**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**ĐỒ ÁN CUỐI KÌ**Môn học : Phân tích dữ liệu kinh doanh

Học kỳ II (2021-2022)

**ĐỀ TÀI: DỰ ĐOÁN KHẢ NĂNG ĐĂNG KÝ TIỀN GỬI TÍN DỤNG CỦA KHÁCH HÀNG**

Sinh viên thực hiện:

1. Trần Nhật Tân (lead) MSSV: 19522177

2. Huỳnh Công Mạnh MSSV: 19521825

3. Ngô Hồng Hải MSSV: 19521463

GVHD: Thầy Nguyễn Đình Thuân  
 Thầy Nguyễn Minh Nhựt

Lớp: IS403.M21.HTCL

**Thành phố Hồ Chí Minh, tháng 6 năm 2022**

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**ĐỒ ÁN CUỐI KÌ**Môn học : Phân tích dữ liệu kinh doanh

Học kỳ II (2021-2022)

**ĐỀ TÀI: DỰ ĐOÁN KHẢ NĂNG ĐĂNG KÝ TIỀN GỬI TÍN DỤNG CỦA KHÁCH HÀNG**

Sinh viên thực hiện:

1. Trần Nhật Tân (lead) MSSV: 19522177

2. Huỳnh Công Mạnh MSSV: 19521825

3. Ngô Hồng Hải MSSV: 19521463

GVHD: Thầy Nguyễn Đình Thuân

Thầy Nguyễn Minh Nhựt

Lớp: IS403.M21.HTCL

**Thành phố Hồ Chí Minh, tháng 6 năm 2022**

**MỤC LỤC**

[**LỜI CẢM ƠN** 4](#_Toc105795197)

[**NHẬN XÉT CỦA GIẢNG VIÊN** 5](#_Toc105795198)

[**CHƯƠNG I – TỔNG QUAN ĐỀ TÀI** 6](#_Toc105795199)

[**1.1** **Lý do chọn đề tài** 6](#_Toc105795200)

[**1.2** **Nội dung đề tài** 6](#_Toc105795201)

[**1.3** **Mô tả dữ liệu gốc** 7](#_Toc105795202)

[**CHƯƠNG II – TRIỂN KHAI THỰC NGHIỆM** 9](#_Toc105795203)

[**2.1** **Công cụ sử dụng** 9](#_Toc105795204)

[**2.2** **Thuật toán Logistic Regression** 9](#_Toc105795205)

[**Định nghĩa** 9](#_Toc105795206)

[**2.3** **Thuật toán cây quyết định** 10](#_Toc105795207)

[**Định nghĩa** 10](#_Toc105795208)

[**2.4** **Thuật toán K-Nearest Neighbors** 14](#_Toc105795209)

[**Định nghĩa** 14](#_Toc105795210)

[**2.5** **Thuật toán Random Forest** 15](#_Toc105795211)

[**Định nghĩa** 15](#_Toc105795212)

[**2.6** **Các bước thực hiện** 16](#_Toc105795213)

[**2.6.2** **Data Scaling** 21](#_Toc105795214)

[**CHƯƠNG IV - PHẦN MỀM DỰ ĐOÁN** 24](#_Toc105795215)

[**4.1 Tổng quan phần mềm** 24](#_Toc105795216)

[**4.1.1 Thuật toán được sử dụng** 24](#_Toc105795217)

[**4.1.2 Các thuộc tính được sử dụng để đưa ra dự đoán** 24](#_Toc105795218)

[**4.1.3 Giao diện và kiểm thử** 24](#_Toc105795219)

[**4.1.3.1 Giao diện** 24](#_Toc105795220)

[**4.1.3.2 Kiểm thử** 24](#_Toc105795221)

[**4.2 Code phần mềm** 25](#_Toc105795222)

[**4.2.1 Code phần giao diện:** 25](#_Toc105795223)

[**CHƯƠNG V - KẾT LUẬN** 25](#_Toc105795224)

[**5.1 Ưu điểm, hạn chế của từng thuật toán** 25](#_Toc105795225)

[***5.1.1 Cây quyết định*** 25](#_Toc105795226)

[***5.1.2. Random Forest*** 27](#_Toc105795227)

[***5.1.3 Logistic Regression*** 27](#_Toc105795228)

[***5.1.4 K-nearest Neighbors*** 27](#_Toc105795229)

[**5.2 Hướng phát triển** 28](#_Toc105795230)

[**Tài liệu tham khảo** 29](#_Toc105795231)

[**Hết.** 29](#_Toc105795232)

# **LỜI CẢM ƠN**

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Với lòng biết ơn sâu sắc nhất, đầu tiên nhóm chúng em xin gởi lời cảm ơn chân thành đến tập thể quý Thầy Cô Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM và quý Thầy Cô khoa Hệ thống thông tin đã giúp cho nhóm có những kiến thức cơ bản làm nền tảng để thực hiện đề tài này.

Đặc biệt nhóm chúng em xin gửi lời cảm ơn chân thành tới thầy Nguyễn Đình Thuân và thầy Nguyễn Minh Nhựt – giảng viên lí thuyết và thực hành môn Phân tích dữ liệu kinh doanh đã tận tình giúp đỡ, trực tiếp chỉ bảo, hướng dẫn nhóm trong suốt quá trình làm đồ án môn học. Nhờ đó, chúng em đã tiếp thu được nhiều kiến thức bổ ích trong việc vận dụng cũng như kỹ năng làm đồ án. Nếu không có những lời hướng dẫn, dạy bảo của thầy thì nhóm chúng em nghĩ đồ án này của nhóm rất khó có thể hoàn thiện được. Một lần nữa, em xin chân thành cảm ơn thầy. Ngoài ra, để đồ án được hoàn thành thì không thể nào cám ơn những người đã làm ra đó, cám ơn các bạn các thành viên trong nhóm đã chăm chỉ và chịu khó hoàn thành nhiệm vụ đúng tiến độ.

Cuối cùng, cảm ơn tất cả các thành viên trong nhóm đã làm việc hết công suất để hoàn thành tốt đề tài của mình. Xin chân thành cảm ơn!

# **NHẬN XÉT CỦA GIẢNG VIÊN**

……………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………

# **CHƯƠNG I – TỔNG QUAN ĐỀ TÀI**

* 1. **Lý do chọn đề tài**

* 1. **Nội dung đề tài**

Đây là Bộ dữ liệu phân loại rủi ro Heart Disease được chia sẻ trên <https://www.kaggle.com/> . Tập dữ liệu này có từ năm 1988 và bao gồm bốn cơ sở dữ liệu: Cleveland, Hungary, Thụy Sĩ và Long Beach V. Nó chứa 76 thuộc tính, bao gồm cả thuộc tính dự đoán, nhưng tất cả các thử nghiệm đã xuất bản đều sử dụng một tập con gồm 12 thuộc tính trong số đó. Trường ***"Target"*** đề cập đến sự hiện diện của bệnh tim ở bệnh nhân. Nó là số nguyên có giá trị 0 = không có bệnh và 1 = bệnh.

Người ta cho rằng những thông tin như vậy đóng một vai trò quan trọng trong việc dự đoán và đưa ra quyết định giảm thiểu bệnh tim.

Nguồn dữ liệu: ‎[Heart Disease Dataset | Kaggle](https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset)

Chủ đề: Y tế.

Dataset gồm : 14 cột thuộc tính và 3235 dòng dữ liệu

* 1. **Mô tả dữ liệu gốc**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Ý nghĩa** |
| 1 |  |  |  |
| 2 |  |  |  |
| 3 |  |  |  |
| 4 |  |  |  |
| 5 |  |  |  |
| 6 |  |  |  |
| 7 |  |  |  |
| 8 |  |  |  |
| 9 |  |  |  |
| 10 |  |  |  |
| 11 |  |  |  |
| 12 |  |  |  |

# **CHƯƠNG II – TRIỂN KHAI THỰC NGHIỆM**

* 1. **Công cụ sử dụng**

Trong quá trình thực hiện, nhóm đã sử dụng một số phần mềm phục vụ cho việc tìm  hiểu và xây dựng đề tài:

* Phần thu thập và phân tích thông tin sử dụng thư viện  và ngôn ngữ lập trình python: python3.7, Jupyter notebook,…
* Dữ liệu:

Tất cả các phần mềm trên được nhóm cài đặt và sử dụng trên Hệ điều hành Microsoft Windows 10. Việc tương thích các phần mềm trên với các hệ điều hành khác không nằm trong phạm vi nghiên cứu của đề tài này.

* 1. **Thuật toán Logistic Regression**

**Định nghĩa**

Phương pháp hồi quy logistic là một mô hình hồi quy nhằm dự đoán giá trị đầu  ra rời rạc (discrete target variable) y ứng với một véc-tơ đầu vào **x**. Việc này tương đương với chuyện phân loại các đầu vào **x** vào các nhóm y tương ứng.

Phân tích hồi qui logistic là một kỹ thuật thống kê để xem xét mối liên hệ giữa biến độc lập (biến số hoặc biến phân loại) với biến phụ thuộc là biến nhị phân. Trong hồi qui tuyến tính đơn, biến độc lập x và phụ thuộc y là biến số liên tục liên hệ qua phương trình:

y = a + bx + e

Trong đó: a là alpha, b là beta, e là epxilon

Trong hồi qui logistic, biến phụ thuộc y chỉ có 2 trạng thái 1 và 0. Muốn đổi ra biến số liên tục người ta tính xác suất của 2 trạng thái này. Nếu gọi p là xác suất để một biến cố xảy ra, thì 1-p là xác suất để biến cố không xảy ra. Phương trình hồi qui logistic phát biểu:A picture containing text, watch, clock

Description automatically generated

Ta tính được xác xuất:Chart

Description automatically generated

* 1. **Thuật toán cây quyết định**

**Định nghĩa**

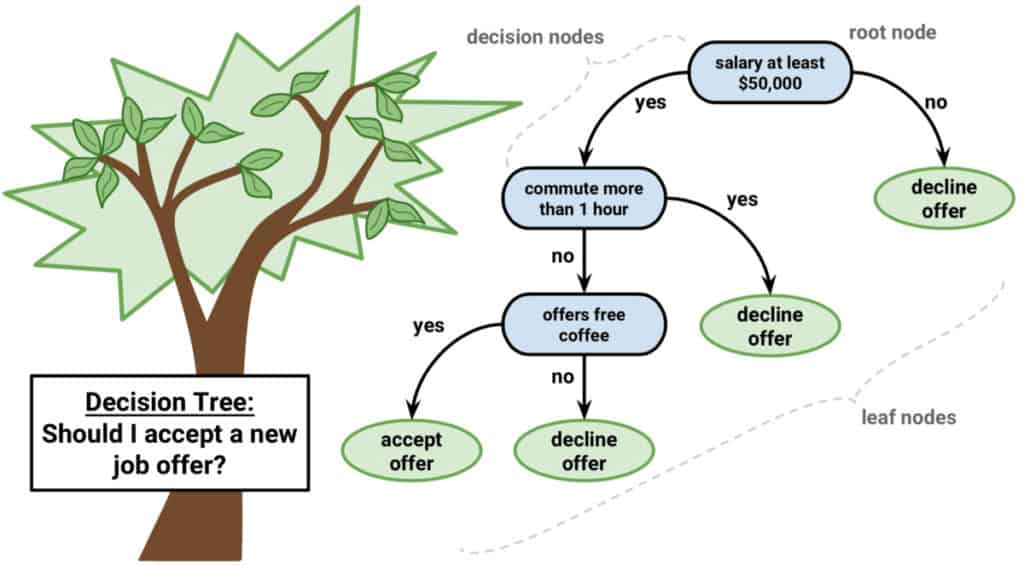
Cây quyết định (Decision Tree) là cấu trúc cây sao cho:

+ Mỗi nút trong ứng với một phép kiểm tra trên một thuộc tính

+ Mỗi nhánh biểu diễn kết quả phép kiểm tra

+ Các nút lá biểu diễn các lớp hay các phân bố lớp

+ Nút cao nhất trong cây là nút gốc. Hình dạng cây quyết định:



Chiến lược cơ bản:

+  Bắt đầu từ nút đơn biểu diễn tất cả các mẫu

+  Nếu các mẫu thuộc về cùng một lớp, nút trở thành nút lá và được gán nhãn bằng lớp đó

+  Ngược lại, dùng độ đo thuộc tính để chọn thuộc tính sẽ phân tách tốt nhất các mẫu vào các lớp

+  Một nhánh được tạo cho từng giá trị của thuộc tính được chọn và các mẫu được phân hoạch theo.

+  Dùng đệ quy cùng một quá trình để tạo cây quyết định

+  Tiến trình kết thúc chỉ khi bất kỳ điều kiện nào sau đây là đúng

* Tất cả các mẫu cho một nút cho trước đều thuộc về cùng một lớp.
* Không còn thuộc tính nào mà mẫu có thể dựa vào để phân hoạch xa hơn.
* Không còn mẫu nào tại nút

**ID3** là một thuật toán được sử dụng trong cây quyết định. Thuật toán này sử dụng độ lợi thông tin (Information Gain) để tiến hành xây dựng cây quyết định**. Thuộc tính Information Gain lớn nhất sẽ được chọn làm nút gốc.**

+ Độ lợi thông tin (Information Gain):

A picture containing text, watch, clock

Description automatically generated

+ Lượng thông tin cần để phân lớp một phần tử trong S dựa trên thuộc tính A: InfoA(S)Icon

Description automatically generated with medium confidence

* Information gain chính là độ sai biệt giữa giá trị thông tin Info(S) ban đầu (trước phân hoạch) và giá trị thông tin mới InfoA(S)(sau phân hoạch với A).



+ Độ bất định (Entropy):

* Entropy đo lường lượng thông tin trong một thuộc tính của một bộ giá trị thu thập được của tập mẫu.
* Entropy được dùng để xác định nút nào được tách kế tiếp trong thuật toán.
* Entropy càng cao thì khả năng cải thiện phân lớp càng cao.
* Công thức:

Text

Description automatically generated

𝑛 là số giá trị khác nhau của thuộc tính A đang xét

𝐴i là số mẫu tương ứng với mỗi giá trị 𝑖 của thuộc tính A

𝑓S(𝐴i) là tỷ lệ của số mẫu có thuộc tính 𝐴𝑖 với S

**CART:** Khác với ID3 dùng công thức Information Cain thì thuật toán Cart sử dụng công thức Gini. **Thuộc tính có giá trị Gini nhỏ nhất sẽ được làm nút gốc**.

+ Chỉ số Gini của tập S

A picture containing text, clock, clipart

Description automatically generated

P(j|S) là tần xuất của j trong S

+ Gini của thuộc tính

A picture containing text

Description automatically generated

Trong đó: ni là số mẫu trong nốt I, n là số mẫu trong nốt A

* 1. **Thuật toán K-Nearest Neighbors**

**Định nghĩa**

* K-nearest Neighbors là một trong những thuật toán học có giám sát đơn giản nhất được sử dụng nhiều trong khai thác dữ liệu. Ý tưởng của thuật toán này là nó không học một điều gì từ tập dữ liệu học (nên KNN được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán nhãn của dữ liệu mới.
* Lớp (nhãn) của một đối tượng dữ liệu mới có thể dự đoán từ các lớp (nhãn) của k hàng xóm gần nó nhất.
* Cho trước tập dữ liệu huấn luyện D với các lớp, phân lớp X vào các lớp dựa vào k phần tử tương tự với X nhất (quy tắc bỏ phiếu: majority vote).
* Các bước thực hiện:

+  Ta có D là tập các điểm dữ liệu đã được gắn nhãn và A là dữ liệu chưa được phân loại.

+  Đo khoảng cách (Euclidian, Manhattan, Minkowski, Minkowski hoặc Trọng số) từ dữ liệu mới A đến tất cả các dữ liệu khác đã được phân loại trong D.

+  Chọn K (K là tham số mà bạn định nghĩa) khoảng cách nhỏ nhất.

+  Kiểm tra danh sách các lớp có khoảng cách ngắn nhất và đếm số lượng của mỗi lớp xuất hiện.

+  Lấy đúng lớp (lớp xuất hiện nhiều lần nhất).

+  Lớp của dữ liệu mới là lớp đã nhận được ở bước 5.

* Phụ thuộc:

+  Độ đo khoảng cách để xác định sự tương tự.

+  Trị k, số phần tử láng giềng => k<=|D|1/2

* Độ đo Euclidean

A picture containing text, watch, gauge

Description automatically generated

* 1. **Thuật toán Random Forest**

**Định nghĩa**

Như chúng ta biết, Random Forest là một tập hợp mô hình (ensemble). Mô hình Random Forest rất hiệu quả cho các bài toán phân loại vì nó huy động cùng lúc hàng trăm mô hình nhỏ hơn bên trong với quy luật khác nhau để đưa ra quyết định cuối cùng. Mỗi mô hình con có thể mạnh yếu khác nhau, nhưng theo nguyên tắc **« wisdom of the crowd »**, ta sẽ có cơ hội phân loại chính xác hơn so với khi sử dụng bất kì một mô hình đơn lẻ nào.

Như tên gọi của nó, Random Forest (RF) dựa trên cơ sở :

1. Random = Tính ngẫu nhiên ;
2. Forest = nhiều cây quyết định (decision tree).

Đơn vị của RF là thuật toán cây quyết định, với số lượng hàng trăm. Mỗi cây quyết định được tạo ra một cách ngẫu nhiên từ việc : Tái chọn mẫu (bootstrap, random sampling) và chỉ dùng một phần nhỏ tập biến ngẫu nhiên (random features) từ toàn bộ các biến trong dữ liệu. Ở trạng thái sau cùng, mô hình RF thường hoạt động rất chính xác, nhưng đổi lại, ta không thể nào hiểu được cơ chế hoạt động bên trong mô hình vì cấu trúc quá phức tạp. RF do đó là một trong số những mô hình hộp đen (black box).

* 1. **Các bước thực hiện**

Code được upload tại :

**Tiền xử lý dữ liệu**

Bước 1: Import các thư viện cần thiết.

Hình 1: import thư viện

Dataset là file /heart\_Disease\_prediction\_new\_final.csv.

Bước 2: Load dữ liệu từ file heart\_Disease\_prediction\_new\_final.csv vào c\_data.

Thống kê sau khi đã thay đổi dữ liệu:

|  |  |
| --- | --- |
| Thuộc tính | Giá trị thuộc tính |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Sau khi tiến hành tiền xử lý dữ liệu, ta thu được tập dữ liệu có đầy đủ các giá trị của các thuộc tính. Các thuộc tính có độ phân tán dữ liệu và tính liên tục cạo được rời rạc hóa thành các khoảng dữ liệu. Các thuộc tính không có độ tương quan cao đồng nghĩa không mang lại độ phức tạp của dữ liệu. Lúc này, tập dữ liệu đã tinh gọn và chính xác hơn, sẵn sàng cho việc khai thác dữ liệu và đưa ra dự đoán.

* + 1. **Phân loại data**
    2. **Data Scaling**

Nhận xét:

Nhìn vào biểu đồ cũng như báo cáo về độ chính xác của lần lượt các thuật toán là

**Random Forest > ID3 > K-Nearest-Neighbors, Cart > Logistic Regression**

* Nếu quan tâm đến độ chính xác toàn mô hình tối ưu, ta có thể chọn thuật toán Random Forest để dự đoán tốt nhất cho bộ dữ liệu này với độ chính xác khoảng 96.23%
* Nếu quan tâm về độ nhanh của mô hình, ta có thể sử dụng ID3 hoặc Cart, tuy nhiên so về tổng thể thì thuật toán Random Forest vẫn là thuật toán tối ưu nhất.

.

# **CHƯƠNG IV - PHẦN MỀM DỰ ĐOÁN**

## **4.1 Tổng quan phần mềm**

### **4.1.1 Thuật toán được sử dụng**

Dựa vào kết quả thu được ở phần trước, nhóm quyết định sử dụng thuật toán Random Forest cho phần mềm này. Theo kết quả so sánh, thuật toán này dù có tốc độ không tốt nhưng nó lại cho độ chính xác cao nhất.

### **4.1.2 Các thuộc tính được sử dụng để đưa ra dự đoán**

=> Sau khi đưa vào tính toán độ tin cậy của các thuộc tính ta thu được 7 thuộc tính là **age, cp, trestbps, chol, thalach, oldpeak,slope** để đưa vào phần mềm. Đây là 7 thuộc tính có độ tin cậy cao nhất.

### **4.1.3 Giao diện và kiểm thử**

#### **4.1.3.1 Giao diện**

#### **4.1.3.2 Kiểm thử**

Bộ dữ liệu kiểm thử:

Kết quả:

Kết quả phần mềm đưa ra khớp với dữ liệu ban đầu.

## **4.2 Code phần mềm**

### **4.2.1 Code phần giao diện:**

# **CHƯƠNG V - KẾT LUẬN**

**5.1 Ưu điểm, hạn chế của từng thuật toán**

***5.1.1 Cây quyết định***

***-* Ưu điểm:**

* Thuật toán đơn giản, trực quan, không quá phức tạp để hiểu ngay lần đầu tiên.
* Bộ dữ liệu training không nhất thiết phải quá lớn để tiến hành xây dựng mô hình phân tích.
* Một số thuật toán cây quyết định có khả năng xử lý dữ liệu bị missing và dữ liệu bị lỗi mà không cần áp dụng phương pháp như “imputing missing values” hay loại bỏ. Ít bị ảnh hưởng bởi các dữ liệu ngoại lệ.
* Không cần phải có các giả định ban đầu về các quy luật phân phối như trong thống kê, và nhờ đó kết quả phân tích có được là khách quan, “tự nhiên” nhất.
* Có thể giúp chúng ta phân loại đối tượng dữ liệu theo biến mục tiêu có nhiều lớp, nhiều nhóm khác nhau (multi-class classification) đặc biệt nếu biến mục tiêu là dạng biến định lượng phức tạp.
* Có thể áp dụng linh hoạt cho các biến target, biến mục tiêu.
* Mang lại kết quả dự báo có độ chính xác cao, dễ dàng thực hiện, nhanh chóng trong việc huấn luyện, không cần phải chuyển đổi các biến.
* Dễ diễn giải hay giải thích đến người nghe, người xem – những người muốn hiểu rõ về kết quả phân tích nhưng không có kiến thức gì về khoa học dữ liệu.
* Nói lên được mối liên hệ giữa các biến, các thuộc tính dữ liệu một cách trực quan nhất.
* Ngoài kinh tế, tài chính, thuật toán cây quyết định có thể được ứng dụng trong lĩnh vực y tế, nông nghiệp, sinh học.

**Khuyết điểm:**

* Thuật toán cây quyết định hoạt động hiệu quả trên bộ dữ liệu đơn giản có ít biến dữ liệu liên hệ với nhau, và ngược lại nếu áp dụng cho bộ dữ liệu phức tạp.
* Khi được áp dụng với bộ dữ liệu phức tạp, nhiều biến và thuộc tính khác nhau có thể dẫn đến mô hình bị overfitting, quá khớp với dữ liệu training dẫn đến vấn đề không đưa ra kết quả phân loại chính xác khi áp dụng cho dữ liệu test, và dữ liệu mới.
* Giá trị phương sai cao, khi có sự thay đổi nhỏ trong bộ dữ liệu có thể gây ảnh hưởng đến cấu trúc của mô hình.
* Thuật toán cây quyết định chỉ áp dụng cho biến định tính (classification tree) nếu phân loại sai có thể dẫn đến sai lầm nghiêm trọng.
* Thuật toán cây quyết định có khả năng “bias” hay thiên vị nếu bộ dữ liệu không được cân bằng.
* Yêu cầu bộ dữ liệu training và test phải được chuẩn bị hoàn hảo, chất lượng tốt phải được cân đối theo các lớp, các nhóm trong biến mục tiêu.
* Không  “hỗ trợ” kỹ thuật hay khả năng “truy vấn ngược”.

***5.1.2. Random Forest***

- Ưu điểm:

* Cải tiến hơn với thuật toán cây quyết định, giải quyết được sự nhiễu khi bộ dữ liệu thay đổi.

- Nhược điểm

* Nhược điểm chính của Random Forest đó chính là khối lượng tính toán lớn, tuy nhiên với năng lực tính toán ngày càng tăng của máy tính (theo cấp lũy thừa) thì hạn chế của Random Forest không phải là vấn đề lớn.

***5.1.3 Logistic Regression***

- Ưu điểm

* Phân lớp tốt khi dữ liệu có khả năng phân tách tuyến tích (linearly separable).
* Dễ triển khai và training.

- Nhược điểm

* Dễ bị ảnh hưởng bởi nhiễu.
* Không thể giải quyết các bài toán phi tuyến tính (non-linear).
* Nhạy cảm với overfitting.

***5.1.4 K-nearest Neighbors***

**- Ưu điểm:**

* Thuật toán đơn giản, dễ dàng triển khai.
* Độ phức tạp tính toán nhỏ.
* Xử lý tốt với tập dữ liệu nhiễu

**- Nhược điểm**

* Với K nhỏ dễ gặp nhiễu dẫn tới kết quả đưa ra không chính xác
* Cần nhiều thời gian để thực hiện do phải tính toán khoảng cách với tất cả các đối tượng trong tập dữ liệu.
* Cần chuyển đổi kiểu dữ liệu thành các yếu tố định tính.

**5.2 Hướng phát triển**

* Nghiên cứu và tìm hiểu thêm về dataset, điều chỉnh lại các thuộc tính và các tiền xử lý để đạt độ chính xác cao hơn.
* Tiếp tục áp dụng các thuật toán phân lớp khác như SVM, Artificial Neuron Network, … để tìm thuật toán tối ưu.

# **Tài liệu tham khảo**

* 1. Slide môn Phân tích dữ liệu kinh doanh trường đại học CNTT – DHQG TPHCM.
  2. Ma trận nhầm lẫn: [What is Precision, Recall and F1-score? - The Conscious's notes (wordpress.com)](https://caihuuthuc.wordpress.com/2020/02/23/precision-recall-va-f1-score-la-gi/)
  3. Các thuật toán: [CART: Classification and Regression Trees for Clean but Powerful Models | by Saul Dobilas | Towards Data Science](https://towardsdatascience.com/cart-classification-and-regression-trees-for-clean-but-powerful-models-cc89e60b7a85)
  4. Các loại data: [What is Ordinal Data? [Definition, Analysis & Examples] (careerfoundry.com)](https://careerfoundry.com/en/blog/data-analytics/what-is-ordinal-data/)
  5. Các loại machine learning: [4 Machine Learning Approaches that Every Data Scientist Should Know | by Orhan G. Yalçın | Towards Data Science](https://towardsdatascience.com/4-machine-learning-approaches-that-every-data-scientist-should-know-e3a9350ec0b9#:~:text=Unsupervised%20learning%20is%20a%20type%20of%20machine%20learning,datasets%2C%20which%20do%20not%20contain%20labels.%20Figure%204.)
  6. Video youtube:
  7. Dataset: [Heart Disease Dataset | Kaggle](https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset)

# **Hết.**