**Dự Đoán Khả Năng Sống Sót Thảm Họa Titanic Bằng Machine Learning**

**- Nguyễn Tuấn Đạt, Châu Hải Đăng, Trần Đại Thắng -**

**Problem Review**

Thuộc dạng bài toán phân lớp (Classification)

**Input**:

- Dữ liệu hành khách gồm thông tin cá nhân và vé tàu từ **Kaggle Titanic Dataset**.

**Output**:

- Khả năng sống sót của của hành khách (0 hay 1).

**Notice**:

- Tập dữ liệu bị thiếu, mất mát và nhiễu. Cần tiền xử lí và phân tích đặc trưng.

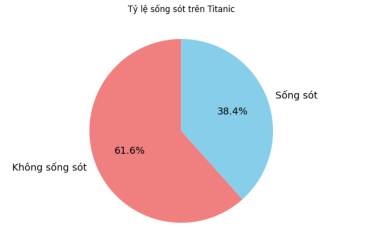
**Dataset**

**Source:**

- Kaggle – Titanic: Machine Learning from Disaster.

<https://www.kaggle.com/competitions/titanic/data>

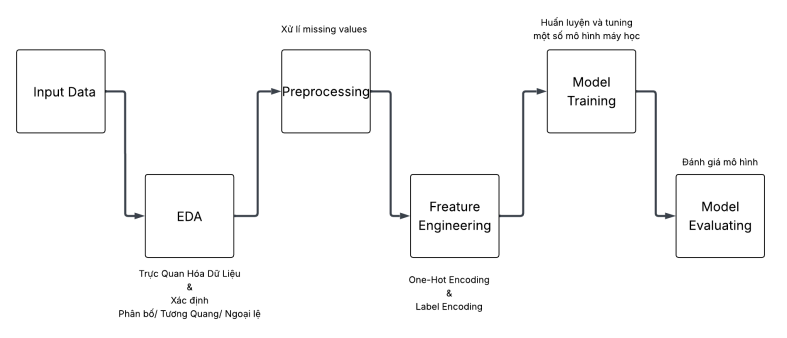
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Embarked | Bến khởi hành | Parch | Cha mẹ, Con cái | SibSp | Anh, chị, em và họ hàng |
| PClass | Hạng khoa | Fare | Giá vé | PassengerId | Mã hành khách |
| Age | Tuổi | Sex | Giới tính | Survived | Sống sót |
| Name | Tên | Cabin | Số cabin | Ticket | Mã vé |



- Gồm 891 bộ train và 418 bộ test.

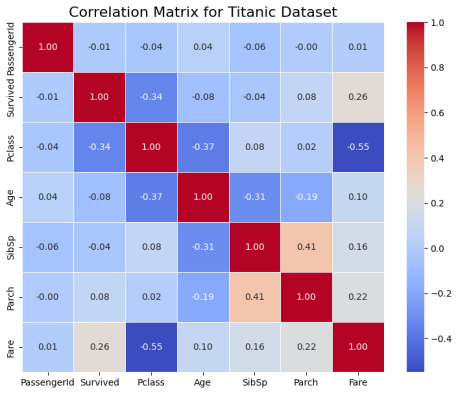
- Tỉ lệ phân lớp của tập train xấp xỉ 1:2

**Propose Method**

****

**Exploring Data Analysis**

|  |  |
| --- | --- |
| Đáng chú ý | |
| Fare | Mật độ outlier dày đặc |
| Age |
| Sibps | Mật thiết với và có thể xử lí chung |
| Parch |



|  |  |  |
| --- | --- | --- |
| Thuộc tính | Thiếu | Tỉ lệ |
| Cabin | 687 | 77.1% |
| Age | 177 | ~20% |
| Embark | 2 | 0.22% |

**Preprocessing & Freature Engineering**

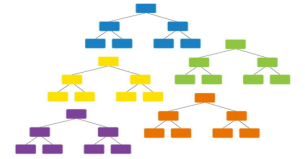
- Thực nghiệm 1: One-hot Encoding:

|  |  |  |
| --- | --- | --- |
| Đặc trưng | Phương Thức xử lí | One-hot encoding |
| PassengerID, Ticket | Đặc trưng không hữu dụng với mô hình. Loại bỏ (drop). |  |
| Name | Trích xuất đặc trưng thành danh hiệu - Tittle. | X |
| Age, Fare | Sử dụng trung vị (median) cho dữ liệu thiếu |  |
| Embarked | Giả sử các giá trị thiếu là cảng phổ biến nhất vì chỉ có 2 dữ liệu thiếu. | X |
| Cabin | Chuyển mã cabin về thành phân chia khu vực trong tàu, các giá trị thiếu cho là unknown. | X |
| Sex | Chỉ One-hot. | X |

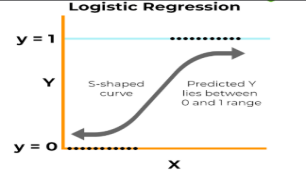
- Thực nghiệm 2: Label Encoding (Phương thức xử lí tương tự thực nghiệm 1, Chỉ khác Encoding)

**Model Training**

**Proposed models:**

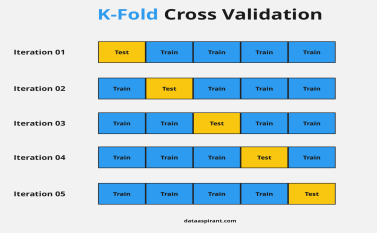


**2.Random Forest**



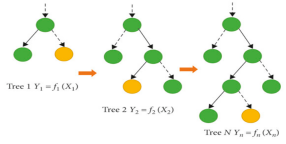
**1.Logistic Regression**

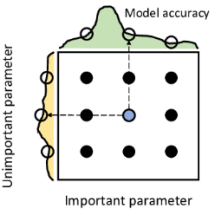
**Evaluating Method:**



- K-Fold Cross validation và kaggle submition score.

**3.Gradient boosting**





**Tuning Method:**

- Kết hợp giữa K-Fold và GridSearch

Cho các hyperparameters của các model.

**Model Evaluating**

**Ex1**: One-Hot Encoding

|  |  |  |
| --- | --- | --- |
| **Baseline** | **Kaggle Score** | **accuracy** |
| LogisticRegression | [0.78708](https://www.kaggle.com/competitions/titanic/submissions) | 79.90% |
| LogisticRegressionCV | [0.78708](https://www.kaggle.com/competitions/titanic/submissions) | 79.10% |
| RandomForestClassifier | [0.73923](https://www.kaggle.com/competitions/titanic/submissions) | 81.58% |
| GradientBoostingClassifier | 0.76555 | 82.37% |

**Ex2:** Label Encoding

|  |  |  |
| --- | --- | --- |
| **Baseline** | **Kaggle Score** | **accuracy** |
| LogisticRegression | [0.75837](https://www.kaggle.com/competitions/titanic/submissions) | 81.48% |
| LogisticRegressionCV | [0.78468](https://www.kaggle.com/competitions/titanic/submissions) | 81.26% |
| RandomForestClassifier | [0.77033](https://www.kaggle.com/competitions/titanic/submissions) | 81.03% |
| GradientBoostingClassifier | [0.76794](https://www.kaggle.com/competitions/titanic/submissions) | 82.71% |

**Result**

**Best Model:**

**-** Logistic Regression.

**Best Fearture Engineering method:**

- One-hot Encoding.

**Prove**:

