|  |  |
| --- | --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **ĐỒ ÁN MÔN HỌC**  **ỨNG DỤNG THUẬT TOÁN NAIVE BAYES ĐỂ PHÂN LOẠI CÁC LOẠI NẤM ĂN ĐƯỢC**    **Sinh viên thực hiện : Võ Quốc Kiệt**  **MSSV : 2000000000625**  **Sinh viên thực hiện : Nguyễn Vũ Nguyên Chương**  **MSSV : 0123456789**  **Chuyên ngành : Trí tuệ nhân tạo**  **Môn học : Khai Phá Dữ Liệu Lớn**  **Khóa : 2020**  **Giảng viên giảng dạy : ThS. Hồ Ngọc Trung Kiên**  **Tp.HCM, tháng 09 Năm 2022** | |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **KHOA CÔNG NGHỆ THÔNG TIN**  🙜 🙜 🙝 | | CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM  **Độc lập – Tự do – Hạnh phúc**  🙜 🙜 🙝 | |

**NHIỆM VỤ ĐỒ ÁN MÔN HỌC**

*(Sinh viên phải đóng tờ này vào cuốn báo cáo)*

Họ và tên:**VÕ QUỐC KIỆT** MSSV:**2000000625**

Họ và tên: **NGUYỄN VŨ NGUYÊN CHƯƠNG** MSSV: **123456**

Chuyên ngành: **Trí tuệ nhân tạo** Lớp: **13CTH** Email: **123456@GMAIL.COM** SĐT: **11111111**

Tên đề tài: **ỨNG DỤNG THUẬT TOÁN NAIVE BAYES ĐỂ PHÂN LOẠI CÁC LOẠI NẤM ĂN ĐƯỢC………………………………………………………...................**

Giảng viên giảng dạy: **ThS. Hồ Ngọc Trung Kiên**

Thời gian thực hiện: **12/06/2022** **đến …./09/2022**

**Nhiệm vụ/nội dung (mô tả chi tiết nội dung, yêu cầu, phương pháp… ):**

* Phân tích yêu cầu bài toán
* Xác định đầu vào và đầu ra
* Xác định các luật
* Phân tích đặc trưng của dữ liệu
* Xây dựng mô hình điều khiển, đánh giá
* Dự báo, biểu diễn kết quả

**Nội dung và yêu cầu đã được thông qua Bộ môn.**

*Tp.HCM, ngày tháng 09 năm 2022*

|  |  |
| --- | --- |
| **TRƯỞNG BỘ MÔN**  **Đặng Như Phú** | **GIẢNG VIÊN GIẢNG DẠY**  **Hồ Ngọc Trung Kiên** |

# LỜI CẢM ƠN

Đặc biệt xin gửi lời cám ơn đến cha mẹ và anh chị em trong gia trình đã động viên tinh thần cũng như chia sẻ những khó khăn trong những năm qua.

Để hoàn thành chuyên đề báo cáo đồ án lần này ,trước tiên em xin gửi đến các qua thầy, cô giáo trường Đại Học Nguyễn Tất Thành lời cảm ơn chân thành và sâu sắc nhất.

Em xin chân thành cảm ơn thầy Hồ Ngọc Trung Kiên giảng viên của trường Đại Học Nguyễn Tất Thành đã hướng dẫn và hỗ trợ cho em cũng như đã giải đáp những thắc mắc của em về môn học trong quá trình học học tập và thực hiện đồ án .Và em cũng xin chân thành cảm ơn những  anh  chị khóa trên đã hướng  dẫn và chỉ bảo cho em.

Báo cáo về đồ án của chúng em không thể tránh được sự thiếu  xót  nên mong các thầy,cô cảm thông cảm cho chúng em , và chúng em cũng rất mong nhận lại được sự đóng góp ý kiến của thầy cô để giúp em hoàn thiện báo cáo này được tốt hơnhơn.

Em xin chân thành cảm ơn.

**Sinh Viên Thực Hiện**

**VÕ QUỐC KIỆT**

**NGUYỄN VŨ NGUYÊN CHƯƠNG**

# LỜI MỞ ĐẦU

-Khai phá dữ liệu (data mining) Là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính. Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp. Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến. Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu"

-Trong báo cáo này chúng em sẽ tiến hành thu thập dữ liện, tiến hành xử lý dữ liệu và cuối cùng sẽ sử dụng giải thuật Naïve Basyes để phân loại các loại nấm có thể ăn được.

# NHẬN XÉT CỦA GIẢNG VIÊN GIẢNG DẠY

*Tp.HCM, Ngày . . . . tháng . . . . năm . . . .*

**Giảng viên giảng dạy**

(Ký tên và ghi rõ họ tên)

**Hồ Ngọc Trung Kiên**

|  |  |
| --- | --- |
| **BẢNG PHÂN CÔNG THỰC HIỆN ĐỒ ÁN MÔN HỌC** | |
| **Võ Quốc Kiệt-MSSV:200000625** | **Nguyễn Vũ Nguyên Chương -MSSV:** |
| * Công việc 1… | * Công việc 1… |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
| **Điểm A:** | **Điểm B:** |
| Nhận xét: | Nhận xét: |

# MỤC LỤC

Table of Contents

[LỜI CẢM ƠN iv](#_Toc120136581)

[LỜI MỞ ĐẦU v](#_Toc120136582)

[NHẬN XÉT CỦA GIẢNG VIÊN GIẢNG DẠY vi](#_Toc120136583)

[MỤC LỤC ix](#_Toc120136584)

[DANH MỤC CÁC BẢNG BIỂU x](#_Toc120136585)

[DANH MỤC CÁC BẢNG HÌNH xi](#_Toc120136586)

[DANH MỤC CÁC TỪ VIẾT TẮT xi](#_Toc120136587)

[CHƯƠNG I: GIỚI THIỆU 1](#_Toc120136588)

[1.1: Lý Thuyết : 1](#_Toc120136589)

[1.1.1:Khái niệm: 1](#_Toc120136590)

[1.1.2:Các bước khai phá dữ liệu: 1](#_Toc120136591)

[1.1.3:Tiền xử lý dữ liệu: 2](#_Toc120136592)

[1.1.4:Làm sạch dữ liệu: 3](#_Toc120136593)

[1.1.5. Tích hợp dữ liệu (data integration): 4](#_Toc120136594)

[1.1.6: Biến đổi dữ liệu (data transformation): 4](#_Toc120136595)

[1.1.7.:Thu giảm dữ liệu (data reduction) : 4](#_Toc120136596)

[1.2:Tổng quan đề tài: 5](#_Toc120136597)

# DANH MỤC CÁC BẢNG BIỂU

# DANH MỤC CÁC BẢNG HÌNH

# DANH MỤC CÁC TỪ VIẾT TẮT

1. CHƯƠNG I: GIỚI THIỆU

## 1.1: Lý Thuyết :

### 1.1.1:Khái niệm:

- Khai phá dữ liệu (data mining) Là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính. Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp.Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến.Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu".

### 1.1.2:Các bước khai phá dữ liệu:

Bước 1: Làm sạch dữ liệu – Trong bước này, dữ liệu được làm sạch sao cho không có tạp âm hay bất thường trong dữ liệu.

+ Bước 2: Tích hợp dữ liệu – Trong quá trình tích hợp dữ liệu, nhiều nguồn dữ liệu sẽ kết hợp lại thành một.

+ Bước 3: Lựa chọn dữ liệu – Trong bước này, dữ liệu được trích xuất từ cơ sở dữ liệu.

+ Bước 4: Chuyển đổi dữ liệu – Trong bước này, dữ liệu sẽ được chuyển đổi để thực hiện phân tích tóm tắt cũng như các hoạt động tổng hợp.

+ Bước 5: Khai phá dữ liệu – Trong bước này, chúng tôi trích xuất dữ liệu hữu ích từ nhóm dữ liệu hiện có.

+ Bước 6: Đánh giá mẫu – Chúng tôi phân tích một số mẫu có trong dữ liệu.

+ Bước 7: Trình bày thông tin – Trong bước cuối cùng, thông tin sẽ được thể hiện dưới dạng cây, bảng, biểu đồ và ma trận

**Diagram, schematic

Description automatically generated**

Hình 1. :Data Mining

### 1.1.3:Tiền xử lý dữ liệu:

-Là một bước rất quan trọng trong lĩnh vực học máy. Đây là bước chúng ta sẽ làm ngay sau khi dữ liệu được làm sạch và việc phân tích số liệu thăm dò được thực hiện. Khi đã hiểu được tập dữ liệu của mình, ta có thể sẽ có ý tưởng về việc mô hình hóa dữ liệu..

**a) Tập dữ liệu :**

- Một tập dữ liệu (dataset) là một tập hợp các đối tượng (object) và các thuộc tính của chúng. - Mỗi thuộc tính (attribute) mô tả một đặc điểm của một đối tượng.

**Graphical user interface, application

Description automatically generated**

Hình 1. :VD DataSet

**b) Các kiểu tập dữ liệu :**

- Bản ghi (record): Các bản ghi trong cơ sở dữ liệu quan hệ. Ma trận dữ liệu. Biểu diễn văn bản. Hay dữ liệu giao dịch.„

- Đồ thị (graph): World wide web. Mạng thông tin, hoặc mạng xã hội - Dữ liệu có trật tự: Dữ liệu không gian (ví dụ: bản đồ). Dữ liệu thời gian (ví dụ: time-series data). Dữ liệu chuỗi (ví dụ: chuỗi giao dịch).

**c) Các kiểu giá trị thuộc tính:**

- Kiểu định danh/chuỗi (norminal): không có thứ tự. Ví dụ: Các thuộc tính như: Name, Profession, …

- Kiểu nhị phân (binary): là một trường hợp đăc biệt của kiểu định danh. Tập các giá trị chỉ gồm có 2 giá trị (Y/N, 0/1, T/F).

- Kiểu có thứ tự (ordinal): Integer, Real, … -lấy giá trị từ một tập có thứ tự giá trị. Ví dụ: -Các thuộc tính lấy giá trị số như : Age, Height ,… Hay lấy một tập xác định, thuộc tính Income lấy giá trị từ tập {low, medium, high}.

-Kiểu thuộc tính rời rạc (discrete-valued attributes): có thể là tập các giá trị của một tập hữu hạn. Bao gồm thuộc tính có kiểu giá trị là các số nguyên, nhị phân.

-Kiểu thuộc tính liên tục (continuous-valued attributes):Các giá trị làsố thực.

**d) Các đặc tính mô tả của dữ liệu**:

- Giúp hiểu rõ về dữ liệu có được: chiều hướng chính/trung tâm, sự biến thiên, sự phân bố.

- Sự phân bố của dữ liệu (data dispersion):

+ Giá trị cực tiểu/cực đại (min/max).

+ Giá trị xuất hiện nhiều nhất (mode).

+ Giá trị trung bình (mean).

+ Giá trị trung vị (median).

+ Sự biến thiên (variance) và độ lệch chuẩn (standard deviation) .

+ Các ngoại lai (outliers).

### 1.1.4:Làm sạch dữ liệu:

-Đối với dữ liệu thu thập được, cần xác định các vấn đề ảnh hưởng là cho nó không sạch. Bởi vì, dữ liệu không sạch (có chứa lỗi, nhiễu, không đầy đủ, có mâu thuẫn) thì các tri thức khám phá được sẽ bị ảnh hưởng và không đáng tin cậy, sẽ dẫn đến các quyết định không chính xác. Do đó, cần gán các giá trị thuộc tính còn thiếu; sửa chữa các dữ liệu nhiễu/lỗi; xác định hoặc loại bỏ các ngoại lai (outliers); giải quyết các mâu thuẫn dữ liệu.

### 1.1.5. Tích hợp dữ liệu (data integration):

-Tích hợp dữ liệu là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu có sẵn cho quá trình khai phá dữ liệu. Khi tích hợp cần xác định thực thể từ nhiều nguồn dữ liệu để tránh dư thừa dữ liệu. Ví dụ: Bill Clinton ≡ B.Clinton. Việc dư thừa dữ liệu là thường xuyên xảy ra, khi tích hợp nhiều nguồn. Bởi cùng một thuộc tính (hay cùng một đối tuợng) có thể mang các tên khác nhau trong các nguồn (cơ sở dữ liệu) khác nhau. Hay các dữ liệu suy ra được như một thuộc tính trong một bảng có thể được suy ra từ các thuộc tính trong bảng khác. Hay sự trùng lắp các dữ liệu. Các thuộc tính dư thừa có thể bị phát hiện bằng phân tích tương quan giữa chúng.

-Phát hiện và xử lý các mâu thuẫn đối với giá trị dữ liệu: Đối với cùng một thực thể trên thực tế, nhưng các giá trị thuộc tính từ nhiều nguồn khác nhau lại khác nhau. Có thể cách biểu diễn khác nhau, hay mức đánh giá, độ do khác nhau.

### 1.1.6: Biến đổi dữ liệu (data transformation):

-Biến đổi dữ liệu là việc chuyển toàn bộ tập giá trị của một thuộc tính sang một tập các giá trị thay thế, sao cho mỗi giá trị cũ tương ứng với một trong các giá trị mới. Các phương pháp biến đổi dữ liệu:

- Làm trơn (smoothing): Loại bỏ nhiễu/lỗi khỏi dữ liệu.

- Kết hợp (aggregation): Sự tóm tắt dữ liệu, xây dựng các khối dữ liệu.

- Khái quát hóa (generalization): Xây dựng các phân cấp khái niệm.

- Chuẩn hóa (normalization): Đưa các giá trị về một khoảng được chỉ định.

### 1.1.7.:Thu giảm dữ liệu (data reduction) :

- Một kho dữ liệu lớn có thể chứa lượng dữ liệu lên đến terabytes sẽ làm cho quá trình khai phá dữ liệu chạy rất mất thời gian, do đó nên thu giảm dữ liệu. Việc thu giảm dữ liệu sẽ thu đuợc một biểu diễn thu gọn, mà nó vẫn sinh ra cùng (hoặc xấp xỉ) các kết quả khai phá như tập dữ liệu ban đầu. Các chiến lược thu giảm:

- Giảm số chiều (dimensionality reduction), loại bỏ bớt các thuộc tính không (ít) quan trọng.

- Giảm lượng dữ liệu (data/numberosity reduction)

+ Kết hợp khối dữ liệu.

+ Nén dữ liệu.

+ Hồi quy.

+ Rời rạc hóa

## 1.2:Tổng quan đề tài: