

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO ĐỒ ÁN**

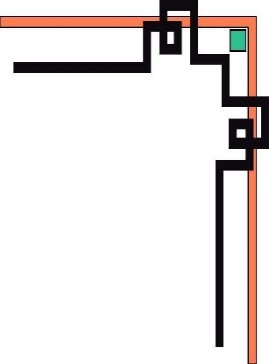
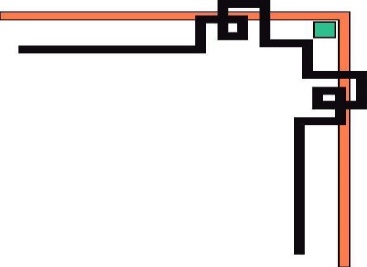
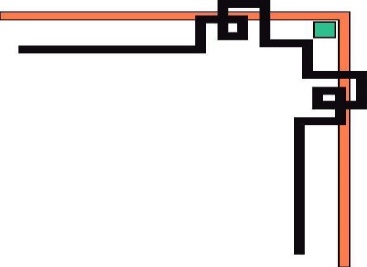
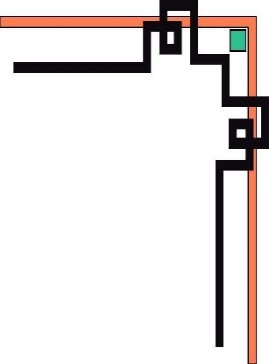


**MÔN HỌC**

**KHAI THÁC DỮ LIỆU**

**ĐỀ TÀI**

**PHÂN LOẠI HOA BẰNG THUẬT TOÁN KNN**



**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO ĐỒ ÁN**



**MÔN HỌC**

**KHAI THÁC DỮ LIỆU**

**GVHD : NGÔ DƯƠNG HÀ**

**Nhóm : 9**

|  |  |
| --- | --- |
| **Họ Và Tên** | **MSSV** |
| **1. Lê Xuân Huy** | **2001190568** |
| **2. Nguyễn Hữu Đông Khanh** | **2001190600** |

LỜI CẢM ƠN

Lời đầu tiên cho phép chúng em gửi lời cảm ơn tới toàn thể các thầy cô giáo trong Khoa Công Nghệ Thông Tin - Trường Đại học Công Nghiệp Thực Phẩm TP.HCM, những người đã hết mình truyền đạt và chỉ dẫn cho chúng em những kiến thức, những bài học quý báu và bổ ích trong suốt thời gian học vừa qua.

Để hoàn thành được đồ án này, đặt biệt chúng em xin được bày tỏ sự tri ân và xin chân thành cảm ơn giảng viên **Ngô Dương Hà** người trực tiếp hướng dẫn, chỉ bảo chúng em trong suốt quá trình học tập và Thực Hành để hoàn thành đồ án này.

Trong quá trình Thực Hành và làm báo cáo do năng lực, kiến thức, trình độ bản thân còn hạn hẹp nên không tránh khỏi những thiếu sót. Chúng em mong nhận được sự thông cảm và những ý kiến đóng góp của quý thầy cô và các bạn

*Chúng em xin chân thành cảm ơn !*

**Mục Lục**

[**CHƯƠNG I : GIỚI THIỆU TỔNG QUAN** 5](#_Toc89119171)

[**I.** **Giới Thiệu** 5](#_Toc89119172)

[**II.** **Mục đích nghiên cứu** 5](#_Toc89119173)

[**III.** **Phạm vi nghiên cứu và đối tượng** 5](#_Toc89119174)

[**IV.** **Nội dung thực hiện** 5](#_Toc89119175)

[**CHƯƠNG II : CƠ SỠ LÝ THUYẾT** 6](#_Toc89119176)

[**I.** **Thuật Toán KNN** 6](#_Toc89119177)

[**1.** **Định nghĩa** 6](#_Toc89119178)

[**2.** **Quy trình làm việc của thuật toán KNN** 6](#_Toc89119179)

[**3.** **Ví dụ về KNN** 7](#_Toc89119180)

[**II.** **Ưu điểm và nhược điểm của thuật toán** 9](#_Toc89119181)

[**CHƯƠNG II : ỨNG DỤNG KNN** 10](#_Toc89119182)

[**I.** **Bộ dữ liệu Iris flower dataset** 10](#_Toc89119183)

[**1.** **Giới thiệu** 10](#_Toc89119184)

[**2.** **Sử dụng tập dữ liệu** 10](#_Toc89119185)

[**3.** **Tập dữ liệu** 11](#_Toc89119186)

[**4.** **Demo ứng dụng** 15](#_Toc89119187)

[**CHƯƠNG III : KẾT LUẬN** 20](#_Toc89119188)

# **CHƯƠNG I : GIỚI THIỆU TỔNG QUAN**

## **Giới Thiệu**

Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ Nhân Tạo), và cụ thể hơn là Machine Learning (Học Máy hoặc Máy Học) nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư (1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin). Trí Tuệ Nhân Tạo đang len lỏi vào mọi lĩnh vực trong đời sống mà có thể chúng ta không nhận ra. Xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix, máy chơi cờ vây AlphaGo của Google DeepMind, …, chỉ là một vài trong vô vàn những ứng dụng của AI/Machine Learning.

Machine learning gây nên cơn sốt công nghệ trên toàn thế giới trong vài năm nay. Trong giới học thuật, mỗi năm có hàng ngàn bài báo khoa học về đề tài này. Trong giới công nghiệp, từ các công ty lớn như Google, Facebook, Microsoft đến các công ty khởi nghiệp đều đầu tư vào machine learning. Hàng loạt các ứng dụng sử dụng machine learning ra đời trên mọi linh vực của cuộc sống, từ khoa học máy tính đến những ngành ít liên quan hơn như vật lý, hóa học, y học, chính trị.

1. **Mục đích nghiên cứu**

* Nghiên cứu, tìm hiểu thuật toán KNN.
* Đánh giá hiệu quả của thuật toán.
* Áp dụng học tập và thực hành tạo ra một ứng dụng mô phỏng thuật toán KNN.
* Giúp đỡ người dùng có thể hiểu rõ hơn về cách thực hiện các thuật toán KNN.
* Xây dựng phần mềm với giao diện đơn giản, dễ dàng sử dụng chương trình
* Phát triển ứng dụng với có thêm nhiều chức năng hữu ích

1. **Phạm vi nghiên cứu và đối tượng**

* Phạm vi nghiên cứu: Thử nghiệm trên Lris flower dataset
* Đối tượng nghiên cứu:Thuật Toán KNN và bộ Lris flower dataset

1. **Nội dung thực hiện**

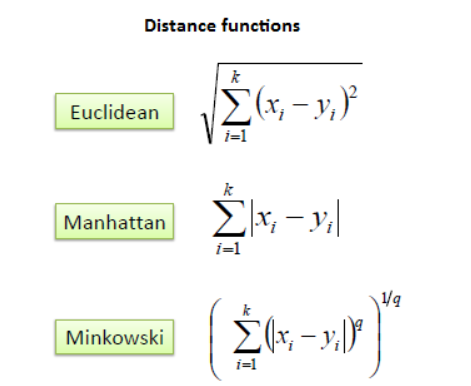
* Tìm hiểu thuật toán KNN
* Sủ dụng ngôn ngữ Python để thực hiện
* Làm quen với bộ dữ liệu lris
* Sử dụng bộ dữ liệu vào thí nghiệm và đánh giá

**CHƯƠNG II : CƠ SỠ LÝ THUYẾT**

1. **Thuật Toán KNN**
2. **Định nghĩa**

Thuật toán K láng giềng gần nhất (tiếng Anh: K-Nearest Neighbor - KNN) là một kĩ thuật học có giám sát (supervised learning) dùng để phân loại quan sát mới bằng cách tìm điểm tương đồng giữa quan sát mới này với dữ liệu sẵn có

Ý tưởng của thuật toán: Thuật toán KNN cho rằng những dữ liệu tương tự nhau sẽ tồn tại **gần nhau** trong một không gian, từ đó công việc của chúng ta là sẽ tìm k điểm gần với dữ liệu cần kiểm tra nhất. Việc tìm khoảng cách giữa 2 điểm củng có nhiều công thức có thể sử dụng, tùy trường hợp mà chúng ta lựa chọn cho phù hợp. Đây là 3 cách cơ bản để tính khoảng cách 2 điểm dữ liệu x, y có k thuộc tính:



1. **Quy trình làm việc của thuật toán KNN**

* Xác định giá trị tham số K (số láng giềng gần nhất)
* Tính khoảng cách giữa đối tượng cần phân lớp(Querty Point) với tất cả các đối tượng trong các mẫu có sẵn (Trainning Data) sử dụng khoảng cách Euclidean)
* Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với Query Point
* Lấy tất cả các lớp của K láng giềng gần nhất đã xác định
* Dựa vào phần lớp của K láng giềng gần nhất để xác định Query Point

**Khoảng cách Euclidean**

và p=(p1,….,pn)

d(p,q)=d(q,p)

=

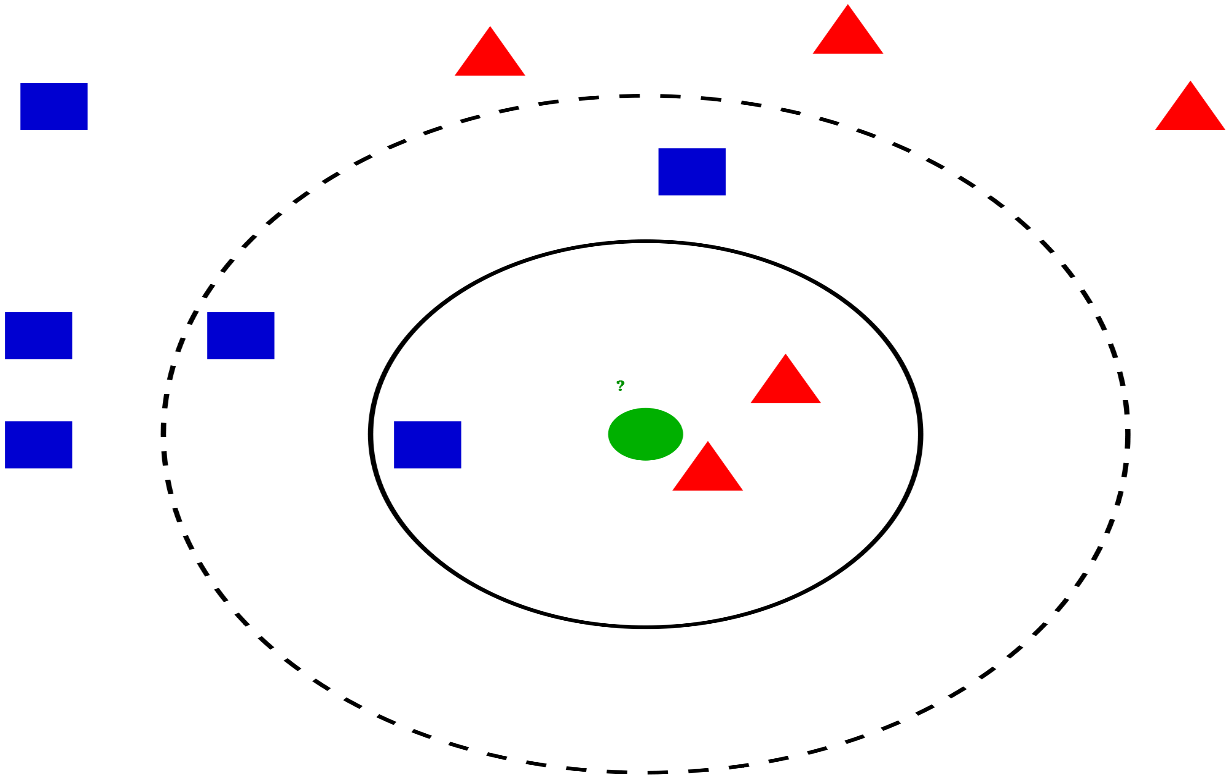
|  |  |  |  |
| --- | --- | --- | --- |
| **phimm** | Điểm tình cảm | Điểm hành động | Khoảng cách |
| **Spider-Man 2** | 8 | 3 | 7.61 |
| **Beauty and the Beast** | 9 | 4 | 7.21 |
| **Harry Potter and the Chamber of Secrets** | 10 | 2 | 9.43 |
| **Deadpool** | s1 | 9 | 4.12 |
| **Mortal Kombat** | 3 | 8 | 2.82 |
| **The Conjuring** | 3 | 9 | 2.23 |
| **The Intouchables** | 5 | 10 |  |

Ví dụ Khoảng cách Euclidean

d(**Spider-Man 2**, **The Intouchables**)=

=7.61

1. **Ví dụ về KNN**

****

K=5

K=3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phim** | **Điểm tình cảm** | **Điểm hành động** | **Khoảng Cách** | **Thể Loại** |
| **Spider-Man 2** | 8 | 3 | 7.61 | Tình Cảm |
| **Beauty and the Beast** | 9 | 4 | 7.21 | Tình Cảm |
| **Harry Potter and the Chamber of Secrets** | 10 | 2 | 9.43 | Tình Cảm |
| **Deadpool** | 1 | 9 | 4.12 | Hành Động |
| **Mortal Kombat** | 3 | 8 | 2.82 | Hành Động |
| **The Conjuring** | 3 | 9 | 2.23 | Hành Động |
| **The Intouchables** | 5 | 10 |  | Chưa xác định |

Chọn k=3 thì láng giềng gần nhất là 3 phim: **The Conjuring** , **Mortal Kombat**, **Deadpool** Thuật toán K-NN sẽ chọn loại phim nào chiếm ưu thế nhiều nhất trong các láng giềng gần nhất để làm loại phim cần xác định thể loại. Vì 3 phim trên đều là thể loại Hành Động ==> Phim cần xác định thuộc ***thể loại phim Hành Động***.

1. **Ưu điểm và nhược điểm của thuật toán**

* **Ưu điểm:**
* Độ phức tạp tính toán của quá trình training là bằng 0.
* Việc dự đoán kết quả của dữ liệu mới rất đơn giản.
* Không cần giả sử gì về phân phối của các class.
* **Nhược Điểm:**
* KNN rất nhạy cảm với nhiễu khi K nhỏ.
* Như đã nói, KNN là một thuật toán mà mọi tính toán đều nằm ở khâu test. Trong đó việc tính khoảng cách tới *từng* điểm dữ liệu trong training set sẽ tốn rất nhiều thời gian, đặc biệt là với các cơ sở dữ liệu có số chiều lớn và có nhiều điểm dữ liệu. Với K càng lớn thì độ phức tạp cũng sẽ tăng lên. Ngoài ra, việc lưu toàn bộ dữ liệu trong bộ nhớ cũng ảnh hưởng tới hiệu năng của KNN.

**CHƯƠNG II : ỨNG DỤNG KNN**

1. **Bộ dữ liệu Iris flower dataset**
2. **Giới thiệu**

Tập dữ liệu hoa Iris hoặc tập dữ liệu Iris của Fisher là tập dữ liệu đa biến được giới thiệu bởi nhà thống kê và nhà sinh vật học người Anh Ronald Fisher trong bài báo năm 1936 Việc sử dụng nhiều phép đo trong các vấn đề phân loại như một ví dụ về phân tích phân biệt tuyến tính. Đôi khi nó được gọi là tập dữ liệu Iris của Anderson vì Edgar Anderson đã thu thập dữ liệu để định lượng sự biến đổi hình thái của hoa Iris của ba loài liên quan. Hai trong số ba loài được thu thập ở Bán đảo Gaspé "tất cả từ cùng một đồng cỏ, và được chọn vào cùng một ngày và được đo cùng lúc bởi cùng một người với cùng một bộ máy".

Bộ dữ liệu bao gồm 50 mẫu từ mỗi ba loài Iris (Iris setosa, Iris virginica và Iris Verscolor). Bốn đặc điểm được đo từ mỗi mẫu: chiều dài và chiều rộng của đài hoa, chiều dài và chiều rộng cánh hoa, tính bằng centimet. Dựa trên sự kết hợp của bốn tính năng này, Fisher đã phát triển một mô hình phân biệt tuyến tính để phân biệt các loài với nhau

1. **Sử dụng tập dữ liệu**

Bộ dữ liệu này đã trở thành trường hợp thử nghiệm điển hình cho nhiều kỹ thuật phân loại thống kê trong học máy như máy vector hỗ trợ. Tuy nhiên, việc sử dụng tập dữ liệu này trong phân tích cụm không phổ biến, vì tập dữ liệu chỉ chứa hai cụm có sự phân tách khá rõ ràng. Một trong những cụm chứa Iris setosa, trong khi cụm còn lại chứa cả Iris virginica và Iris Versolor và không thể tách rời nếu không có thông tin về loài mà Fisher sử dụng. Điều này làm cho dữ liệu trở thành một ví dụ tốt để giải thích sự khác biệt giữa các kỹ thuật được giám sát và không giám sát trong khai thác dữ liệu: Mô hình phân biệt tuyến tính của Fisher chỉ có thể thu được khi biết các loài đối tượng: nhãn lớp và cụm không nhất thiết giống nhau. Tuy nhiên, cả ba loài Iris đều có thể tách rời trong hình chiếu trên thành phần chính phân nhánh phi tuyến. Tập dữ liệu được xấp xỉ bởi cây gần nhất với một số hình phạt cho số lượng nút, uốn cong và kéo dài quá mức. Các điểm dữ liệu được chiếu vào nút gần nhất. Đối với mỗi nút, sơ đồ hình tròn của các điểm được chiếu được chuẩn bị. Diện tích của chiếc bánh tỷ lệ thuận với số lượng điểm được chiếu. Rõ ràng từ sơ đồ (bên dưới) rằng phần lớn tuyệt đối các mẫu của các loài Iris khác nhau thuộc về các nút khác nhau. Chỉ một phần nhỏ Iris-virginica được trộn với Iris- Versolor (các nút màu xanh lam hỗn hợp trong sơ đồ). Do đó, ba loài Iris (Iris setosa, Iris virginica và Iris Verscolor) có thể được phân tách bằng các thủ tục không giám sát trong phân tích thành phần chính phi tuyến. Để phân biệt chúng, chỉ cần chọn các nút tương ứng trên cây chính.

1. **Tập dữ liệu**

Tập dữ liệu Bộ dữ liệu chứa một bộ 150 bản ghi bao gồm các thuộc tính – chiều dài và chiều rộng của đài hoa, chiều dài và chiều rộng của cánh hoa.

 Bảng 1: Thông tin loài Setosa

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Thứ tự | Chiều dài đài hoa | Chiều rộng đài hoa | Chiều dài cánh hoa | Chiều rộng cánh hoa |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3 | 1.4 | 0.2 |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 |
| 5 | 5 | 3.6 | 1.4 | 0.2 |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 |
| 8 | 5 | 3.4 | 1.5 | 0.2 |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 |
| 13 | 4.8 | 3 | 1.4 | 0.1 |
| 14 | 4.3 | 3 | 1.1 | 0.1 |
| 15 | 5.8 | 4 | 1.2 | 0.2 |
| 16 | 5.7 | 4.4 | 1.5 | 0.4 |
| 17 | 5.4 | 3.9 | 1.3 | 0.4 |
| 18 | 5.1 | 3.5 | 1.4 | 0.3 |
| 19 | 5.7 | 3.8 | 1.7 | 0.3 |
| 20 | 5.1 | 3.8 | 1.5 | 0.3 |
| 21 | 5.4 | 3.4 | 1.7 | 0.2 |
| 22 | 5.1 | 3.7 | 1.5 | 0.4 |
| 23 | 4.6 | 3.6 | 1 | 0.2 |
| 24 | 5.1 | 3.3 | 1.7 | 0.5 |
| 25 | 4.8 | 3.4 | 1.9 | 0.2 |
| 26 | 5 | 3 | 1.6 | 0.2 |
| 27 | 5 | 3.4 | 1.6 | 0.4 |
| 28 | 5.2 | 3.5 | 1.5 | 0.2 |
| 29 | 5.2 | 3.4 | 1.4 | 0.2 |
| 30 | 4.7 | 3.2 | 1.6 | 0.2 |
| 31 | 4.8 | 3.1 | 1.6 | 0.2 |
| 32 | 5.4 | 3.4 | 1.5 | 0.4 |
| 33 | 5.2 | 4.1 | 1.5 | 0.1 |
| 34 | 5.5 | 4.2 | 1.4 | 0.2 |
| 35 | 4.9 | 3.1 | 1.5 | 0.1 |
| 36 | 5 | 3.2 | 1.2 | 0.2 |
| 37 | 5.5 | 3.5 | 1.3 | 0.2 |
| 38 | 4.9 | 3.1 | 1.5 | 0.1 |
| 39 | 4.4 | 3 | 1.3 | 0.2 |
| 40 | 5.1 | 3.4 | 1.5 | 0.2 |
| 41 | 5 | 3.5 | 1.3 | 0.3 |
| 42 | 4.5 | 2.3 | 1.3 | 0.3 |
| 43 | 4.4 | 3.2 | 1.3 | 0.2 |
| 44 | 5 | 3.5 | 1.6 | 0.6 |
| 45 | 5.1 | 3.8 | 1.9 | 0.4 |
| 46 | 4.8 | 3 | 1.4 | 0.3 |
| 47 | 5.1 | 3.8 | 1.6 | 0.2 |
| 48 | 4.6 | 3.2 | 1.4 | 0.2 |
| 49 | 5.3 | 3.7 | 1.5 | 0.2 |
| 50 | 5 | 3.3 | 1.4 | 0.2 |

Bảng 2: Thông tin loài Versicolor

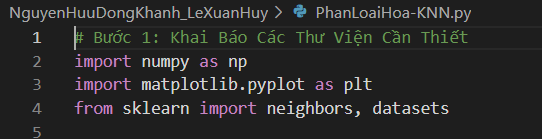
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Thứ tự | Chiều dài đài hoa | Chiều rộng đài hoa | Chiều dài cánh hoa | Chiều rộng cánh hoa |
| 51 | 7 | 3.2 | 4.7 | 1.4 |
| 52 | 6.4 | 3.2 | 4.5 | 1.5 |
| 53 | 6.9 | 3.1 | 4.9 | 1.5 |
| 54 | 5.5 | 2.3 | 4 | 1.3 |
| 55 | 6.5 | 2.8 | 4.6 | 1.5 |
| 56 | 5.7 | 2.8 | 4.5 | 1.3 |
| 57 | 6.3 | 3.3 | 4.7 | 1.6 |
| 58 | 4.9 | 2.4 | 3.3 | 1 |
| 59 | 6.6 | 2.9 | 4.6 | 1.3 |
| 60 | 5.2 | 2.7 | 3.9 | 1.4 |
| 61 | 5 | 2 | 3.5 | 1 |
| 62 | 5.9 | 3 | 4.2 | 1.5 |
| 63 | 6 | 2.2 | 4 | 1 |
| 64 | 6.1 | 2.9 | 4.7 | 1.4 |
| 65 | 5.6 | 2.9 | 3.6 | 1.3 |
| 66 | 6.7 | 3.1 | 4.4 | 1.4 |
| 67 | 5.6 | 3 | 4.5 | 1.5 |
| 68 | 5.8 | 2.7 | 4.1 | 1 |
| 69 | 6.2 | 2.2 | 4.5 | 1.5 |
| 70 | 5.6 | 2.5 | 3.9 | 1.1 |
| 71 | 5.9 | 3.2 | 4.8 | 1.8 |
| 72 | 6.1 | 2.8 | 4 | 1.3 |
| 73 | 6.3 | 2.5 | 4.9 | 1.5 |
| 74 | 6.1 | 2.8 | 4.7 | 1.2 |
| 75 | 6.4 | 2.9 | 4.3 | 1.3 |
| 76 | 6.6 | 3 | 4.4 | 1.4 |
| 77 | 6.8 | 2.8 | 4.8 | 1.4 |
| 78 | 6.7 | 3 | 5 | 1.7 |
| 79 | 6 | 2.9 | 4.5 | 1.5 |
| 80 | 5.7 | 2.6 | 3.5 | 1 |
| 81 | 5.5 | 2.4 | 3.8 | 1.1 |
| 82 | 5.5 | 2.4 | 3.7 | 1 |
| 83 | 5.8 | 2.7 | 3.9 | 1.2 |
| 84 | 6 | 2.7 | 5.1 | 1.6 |
| 85 | 5.4 | 3 | 4.5 | 1.5 |
| 86 | 6 | 3.4 | 4.5 | 1.6 |
| 87 | 6.7 | 3.1 | 4.7 | 1.5 |
| 88 | 6.3 | 2.3 | 4.4 | 1.3 |
| 89 | 5.6 | 3 | 4.1 | 1.3 |
| 90 | 5.5 | 2.5 | 4 | 1.3 |
| 91 | 5.5 | 2.6 | 4.4 | 1.2 |
| 92 | 6.1 | 3 | 4.6 | 1.4 |
| 93 | 5.8 | 2.6 | 4 | 1.2 |
| 94 | 5 | 2.3 | 3.3 | 1 |
| 95 | 5.6 | 2.7 | 4.2 | 1.3 |
| 96 | 5.7 | 3 | 4.2 | 1.2 |
| 97 | 5.7 | 2.9 | 4.2 | 1.3 |
| 98 | 6.2 | 2.9 | 4.3 | 1.3 |
| 99 | 5.1 | 2.5 | 3 | 1.1 |
| 100 | 5.7 | 2.8 | 4.1 | 1.3 |

Bảng 3: Thông tin loài Virginica

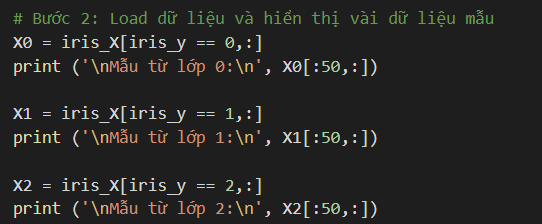
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Thứ tự | Chiều dài đài hoa | Chiều rộng đài hoa | Chiều dài cánh hoa | Chiều rộng cánh hoa |
| 101 | 6.3 | 3.3 | 6 | 2.5 |
| 102 | 5.8 | 2.7 | 5.1 | 1.9 |
| 103 | 7.1 | 3 | 5.9 | 2.1 |
| 104 | 6.3 | 2.9 | 5.6 | 1.8 |
| 105 | 6.5 | 3 | 5.8 | 2.2 |
| 106 | 7.6 | 3 | 6.6 | 2.1 |
| 107 | 4.9 | 2.5 | 4.5 | 1.7 |
| 108 | 7.3 | 2.9 | 6.3 | 1.8 |
| 109 | 6.7 | 2.5 | 5.8 | 1.8 |
| 110 | 7.2 | 3.6 | 6.1 | 2.5 |
| 111 | 6.5 | 3.2 | 5.1 | 2 |
| 112 | 6.4 | 2.7 | 5.3 | 1.9 |
| 113 | 6.8 | 3 | 5.5 | 2.1 |
| 114 | 5.7 | 2.5 | 5 | 2 |
| 115 | 5.8 | 2.8 | 5.1 | 2.4 |
| 116 | 6.4 | 3.2 | 5.3 | 2.3 |
| 117 | 6.5 | 3 | 5.5 | 1.8 |
| 118 | 7.7 | 3.8 | 6.7 | 2.2 |
| 119 | 7.7 | 2.6 | 6.9 | 2.3 |
| 120 | 6 | 2.2 | 5 | 1.5 |
| 121 | 6.9 | 3.2 | 5.7 | 2.3 |
| 122 | 5.6 | 2.8 | 4.9 | 2 |
| 123 | 7.7 | 2.8 | 6.7 | 2 |
| 124 | 6.3 | 2.7 | 4.9 | 1.8 |
| 125 | 6.7 | 3.3 | 5.7 | 2.1 |
| 126 | 7.2 | 3.2 | 6 | 1.8 |
| 127 | 6.2 | 2.8 | 4.8 | 1.8 |
| 128 | 6.1 | 3 | 4.9 | 1.8 |
| 129 | 6.4 | 2.8 | 5.6 | 2.1 |
| 130 | 7.2 | 3 | 5.8 | 1.6 |
| 131 | 7.4 | 2.8 | 6.1 | 1.9 |
| 132 | 7.9 | 3.8 | 6.4 | 2 |
| 133 | 6.4 | 2.8 | 5.6 | 2.2 |
| 134 | 6.3 | 2.8 | 5.1 | 1.5 |
| 135 | 6.1 | 2.6 | 5.6 | 1.4 |
| 136 | 7.7 | 3 | 6.1 | 2.3 |
| 137 | 6.3 | 3.4 | 5.6 | 2.4 |
| 138 | 6.4 | 3.1 | 5.5 | 1.8 |
| 139 | 6 | 3 | 4.8 | 1.8 |
| 140 | 6.9 | 3.1 | 5.4 | 2.1 |
| 141 | 6.7 | 3.1 | 5.6 | 2.4 |
| 142 | 6.9 | 3.1 | 5.1 | 2.3 |
| 143 | 5.8 | 2.7 | 5.1 | 1.9 |
| 144 | 6.8 | 3.2 | 5.9 | 2.3 |
| 145 | 6.7 | 3.3 | 5.7 | 2.5 |
| 146 | 6.7 | 3 | 5.2 | 2.3 |
| 147 | 6.3 | 2.5 | 5 | 1.9 |
| 148 | 6.5 | 3 | 5.2 | 2 |
| 149 | 6.2 | 3.4 | 5.4 | 2.3 |
| 150 | 5.9 | 3 | 5.1 | 1.8 |

1. **Demo ứng dụng**

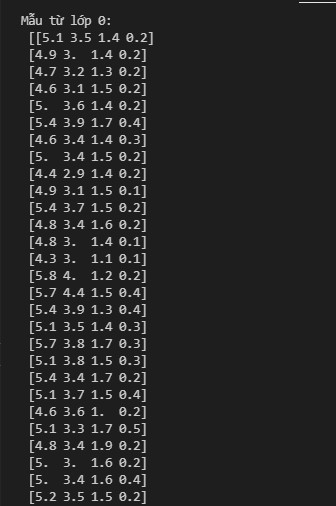
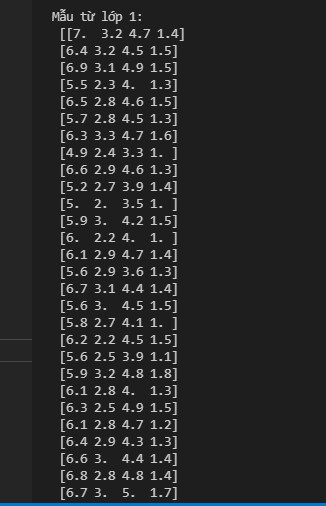
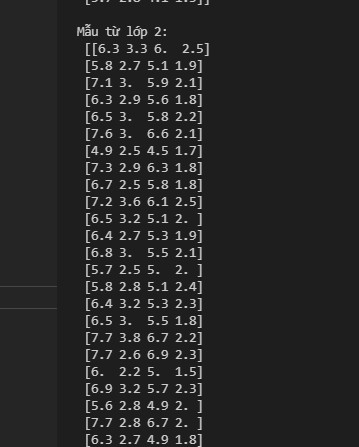
**Bước 1: Khai báo các thư viện cần thiết**



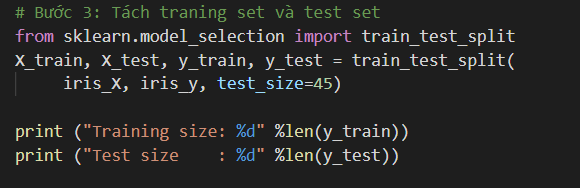
**Bước 2: Load dữ liệu và hiển thị vài dữ liệu mẫu**



**Kết quả:**

**Bước 3: Tách traning set và test set**



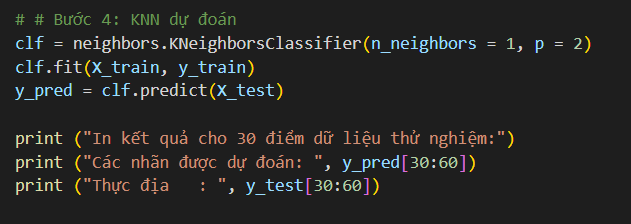
Giả sử chúng ta muốn dùng 45 điểm dữ liệu cho test set và 105 (theo tỉ lệ 70:30) điểm còn lại traning set. Thư viện sklearn có một hàm số cho phép chúng ta ngẫu nhiên lựa chọn:

**Kết quả:**

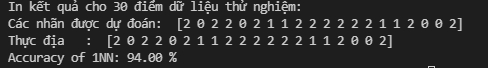


**Bước 4:KNN dự đoán**

Xét trường hợp K=1, tức là với thời điểm dữ liệu test ta chỉ xét1 điểm dữ liệu traning gần nhất và lấy nhãn cã điểm đó để dự đoán cho điểm dữ liệu test

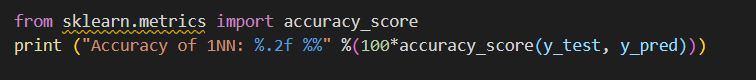


Kết quả:



Kết quả cho thấy nhãn dự đoán giống với label thật của dữ liệu test

**Bước 5: Đánh Giá**

Để đánh giá độ chính xác của thuật toán KNN, chúng ta xem có bao nhiêu điểm trong dữ liệu test được dự đoán đúng. lấy số lượng đó chia cho tổng số lượng trong tập dữ liệu test sẽ ra độ chính xác. Thư viện sklearn cung cấp cho ta hàm số accuracy\_score để thực hiện công việc này

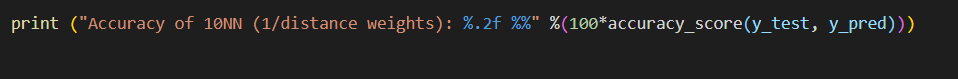
Kết quả:



Với k=1 đã cho ra kết quả là 94%. Nhận thấy rằng nếu chỉ xét 1 điểm gần nhất có thể dẫn đến kết quả sai nếu điểm đó là nhiễu. Vậy nên ta thử xét với k=9 thì kết quả



Kết quả đã tăng lên 98%. Như vậy chúng ta đã giải quyết được bài toán đặt ra với kết quả thu được gần như hoàn hảo. Trong kỹ thuật majoring voting trên, mỗi trong 9 điểm gần nhất được xem có vai trò và giá trị phiếu như nhau. Xét trên thực tế điều này là không đúng, vì rõ ràng những điểm gần hơn sẽ phải có trọng số cao hơn. Vậy nên ta sẽ đánh trọng số khác nhau cho 9 điểm đang xét nhưng phải thỏa mãn điều kiện điểm gần test data thì phải có trọng số càng cao. Với giá trị mặc định weights=” uniform” tương ứng với các điểm lân cận có giá trị như nhau, ta gán giá trị của weights=” distance” lúc đó các điểm gần với test data sẽ có trọng số cao hơn



**CHƯƠNG III : KẾT LUẬN**

* **Kết luận**

Qua bài báo cáo, thuật toán cho ta thấy được cách làm việc của k-NN (K Láng giềng gần nhất). KNN là một mô hình đơn giản và trực quan nhưng vẫn có hiệu quả cao vì nó không tham số; mô hình không đưa ra giả định nào về việc phân phối dữ liệu. Hơn nữa, nó có thể được sử dụng trực tiếp để phân loại đa lớp.

Thuật toán KNN có nhiều ứng dụng nhiều trong ngành đầu tư, bao gồm dự đoán phá sản, dự đoán giá cổ phiếu, phân bổ xếp hạng tín dụng trái phiếu doanh nghiệp, tạo ra chỉ số vốn và trái phiếu tùy chỉnh

* **Hướng phát triển**

Vì thời gian có hạn và kiến thức còn hạn chế, chúng em sẽ cố gắng phát triển mô hình này tốt hơn trong tương lai và ứng dụng vào thực tế. Nhóm chúng em xin chân thành cảm ơn.

**TÀI LIỆU THAM KHẢO**

[**https://machinelearningcoban.com/2017/01/08/knn/**](https://machinelearningcoban.com/2017/01/08/knn/)

[**https://tailieu.vn/doc/do-an-tot-nghiep-nghien-cuu-thuat-toan-k-nearest-neighbor-va-su-dung-iris-flowers-dataset-danh-gia--2150667.html**](https://tailieu.vn/doc/do-an-tot-nghiep-nghien-cuu-thuat-toan-k-nearest-neighbor-va-su-dung-iris-flowers-dataset-danh-gia--2150667.html)