

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

**BÁO CÁO ĐỒ ÁN**

**MÔN HỌC: MÁY HỌC**

*HỌC KÌ 2 – NĂM HỌC: 2019-2020*



**ĐỀ TÀI**

ỨNG DỤNG MÁY HỌC VÀO CHẨN ĐOÁN BỆNH TIỂU ĐƯỜNG

**Lớp:** CS114.K22.KHCL

**Giảng viên hướng dẫn:** Mai Tiến Dũng

**Sinh viên thực hiện:**

1. Nguyễn Ngọc Bình 18520506
2. Nguyễn Huỳnh Nhi 18521204

THÀNH PHỐ HỒ CHÍ MINH, NGÀY 17 THÁNG 06 NĂM 2020

Mục Lục

[Phần 1: THÔNG TIN NHÓM 1](file:///C:\Users\Dell\Desktop\Đồ%20Án%20Máy%20Học\Nội%20Dung%20Máy%20Học.docx#_Toc45301978)

[1. Thành viên 1](#_Toc45301979)

[2. Phân công nhiệm vụ 1](#_Toc45301980)

[3. Đánh giá chung 2](#_Toc45301981)

[Phần 2: THÔNG TIN ĐỒ ÁN 3](file:///C:\Users\Dell\Desktop\Đồ%20Án%20Máy%20Học\Nội%20Dung%20Máy%20Học.docx#_Toc45301982)

[1. Giới thiệu chung 3](#_Toc45301983)

[**1.1** **Bộ dữ liệu sử dụng trong đồ án** 3](#_Toc45301984)

[**1.2** **Lý do chọn đề tài** 3](#_Toc45301985)

[2 Sử dụng Naive Bayes để chuẩn đoán bệnh 3](#_Toc45301986)

[**2.1** **Định nghĩa** 3](#_Toc45301987)

[**2.2** **Phân loại** 4](#_Toc45301988)

[**2.3** **Ứng dụng của thuật toán Naive Bayes** 5](#_Toc45301989)

[**2.4** **Các bước áp dụng Gaussian Naive Bayes vào các bài toán bằng Python** 6](#_Toc45301990)

[**2.5** **Áp dụng Gaussian Naive Bayes vào chuẩn đoán bệnh tiểu đường.** 6](#_Toc45301991)

[**2.6** **Ưu, nhược điểm của thuật toán Naive Bayes** 7](#_Toc45301992)

[3. Sử dụng thuật toán KNN để chuẩn đoán bệnh 8](#_Toc45301993)

[**3.1** **Định nghĩa** 8](#_Toc45301994)

[**3.2** **Cách thực hiện** 8](#_Toc45301995)

[**3.3** **Cách tính khoảng cách giữa các mẫu.** 9](#_Toc45301996)

[**3.4** **Ứng dụng của thuật toán KNN** 9](#_Toc45301997)

[**3.5** **Các bước để áp dụng KNN** 10](#_Toc45301998)

[**3.6** **Áp dụng KNN vào chuẩn đoán bệnh tiểu đường** 10](#_Toc45301999)

[**3.7** **Ưu, nhược điểm của thuật toán KNN** 11](#_Toc45302000)

[4. So sánh kết quả giữa 2 thuật toán khi chạy trên bộ dữ liệu 12](#_Toc45302001)

[5. Ý kiến, quan điểm của nhóm 12](#_Toc45302002)

[6. Tài liệu tham khảo 13](#_Toc45302003)

# Phần 1: THÔNG TIN NHÓM

## **Thành viên**

|  |  |  |
| --- | --- | --- |
| Họ và tên | Mã số sinh viên | Lớp |
| Nguyễn Ngọc Bình | 18520506 | CS114.K22.KHCL |
| Nguyễn Huỳnh Nhi | 18521204 | CS114.K22.KHCL |

## **Phân công nhiệm vụ**

|  |  |
| --- | --- |
| Phân công nhiệm vụ | Sinh viên phụ trách |
| Tìm và chọn đề tài | Nguyễn Huỳnh Nhi  Nguyễn Ngọc Bình |
| Đăng kí nhóm | Nguyễn Huỳnh Nhi |
| Thu thập thông tin, tìm tài liệu về các thuật toán | Nguyễn Ngọc Bình  Nguyễn Huỳnh Nhi |
| Viết code cho giải thuật Naive Bayes | Nguyễn Huỳnh Nhi |
| Viết code cho giải thuật KNN | Nguyễn Ngọc Bình |
| Viết báo cáo đồ án | Nguyễn Ngọc Bình |
| Làm PowerPoint thuyết trình | Nguyễn Huỳnh Nhi  Nguyễn Ngọc Bình |
| Báo cáo đồ án cuối kỳ | Nguyễn Ngọc Bình  Nguyễn Huỳnh Nhi |

## **Đánh giá chung**

* Nhìn chung, mỗi thành viên đều đóng góp nhiều và hoàn thành tốt nhiệm vụ được phân công trong quá trình làm đồ án
* Các thành viên đều cố gắng, tích cực hoàn thành tốt nhiệm vụ của mình, thúc đẩy thực hiện đồ án
* Nâng cao được kĩ năng làm việc nhóm cũng như phát huy những điểm mạnh của từng cá nhân. Hỗ trợ nhau tích cực trong những mặt hạn chế để đạt kết quả cao nhất. Ngoài ra còn nâng cao kĩ năng tư duy, lập trình của từng cá nhân, làm cơ sở định hướng chuyên ngành tương lai.
* Các thành viên đảm bảo hoàn thành tốt đồ án cuối kì

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| HỌ VÀ TÊN | MSSV | ĐÓNG GÓP | MỨC ĐỘ HOÀN THÀNH | NHẬN XÉT VÀ ĐÁNH GIÁ |
| Nguyễn Ngọc Binh | 18520506 | 50% | 95% | Tốt |
| Nguyễn Huỳnh Nhi | 18521204 | 50% | 95% | Tốt |

# Phần 2: THÔNG TIN ĐỒ ÁN

## **Giới thiệu chung**

### **Bộ dữ liệu sử dụng trong đồ án**

* Bộ dữ liệu này có nguồn gốc từ Viện tiểu đường và bệnh tiêu hóa và thận quốc gia Ấn Độ.
* Tập dữ liệu bao gồm các cột: Mang thai, Glucoso, Huyết áp, Độ dày da, Insulin, BMI, Chức năng phả hệ, Tuổi, Kết quả.

A picture containing bird, tree, flower

Description automatically generated

### **Lý do chọn đề tài**

* Tìm hiểu về thuật toán Naive Bayes và KNN, cách thức hoạt động của thuật toán
* Áp dụng các thuật toán khác nhau trên bộ dữ liệu để so sánh kết quả

## **Sử dụng Naive Bayes để chuẩn đoán bệnh**

### **Định nghĩa**

* Naive Bayes là giải thuật học có giảm sát, được dựa theo lý thuyết Bayes và được dùng trong giải quyết các bài toán phân lớp.
* Thường được sử dụng trong phân loại văn bản mà có bội dữ liệu huấn luyện nhiều chiều.
* Được gọi là Naive Bayes vì:
* Naive: Bởi vì nó giả định rằng sự xuất hiện của một tính năng nhất định là độc lập với sự xuất hiện của các tính năng khác. Chẳng hạn như nếu trái cây được xác định trên cơ sở của màu sắc, hình dạng và mùi vị, thì trái cây màu đỏ, hình cầu và ngọt được công nhận là một quả táo. Do đó, mỗi tính năng đều góp phần xác định rằng đó là một quả táo mà không phụ thuộc vào nhau
* Bayes: Dựa theo lý thuyết xác xuất của Bayes được dùng để quyết định xác xuất của 1 giả thuyết với kiến thức cho trước. Nó phụ thuộc vào xác xuất có điều kiện.
* Công thức Bayes:

A screenshot of a cell phone

Description automatically generated

* + - * P(A|B): Xác suất của giá thuyết A biết rằng B xảy ra.
      * P(B|A): Xác suất của giả thuyết B mà A đã xảy ra.
      * P(A): Xác suất giả thuyết A xảy ra.
      * P(B): Xác suất giả thuyết B xảy ra.

### **Phân loại**

* Có 3 loại chính:
* Gaussian Naive Bayes: Mô hình này được sử dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến liên tục. Công thức:



* Multinomial Naive Bayes: Mô hình chủ yếu được sử dụng trong phân loại văn bản mà vector đặc trưng được tính bằng Bags of Words(BOW). Lúc này mỗi văn bản sẽ được biểu diễn với độ dài D là số từ trong BOW. Giá trị của thành phần thứ i trong mỗi vector chính là số lần từ thứ i xuất hiện trong văn bản đó. Khi đó **p(Xi|C)** tỉ lệ với tần suất từ thứ i xuất hiện trong văn bản của class C. Giá trị được tính bằng cách:

A picture containing clock

Description automatically generated

* + Với **Nci** là số lần từ thứ i xuất hiện trong văn bản của lớp **c**. được tính bằng tổng tất cả các phần tử thứ i của vector đặc trưng với lớp **c**.
  + Với **Nc** là tổng số từ xuất hiện trong lớp **c**, bằng độ dài của toàn bộ văn bản thuộc lớp **c**.
* Bernoulli Naive Bayes: Mô hình được áp dụng cho các loại dữ liệu mà mỗi thành phần là 1 giá trị nhị phân bằng 0 hoặc 1. Công thức tính:



* + - Với **P(i|c)** là xác xuất từ thứ **i** xuất hiện trong văn bản của lớp **c.**

### **Ứng dụng của thuật toán Naive Bayes**

* Real-time Prediction : Do Naive Bayes rất nhanh, nên có thể sử dụng để dự đoán trong thời gian thực.
* Multi-class Prediction : Giải thuật này có thể dự đoán xác xuất của nhiều lớp của một biến mục tiêu.
* Text Classification / Spam Filtering / Sentiment Analysis : Naive Bayes được sử dụng nhiều trong phân lớp văn bản do kết quả tốt hơn của nó trong vấn đề nhiều lớp và các luật độc lập so với các giải thuật khác.
* Recommendation System : Naive Bayes Classifier cùng với giải thuật như Cllaborative Filtering tạo nên một Recommendation System mà có thể sử dụng máy học và khai thác dữ liệu để lọc thông tin không nhìn thấy và chuẩn đoán liệu người dùng có muốn một tài nguyên nhất định hay không.

### **Các bước áp dụng Gaussian Naive Bayes vào các bài toán bằng Python**

* Bước 1: Nhận dữ liệu từ dataset, chia các cột riêng trong giá trị độc lập và giá trị không độc lập
* Bước 2: Chia dataset thành training và test set.
* Bước 4: Huấn luyện dữ liệu trên training sets
* Bước 5: Xuất ra kết quả khi áp dụng lên test set

### **Áp dụng Gaussian Naive Bayes vào chuẩn đoán bệnh tiểu đường.**

* Bước 1: Nạp bộ dữ liệu từ file
* Bước 2: Xuất ra 5 phần tử đầu tiên của bộ dữ liệu và miêu tả về bộ dữ liệu

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

* Bước 3: Sử dụng GaussianNB từ thư viện sklearn để huấn luyện dữ liệu trên tập dữ liệu và đưa ra độ chính xác.

A screenshot of a cell phone

Description automatically generated

### **Ưu, nhược điểm của thuật toán Naive Bayes**

1. *Ưu điểm:*

* Nhanh trong việc huấn luyện và xử lý.
* Không nhạy cảm với các tính năng không liên quan
* Xử lý dữ liệu thực và rời rạc.
* Xử lý dữ liệu phát trực tuyến ( streaming data) tốt.
* Đặc biệt chạy tốt trong những trường hợp như :
  + Khi các giả định ngây thơ thực sự khớp với dữ liệu( điều rất hiếm trong thực tế).
  + Đối với các danh mục được phân tách rất tốt, khi độ phức tạp của mô hình ít qua trọng hơn.
  + Đối với dữ liệu nhiều chiều, khi độ phức tạp của mô hình ít quan trọng hơn*.*

1. *Nhược điểm, hạn chế và cách giải quyết:*

* Thuật toán giả định các thuộc tính là độc lập. Điểm yếu này có thể giải quyết bằng cách thực hiện một số phân tích thống kê trước khi sử dụng Naive Bayes để đo mức độ tương quan giữa các tính năng và sau đó chọn những điểm không tương thích nhất.
* Thuật toán đối xử với tất cả các thuộc tính là như nhau. Vì vậy, một số trọng số có thể được thêm vào các thuộc tính quan trọng để tăng đóng góp của chúng trong quyết định cuối cùng.
* Vấn đề xác xuất bằng 0. Vấn đề này được giải quyết bằng cách thêm giá trị một vào tần số của từng thuộc tính hoặc sử dụng phương thức phân phối Gaussian.
* Vấn đề thuộc tính giá trị liên tục. Chuyển đổi từ giá trị liên tục sang giá trị rời rạc là cách giải quyết vấn đề

## **3. Sử dụng thuật toán KNN để chuẩn đoán bệnh**

### **Định nghĩa**

* Thuật toán K-nearest neighbors (KNN) là 1 trong những giải thuật học có giám sát của máy học.
* Thuật toán có thể dùng cho cả việc dự báo phân loại cũng như hồi quy
* Hai tính chất đăc trưng của KNN là:
  + Lazy learning algorithm – KNN: là 1 thuật toán học lười bởi vì nó không có 1 giai đoạn đào tạo đặc biên và sử dụng toàn bộ dữ liệu cho việc huấn luyện trong khi phân lớp.
  + Non-parametic learning algorithm – KNN : là một thuật toán học không tham số vì nó không nhận bất cứ điều gì về dữ liệu nền tảng.
* Ý tưởng của thuật toán: Để dự đoán mẫu dữ liệu mới z thuộc về lớp nào thì ta dựa vào số k dữ liệu gần nó nhất.
  + VD: k = 3, nghĩa là gần z có 3 điểm dữ liệu. Giả sử trong d dó có 2 điểm dữ liệu thuộc về lớp B và 1 điểm dữ liệu thuộc về lớp A thì ta gán z thuộc về lớp B.
* Là 1 trong những thuật toán máy học đơn giản nhất. Không đào tạo hoặc mô hình rõ ràng. Bất cứ khi nào có 1 điểm mới để phân loại, thì chúng ta tìm K điểm gần nhất của nó trong bộ dữ liệu huấn luyện.

### **Cách thực hiện**

* Bước 1: Tính khoảng cách giữa z và tất cả N mẫu dữ liệu x1,x2,x3….xn trong tập huấn luyện.
* Bước 2: Chọn k mẫu gần nhất với z trong tập huấn luyện.
* Bước 3: Gán z vào lớp có nhiều mẫu nhất trong số k mẫu láng giềng đó.

### **Cách tính khoảng cách giữa các mẫu.**

* Có 3 cách tính khoảng cách:
  + Euclidean

A picture containing clock

Description automatically generated

* + Manhattan

A picture containing object, clock

Description automatically generated

* + Minkowski

A close up of a logo

Description automatically generated

### **Ứng dụng của thuật toán KNN**

* Hệ thống ngân hàng : giải thuật KNN có thể được dùng trong hệ thống ngân hàng để dự đoán xem một cá nhân có thể phù hợp cho chấp nhận vay hay không? Cá nhân đó có những đặc điểm giống với những tiêu chuẩn hay không?
* Tính toán xếp hạng tín dụng:giải thuật KNN có thể được dùng để tìm xếp hạng tín dụng của một cá nhân dựa vào việc so sánh với những người có cùng đặc điểm.
* Chính trị: Với sự giúp đỡ của giải thuật KNN, chúng ta có thể phân loại một người bầu cử tiềm năng vào các lớp như “Sẽ bầu”, “Sẽ không bầu”.
* Ngoài ra, các lĩnh vực mà giải thuật KNN có thể được áp dụng gồm có nhận diện giọng nói, nhận diện chữ viết tay, nhận diện ảnh và video.

### **Các bước để áp dụng KNN**

* Bước 1: Nạp dữ liệu test và dữ liệu huấn luyện.
* Bước 2: Chọn giá trị của K phần tử gần nhất. K có thể là số nguyên bất kì.
* Bước 3: Với mỗi 1 điểm trong dữ liệu test thì :
  + Tính toán khoảng cách giữa giữ liệu test và mỗi hàng của dữ liệu huấn luyện = các hàm tính khoảng cách.
  + Dựa vào giá trị khoảng cách, sắp xếp theo thứ tự tăng dần.
  + Chọn K hàng từ mảng đã sắp xếp.
  + Phần tử đang xét sẽ được gán vào lớp dựa theo số lần xuất hiện nhiều nhất của các hàng.
* Bước 4: Kết thúc.

### **Áp dụng KNN vào chuẩn đoán bệnh tiểu đường**

* Bước 1: Nạp dữ liệu từ file.
* Bước 2: Xuất ra 5 phần tử đầu tiên của bộ dữ liệu và mô tả về bộ dữ liệu

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

* Bước 3: Sử dụng thư viên sklearn.neighbors để huấn luyện KNN trên bộ dữ liệu huấn luyện
* Bước 4: Đưa ra độ chính xác

A screenshot of a cell phone

Description automatically generated

### **Ưu, nhược điểm của thuật toán KNN**

1. *Ưu điểm*

* KNN là thuật toán dễ để hiểu và dễ diễn tả.
* KNN rất hữu dụng cho dữ liệu phi tuyết tình bởi vì không có giả định về dữ liệu trong thuật toán này.
* Đây là một thuật toán linh hoạt vì có thể sử dụng chúng để phân loại cũng như là hồi quy.
* KNN có độ chính xác tương đối cao nhưng cũng có các mô hình học tập có giám sát tốt hơn nhiều KNN.
* Chỉ có sự tương đồng giữa các đối tượng là cần thiết, không phải chính xác giá trị tính năng. Do đó nó có thể được áp dụng chô các đối tượng với mô tả tính năng phức tạp tùy ý.

1. *Nhược điểm*

* Là một thuật toán hơi đắt vì nó lưu trữ tất cả các dữ liệu huấn luyện.
* Cần bộ nhớ cao so với các thuật toán học có giám sát khác.
* Dự đoán chậm trong trường hợp N lớn.
* Thuật toán rất nhạy cảm với quy mô dữ liệu cũng như các tính năng không liên quan.
* Độ chính xác giảm với sự gia tăng của chiều không gian đặc trưng ( Cure of dimensionality).

## **So sánh kết quả giữa 2 thuật toán khi chạy trên bộ dữ liệu**

* Kết quả cuối cùng của thuật toán KNN:

A screenshot of a cell phone

Description automatically generated

* Kết quả cuối cùng của thuật toán Naive Bayes:

A screenshot of a cell phone

Description automatically generated

* Kết luận :
  + Độ chính xác của KNN lớn hơn so với Naive Bayes trong bộ dữ liệu này.
  + Macro avg, weighted avg của KNN luôn lớn hơn hoặc bằng so với Naive Bayes

## **Ý kiến, quan điểm của nhóm**

* Thuật toán KNN trong bộ dữ liệu này đã đưa ra kết quả chính xác hơn so với thuật toán Naive Bayes.
* Điểm khác biệt cơ bản giữa KNN và Naive Bayes là : KNN là thuật toán phân lớp có phân biệt đối xử còn Naive Bayes là phân lớp thế hệ.
* Naive Bayes chạy ra kết quả nhanh hơn nhiều KNN khi bộ dữ liệu trở nên nhiều hơn.

## **Tài liệu tham khảo**

[1] Thư viện Sklearn, Naive Bayes:

<https://scikit-learn.org/stable/modules/naive_bayes.html>

[2] Thư viện Sklearn, KneighborsClassifer:

<https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

[3] Các slide bài giảng của giáo viên hướng dẫn về KNN, Bayes.