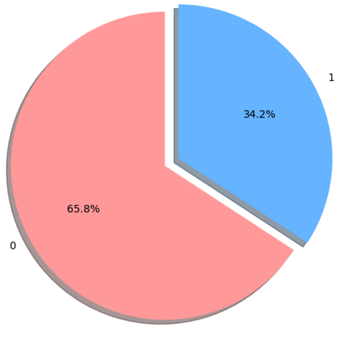
huấn luyện và 419 mẫu để kiểm tra

- Nguồn dữ liệu: Kaggle

https://colab.research.google.com/drive/1y-KsXtdawv-C91GqR4dQRdxBl6uqQj7G?usp=drive_link

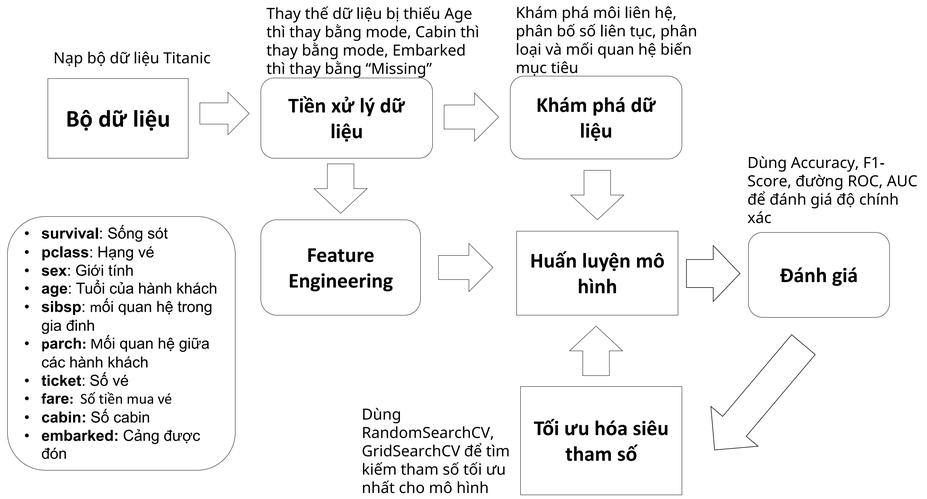


Phân bố của cột **Survival** (đặc trưng mục tiêu) trong tập dữ liệu được sử dụng để huấn luyện thuật toán học máy với nhãn **1** được xem là tồn tại sống sót vơi nhãn **0** được xem là không sống sót



Phân bố của cột **Survival** (đặc trưng mục tiêu) trong tập dữ liệu được sử dụng để kiểm tra thuật toán học máy và là nguồn dữ liệu để nộp bài với nhãn **1** được xem là tồn tại sống sót vơi nhãn **0** được xem là không sống sót

Phương pháp thực hiện



Tổng quan quá trình thực hiện tiền xử lý dữ liệu, khám phá dữ liệu huấn luyện mô hình để cho ra đầu ra hợp lý

1. Tiền xử lý dữ liệu

- Đối với các đặc trưng có giá trị bị thiếu ta sử dụng phương pháp thay thê bằng giá trị mode, KNN hay thay thế "Missing"
- Chăng hạn như cột **Age** thay thế bằng mô hình KNN. Lựa chọn phương pháp tính theo khoảng cách để thay thế giá trị
- Cột **Cabin** thực hiện điền giá trị mode đối với các giá trị bị thiếu
- Cột **Ebarked** thay thế bằng dữ liệu "Missing"

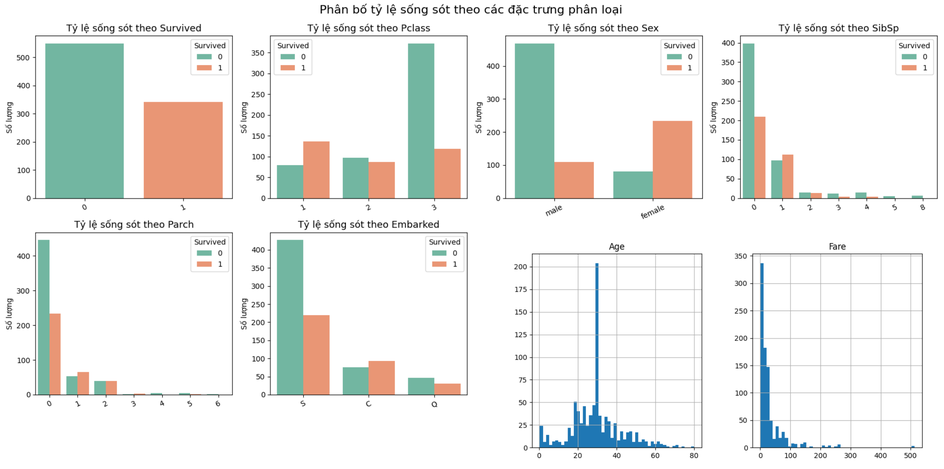
Sau khi đã thực hiện tiền sử lý dữ liệu mã hóa các biến phân loại bằng phương pháp (OneHotEncoder) để phân chia thành nhiều đặc trưng nhỏ với đặc trưng nào xuất hiện thì điền giá trị 1

Chuẩn hóa tuổi và giá vé về chung một thang đo ddercos thể tránh sai lệch thang đo với nhau sử dụng StandardScaler để chuẩn hóa phục vụ cho quá trình huấn luyện và kiểm tra mô hình

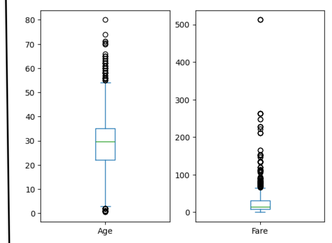
Tất cả đều dược thực hiện dựa trên thư viện sklearn.prepocessing

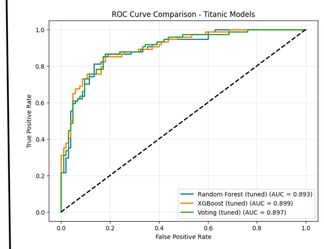
2. Khám phá dữ liệu

Mối quan hệ giữa biến đặc trưng mục tiêu với những biến khác



Phân bố giữa các biến số tìm kiếm giá trị ngoại lai trong bộ dữ liệu





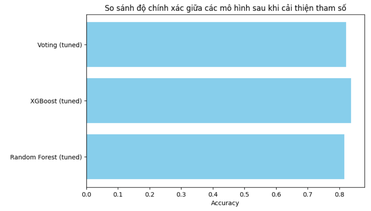
3. Huấn luyện và đánh giá mô hình

.

Thực hiện huấn luyện tập dữ liệu dựa trên 4 mô hình tất cả bao gồm:

- Logistic Regression: Mô hình sử dụng hàm sigmoid làm hàm khởi động để thực hiện phân loại
- Support Vector Machine: Mô hình sử dụng mặt phẳng để làm phân cách các nhãn
- Random Forest: Sử dụng chiều sâu của các nốt và lá để đưa ra lựa chọn
- XGBoost: Sử dụng khi huấn luyện lại nhiều lần cho đến khi đạt kết quả tốt nhất

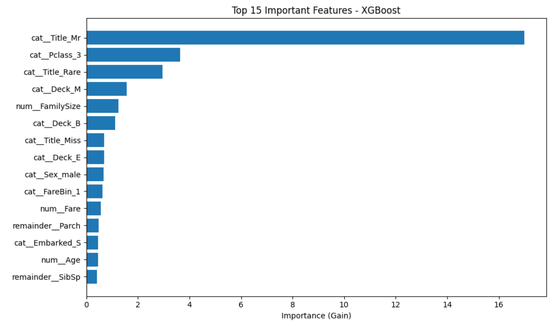
Sử dụng các thông số đánh giá Acurracy để đánh giá độ chính xác cho mô hình cũng như sơ đồ hình ROC để đánh giá tránh overfitting cũng như xem mô hình nào là chính xác cao



Biểu đồ đánh giá độ chính xác mô hình thông qua Acurracy và ROC

Mô hình	Độ chính xác (Acurracy)	Siêu tham số
Random Forest	81%	n_estimators=100, max_depth=10, max_features='10'
Logistic Regression	81%	n_estimators=200
Support Vector Machine	81%	n_estimators=100
Voting Classifier	82%	Theo Random Forest và XGBoost
XGBoost	83%	earning_rate=0.01, max_depth=7, n_estimators=500, subsample=0.7

Kết luận và đánh giá



Mô hình đạt hiệu suất cao nhất chính là mô hình **XGBoost** với độ chính xác là 83% trên toàn bộ tập dữ liệu huấn luyện với các siêu tham số được ghi ở bảng trên

Các yếu tố ảnh hưởng mạnh nhất đến khả năng sống sót:

- Giới tính (nữ có xác suất sống cao hơn),
- Hạng vé (hạng nhất an toàn hơn),
- Tuổi (người trẻ có cơ hội sống sót cao hơn).

Mô hình học máy có thể **hỗ trợ hiểu rõ các yếu tố ảnh hưởng đến hành vi sống sót** trong các tình huống khẩn cấp tương tự.

Hướng phát triển tiếp theo:

Mở rộng tập dữ liệu (tích hợp thêm dữ liệu nhân khẩu học, thời tiết, vị trí cabin).

Thử nghiệm mô hình nâng cao như **Neural Network**.