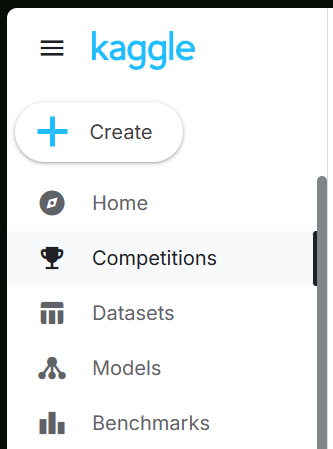
Họ tên: Trần Đại Thắng

MSSV: 3123410346

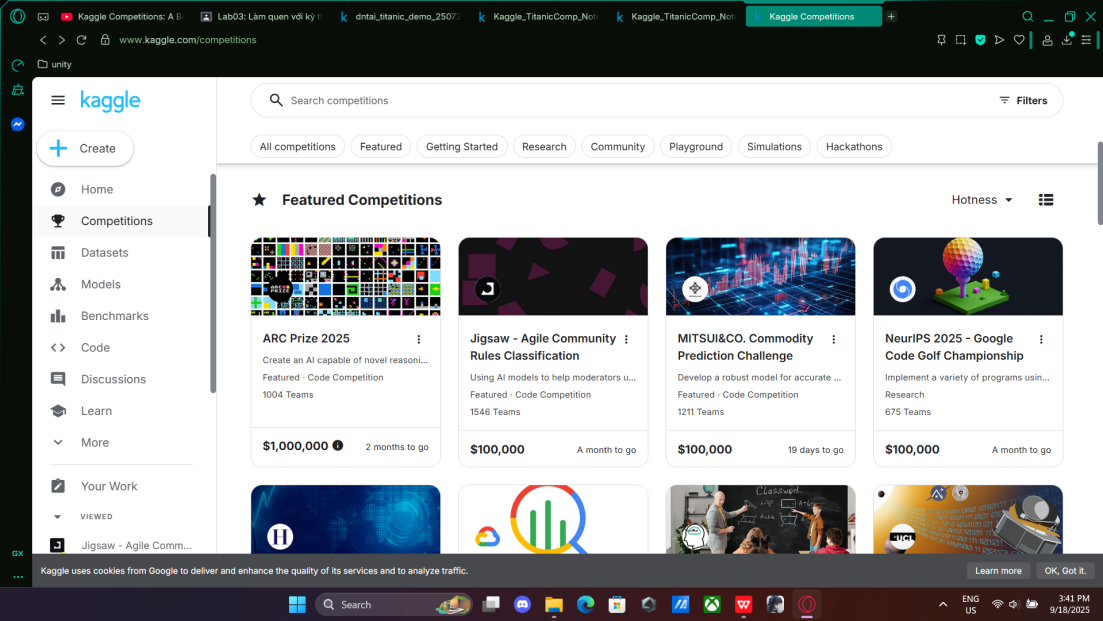
**Lab03: làm quen kì thi Kaggle**

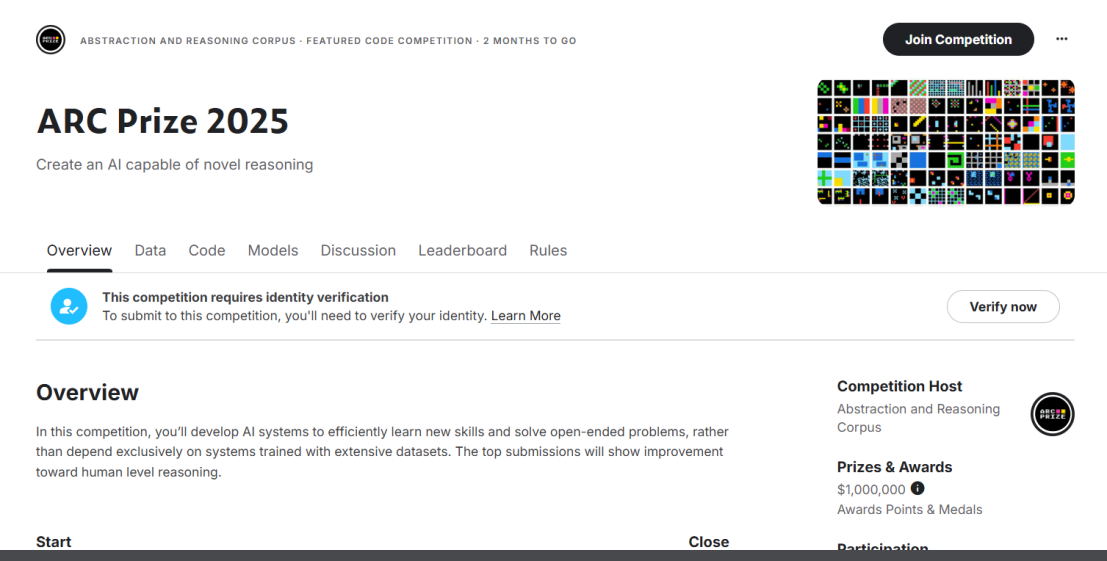
1. Tham dự cuộc thi Kaggle:

B1. Di chuyển đến mục competitions trên kaggle.



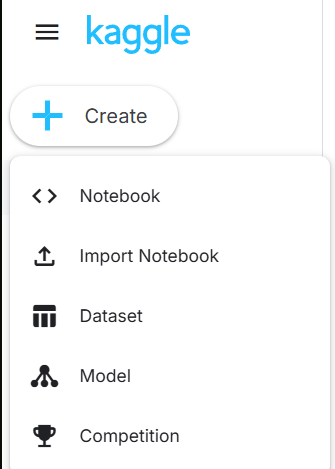
B2. Chọn một trong những competitions hiện có và chọn “Join competition” trên góc phải để tham gia cuộc thi.



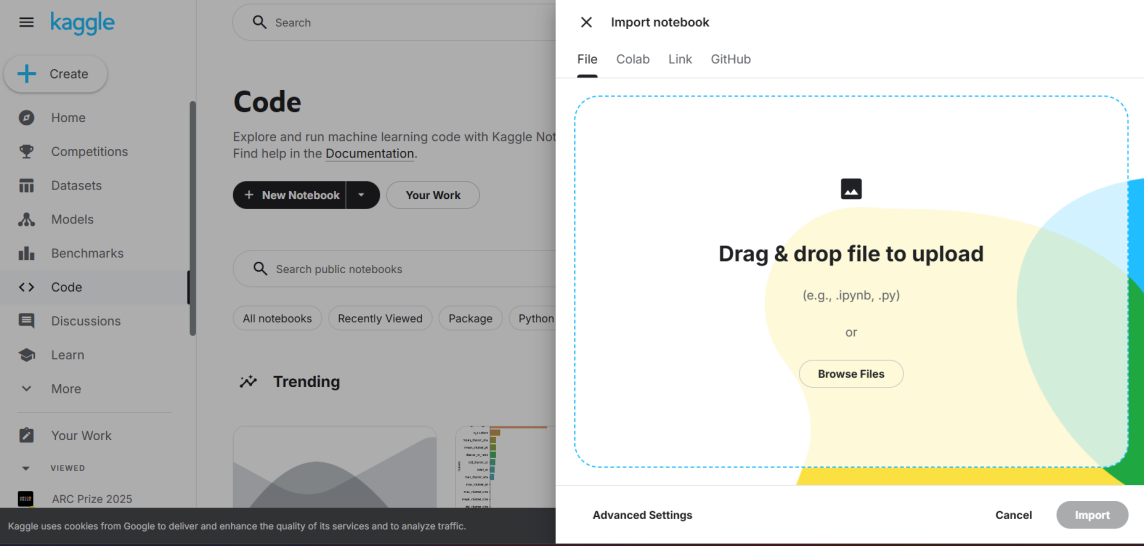


1. Khởi tạo NoteBook trên kaggle:

Cách 1: Chọn Create -> Notebook để tạo một notebook mới.

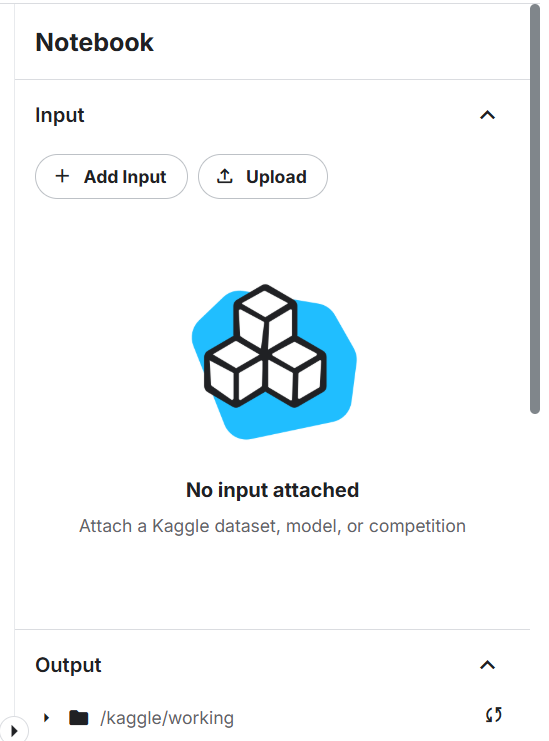


Cách 2: Chọn Import Notebook để tạo notebook mới từ file có sẵn.

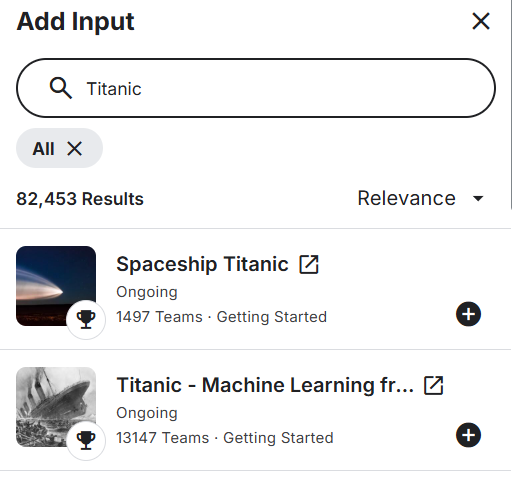


1. Thực hiện thêm dữ liệu input vào Notebook:

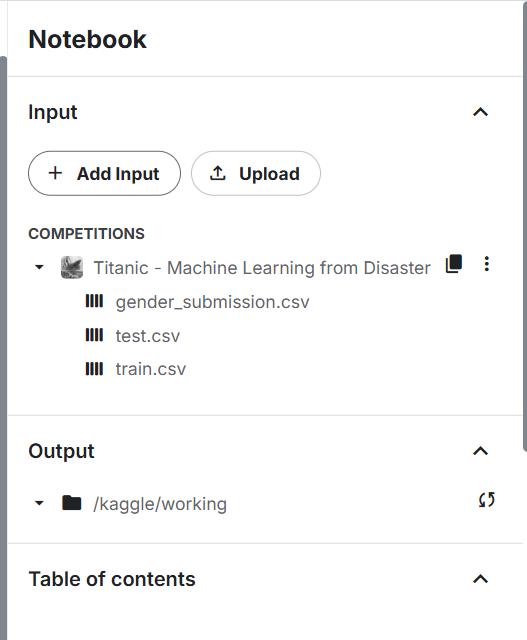
B1. Chọn Add Input bên trong notebook ở phía bên phải (Hoặc upload nếu bạn có sẵn tập Dataset trên máy).



B2. Thực hiện tìm kiếm tập Dataset bạn cần (các tập Dataset của competitions sẽ có icon cúp cùng với số liệu teams và trạng thái).



B3. Thực hiện thêm vào Notebook bằng dấu cộng kế bên (Bạn có thể thấy dataset đã đc thêm vào Notebook bên tay phải).



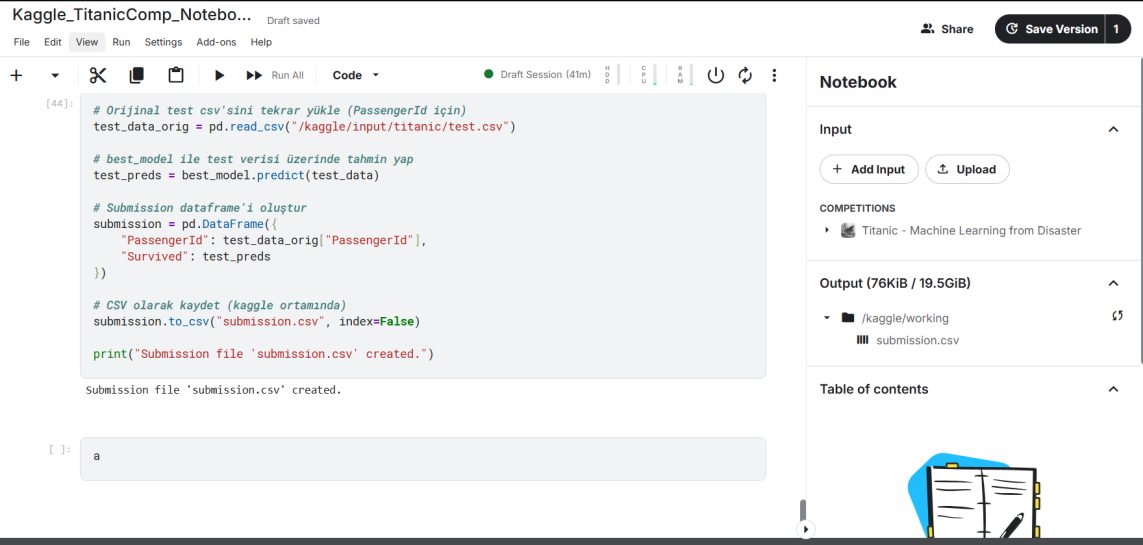
1. Thực hiện chạy code:

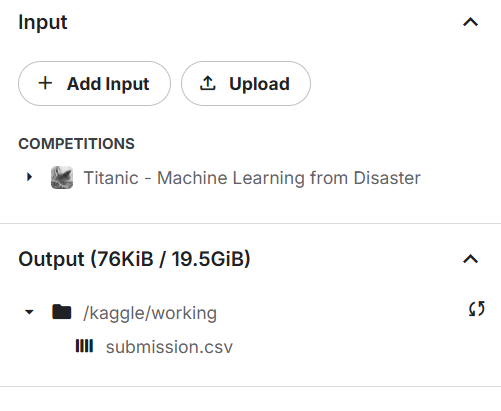
\*Code được cung cấp sẵn:

<https://www.kaggle.com/code/dntai1983/dntai-titanic-demo-250731>

- Đảm bảo các cells chạy ổn định theo notebook mẫu.

- Sau khi chạy xong thì ta sẽ thấy 1 file output từ tập dữ liệu train.





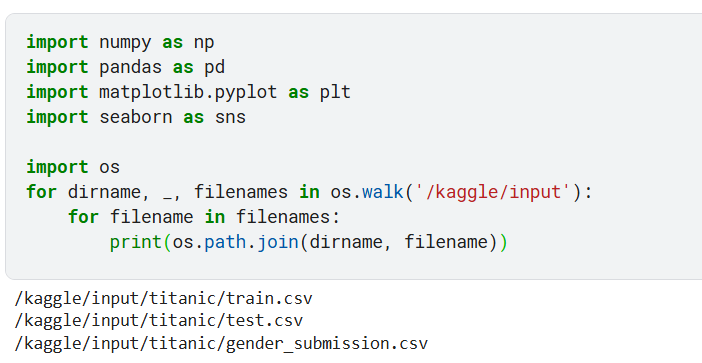
**Lưu ý thực hiện tải tệp dữ liệu dự đoán về và save Notebook lại.**

*\*Link đến Notebook:*

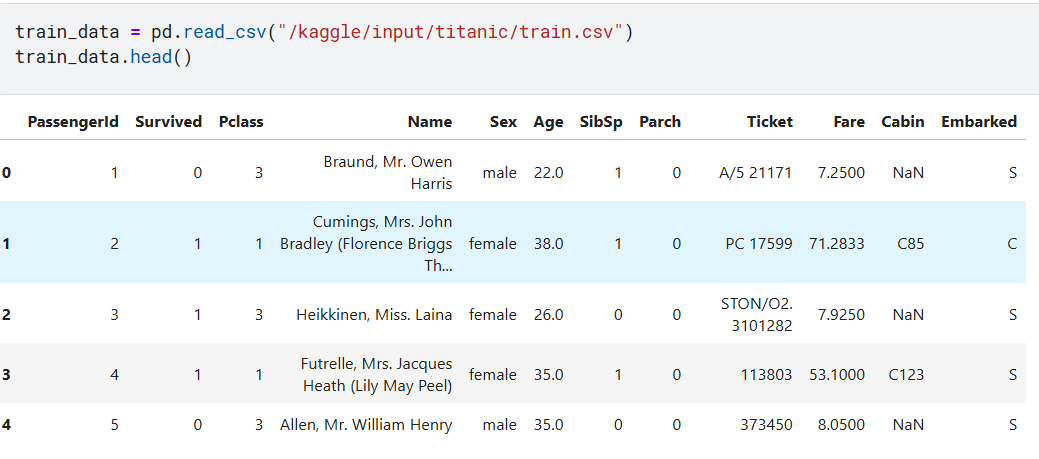
*<https://www.kaggle.com/code/trnithng/kaggle-titaniccomp-notebook>*

* 1. *phân tích code:*

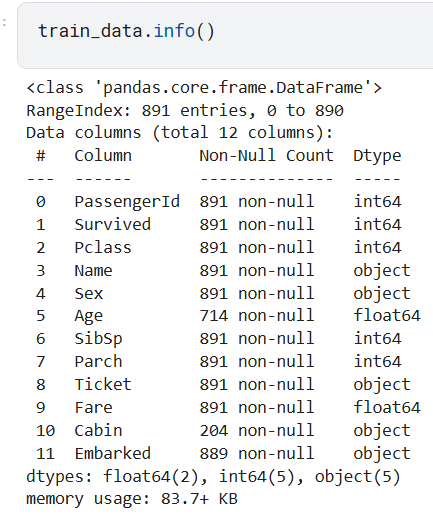
*-Import thư viện và kiểm tra đường dẫn tập dữ liệu.*



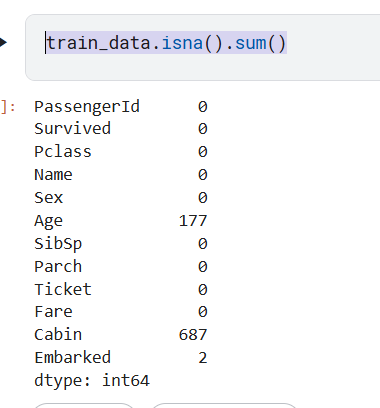
-Chuyển đổi file csv thành mảng train và test data trong python, đồng thời kiểm tra dữ liệu.



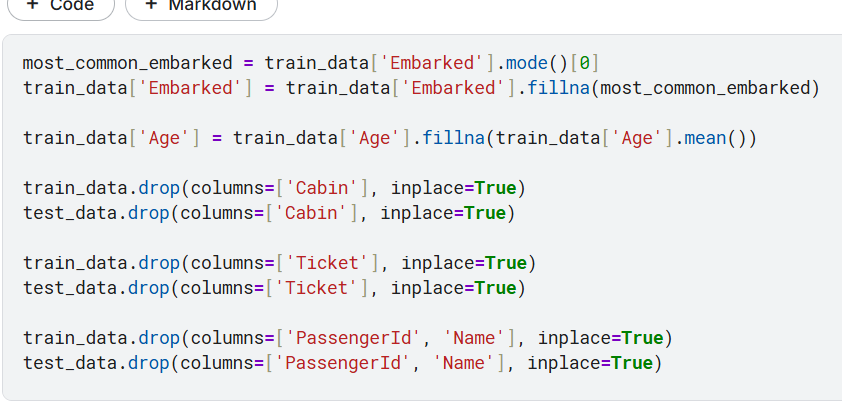
- Các thông tin về tập dữ liệu.



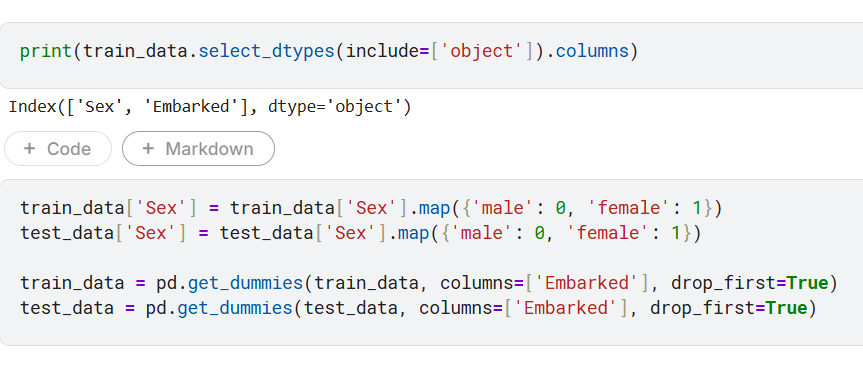
- Kiểm tra tính đầy đủ của các thuộc tính trong tập dữ liệu.



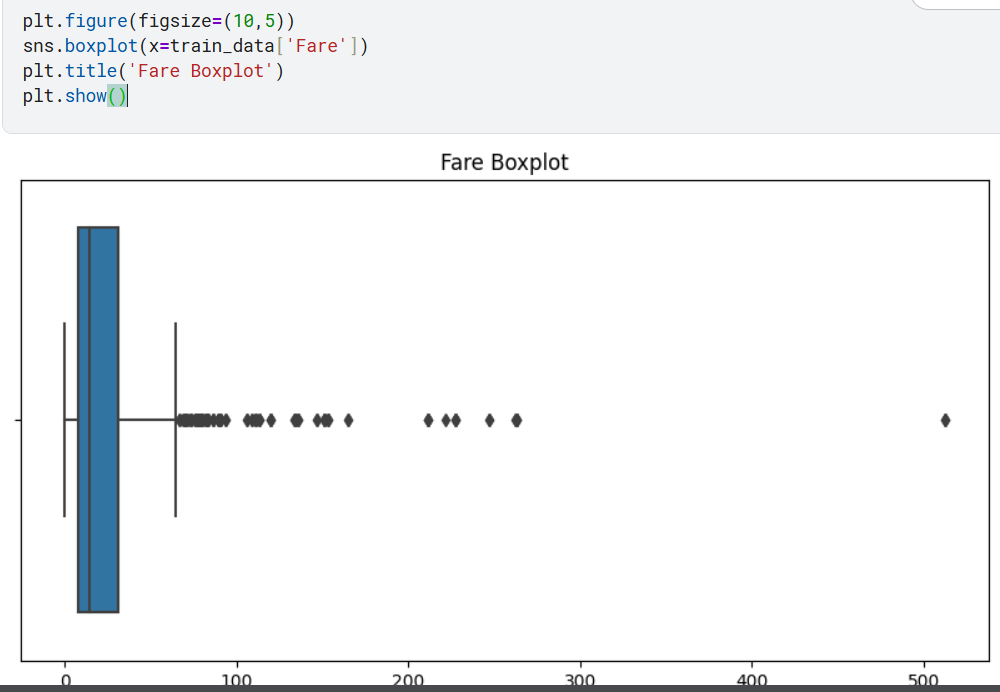
- Thêm vào các dữ liệu xuất hiện thường xuyên ở các thuộc tính cần thiết như Embarked (Tham gia), các giá trị trung bình vào Age (tuổi). Loại bỏ các thuộc tính thiếu quá nhiều dữ liệu để có thể dùng và các thuộc tính không cần thiết như Ticket, Cabin, Passenger ID/ Name. Tương tự với tệp test.



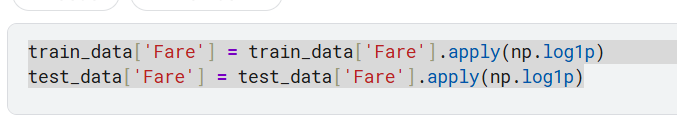
- Thay đổi kiểu dữ liệu về các các kiểu dữ liệu cơ bản để sử dụng. Embarked có 3 giá trị S, C, Q tượng trưng cho 3 địa điểm, việc ta binary hóa thành 3 cột riêng biệt và bỏ 1 cột để tiết kiệm dữ liệu và vẫn giữ tính đúng đoán cho dữ liệu vì nếu cột a và b đều 0 thì nó ám chỉ đến giá trị còn lại.



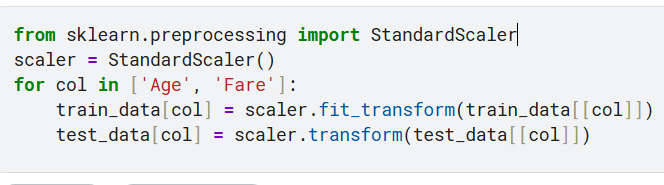
- Biểu đồ thể hiến sự phân phối trong giá vé (phân biệt toa hạng trên tàu).



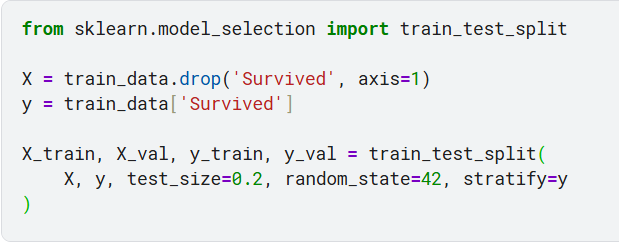
- Sử dụng hàm Log(1 + x) để giảm giá trị lớn của các vé, đồng thời giảm tính biến thiên cho tập giá trị. Việc này đem lại ý nghĩa hơn cho tập giá trị khi train và nâng hiệu suất mô hình.



- Giảm tỉ lệ biến thiên trên tập các giá trị có khoảng cách lớn bằng cách sử dụng phân phối chuẩn.



- Tách biệt giá trị cần dự đoán ra khỏi tập dữ liệu và thực hiện chia thành 2 tập train và validate (đánh giá). Random state chỉ đơn giản là seed random được cố định để luôn có các tập dữ liệu như vậy trong mỗi lần chạy.



- Thực hiện khởi tạo các models, train và xem hiệu suất.

+Accuracy: độ chuẩn xác trong việc dự đoán.

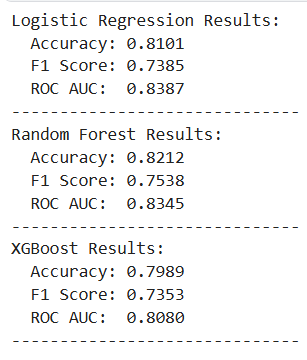
+F1 Score: là tỉ lệ của precision và recall

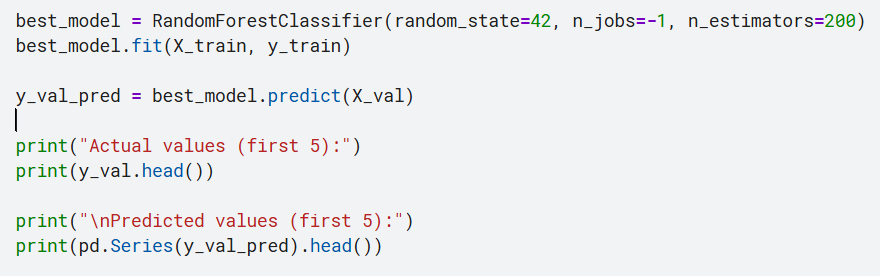
(VD: số người sống sót / số người dự doán).

+ROC AUC: đo lường hiệu suất phân biệt các lớp của mô hình.

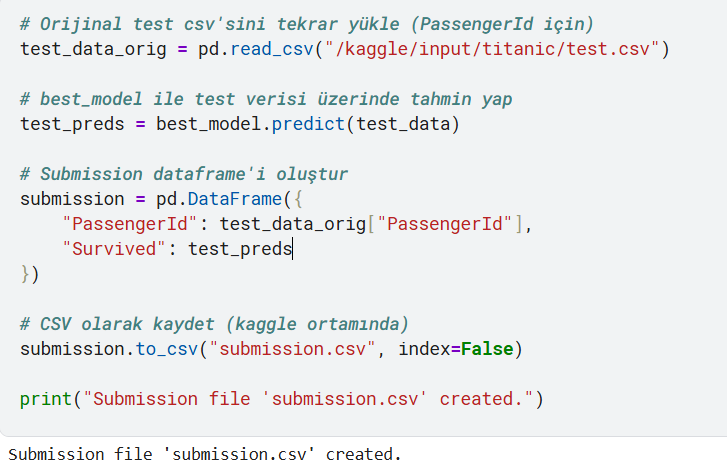


- Trong số các mô hình Random Forest có hiệu suất tốt nhất nên ta sẽ thực hiện train và dùng nó để dự đoán tập test cho cuộc thi. So sánh 5 dòng đầu của dự đoán và sự thật, ta thấy được 4/5 hay 80% dự đoán đã chính xác.



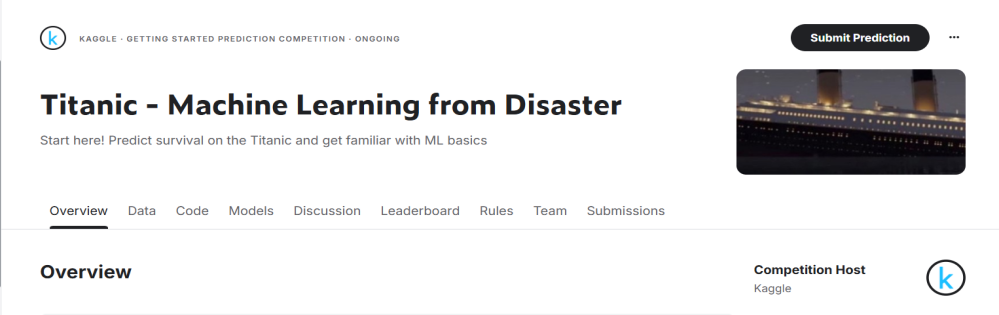


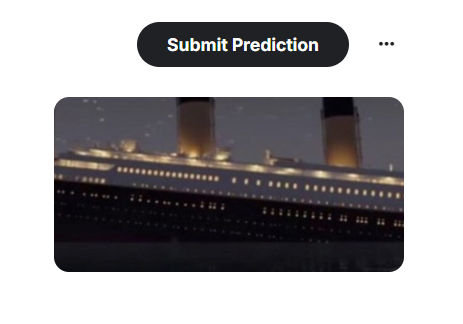
- Lưu trữ file csv dự đoán về để nộp cho cuộc thi kaggle.



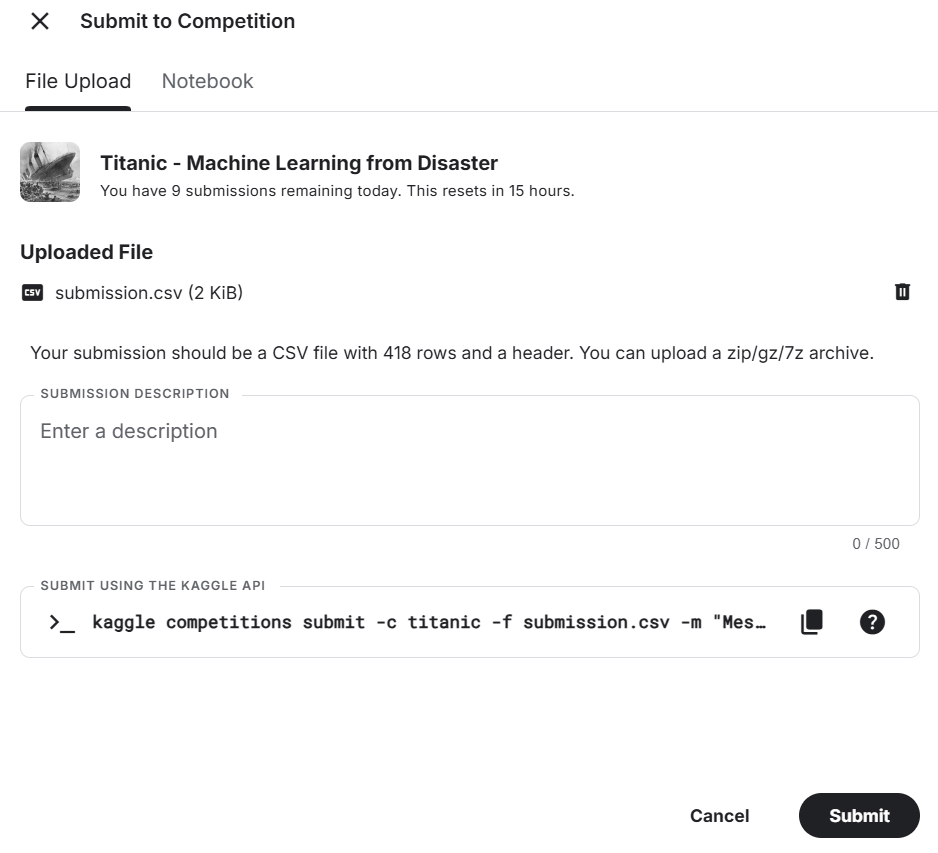
1. Thực hiện nộp kết quả dự đoán:

- Quay trở về trang competitions chọn cuộc thi cần nộp kết quá. Thực hiện “Submit Prediction” trên góc phải.

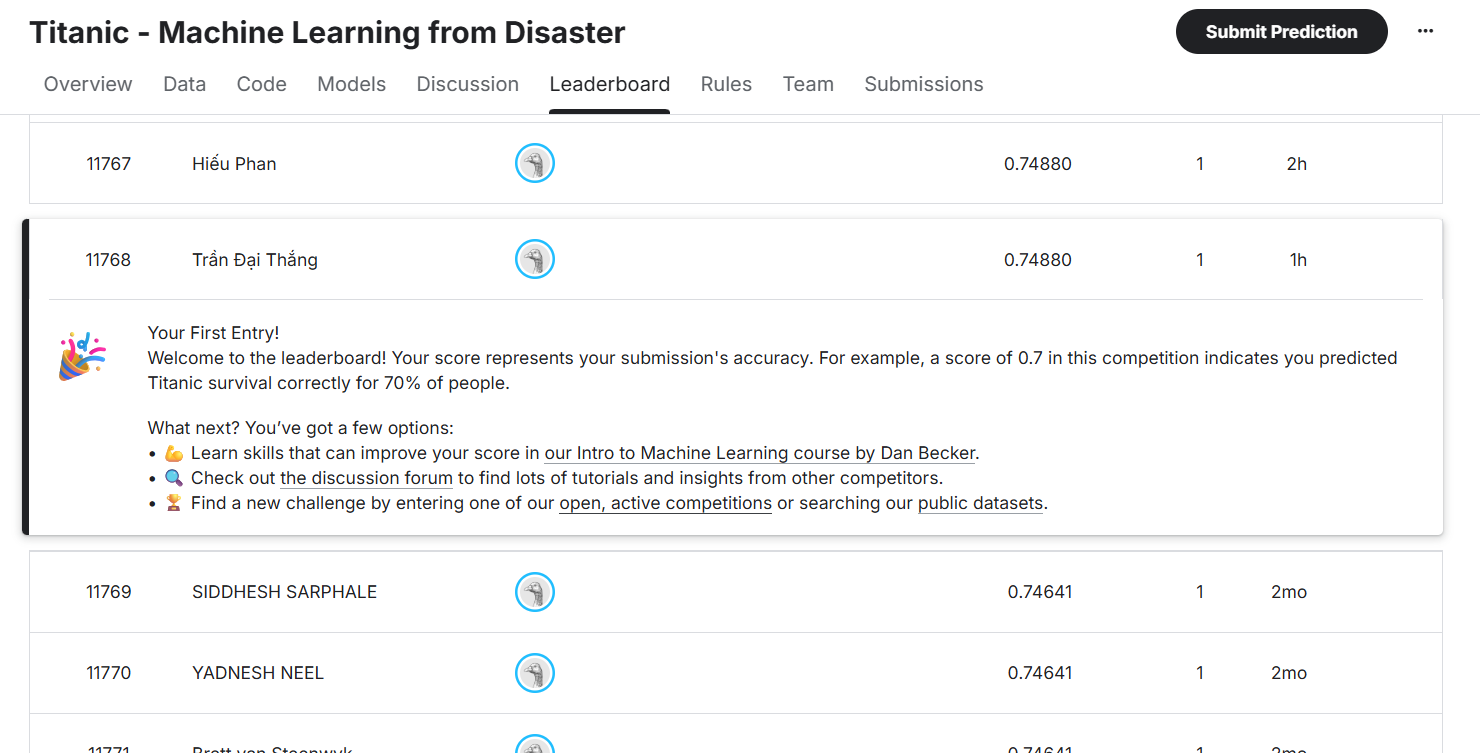


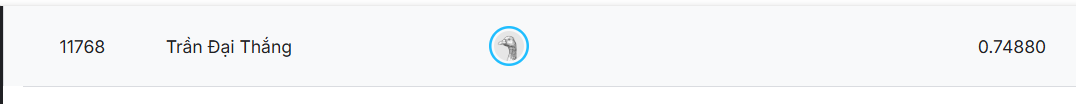


- Tải tệp dự đoán của bạn lên và thực hiện “Submit”.



- Kiểm tra kết quá dự đoán model bạn đưa ra trong Global Leaderboard.





**- Kết Thúc -**