A logo with a red and blue letter and a drop of water

Description automatically generated

**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BÁO CÁO BÀI TẬP NHÓM MÔN LẬP TRÌNH PHÂN TÍCH DỮ LIỆU 2**

**ĐỀ TÀI :** DỰ ĐOÁN NGUY CƠ MẮC BỆNH TIM DỰA TRÊN CÁC CHỈ SỐ CƠ THỂ VÀ THÓI QUEN SINH HOẠT

**NHÓM :** Dữ Liệu Tim

|  |  |  |
| --- | --- | --- |
| Họ Tên | MSSV | Ghi Chú |
| Bùi Trí Thức | 20088361 | Nhóm trưởng |
| Nguyễn Thái Bảo | 21031631 |  |
| Trần Văn Lợi | 21030831 |  |
| Nguyễn Bảo Kha | 21041651 |  |

# Phân tích dataset và lựa chọn mô hình

## Phân tích tập dữ liệu

* **dataset : heart\_2020\_cleaned.csv – dữ liệu có hình dạng (319795, 18)**

|  |  |  |  |
| --- | --- | --- | --- |
| Tên cột | Mô tả | Kiểu | Miền Giá Trị |
| HeartDisease | Người trả lời có bị bệnh tim hay không. | object | Yes/No |
| BMI | Chỉ số khối cơ thể của người trả lời. | float64 | [12.02, 94.85] |
| Smoking | Người trả lời có hút thuốc hay không. | object | Yes/No |
| AlcoholDrinking | Người trả lời có phải là người nghiện rượu nặng hay không. | object | Yes/No |
| Stroke | Người trả lời đã từng bị đột quỵ hay chưa. | object | Yes/No |
| PhysicalHealth | Số ngày trong 30 ngày qua mà sức khỏe thể chất của người trả lời không tốt. | float64 | [0,30] |
| MentalHealth | Số ngày trong 30 ngày qua mà sức khỏe tinh thần của người trả lời không tốt. | float64 | [0,30] |
| DiffWalking | Người trả lời có gặp khó khăn khi đi bộ hoặc leo cầu thang hay không. | object | Yes/No |
| Sex | Giới tính của người trả lời. | object | Male/Female |
| AgeCategory | Nhóm tuổi của người trả lời. | object | 18-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64  65-69  70-74  75-79  80 or older |
| Race | Chủng tộc của người trả lời. | object | -American Indian/Alaskan Native  -Asian  -Black  -Hispanic  -Other  -White |
| Diabetic | Người trả lời có bị tiểu đường hay không. | object | Yes No No, borderline diabete Yes (during regnancy) |
| PhysicalActivity | Người trả lời có tập thể dục trong 30 ngày qua hay không. | object | Yes/No |
| GenHealth | Người trả lời tự đánh giá sức khỏe tổng quát của họ như thế nào. | object | Excellent  Fair  Good  Poor  Very good |
| SleepTime | Người trả lời ngủ bao nhiêu giờ mỗi ngày. | float64 | [0,24] |
| Asthma | Người trả lời có bị hen suyễn hay không. | object | Yes/No |
| KidneyDisease | Người trả lời có bị bệnh thận hay không. | object | Yes/No |
| SkinCancer | Người trả lời có bị ung thư da hay không. | object | Yes/No |

1. **Phân tích cột target ( )**

Nhận xét : - Số lượng người không mắc bệnh tim chiếm phần lớn: Phần lớn các mẫu dữ liệu đều thuộc nhóm người không mắc bệnh tim, biểu thị bằng cột cao hơn nhiều ở vị trí "No".

Số lượng người mắc bệnh tim chiếm tỷ lệ nhỏ: Số lượng người mắc bệnh tim chỉ chiếm một phần rất nhỏ so với tổng số mẫu dữ liệu, thể hiện bằng cột thấp hơn ở vị trí "Yes".

**A graph showing distribution of heart disease

Description automatically generated**

1. Phân tích các cột khác

A graph of a number of different colored bars

Description automatically generated

Nhận xét :

***Tập trung ở các mức BMI cao***: Phần lớn bệnh nhân mắc bệnh tim rơi vào các nhóm cân nặng thừa và béo phì (Overweight, Obesity I, Obesity II). Điều này cho thấy mối tương quan dương giữa BMI cao và nguy cơ mắc bệnh tim.

***Ít bệnh nhân gầy:*** Số lượng bệnh nhân thuộc nhóm gầy (Underweight) là rất thấp, cho thấy việc thiếu cân không phải là yếu tố nguy cơ chính cho bệnh tim trong nhóm đối tượng này.

***Mức độ béo phì I là phổ biến nhất:*** Nhóm bệnh nhân có chỉ số BMI thuộc mức béo phì I (Obesity I) chiếm tỷ lệ cao nhất, cho thấy béo phì ở mức độ vừa phải là một yếu tố nguy cơ quan trọng.

A graph of different colored bars

Description automatically generated

**Nhận xét :**

**Độ tuổi cao, nguy cơ cao**: Phần lớn bệnh nhân mắc bệnh tim tập trung ở các nhóm tuổi lớn, đặc biệt là nhóm 80 tuổi trở lên và các nhóm tuổi từ 55 trở lên. Điều này cho thấy nguy cơ mắc bệnh tim tăng theo tuổi tác.

**Giảm dần theo độ tuổi**: Số lượng bệnh nhân giảm dần khi độ tuổi giảm. Nhóm tuổi trẻ (dưới 45 tuổi) có số lượng bệnh nhân rất ít, cho thấy bệnh tim ít gặp ở người trẻ tuổi.

**Đỉnh cao ở độ tuổi trung niên**: Nhóm tuổi từ 55-79 có số lượng bệnh nhân mắc bệnh tim khá cao, cho thấy đây là giai đoạn mà nguy cơ mắc bệnh tim tăng đáng kể.

A graph of distribution of heart disease

Description automatically generated

**\* Nhận xét:**

***Nam giới có tỷ lệ mắc bệnh tim cao hơn nữ giới:*** Số lượng nam giới mắc bệnh tim cao hơn đáng kể so với nữ giới. Điều này cho thấy nam giới có nguy cơ mắc bệnh tim cao hơn.

***Phụ nữ có tỷ lệ mắc bệnh tim thấp hơn:*** Số lượng nữ giới mắc bệnh tim thấp hơn so với nam giới, tuy nhiên vẫn có một lượng đáng kể phụ nữ mắc bệnh tim.

A graph of a number of people with different colored squares

Description automatically generated with medium confidence

**\* Nhận xét :**

***Tình trạng sức khỏe kém liên quan đến nguy cơ bệnh tim cao:*** Người có đánh giá tình trạng sức khỏe ở mức "Kém" và "Khá" có tỷ lệ mắc bệnh tim cao hơn so với những người đánh giá tình trạng sức khỏe ở mức "Rất tốt" và "Tuyệt vời". Điều này cho thấy tình trạng sức khỏe tổng quan là một yếu tố dự báo quan trọng cho nguy cơ mắc bệnh tim.

***Tình trạng sức khỏe tốt liên quan đến nguy cơ bệnh tim thấp:*** Ngược lại, những người đánh giá tình trạng sức khỏe của mình ở mức "Rất tốt" và "Tuyệt vời" có tỷ lệ mắc bệnh tim thấp nhất. Điều này cho thấy việc duy trì một lối sống lành mạnh và có tình trạng sức khỏe tốt có thể giúp giảm nguy cơ mắc bệnh tim.

A graph showing the number of people in the heart disease

Description automatically generated

\* **Nhận xét :**

***Người da trắng có tỷ lệ mắc bệnh tim cao nhất:*** Nhóm người da trắng có số lượng mắc bệnh tim cao nhất so với các nhóm chủng tộc khác. Điều này cho thấy người da trắng có nguy cơ mắc bệnh tim cao hơn.

***Người da đen và Hispanic có tỷ lệ mắc bệnh tim cao hơn so với trung bình:*** Nhóm người da đen và Hispanic cũng có tỷ lệ mắc bệnh tim cao hơn so với các nhóm thiểu số khác.

***Người châu Á và người Mỹ bản địa có tỷ lệ mắc bệnh tim thấp nhất:*** Nhóm người châu Á và người Mỹ bản địa có tỷ lệ mắc bệnh tim thấp nhất trong số các nhóm chủng tộc được khảo sát.

A group of blue and orange bars

Description automatically generated

**\* Nhận xét :**

***Đột quỵ:*** Người đã từng bị đột quỵ có nguy cơ mắc bệnh tim cao hơn đáng kể so với người chưa từng bị đột quỵ. Điều này cho thấy đột quỵ không chỉ là một hậu quả của bệnh tim mà còn là một yếu tố dự báo nguy cơ mắc bệnh tim trong tương lai.

***Tiểu đường:*** Tương tự, người bị tiểu đường cũng có nguy cơ mắc bệnh tim cao hơn đáng kể so với người không bị tiểu đường. Điều này cho thấy tiểu đường là một yếu tố nguy cơ độc lập và mạnh mẽ cho bệnh tim.

***Bệnh hen suyễn:*** Người mắc bệnh hen suyễn có nguy cơ mắc bệnh tim cao hơn đáng kể so với người không mắc bệnh hen suyễn. Điều này cho thấy bệnh hen suyễn không chỉ ảnh hưởng đến đường hô hấp mà còn có thể liên quan đến các vấn đề về tim mạch.

***Bệnh thận:*** Bệnh nhân mắc bệnh thận mãn tính cũng có nguy cơ mắc bệnh tim cao hơn. Điều này cho thấy mối liên hệ chặt chẽ giữa sức khỏe tim mạch và sức khỏe thận.

***Ung thư da và bệnh tim:*** Biểu đồ cho thấy không có sự khác biệt đáng kể về tỷ lệ mắc bệnh tim giữa những người có và không có tiền sử ung thư da. Điều này có thể cho thấy, ít nhất dựa trên dữ liệu này, ung thư da không phải là một yếu tố nguy cơ trực tiếp và mạnh mẽ cho bệnh tim. Tuy nhiên, cần lưu ý rằng đây chỉ là một quan sát ban đầu và cần có thêm nghiên cứu để xác định mối liên hệ cụ thể giữa hai bệnh này.

***Uống rượu và bệnh tim:*** Biểu đồ cho thấy một xu hướng tương tự giữa những người uống rượu và bệnh tim. Không có sự khác biệt rõ rệt về tỷ lệ mắc bệnh tim giữa hai nhóm này. Điều này có thể là do nhiều yếu tố khác nhau, bao gồm loại rượu, lượng rượu tiêu thụ, và các yếu tố nguy cơ khác đi kèm.

## Lựa chọn và Traning Model

### 2.1 Logistic Regression

#### Logistic Regression

* Với dữ liệu được phân tích như trên, đòng thời dữ liệu đầu ra thuộc về dạng Binary Class.
* Kết luận : Mô hình có thể phù hợp là ***Logistic Regression***

#### Tiến hành huấn luyện model

* Chuyển hết các feature không phải số về dạng Binary, sau đó tiến hành tạo tập train và test theo tỉ lệ 80/20, huấn luyện mô hình

A screen shot of a computer program

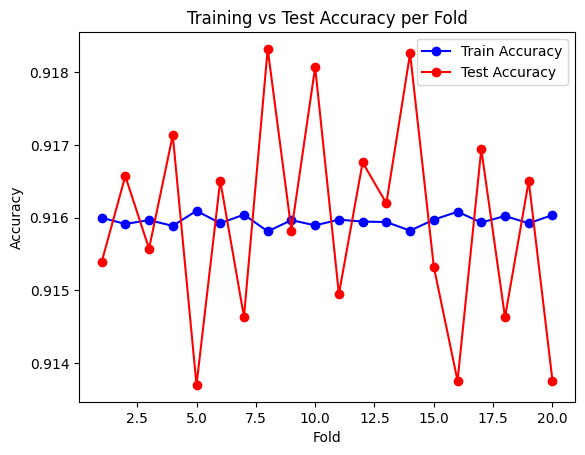
Description automatically generated

* Tính toán thông số test của mô hình sau khi huấn luận :

A computer screen shot of text

Description automatically generated

#### Visual Model và Testing với UI



- **Hiện tượng Overfitting:** Mặc dù không có sự khác biệt quá lớn giữa độ chính xác trên tập huấn luyện và kiểm tra, nhưng sự dao động lớn trên tập kiểm tra có thể ám chỉ mô hình đang có dấu hiệu overfitting nhẹ. Mô hình có thể hoạt động rất tốt trên tập huấn luyện nhưng không tổng quát hóa được tốt trên tập kiểm tra. Khi đó, mô hình cố gắng "học thuộc" dữ liệu huấn luyện, dẫn đến không thể xử lý các biến động trên dữ liệu mới.

Cần cải thiện tính ổn định: Sự dao động mạnh của đường Test Accuracy cho thấy cần có các biện pháp cải thiện như:

- Tăng số lượng dữ liệu: Đảm bảo dữ liệu phong phú hơn để mô hình có khả năng học tốt hơn và tổng quát hóa cao hơn.

- Regularization: Bạn có thể thử các biện pháp giảm overfitting như thêm regularization (L1, L2).

- Điều chỉnh hyperparameter: Thử điều chỉnh các tham số của mô hình để tìm kiếm sự cân bằng tốt hơn giữa huấn luyện và kiểm tra.

A screenshot of a black screen

Description automatically generated

A screenshot of a graph

Description automatically generated

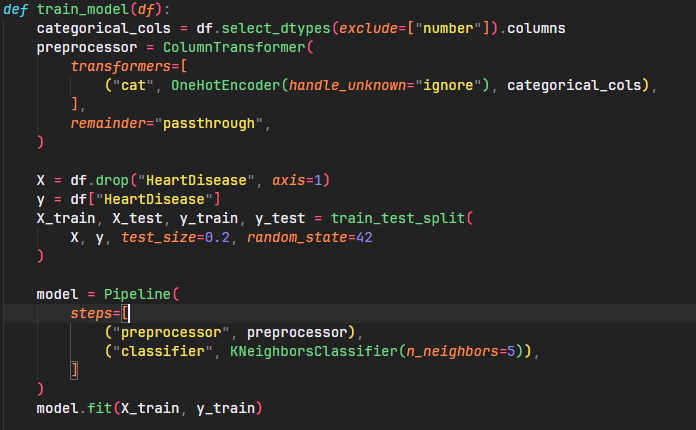
### KNN

#### **KNN Model**

* Mô hình trước đó có hiệu suất thấp, tiếp theo lựa chọn một model khác để so sánh hiệu suất.
* KNN là một trong những mô hình có khả năng đưa ra các sư đoán phân lớp ổn định

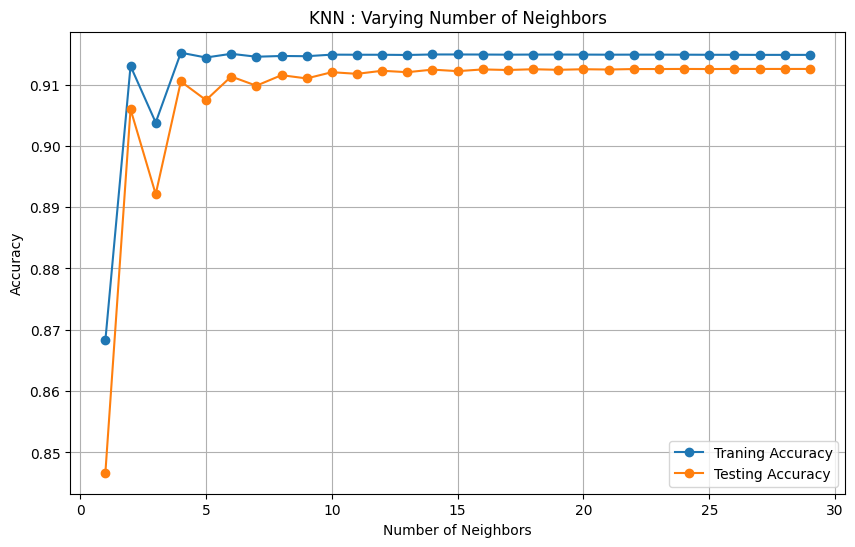
#### Huấn luyện mô hình

Mô hình đươc huấn luyện với tập dữ liệu tương đương và k = 5 ( default )



#### Visual model và kiểm thử, đánh giá

* Over/Under – fitting :



**Nhận xét**:

- Độ chính xác trên tập huấn luyện:

    - Tăng dần: Khi số lượng neighbors tăng, độ chính xác trên tập huấn luyện cũng tăng dần và đạt mức ổn định ở khoảng k = 10. Điều này cho thấy mô hình ngày càng "học thuộc lòng" dữ liệu huấn luyện khi k tăng.

- Độ chính xác trên tập kiểm thử:

    - Tăng ban đầu, sau đó giảm: Độ chính xác trên tập kiểm thử tăng ban đầu khi k tăng từ 1 đến khoảng 5, đạt đỉnh ở khoảng này. Tuy nhiên, khi k tiếp tục tăng, độ chính xác trên tập kiểm thử lại giảm dần.

- Overfitting:

    - Khi k quá nhỏ (ví dụ k = 1), mô hình dễ bị overfitting, tức là mô hình học quá kỹ dữ liệu huấn luyện đến mức không thể generalize tốt trên dữ liệu mới (tập kiểm thử). Điều này dẫn đến độ chính xác trên tập huấn luyện cao nhưng độ chính xác trên tập kiểm thử lại thấp.

    - Ngược lại, khi k quá lớn, mô hình trở nên quá đơn giản, không thể bắt được các pattern phức tạp trong dữ liệu. Điều này dẫn đến cả độ chính xác trên tập huấn luyện và tập kiểm thử đều thấp.

**Kết luận**:

- Giá trị k tối ưu: Dựa trên biểu đồ, giá trị k tối ưu bắt đầu từ khoảng 5-7,Tuy nhiên, đến khoảng k = 30, mô hình hoàn toàn không còn thay đổi, tại đây, nó cho ra kết quả chính xác trên tập kiểm thử đạt mức cao nhất và mô hình có khả năng tổng quát hóa tốt nhất.

- Hiện tượng overfitting và underfitting: Biểu đồ minh họa rõ ràng hiện tượng overfitting và underfitting khi thay đổi giá trị k.

- Cân bằng giữa độ phức tạp của mô hình và khả năng tổng quát hóa: Việc chọn giá trị k phù hợp là một trade-off giữa độ phức tạp của mô hình và khả năng tổng quát hóa. Một mô hình quá phức tạp dễ bị overfitting, trong khi một mô hình quá đơn giản lại không thể bắt được các pattern phức tạp trong dữ liệu.

**Nhận Xét tổng quát mô hình :**

**A blue bar graph with white text

Description automatically generated**

**-** Thông qua các chỉ số độ đo cho thấy :

**\*\*Nhận xét\*\***:

- Độ chính xác (Accuracy) đạt 91.28%: Điều này cho thấy mô hình đã dự đoán đúng một tỷ lệ lớn các mẫu dữ liệu. Tuy nhiên, đây không phải là chỉ số đánh giá hiệu suất duy nhất và có thể gây hiểu nhầm trong một số trường hợp, đặc biệt khi dữ liệu không cân bằng.

- Độ chính xác (Precision) thấp (62.3%) và độ nhạy (Recall) rất thấp (0.68%): Điều này cho thấy mô hình có thể đang gặp vấn đề trong việc phân loại đúng các mẫu dương. Cụ thể:

- Precision thấp: Mô hình thường dự đoán sai các mẫu dương thành mẫu âm.

- Recall thấp: Mô hình bỏ sót nhiều mẫu dương thực tế.

- F1-score rất thấp (1.34%): F1-score là trung bình hài hòa của Precision và Recall, giá trị thấp cho thấy cả Precision và Recall đều thấp, mô hình không đạt hiệu suất tốt trong việc phân loại.

#### **Tổng quan đánh giá mô hình:**

* Mặc dù độ chính xác tổng thể cao nhưng các chỉ số Precision, Recall và F1-score lại rất thấp cho thấy mô hình hiện tại chưa đạt được hiệu suất mong muốn. Cần tiến hành phân tích sâu hơn để tìm ra nguyên nhân chính xác và đưa ra các giải pháp cải thiện.

### 2.3 : Random Forest

A graph of blue rectangular bars

Description automatically generated

**Nhận xét mô hình :**

1. Accuracy : (86%) Mô hình có độ chính xác ở mức tương đối, điều này cho thấy nó có khả năng dự đoán đúng một cách tương đối. Tuy nhiên cần xem xét thêm các độ đo khác để kết luận

2. Precision : (25%) : cho thấy trong số những trường hợp được dự đoán là mắc bệnh tim, chỉ có 25% là thực sự mắc bệnh. Điều này có nghĩa là mô hình có xu hướng đưa ra nhiều dự đoán dương tính giả (false positive). Trong ngữ cảnh dự đoán bệnh tim, điều này có thể dẫn đến việc nhiều người khỏe mạnh bị chẩn đoán nhầm là mắc bệnh, gây ra lo lắng và tốn kém cho các xét nghiệm y tế không cần thiết.

3. Recall (44%) : Đây là một con số tương đối tốt, nó cho thấy mô hình thực sự có hiệu quả trong việc phát hiện sớm các trường hợp dương tính chính xác

4. F1 Score (32%) : Các giá trị này ở mức trung bình cho thấy mô hình chưa thực sự tốt trong việc xác định những người mắc bệnh tim. Tuy nhiên nó cũng sẽ là một tín hiệu tốt cho thấy mô hình cũng có độ phản ứng tương đối đối với các dữ liệu được dụ đoán

# Kết luận lựa chọn mô hình :

* Qua 3 mô hình, khi so sánh :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F1** |
| **Logistic Regression** | 91% | 54% | 10% | 17% |
| **KNN** | 91% | 62.30% | 0.68% | 1.34% |
| **Random Forest** | 86% | 25% | 44% | 32% |

* **Kết luận : mô hình Random Forest mang lại hiệu quả cao nhất trong 3 mô hình**