TITANIC-MACHINE LEARNING FROM DISASTER

Biện Hữu Khang, Nguyễn Thị Thùy Linh, Mai Hoàng Hải Yến

1.GIỚI THIỆU

1.1. Định nghĩa vấn đề:

- **Input:** dựa vào tập train lựa chọn các đặc trưng để huấn luyện mô hình.
- **Ouput:** Dự đoán khả năng sống sót ở tập test.

1.2. Thử thách:

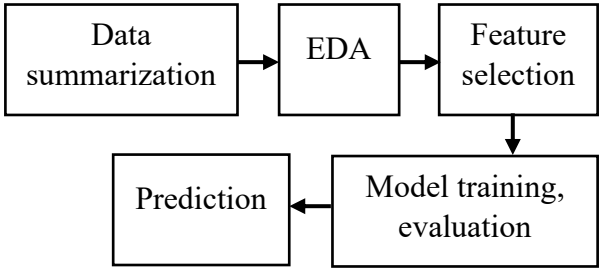
- Dữ liệu có vài đặc trưng thiếu giá trị
- Một số đặc trưng có kiểu dữ liệu không đúng
- Mô hình cần cân bằng giữa độ chính xác và khả năng tổng quát, tránh overfitting

1.3. Mục tiêu báo cáo: Xây dựng và đánh giá mô hình dự đoán khả năng sống sót của hành khách trên tàu Titanic dựa trên các đặc trưng cá nhân và thông tin vé tàu.

2. DATASET:

- Bộ dữ liệu gồm 2 tập (train, test):
  - Tập train gồm có 891 mẫu và 12 đặc trưng
  - Tập test gồm có 418 mẫu và 12 đặc trưng
- Data source: https://colab.research.google.com/drive/1YP29RvhKPuBYc4vUO87fD4JLkQFc2lmK

3. PHƯƠNG PHÁP ĐỀ XUẤT



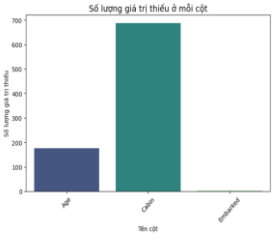
3.1. Data summarization:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None
```

→ Thông tin về bộ dữ liệu tập Train

3.2. EDA

3.2.1. Giá trị thiếu

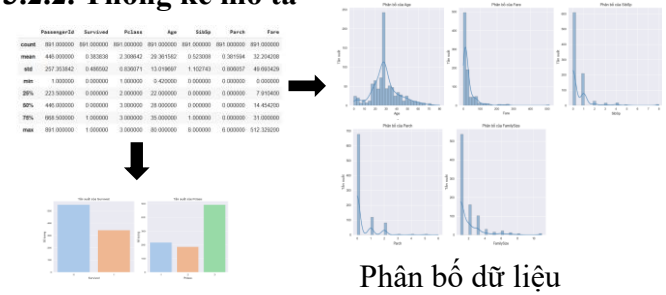


Xử lý giá trị thiếu →

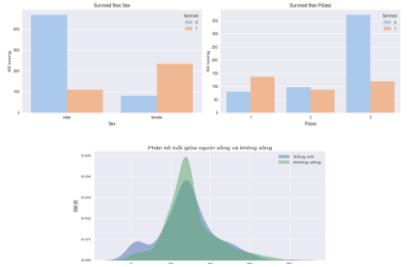
```
train_df.isnull().su
train_df.isna().sum(

PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
dtype: int64
```

3.2.2. Thống kê mô tả

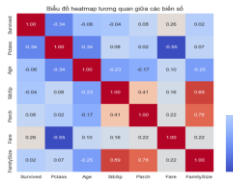


Phân bố dữ liệu



So sánh khả năng sống sót với nam nữ, hạng vé, tuổi

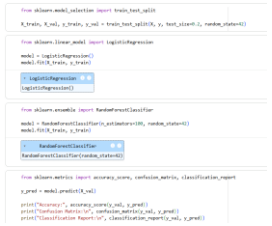
3.3.Feature selection



Chọn đặc trưng →

Age', 'Fare', 'FamilySize', 'Sex', 'Pclass_2', 'Pclass_3']

4. Model training, evaluation



```
Accuracy: 0.8100558659217877
Confusion Matrix:
 [[89 16]
 [18 56]]
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.85      0.84       105
           1       0.78      0.76      0.77        74

    accuracy                           0.81       179
   macro avg       0.80      0.80      0.80       179
weighted avg       0.81      0.81      0.81       179
```

81% mẫu validation được dự đoán đúng. Hàng đầu tiên = class 0 (không sống sót). True Negative = 89 (dự đoán đúng là 0). False Positive = 16 (dự đoán sống sót nhầm). Hàng thứ hai = class 1 (sống sót). False Negative = 18 (dự đoán không sống sót nhầm). True Positive = 56 (dự đoán đúng là sống sót). Precision (1) = 0.78 số dự đoán sống sót. Recall (1) = 0.76 số người thật sự sống sót. F1-score (1) = 0.77 → kết hợp precision & recall. Class 0 có hiệu suất nhỉnh hơn → phù hợp với mất cân bằng nhãn (số người không sống sót > số người sống sót).

5. Prediction

Output →

PassengerId	Survived
892	0
893	0
894	1
895	1
896	1
897	0
898	1
899	0
900	0
901	0