

## 1.GIỚI THIỆU

### 1.1. Định nghĩa vấn đề:

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

### 1.2. Thủ thách:

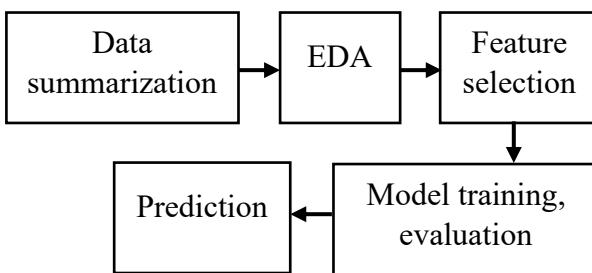
- Dữ liệu có vài đặc trưng thiếu giá trị
- Một số đặc trưng có kiểu dữ liệu không đúng
- Mô hình cần cân bằng giữa độ chính xác và khả năng tổng quát, tránh overfitting

**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



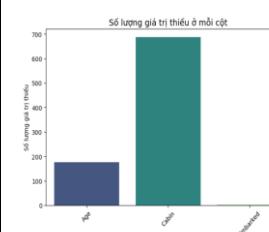
### 3.1. Data summarization:

```

<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
  
```

### 3.2. EDA

#### 3.2.1. Giá trị thiếu



#### Xử lý giá trị thiếu

```

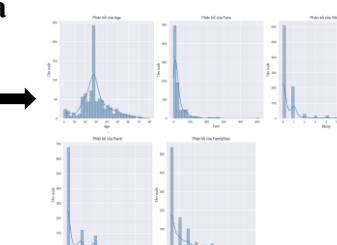
train_df.isnull().sum()
train_df.isna().sum()

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

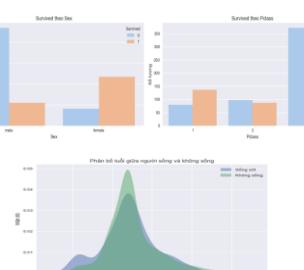
#### 3.2.2. Thống kê mô tả

```

PassengerId  Survived  Pclass  Age  SibSp  Parch  Fare
Count      891.000000  891.000000  12.000000  12.000000  12.000000  83.700000
mean      445.000000  0.383838  2.39942  29.19587  0.523200  32.39450
std       257.35142  0.466502  1.00000  13.96967  1.15740  49.90450
min       1.000000  0.000000  1.00000  0.400000  0.00000  0.00000
25%      223.000000  0.000000  2.00000  22.00000  0.00000  7.91000
50%      445.000000  0.000000  3.00000  28.00000  0.00000  14.45400
75%      693.000000  1.000000  3.00000  35.00000  1.00000  31.00000
max      891.000000  1.000000  3.00000  80.00000  0.00000  512.32600
  
```

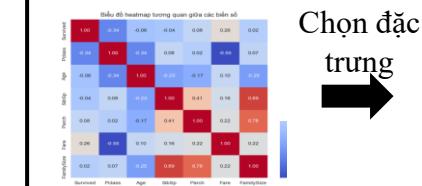


#### Phân bố dữ liệu



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

### 3.3. Feature selection



Age', 'Fare', 'FamilySize',  
'Sex', 'Pclass\_2', 'Pclass\_3']

### 4. Model training, evaluation

```

from sklearn.model_selection import train_test_split
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2, random_state=42)

from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X_train, y_train)
model.score(X_val, y_val)

from sklearn.metrics import classification_report
print(classification_report(y_val, model.predict(X_val)))

accuracy: 0.83
precision: 0.85
recall: 0.84
f1-score: 0.85
support: 105

accuracy: 0.78
precision: 0.76
recall: 0.77
f1-score: 0.76
support: 74
  
```

```

Accuracy: 0.8389558659217877
Confusion Matrix:
[[16 16]
 [18 96]]
Classification Report:
precision    recall    f1-score   support
0           0.83      0.85      0.84      105
1           0.78      0.76      0.77      74
accuracy                           0.80      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 môt cân bằng nhẫn (số người không sống sót > số người sống sót).

### 5. Prediction

PassengerId	Name	Sex	Age	SibSp	Parch	Fare	Ticket	Embarked	Pclass_2	Pclass_3	Survived
0	Allen, Mr. William Henry	Male	30	1	0	33.3333	C. 132	Q	3	1	0
1	Allen, Mrs. William Henry	Female	31	1	0	32.3333	130.0000	S	3	1	1
2	Anderson, Mr. Charles Froome	Male	32	0	0	24.5000	310.3000	Q	3	1	1
3	Anderson, Mrs. Charles Froome	Female	37	0	0	23.0000	310.3000	S	3	1	1
4	Anderson, Miss. Heloise	Female	22	1	1	32.3333	132.0000	S	3	1	1
...	...	...	...	...	...	...	...	...	...	...	...
403	Brickell, Mr. William Franklin	Male	37	0	0	15.3333	132.0000	S	3	1	1
404	Brickell, Mrs. William Franklin	Female	36	0	0	30.3333	130.0000	C	3	1	1
405	Brickell, Mr. Charles Franklin	Male	33	0	0	23.3333	310.3000	S	3	1	1
406	Brickell, Mrs. Charles Franklin	Female	34	0	0	23.3333	310.3000	C	3	1	1
407	Brickell, Miss. Florence Franklin	Female	22	1	1	32.3333	132.0000	S	3	1	1
408	Brickell, Mr. James Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
409	Brickell, Mrs. James Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
410	Brickell, Miss. Mary Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
411	Brickell, Mr. Thomas Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
412	Brickell, Mrs. Thomas Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
413	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
414	Brickell, Mr. John Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
415	Brickell, Mrs. John Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
416	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
417	Brickell, Mr. Charles Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
418	Brickell, Mrs. Charles Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
419	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
420	Brickell, Mr. John Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
421	Brickell, Mrs. John Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
422	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
423	Brickell, Mr. Charles Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
424	Brickell, Mrs. Charles Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
425	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
426	Brickell, Mr. John Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
427	Brickell, Mrs. John Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
428	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
429	Brickell, Mr. Charles Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
430	Brickell, Mrs. Charles Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
431	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
432	Brickell, Mr. John Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
433	Brickell, Mrs. John Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
434	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
435	Brickell, Mr. Charles Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
436	Brickell, Mrs. Charles Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
437	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
438	Brickell, Mr. John Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
439	Brickell, Mrs. John Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
440	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
441	Brickell, Mr. Charles Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
442	Brickell, Mrs. Charles Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
443	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
444	Brickell, Mr. John Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
445	Brickell, Mrs. John Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
446	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
447	Brickell, Mr. Charles Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
448	Brickell, Mrs. Charles Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
449	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
450	Brickell, Mr. John Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
451	Brickell, Mrs. John Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
452	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
453	Brickell, Mr. Charles Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
454	Brickell, Mrs. Charles Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
455	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
456	Brickell, Mr. John Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
457	Brickell, Mrs. John Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
458	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
459	Brickell, Mr. Charles Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
460	Brickell, Mrs. Charles Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
461	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
462	Brickell, Mr. John Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
463	Brickell, Mrs. John Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
464	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
465	Brickell, Mr. Charles Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
466	Brickell, Mrs. Charles Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
467	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
468	Brickell, Mr. John Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
469	Brickell, Mrs. John Franklin	Female	34	0	0	30.3333	130.0000	C	3	1	1
470	Brickell, Miss. Anna Franklin	Female	22	1	1	32.3333	132.0000	C	3	1	1
471	Brickell, Mr. Charles Franklin	Male	35	0	0	15.3333	132.0000	C	3	1	1
472	Brickell, Mrs. Charles Franklin	Female	34	0	0	30.3333	130.0000	C	3	1</	