Họ tên: Dương Tùng Thiện

MSSV: 3123410350

Báo cáo bài tập Titanic

1. **Giới thiệu:**

Bài tập Titanic là một thử thách ở trên nền tảng Kaggle, dành cho những người dùng mới lần đầu tiếp xúc với nền tảng. Do đó, đây là một thử thách khá nổi tiếng.

Việc giải thử thách Titanic sẽ giúp người dùng làm quen với cách thức hoạt động, giao diện của nền tảng Kaggle. Ngoài ra giới thiệu người dùng đến những bước đầu tiên trong chặng đường xây dựng các mô hình máy học của mình.

1. **Quy trình xây dựng mô hình máy học:**

Để có thể xây dựng một mô hình máy học, ta phải bám sát vào quy trình xây dựng mô hình ML, bao gồm sáu bước:

* Thu thập dữ liệu: Cần tìm dữ liệu chuẩn xác, được cung cấp bởi những tổ chức có uy tín.
* Tìm hiểu, phân tích dữ liệu: Cần hiểu dữ liệu được thu thập, tiến hành việc phân tích tập dữ liệu để có thể tìm được những đặc trưng chi tiết quan trọng, đem lại hiệu quả huấn luyện mô hình cao hơn.
* Lựa chọn mô hình: Lựa chọn mô hình phù hợp với tập dữ liệu, để đem lại kết quả dự đoán có tỉ lệ chính xác cao nhất bằng cách so sánh kết quả dự đoán của các mô hình với nhau.
* Huấn luyện mô hình.
* Đánh giá mô hình: Đánh giá mô hình để xem xét độ chính xác của mô hình.
* Sử dụng mô hình: Sử dụng mô hình để thực hiện thử thách Titanic.

1. **Quy trình thực hiện thử thách Titanic:**

Bước 1: Thu thập dữ liệu: Trên hệ thống Kaggle có sẵn 3 tệp tin dữ liệu:

* Train.csv: Chứa các đặc trưng của hành khách đã đi tàu Titanic, chứa luôn cả đặc trưng *survived* - Cho biết hành khách có sống sót sau vụ đắm tàu hay không.
* Test.csv: Chứa các đặc trưng như file Train.csv. Không bao gồm trường dữ liệu *survived*.
* Gender\_submission.csv: Nhằm để hiển thị file nộp nên có cấu trúc như thế nào.

Vì mục đích huấn luyện mô hình máy học, ta sẽ sử dụng hai file *train.csv* và *test.csv* là chủ yếu.

Bước 2: Phân tích dữ liệu:

Phân tích các đặc trưng chi tiết của file *train.csv*:

* Survival: Phản ánh việc hành khách có sống sót sau vụ đắm tàu Titanic hay không (0 = không; 1 = có).
* Pclass: Hạng vé của hành khách (1 = 1st, 2 = 2nd, 3 = 3rd).
* Name: Tên hành khách.
* Sex: Giới tính.
* SibSp: Số lượng anh chị em/vợ chồng của hành khách cùng ở trên tàu.
* Parch: Số lượng cha mẹ/trẻ em của hành khách cùng ở trên tàu.
* Ticket: Số vé.
* Fare: Phí mà hành khách phải thanh toán để có thể lên tàu Titanic.
* Cabin: Cabin hành khách ở.
* Embarked: Cảng bắt đầu (C, Q, S).
* Boat: Nếu hành khách có sống sót thì được cứu trợ ở thuyền cứu sinh nào.
* Body: Nếu hành khách không sống sót được thì liệu có tìm thấy được thi thể hay không.
* Loại bỏ các trường dữ liệu không cần thiết (cabin, ticket, passengerId)
  + Cabin: Được phản ánh qua trường dữ liệu class (thứ hạng vé của khách)
  + Ticket: Mỗi người có được một số vé khác nhau => Không liên quan đến việc dự đoán tỉ lệ sống sót của hành khách
  + passengerId: Chỉ được dùng để theo dõi dữ liệu trong dataset dễ dàng hơn, không ảnh hưởng đến việc dự đoán khả năng sống sót của khách.
* Lấp đầy thông tin bị bỏ trống của các trường dữ liệu quan trọng (embarked, age, fare,...)
* Kiểm tra và chuyển đổi những trường dữ liệu có kiểu dữ liệu sao cho máy có thể hiểu được (number)
* Sau đó làm một đống gì đó không biết nhưng mà có sử dụng thư viện sklearn.
* Tách dữ liệu label và train data thành 2 bộ dữ liệu riêng.
* Tách dữ liệu train và label thành 2 tập file nhỏ hơn: X\_train, X\_val, y\_train, y\_val.
  + X\_train, y\_train: dùng để train model.
  + X\_val, y\_val: dùng để đánh giá model.

Bước 3: Lựa chọn mô hình:

* Sử dụng sklearn để import các model (trong trường hợp này, code sử dụng 3 model (logistic regression, random forest, xgboost).
* Sử dụng train\_data (X\_train, y\_train) để train cho cả 3 model.
* Dùng 3 model vừa train để dự đoán kết quả (khả năng sống sót) dựa vào data X\_val.
* Sử dụng kết quả đoán được của cả ba model (y\_pred), đem so sánh với kết quả. thực tế (y\_train), cho ra tỉ lệ chuẩn xác (accuracy\_rate).
* Sử dụng accuracy\_rate để lựa chọn model tốt nhất.

Bước 4: Huấn luyện mô hình: Huấn luyện mô hình sử dụng model tốt nhất cho dataset huấn luyện.

Bước 5: Kiểm tra, đánh giá mô hình:

* Sử dụng mô hình vừa train để dự đoán kết quả y\_pred dựa vào data của X\_val.
* So sánh kết quả giữa y\_pred (dự đoán của mô hình) và y\_val (kết quả thực tế) để đánh giá mô hình.

Bước 6: Sử dụng mô hình:

* Load data test gốc.
* Sử dụng mô hình để dự đoán kết quả sống sót dựa vào thông tin trong test\_data bộ dữ liệu.
* Gán kết quả dự đoán với passengerId trong data test gốc.
* Tạo file submission.
* Nộp file submission lên Kaggle.

Kết quả: