**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

🙤🙧🟍🙥🙦



**PHÂN TÍCH DỮ LIỆU TITANIC**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện: | | |
| STT | Họ tên | MSSV |
| 1 | Nguyễn Huỳnh Bảo Quốc | 20520292 |
| 2 | Châu Đức Hiệp | 20520499 |

**TP. HỒ CHÍ MINH – 04/2022**

# GIỚI THIỆU

Bài báo cáo này nhằm phân tích tỉ lệ sống sót của hành khách trên tàu Titanic, thông qua việc sử dụng Python và các thư viện hỗ trợ như Pandas, NumPy, Matplotlib, Seaborn và Plotly.

Trong bối cảnh này, chúng em áp dụng kỹ thuật Machine Learning để dự đoán khả năng sống sót của hành khách dựa trên các thuộc tính như tuổi, giới tính, hạng ghế, số lượng người thân đi cùng, và nhiều yếu tố khác có trong bộ dữ liệu.

Quy trình phân tích dữ liệu của chúng em bao gồm các bước chính: khám phá dữ liệu, tiền xử lý dữ liệu, xây dựng và tối ưu hóa mô hình, đánh giá mô hình và trực quan hóa kết quả. Mỗi bước trong quy trình được thực hiện thông qua việc sử dụng các thư viện Python phù hợp, giúp tăng cường hiệu quả và chính xác trong việc phân tích dữ liệu.

Thông qua báo cáo này, chúng tôi mong muốn cung cấp một giải pháp toàn diện và trực quan cho việc phân tích dữ liệu, phù hợp cho cả người mới học và chuyên gia dữ liệu, giúp tiết kiệm thời gian và công sức, đồng thời tăng cường hiệu quả trong công việc kinh doanh, nghiên cứu và phát triển..

# NỘI DUNG

## Tổng quan

Báo cáo phân tích này nhằm phân tích khả năng sống sót của hành khách trên con tàu Titanic, một trong những thảm họa đắm tàu nổi tiếng nhất trong lịch sử. Đây là một phần của cuộc thi Machine Learning trên nền tảng Kaggle.

Phương pháp và công cụ: Phân tích được thực hiện bằng các công cụ phân tích dữ liệu khác nhau, bao gồm: Python và các thư viện hỗ trợ như Pandas, Matplotlib, Seaborn và Scikit-Learn. Dữ liệu được phân tích là dữ liệu về hành khách trên con tàu Titanic, được lấy từ cuộc thi Kaggle.

Mô tả dữ liệu: Dữ liệu được thu thập từ hai tập dữ liệu: train.csv và test.csv. Train.csv chứa thông tin chi tiết về một phần hành khách trên tàu (891 người) và quan trọng nhất là thông tin về việc họ có sống sót sau thảm họa hay không - được gọi là "ground truth". Trong khi đó, test.csv chứa thông tin tương tự nhưng không tiết lộ "ground truth" cho mỗi hành khách. Nhiệm vụ của chúng ta là dự đoán những kết quả này.

Thông tin về hành khách bao gồm: tên, tuổi, giới tính, tầng lớp kinh tế xã hội, v.v

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Mô tả** | **Khoá** |
| 1 | PassengerId | Mã hành khách |  |
| 2 | Pclass | Hạng vé | 1 = 1st ,  2 = 2nd ,  3 = 3rd |
| 3 | Name | Tên hành khách |  |
| 4 | Sex | Giới tính |  |
| 5 | Age | Tuổi |  |
| 6 | SibSp | Số lượng anh chị em hoặc vợ/chồng đi cùng |  |
| 7 | Parch | Số lượng cha mẹ hoặc con cái đi cùng |  |
| 8 | Ticket | Mã vé |  |
| 9 | Fare | Giá vé của hành khách |  |
| 10 | Embarked | Cảng lên đường | C = Cherbourq,  Q = Queenstown,  S = Southampton |

Mục tiêu của báo cáo là sử dụng các mô hình học máy để dự đoán loại người nào có khả năng sống sót cao hơn sau thảm họa. Mục tiêu cuối cùng là xác định các mô hình dự đoán có độ chính xác cao nhất và đưa ra những insight hữu ích từ việc phân tích dữ liệu.

## Quy trình xử lý dữ liệu

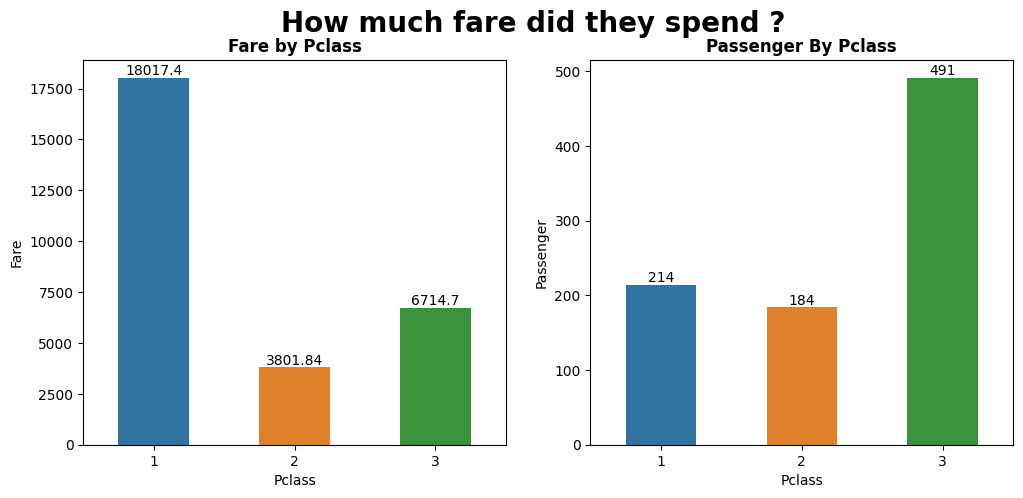
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Xử lý** | **Mô tả chi tiết** | **Công cụ** | **Thời gian** |
| 1 | Tổng hợp dữ liệu | Đọc dữ liệu từ file csv | MS Excel, Python | 13:10 24/04/2023 |
| 2 | Kiểm tra dữ liệu | Kiểu tra dữ liệu null: cột Age, Cabin, Embraked bị null. | Python | 19:47 24/04/2023 |
| 3 | Lọc dữ liệu | Kiểm tra các dữ liệu ở cột Age, Cabin, Embraked bị null có ảnh hưởng đến quá trình phân tích không | Python | 20:06 24/04/2023 |
| 4 | Xử lý dữ liệu để phân tích | Cột Age: lấy tuổi trung vị của mỗi Pclass, sau đó fill các dữ liệu tuổi bị null. | Python,  MS Excel | 20:10 24/04/2023 |
| 5 | Lọc dữ liệu | Kiểu tra dữ liệu null: cột Cabin, Embraked bị null.  Cột Age đã hết null | Python,  MS Excel | 20:15 24/04/2023 |
| 6 | Xóa dữ liệu | Xoá cột Cabin vì có quá nhiều dữ liệu bị null sẽ ảnh hưởng đến quá trình phân tích | Python,  MS Excel | 20:27 24/04/2023 |
| 7 | Xóa dữ liệu | Xoá 2 dòng có dữ liệu Embarked bị null. | Python,  MS Excel | 21:02 24/04/2023 |
| 8 | Lọc dữ liệu | Không còn cột dữ liệu nào bị null. | Python,  MS Excel | 14:30 25/04/2023 |
| 9 | Kiểm tra kiểu dữ liệu | Kiểm tra cụ thể kiểu dữ liệu của các cột | Python,  MS Excel | 16:00 25/04/2023 |
| 10 | Chuyển kiểu dữ liệu | Chuyển kiểu dữ liệu của cột Age từ float64 sang int | Python,  MS Excel | 17:00 25/04/2023 |

Kiểm tra dữ liệu null:

* Có 177 dòng dữ liệu trường Age bị null
* Có 687 dòng dữ liệu trường Cabin bị null
* Có 2 dòng dữ liệu Embraked bị null

## Phân tích dữ liệu

1. **Số tiền mà các hành khách phải chi để lên tàu Titanic**

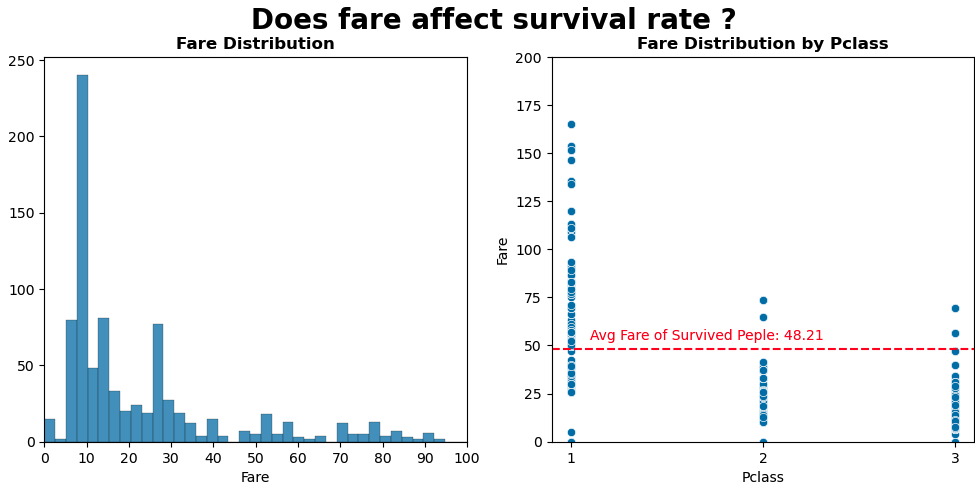


*Hình 1: Biểu đồ Fare by Pclass và Passenger by Pclass.*

Từ biểu đồ Fare by Pclass và Passenger by Pclass, có thể đưa ra nhận xét sau:

* Tàu Titanic gồm ba hạng ghế khác nhau: Hạng Thương Gia (1), Hạng Trung Lưu (2) và Hạng Phổ Thông (3).
* Từ biểu đồ, chúng ta thấy rằng phần lớn hành khách (55.23% tổng số) chọn hạng Phổ Thông, tuy nhiên, hạng Thương Gia, dù ít người hơn, đã chi trả một tổng số tiền vượt trội cho vé (63.1% tổng số). Điều này cho thấy, mặc dù hạng Phổ Thông hấp dẫn đại đa số người đi, nhưng hạng Thương Gia lại chiếm ưu thế về tổng thu từ vé.

Tiếp theo, ta cùng phân tích trả lời câu hỏi: “Liệu vé có ảnh hưởng gì đến tỉ lệ sống sót không?”



*Hình 2: Biểu đồ Fare Distribution và Fare Distribution by Pclass.*

Từ biểu đồ “Fare Distribution” và “Fare Distribution by Pclass”, ta có nhận xét sau:

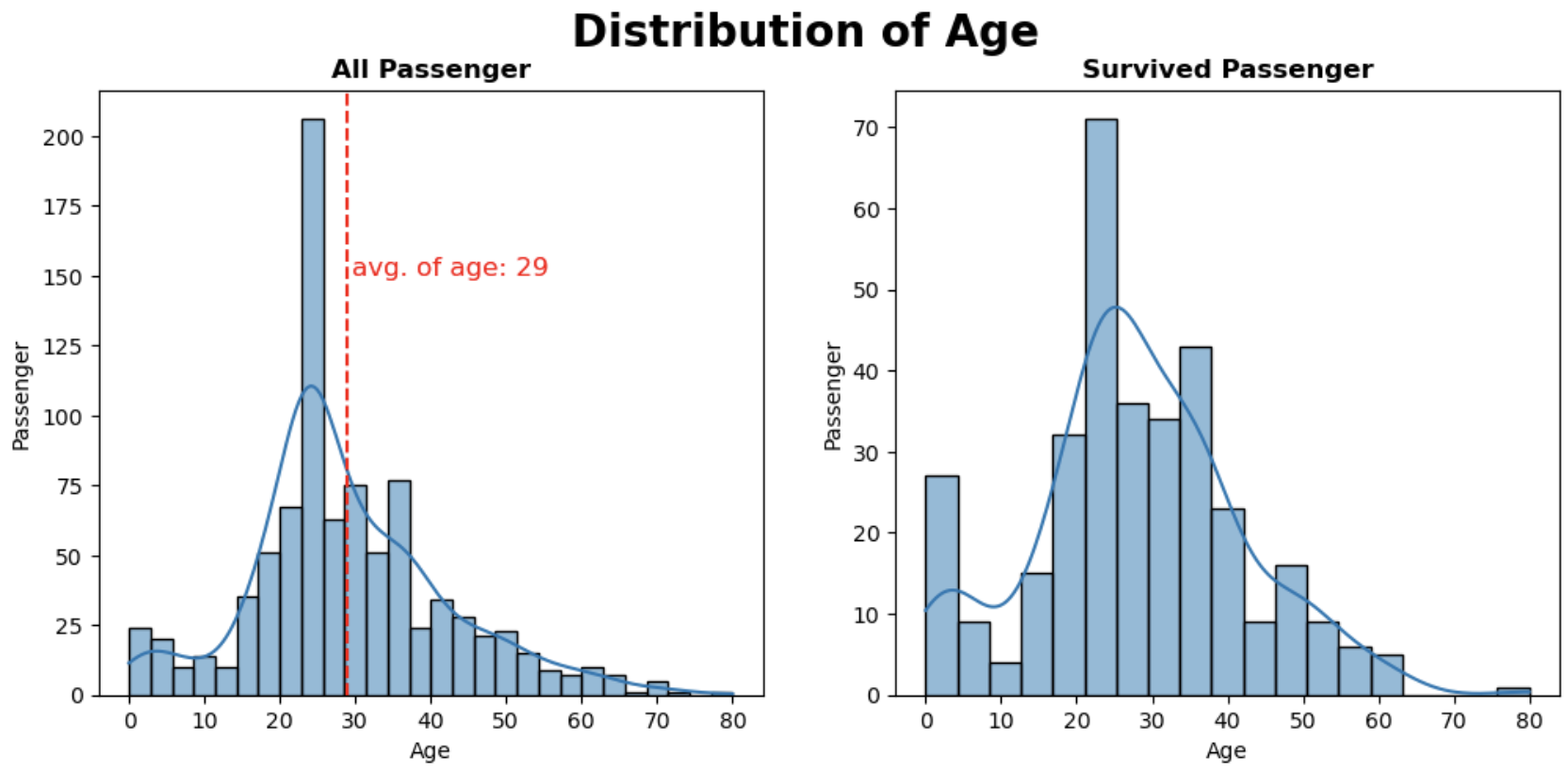
* Dù có 3 hạng vé trên Titanic, giá vé lại rất đa dạng.
* Phần lớn hành khách chi trả từ 5$-35$ cho vé, đặc biệt là từ 8$-10$.
* Vẫn có những hành khách bỏ ra số tiền lớn cho chuyến đi (>50$).
* Số tiền trung bình của những hành khách sống sót trên con tàu là 48.21$.
* Chỉ có 4 người không thuộc Pclass 1 trả số tiền cao hơn mức trung bình này
* Đa số vé từ 0$ ~10$ danh cho hạng Pclass 3.

Ta có kết luận:

* Mặc dù giá vé không hoàn toàn nhất quán giữa các hạng ghế, giá trung bình vẫn tuân theo hạng ghế, phản ánh mức độ chi tiêu sẵn lòng của hành khách và có thể ảnh hưởng đến tỷ lệ sống sót của họ
* Điều này cho thấy phân loại vé Titanic không chỉ dựa trên tầng lớp xã hội, mà còn liên quan đến sự không nhất quán trong giá cả và chi tiêu sẵn lòng của hành khách.

1. **Độ tuổi trung bình của các hành khách trong sự kiện Titacnic:**

Quan sát biểu đồ sau:



*Hình 3: Biểu đồ All Passenger và Survived Passenger.*

Dựa trên biểu đồ “All Passenger” và “Survived Passenger”, có một vài điểm đáng lưu ý sau đây:

* Có hầu hết tất cả các lứa tuổi đi tàu Titanic.
* Độ tuổi phổ biến của các hành khách là khoảng 18 - 40 tuổi, trong đó đông nhất là độ tuổi 24 - 26 tuổi.
* Độ tuổi trung bình của các hành khách là 29 tuổi.
* Đối với những hành khách còn sống, phân bố về độ tuổi gần như tương tự như phân bố của toàn bộ khách trên tàu, chỉ có một chút khác biệt đáng lưu ý ở độ tuổi từ 0 - 5 tuổi.

Có thể thấy, tuổi tác không ảnh hưởng đến khả năng sống sót của các hành khách trên tàu Titanic trrừ các em nhỏ dưới 5 tuổi hầu hết được cứu sống.

1. **Sự ảnh hưởng của việc đi cùng người thân lên tàu Titanic:**

Tiếp theo ta cùng xem sự ảnh hưởng của việc đi cùng người thân lên tàu Titanic:

A picture containing text, plot, screenshot, line

Description automatically generated

*Hình 4: Biểu đồ Sibling or Spouse và Parent or Children.*

Nhận xét:

* So sánh 2 biểu đồ trên, ta thấy nét tương đồng giữa việc đi với anh chị em, vợ chồng so với việc đi với cha mẹ hoặc con cái.
* Đối với những người đi với người thân (từ 1 người trở lên), tỉ lệ sống và chết gần như là như nhau.
* Đối với những người đi một mình, số người sống sót chỉ bằng một nữa số người chết.
* Không có bất kì người nào đi với 4 anh chị em trở lên còn sống, trong khi đó nếu đi với số lượng lớn cha mẹ hoặc con cái tỉ lệ sống sót hoàn toàn cao hơn.

Kết luận:

* Việc đi cùng người thân có thể ảnh hưởng đến tỷ lệ sống sót trên Titanic , khi những người đi cùng người thân có cơ hội cao hơn để lên tàu cứu hộ. Ngược lại, những người đi một mình thường phải tự cứu mình trong tình huống hạn chế về tàu cứu hộ.

1. **Số lượng người sống sót sau sự kiện Titanic:**

**A picture containing text, screenshot, diagram, plot

Description automatically generated**

*Hình 5: Biểu đồ tể hiện sự sống sót trong: Pclass, Sex, Age\_range, Embarked.*

Dựa trên 4 biểu đồ trên ta đưa ra các nhận xét sau:

* Theo hạng ghế:
* Chỉ có hạng thương gia là có số người sống sót nhiều hơn số người thiệt mạng đồng thời là hạng ghế có số lượng người sống sót lớn nhất.
* Số lượng người thiệt mạng của hạng phổ thông nhiều gấp 3 lần số người sống sót ở hạng này và cũng là hạng ghế có số lượng người thiệt mạng lớn nhất.
* Theo giới tính:
* Số người sống sót là nữ nhiều gấp gần 3 lần số người thiệt mạng và nhiều gấp 2 lần số người sống sót là nam giới.
* Ngược lại, số lượng người tử vong ở nam nhiều gấp 4 lần số người sống sót và gấp gần 6 lần số người tử vong là nữ.
* Theo độ tuổi:
* Chỉ có độ tuổi dưới 18 tuổi là có số lượng người sống sót cao hơn số lượng người tử vong.
* Như đã phân tích ở trên, hầu hết trẻ em ở độ tuổi 0 - 5 tuổi đều sẽ được sống sót.
* Theo cảng khởi hành:
* Cảng xuất phát không trực tiếp ảnh hưởng đến tỷ lệ sống sót của hành khách. Tuy nhiên, những hành khách khởi hành từ cảng C cho thấy tỷ lệ sống sót cao hơn, có thể do đa số họ đi hạng Thương gia, là phụ nữ, hoặc có nhiều trẻ nhỏ trong gia đình.

## Ứng dụng Machine Learning để dự đoán tỉ lệ sống sót sau sự kiện Titanic:

**2.4.1 K-Means**

**A picture containing text, screenshot, number, font

Description automatically generated**

*Hình 6: Confustion matrix K-Means*

Báo cáo phân loại:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Casual | 0.31 | 0.03 | 0.06 | 442 |
| Member | 0.36 | 0.88 | 0.75 | 269 |
| Accuracy |  |  | 0.35 | 771 |
| Marco avg | 0.33 | 0.46 | 0.28 | 771 |
| Weighted avg | 0.33 | 0.35 | 0.23 | 771 |

**2.4.2 Logistic Regression**

A picture containing text, screenshot, rectangle, diagram

Description automatically generated

*Hình 7: Confustion matrix before, after drop PARCH*

Báo cáo phân loại:

* Mặc định

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Casual | 0.81 | 0.92 | 0.86 | 107 |
| Member | 0.84 | 0.68 | 0.75 | 71 |
| Accuracy |  |  | 0.82 | 178 |
| Marco avg | 0.83 | 0.80 | 0.80 | 178 |
| Weighted avg | 0.82 | 0.82 | 0.82 | 178 |

* Loại bỏ cột Parch, SipSp và sử dụng Smote Over Sampling

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Casual | 0.84 | 0.83 | 0.83 | 110 |
| Member | 0.83 | 0.85 | 0.84 | 110 |
| Accuracy |  |  | 0.84 | 220 |
| Marco avg | 0.84 | 0.84 | 0.84 | 220 |
| Weighted avg | 0.84 | 0.84 | 0.84 | 220 |

# KẾT LUẬN

So sánh 2 mô hình K-Means và Logistic Regression thấy kết quả phân loại của mô hình Logistic Regression có độ chính xác tốt hơn và phù hợp cho việc dự đoán.

Sau khi áp dụng các thuật toán học máy để phân tích tỷ lệ sống sót trên Titanic, kết quả cho thấy:

* Những hành khách có khả năng sống sót cao nhất là những người thuộc nhóm sau: phụ nữ, trẻ em và người đi hạng thương gia.
* Trên thực tế, việc hành khách còn sống nhưng mô hình dự đoán là thiệt mạng mang lại hậu quả xấu hơn vì vậy cần giảm lỗi loại 1 nhiều nhất có thể.
* Sau quá trình đánh giá dữ liệu và áp dụng các kĩ thuật giúp làm giảm độ thiên lệch dữ liệu, khả năng phân loại của mô hình đã được cải thiện khá tốt và có thể dùng để dự đoán tỉ lệ sống của các hành khách chưa rõ khả năng sống/chết dựa trên các thông số tương đương.

**TÀI LIỆU THAM KHẢO**

Kaggle, Titanic: Machine Learning from Disaster. Link: [Titanic - Machine Learning from Disaster | Kaggle](https://www.kaggle.com/competitions/titanic/data?select=train.csv) (Ngày truy cập: 20/04/2023)

Keras API Documentation. Link: [Keras API reference](https://keras.io/api/)(Ngày truy cập: 30/04/2023)

pandas API Documentation. Link: [pandas documentation — pandas 2.0.1 documentation (pydata.org)](https://pandas.pydata.org/docs/) (Ngày truy cập: 30/04/2023)

Scikit-learn API Documentation. Link: [API Reference — scikit-learn 1.2.2 documentation](https://scikit-learn.org/stable/modules/classes.html) (Ngày truy cập: 01/05/2023)

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **Nhiệm vụ** |
| 1 | Châu Đức Hiệp | Nghiên cứu thư viện |
| 2 | Châu Đức Hiệp | Nghiên cứu và thu thập dữ liệu |
| 3 | Châu Đức Hiệp | Tiền xử lý dữ liệu |
| 4 | Châu Đức Hiệp | Viết báo cáo |
| 5 | Nguyễn Huỳnh Bảo Quốc | Phân tích dữ liệu |
| 6 | Nguyễn Huỳnh Bảo Quốc | Ứng dụng Machine Learning để dự đoán tỉ lệ sống/chết |