**Trường Đại Học Công Nghiệp Thực Phẩm TP.HCM**

**Khoa Công Nghệ Thông Tin**



**BÁO CÁO MÔN HỌC**

**KHAI THÁC DỮ LIỆU**

**CHUẨN ĐOÁN BỆNH TIỂU ĐƯỜNG SỬ DỤNG THUẬT TOÁN**

**KNN**

Giảng viên hướng dẫn: Trần Như Ý

Sinh viên thực hiện: Nguyễn Trọng Liêm MSSV: 2001190646

Nguyễn Duy Khang MSSV: 2001190595

Trần Phương Bình MSSV: 2001191021

Nguyễn Cao Trí MSSV: 2001190553

Ngô Quang Hùng MSSV: 2001190885

TP. Hồ Chí Minh, ngày 10 tháng 6 năm 2016

Ngày nay, cùng với sự vươn lên của đất nước, hòa vào đó Công nghệ thông tin là một ngành đã có những bước phát triển mạnh mẽ không ngừng và nó đã trở thành chiếc chìa khóa dẫn đến thành công cho nhiều cá nhân, tổ chức. Với những ứng dụng của mình, công nghệ thông tin đã góp phần mang lại nhiều lợi ích mà không ai có thể phủ nhận được. Đặc biệt là trong lĩnh vực quản lý kinh tế, công nghệ thông tin đã góp phần tạo ra sự thay đổi nhanh chóng cho bộ mặt xã hội.

Trong quá trình thực hiện đề tài nhóm đã gặp phải không ít khó khăn và trở ngại, tuy nhiên, với sự giúp đỡ của bạn bè, cùng sự hướng dẫn tận tình, chỉ bảo cặn kẽ của giảng viên **Trần Như Ý**, cuối cùng chúng em cũng đã cố gắng hoàn thành đề tài này với tất cả sự cố gắng và nỗ lực. Chúng em xin được gửi lời cảm ơn tới những người bạn đã góp ý và giúp đỡ nhóm, và đặc biệt là cảm ơn cô, người đã giúp chúng tôi xác định được những yêu cầu và mục tiêu đề tài, truyền dạy các kĩ năng nghiệp vụ cũng như hướng dẫn cho chúng tôi cách thức thực hiện đề tài..

Do kiến thức thực tế và hiểu biết còn hạn chế nên trong quá trình hoàn thành báo cáo không tránh khỏi sai sót, nhóm chúng em rất mong sự góp ý đóng góp của thầy và các bạn để nhóm có thể hoàn thiện bài tập lớn được tốt hơn.

Chúng em xin chân thành cảm ơn!

MỤC LỤC

1. **GIỚI THIỆU THUẬT TOÁN**
2. **GIỚI THIỆU ĐỀ TÀI**
3. **KẾT LUẬN**
   1. **Giới thiệu thuật toán**

K-Nearest Neighbor (KNN) là một thuật toán phân loại, không nên nhầm lẫn với k-Means, chúng là hai thuật toán rất khác nhau với cách sử dụng rất khác nhau. k-Means là một thuật toán phân cụm không giám sát, với một số dữ liệu k-Means sẽ phân cụm dữ liệu đó thành k nhóm trong đó k là số nguyên dương. k-Nearest Neighbor là một thuật toán phân loại có giám sát, lưu ý - một thuật toán được giám sát sử dụng dữ liệu đào tạo trong khi thuật toán không giám sát không có dữ liệu đào tạo. Chúng tôi sử dụng KNN để dự đoán nhóm điểm dữ liệu thuộc về dự đoán xem một người có mắc bệnh tiểu đường hay không dựa trên một loạt các chỉ số sức khỏe khác.

Trong một tình huống mà bạn cần phân loại dữ liệu, thuật toán k-Nearest Neighbor là một khả năng, có những thuật toán khác, ví dụ:

* Naive Bayes
* Cây quyết định
* Stochastic Gradient Descent

**Cách hoạt động của KNN:**

Dữ liệu được chia thành hai nhóm, ba điểm dữ liệu màu xanh lam ở dưới cùng bên trái (nhóm 1) và ba điểm dữ liệu màu xanh lam ở trên cùng bên phải (nhóm 2). Nếu chúng ta thêm một điểm mới, chấm đỏ, nó thuộc nhóm 1 hay nhóm 2?

Chart, scatter chart

Description automatically generated

Nếu chúng ta sử dụng KNN để giải quyết vấn đề này và đặt k = 5, thuật toán sẽ tìm kiếm năm điểm dữ liệu gần nhất với chấm đỏ. Hãy tưởng tượng bạn vẽ một vòng tròn với chấm đỏ ở tâm, bán kính của hình tròn này là khoảng cách đến chấm xanh gần thứ năm. Vòng tròn sẽ bao gồm 3 điểm ở nhóm 1 và 2 điểm ở nhóm 2. Thuật toán đếm 'phiếu bầu' và quyết định rằng điểm dữ liệu màu đỏ thuộc nhóm 1. Có một số cách khác nhau để tính khoảng cách, ví dụ:

* [Khoảng cách Manhattan](https://en.wikipedia.org/wiki/Taxicab_geometry)
* [Khoảng cách Minkowski](https://en.wikipedia.org/wiki/Minkowski_distance)
* [Khoảng cách Euclide](https://en.wikipedia.org/wiki/Euclidean_distance#:~:text=In%20mathematics%2C%20the%20Euclidean%20distance,being%20called%20the%20Pythagorean%20distance.)

Có những thước đo khoảng cách khác như Khoảng cách Cosine và Khoảng cách Jaccard nhưng chúng được sử dụng phổ biến hơn nói rằng NLP để đo sự giống hoặc khác nhau giữa hai phần văn bản.

**Bài toán phân loại là gì?**

Một vấn đề khoa học dữ liệu phổ biến đòi hỏi phải phân loại hoặc dán nhãn dữ liệu, chẳng hạn như trong trường học hoặc trường cao đẳng, chúng tôi có thể xác định những sinh viên có nguy cơ trượt cao hơn hoặc cung cấp một số dữ liệu bệnh nhân, chúng tôi có thể xác định những bệnh nhân có khả năng phát triển bệnh tiểu đường hay không. Trong những trường hợp này, chúng tôi gặp vấn đề về phân loại nhị phân. Trong trường hợp học sinh, lớp A là những học sinh có nguy cơ trượt và lớp B gồm những học sinh còn lại. Đôi khi các bài toán phân loại có thể có nhiều hơn hai lớp.

**Ứng dụng thực tế của KNN**

Bài báo nghiên cứu [1] được liệt kê dưới đây trong phần Tài liệu tham khảo là một ví dụ về việc sử dụng KNN để xác định học sinh không đạt. Trong trường hợp này, các sinh viên đã đăng ký một khóa học đánh máy bằng cảm ứng trực tuyến. Đối với dữ liệu này, các nhà nghiên cứu nhận thấy rằng thuật toán KNN khái quát tốt và mang lại kết quả tốt.  Sự tồn tại của các biến thể KNN này một phần là do những thiếu sót của thuật toán. Tất cả các thuật toán phân loại đều có ưu và nhược điểm. Trong trường hợp của KNN, các khuyết điểm bao gồm:

* Chi phí tính toán cao,
* Yêu cầu bộ nhớ lớn,
* Quyết định một giá trị thích hợp của tham số đầu vào k
  1. **Giới thiệu đề tài**

Tập dữ liệu này bao gồm dữ liệu của 768 tình nguyện viên bao gồm những người bị tiểu đường và những người không bị tiểu đường. Tập dữ liệu này bao gồm các thuộc tính như sau:

1. Số lần mang thai
2. Nồng độ đường huyết trong 2 giờ
3. Huyết áp âm trương (mm Hg)
4. Độ dày nếp gấp da cơ tam đầu (mm)
5. Isulin huyết thanh trong 2 giờ (mu U/ml)
6. Khối lượng cơ thể (kg/(height in m)^2)
7. Chức năng phả hệ tiểu đường
8. Tuổi (năm)

Với mỗi tình nguyện viện, dữ liệu bao gồm tập hợp các chỉ số kể trên và tình trạng bị bênh *tức class 1* hay không bị bệnh *tức class 0*.

Về bản chất đây là một bài toán phân loại 2 lớp và chúng ta có thể sử dụng các phương pháp phân loại khác như *SVM, Random Forest, Naïve Bayes...* để phân loại cũng cho kết quả khá tốt.

Table

Description automatically generated

Nếu chúng ta xem xét các giá trị tối thiểu cho các tính năng khác nhau, chúng có ý nghĩa không? Trong một số tính năng, câu trả lời là không, ví dụ huyết áp bằng 0 có vẻ không hợp lý, BMI = 0 cũng không có ý nghĩa gì. Vì vậy, rõ ràng đã có một số vấn đề với việc thu thập dữ liệu và nhập dữ liệu khi biên dịch tập dữ liệu này. Bước tiếp theo là xác định có bao nhiêu hàng có dữ liệu sự cố. Chúng tôi sẽ xác định dữ liệu sự cố là BloodPressure = 0 hoặc BMI = 0 hoặc SkinThickness = 0

Table

Description automatically generated

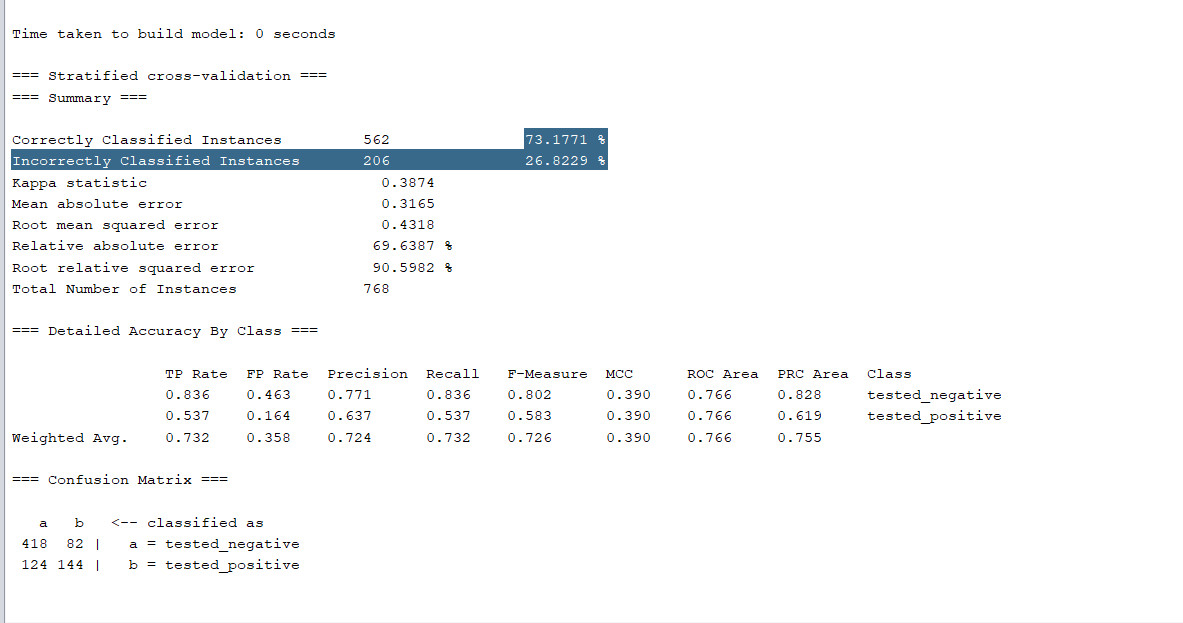
Chúng tôi có 231 hàng sự cố, đó là 30% số hàng của tập dữ liệu gốc

Bây giờ chúng ta đã xác định được những vấn đề cần thiết để quyết định phải làm gì với chúng. Có những khả năng khác nhau:

* bỏ qua vấn đề
* loại bỏ tất cả các hàng vấn đề
* bằng cách nào đó sửa chữa dữ liệu trong các hàng sự cố.

Bỏ qua các vấn đề, đặc biệt là khi chúng ảnh hưởng đến gần một phần ba số hàng có thể là lựa chọn tồi tệ nhất. Trong số các tùy chọn còn lại: loại bỏ tất cả các hàng vấn đề sẽ loại bỏ các vấn đề nhưng cũng làm giảm đáng kể lượng dữ liệu đào tạo có sẵn. Tùy chọn thứ ba yêu cầu thay thế các số không bằng các giá trị khác 0, ví dụ: sử dụng giá trị trung bình cho một cột.

Hình ảnh sau khi chạy dữ liệu trong weka:



ảnh Demo:

Diagram

Description automatically generated with medium confidence

**3. Kết luận**

Vấn đề của chúng ta rõ ràng vẫn là thống kê được những mối liện hệ và tìm được những xác suất của nó với vấn đề mà chúng ta quan tâm (ở đây chính là việc có bị tiểu đường hay không). Tuy độ chính xác còn chưa cao do bản chất cuả phương pháp cũng như tập dữ liệu chưa đủ lớn tuy nhiên nó cũng giúp bạn hình dung được cách cài đặt thuật toán *KNN*

Tỷ lệ chính xác của bài toán là 73.1771 %

Tỷ lệ ko chính xác của bài toán là 26.8229 %