**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**---------------**

****

**ĐỒ ÁN TỐT NGHIỆP**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**TÊN ĐỀ TÀI**

**ỨNG DỤNG CỦA THUẬT TOÁN K-NN**

**TRONG XÂY DỰNG DỮ LIỆU ĐỒ THỊ**

**VÀ BÀI TOÁN DỰ ĐOÁN**

**CBHD: PGS.TS. Nguyễn Thị Kim Sơn**

**Sinh viên: Nguyễn Quang Ninh**

**Mã sinh viên: 2020600419**

**Hà Nội - 2024**

**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

------------------

****

**ĐỒ ÁN TỐT NGHIỆP**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**TÊN ĐỀ TÀI**

**ỨNG DỤNG CỦA THUẬT TOÁN K-NN**

**TRONG XÂY DỰNG DỮ LIỆU ĐỒ THỊ**

**VÀ BÀI TOÁN DỰ ĐOÁN**

**CBHD: PGS.TS. Nguyễn Thị Kim Sơn**

**Sinh viên: Nguyễn Quang Ninh**

**Mã sinh viên: 2020600419**

**Hà Nội - 2024**

**LỜI CẢM ƠN**

Lời đầu tiên, em xin bày tỏ lòng biết ơn sâu sắc tới các thầy cô Trường Đại học Công nghiệp Hà Nội, đặc biệt là các thầy, cô khoa Công nghệ thông tin đã tận tình giảng dạy và hướng dẫn em trong suốt quãng thời gian học tập tại trường. Sự quan tâm, giúp đỡ của các thầy, cô chính là nguồn động viên to lớn cho em trong quá trình chinh phục con đường học tập.

Đặc biệt, em xin gửi lời cảm ơn chân thành nhất tới cô Nguyễn Thị Kim Sơn, người đã giúp em giải đáp các thắc mắc, tận tình chỉ bảo em trong suốt quãng thời gian làm đồ án tốt nghiệp.

Quá trình thực hiện đồ án không chỉ là thách thức còn là cơ hội để em có thể áp dụng những kiến thức đã học vào thực tế. Em đã cố gắng hết mình để có thể hoàn thiện đồ án, song em vẫn luôn trân trọng và mong nhận được những sự đóng góp ý kiến quý báu của các thầy cô để bài làm của em được hoàn thiện hơn.

Em xin chân thành cảm ơn!

Sinh viên thực hiện

Nguyễn Quang Ninh

**MỤC LỤC**

[DANH MỤC CÁC TỪ VIẾT TẮT 3](#_Toc11889)

[DANH MỤC CÁC BẢNG 4](#_Toc2383)

[DANH MỤC CÁC HÌNH 5](#_Toc1605)

[MỞ ĐẦU 6](#_Toc3233)

[1. Giới thiệu 6](#_Toc20482)

[2. Lý do chọn đề tài 7](#_Toc22508)

[3. Tổng quan nghiên cứu 8](#_Toc2491)

[4. Mục tiêu của đề tài 9](#_Toc6300)

[5. Nội dung nghiên cứu 9](#_Toc15423)

[6. Cấu trúc đồ án 9](#_Toc26863)

[7. Kết quả dự kiến 9](#_Toc17090)

[Chương 1: CƠ SỞ LÝ THUYẾT 10](#_Toc24978)

[1.1 Khái niệm và nguyên lý hoạt động của thuật toán K-nearest neighbor 10](#_Toc23670)

[1.1.1 Khái niệm 10](#_Toc4415)

[1.1.2 Nguyên lý hoạt động 10](#_Toc16441)

[1.2 Ưu điểm, nhược điểm của thuật toán 11](#_Toc21052)

[1.3 Tăng tốc cho thuật toán 12](#_Toc27280)

[1.4 Khoảng cách trong không gian vector 13](#_Toc29933)

[Kết luận chương 1: 15](#_Toc24699)

[Chương 2: XÂY DỰNG DỮ LIỆU ĐỒ THỊ VÀ BÀI TOÁN DỰ ĐOÁN 16](#_Toc12245)

[2.1 Xây dựng dữ liệu đồ thị từ tập dữ liệu 16](#_Toc13800)

[2.2 Áp dụng thuật toán k-NN vào dữ liệu đồ thị 34](#_Toc2233)

[2.3 Bài Toán Dự Đoán Trên Dữ Liệu Đồ Thị 34](#_Toc23766)

[Kết luận chương 2: 36](#_Toc8314)

[Chương 3: THỰC NGHIỆM 37](#_Toc19879)

[3.1 Môi trường triển khái 37](#_Toc9460)

[3.2 Quy trình triển khái 40](#_Toc31811)

[3.3 Thiết lập thực nghiệm 40](#_Toc12702)

[3.4 So sánh và đánh giá 43](#_Toc9959)

[KẾT LUẬN 46](#_Toc29045)

[TÀI LIỆU THAM KHẢO 47](#_Toc4929)

[PHỤ LỤC 48](#_Toc26471)

# DANH MỤC CÁC TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| STT | Từ viết tắt | Ý nghĩa |
| 1 | AI | Artificial Intelligence |
| 2 | ML | Machine learning |
| 3 | k-NN | K- nearest neighbor |

# DANH MỤC CÁC BẢNG

|  |  |  |
| --- | --- | --- |
| STT | Tên bảng | Nội dung |
| 1 | Bảng 1 | Tập dữ liệu thu tập được |
| 2 | Bảng 2 | Tập dữ liệu đầu vào |

# DANH MỤC CÁC HÌNH

|  |  |  |
| --- | --- | --- |
| STT | Tên hình ảnh | Nội dung |
| 1 | Hình 1 | Mối liên hệ của Ai, ML và Deep Learning |
| 2 | Hình 2 | Bản đồ của 1-NN |
| 3 | Hình 3 | Biểu đồ biểu diễn điểm dữ liệu |
| 4 | Hình 4 | Mô hình bài toán |
| 5 | Hình 5 | Đồ thị dữ liệu với k = 9 |

# MỞ ĐẦU

## Giới thiệu

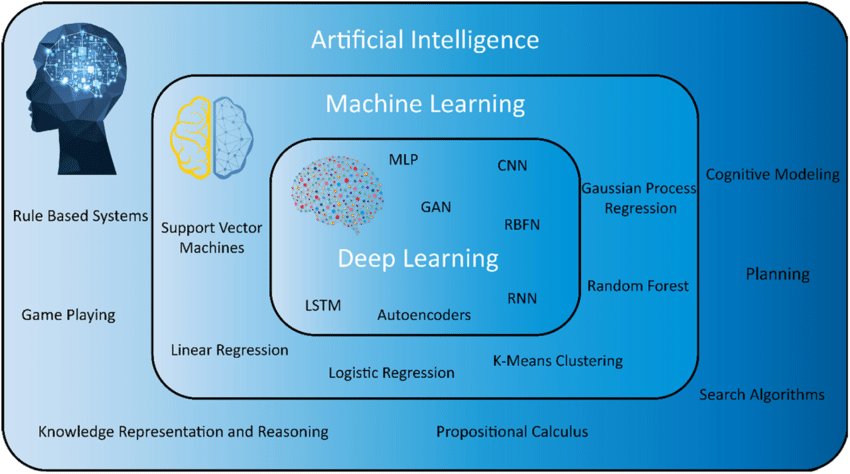
Trong những năm gần đây, AI nổi lên trở thành một bằn chứng của cuộc cách mạng công nghiệp lần thứ tư. Trí tuệ nhân tạo có thể được nhìn nhận như một ngành của khoa học máy tính liên quan đến việc tự động hóa các hành vi thông minh. Trí tuệ nhận tạo là một bộ phận của khoa học máy tính do đó nó phải được dựa trên những nguyên lý lý thuyết vững chắc, có khả năng dược ứng dụng. Ở thời điển hiện nay, ý niệm này thường được dùng như để nhắc đến các máy tính có mục đích không nhất định và ngành khoa học nghiên cứu về các lý thuyết và các ứng dụng có thể có của trí tuệ nhân tạo.

Cùng với đà phát triển như vũ bão của công nghệ, các ứng dụng của trí tuệ nhân tạo luôn là xu hướng công nghệ tương lai mà các công ty công nghệ hàng đầu trên thế giới luôn đua nhau sáng tạo phát triển, điều này khiến nó trở thành cốt lõi của cuộc cách mạng 4.0.

ML là một lĩnh vực của trí tuệ nhân tạo, dược tạo ra từ khả năng nhận diện mẫu và từ lý thuyết các máy tính có thể học mà không cần phải lập trình để xử lý các nhiệm vụ cụ thể nào.

Đại đa số các ngành công nghiệp đang làm việc với hàm lượng lớn dữ liệu đều nhận ra tầm quan trọng của công nghệ ML. Những cái nhìn sáng suốt từ nguồn dữ liệu này chủ yếu trong dạng thời gian thực sẽ giúp các tổ chức vận hành hiệu quả hơn và tạo lợi thế cạnh tranh so với các đối thủ cùng lĩnh vực.

Các ứng dụng của ML đã quá quen thuộc đối với chúng ta như: hệ thống tự tag khuôn mặt trên Facebook, hệ thống gợi ý sản phẩm của Shoppe, hệ thống gợi ý phim của IQIYI, xe tự hành của BYD và TESLA,… chỉ là 1 vài trong vô vàn những ứng dụng của trí tuệ nhận tạo.



*Hình 1: Mối liên hệ giữa AI, ML, và DL*

Xu thế tất yếu là sự phát triển công nghệ thông tin ngày càng tăng, cùng với đó là lượng dữ liệu được sinh ra ngày một lớn dần. Vì vậy nhu cầu xử lý dữ liệu trở lên cấp thiết hơn, ML đang trở thành một lựa chọn để giải quyết vấn đề này. Một trong những thuật toán thường dùng trong ML là thuật toán k-NN.

Ứng dụng của thuật toán này đang được sử dụng rất rộng rãi trong các bài toán phân lớp.

## Lý do chọn đề tài

Đề tài "Ứng dụng thuật toán k-NN trong xây dựng dữ liệu đồ thị và bài toán dự báo" đề xuất một hướng nghiên cứu hấp dẫn và tiềm năng. Thuật toán k-NN, với tính đơn giản và hiệu quả, đã thu hút sự quan tâm rộng rãi trong cộng đồng máy học. Việc áp dụng thuật toán này vào việc xây dựng dữ liệu đồ thị giúp hiểu rõ hơn về mối quan hệ giữa các đối tượng, mở ra cánh cửa cho việc tạo ra các mô hình biểu diễn dữ liệu phức tạp một cách hiệu quả.

Bên cạnh đó, trong lĩnh vực bài toán dự báo, việc sử dụng k-NN có thể cung cấp kết quả dự báo chính xác và dễ hiểu, đặc biệt là đối với các bài toán dự báo có cấu trúc đơn giản. Đề tài này không chỉ tập trung vào việc nghiên cứu thuật toán mà còn đặt ra cơ hội ứng dụng thực tế trong nhiều lĩnh vực như tài chính, y tế, và marketing..

Với tiềm năng và tính ứng dụng cao, đề tài này không chỉ có khả năng đóng góp vào sự phong phú về kiến thức mà còn hỗ trợ trong việc phát triển các phương pháp và công nghệ mới trong lĩnh vực máy học và khoa học dữ liệu. Từ đó, việc nghiên cứu về ứng dụng của thuật toán k-NN trong xây dựng dữ liệu đồ thị và bài toán dự báo hứa hẹn mang lại những giá trị lớn cho cả cộng đồng nghiên cứu và thực tiễn ứng dụng.

## Tổng quan nghiên cứu

Thuật toán k-NN là một thuật toán phổ biến được ứng dụng nhiều trong các bài toán như bài toán “Dự báo mặn trên sông Hàm Luông bằng thuật toán K-nearest neighbors” trình bày về vấn đề xâm nhập mặn, các tác động của biến đổi khí hậu ở vùng đồng bằng sông Cửu Long hay bài toán “Phân lớp dữ liệu hoa Iris” của tác giả Nguyễn Văn Núi, Trường Đại học Công nghệ Thông tin và Truyền thông – Đại Học Thái Nguyên trình bày về việc phân lớp, dự đoán chính xác loài hoa Iris mang lại nhiều ý nghĩa quan trọng trong thực tiễn vì hoa Diên Vĩ là một loài hoa đẹp, đại diện cho sự may mắn, tình yêu, lòng dũng cảm, trung thành và sự khôn ngoan.

Đề tài tập trung vào việc áp dụng thuật toán k-NN trong bài toán dự báo trên dữ liệu đồ thị. Nghiên cứu chủ yếu là triển khái thuật toán k-NN để dự báo kết quả tốt nghiệp của sinh viên dựa trên dữ liệu đầu vào được đồ thị hóa.

Nghiên cứu sẽ bắt đầu bằng việc cung cấp một cái nhìn tổng quan về thuật toán k-NN và cách nó được áp dụng trong việc xây dựng mô hình dự báo. Sau đó, sẽ tập trung vào cách xử lý dữ liệu đồ thị và cách biểu diễn dữ liệu dưới dạng đồ thị để tận dụng thông tin cấu trúc và mối quan hệ giữa các điểm dữ liệu.

Qua quá trình nghiên cứu thực nghiệm, mô hình k-NN sẽ được phát triển và điều chỉnh để phù hợp với bài toán dự báo kết quả tốt nghiệp của sinh viên.

## Mục tiêu của đề tài

Triển khái thuật toán k-NN (k-Nearest Neighbors) để giải quyết bài toán dự báo trên tập dữ liệu đồ thị. Đồ án nhằm phát triển các mô hình k-NN để dự báo kết quả tốt nghiệp của sinh viên dựa trên dữ liệu đầu vào được đồ thị hóa.

## Nội dung nghiên cứu

Tìm hiểu các kiến thức cơ bản về thuật toán k-NN.

Xây dựng bộ dữ liệu cho bài toán dự đoán kết quả tốt nghiệp.

Sử dụng bộ dữ liệu vào giải quyết bài toán dự báo và đánh giá.

## Cấu trúc đồ án

* Mở đầu
* Chương 1: Cơ sở lý thuyết
* Chương 2: Xây dựng dữ liệu đồ thị và bài toán dự đoán
* Chương 3: Triển khái và thử nghiệm
* Kết luận

## Kết quả dự kiến

- Một mô hình xây dựng dữ liệu đồ thị dựa trên thuật toán k-NN.

- Áp dụng mô hình k-NN để dự báo kết quả tốt nghiệp của sinh viên dựa trên dữ liệu đầu vào được đồ thị hóa.

- Đánh giá hiệu suất của các mô hình dự báo.

# CƠ SỞ LÝ THUYẾT

## Khái niệm và nguyên lý hoạt động của thuật toán K-nearest neighbor

### Khái niệm

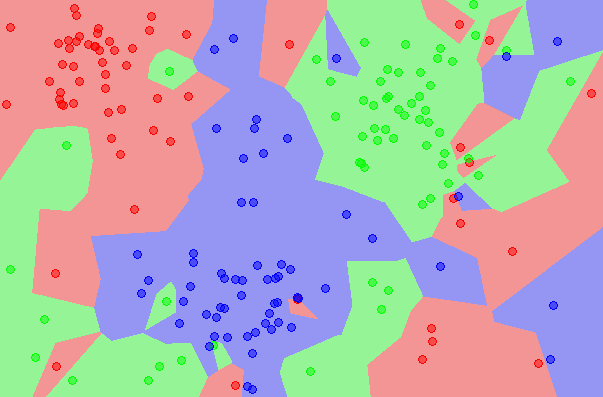
k-NN là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần phân lớp và tất cả các đối tượng trong dữ liệu huấn luyện.

### Nguyên lý hoạt động

Trong k-NN, với các bài toán Classification, label(nhãn) của một điểm dữ liệu mới được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set(tập huấn luyện). Label của một test data(dữ liệu kiểm thử) có thể được quyết định bằng major voting(quyết định bằng số phiếu) trong các điểm gần nhất, hoặc nó có thể được suy ra bằng cách đánh trong số khác nhau cho mỗi trong các điểm gần nhất đó rồi suy ra label cho điểm dữ liệu.

Với các bài toán Regression, đầu ra của điểm dữ liệu sẽ bằng chính đầu ra của điểm dữ liệu đã biết gần nhất (dối với trường hợp K = 1), hoặc là trung bình có trọng số của những điểm gần nhất, hoặc bằng một mối quan hệ nào đó dựa trên khoảng cách tới các điểm gần nhất đó.

Một cách ngắn gọn, k-NN là thuật toán đi tìm đầu ra của một điểm dữ liệu mới bằng cách chỉ dựa trên thông tin của K điểm dữ liệu trong training set gần nó nhất (K - lân cận), không quan tâm đến việc có một vài điểm dữ liệu trong những điểm gần nhất này là nhiễu. Hình dưới đây là một ví dụ về k-NN trong classification với K = 1.

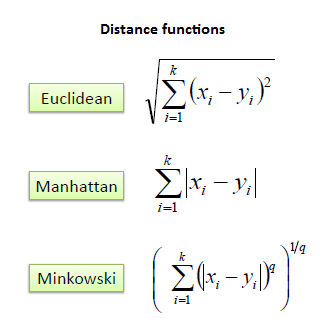


*Hình 2: Bản đồ của 1-NN*

Ví dụ trên đây là bài toán Classification với 3 classes: Đỏ, Lam, Lục. Mỗi điểm dữ liệu mới sẽ được gán nhãn theo màu của điểm mà nó thuộc về. Trong hình này, có một vài vùng nhỏ xen lẫn vào các vùng lớn hơn khác màu. Ví dụ có một điểm màu lục ở gần góc hướng 11 giờ nằm giữa hai vùng lớn hơn với nhiều dữ liệu màu đỏ và lam. Điểm này rất có thể là nhiễu. Dẫn đến nếu dữ liệu test rơi vào vùng này có khả năng cao cho kết quả không chính xác.

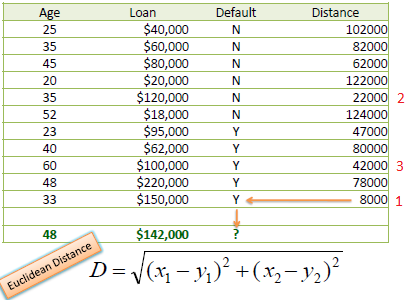
## Khoảng cách trong không gian vector

Thuật toán k-NN 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:

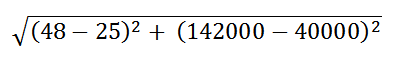


*Hình 3: Các công thức thường dùng để tính khoảng cách*

Ví dụ chúng ta có dữ liệu là tuổi, khoản vay và khả năng vở nợ như hình:

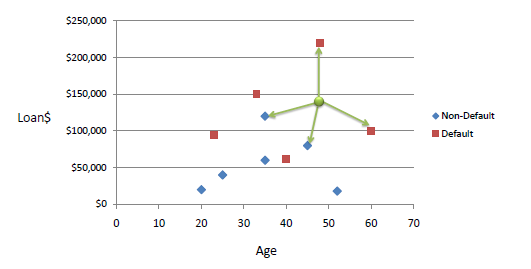


Dữ liệu cần phân loại là {age: 48, loan: 142000}. Đây dữ liệu 2 chiều và chúng ta cần dự đoán người này thuộc nguy cơ vỡ nợ hay không. Chúng ta sẽ dùng một cách khá phổ biến để tính khoảng cách là Euclidean. Ví dụ ở hàng đầu tiên khoảng cách sẽ được tính:



Thực hiện tương tự, ta sẽ tính được khoảng cách ở cột Distance, từ đó chọn ra k = 3 khoảng cách nhỏ nhất (gần với dữ liệu vào nhất). Với 3 khoảng cách này chúng ra nhận được 3 label là (Yes, No, Yes). Trong 3 label này Yes xuất hiện nhiều hơn nên chúng ta sẽ đưa ra dự đoán người này có khả năng vở nợ.

Vì đây là dữ liệu 2 chiều nên chúng ta củng có thể biểu diễn dữ liệu trong hệ tọa độ như hình:



*Hình 3: Biểu đồ biểu diễn điểm dữ liệu*

Trên hệ tọa độ này chúng ta thể dễ dàng nhận thấy cách chúng ta chọn k điểm gần nhất. Nhưng với dữ liệu lớn, nhiều chiều thì việc biểu diễn dữ liệu trên một không gian là không hề dễ dàng.

## Ưu điểm, nhược điểm của thuật toán

* Ưu điểm:

- Khi huấn luyện, thực tế K-NN chỉ lưu trữ toàn bộ dữ liệu huấn luyện mà không thực hiện việc tính toán phức tạp nào. Do đó:

Độ phức tạp về thời gian: O(1), vì chỉ cần lưu trữ dữ liệu.

Độ phức tạp về không gian: O(n), với n là số lượng mẫu trong tập huấn luyện, vì cần lưu toàn bộ dữ liệu.

- Việc dự đoán kết quả của dữ liệu mới rất đơn giản.

- Không cần giả sử gì về phân phối của các class.

- Khả năng làm việc với dữ liệu phi cấu trúc hoặc dữ liệu có số chiều cao.

- Hiệu suất tốt với dữ liệu nhỏ.

- Không cần huấn luyện.

- Dễ sử dụng và cài đặt.

* Nhược điểm:

- k-NN rất nhạy cảm với nhiễu khi K nhỏ.

- Đòi hỏi lựa chọn K phù hợp

- k-NN 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 k-NN.

## Tăng tốc cho thuật toán

Để tăng tốc độ cho thuật toán k-Nearest Neighbors (k-NN), chúng ta có thể áp dụng một số ký thuật và chiến lược sau:

**Chiết lược dữ liệu (Data Reduction):** Sử dụng kỹ thuật chiết lược dữ liệu như Principal Component Analysis để giảm số thuộc tính của dữ liệu. Điều này giúp tăng tốc độ của thuật toán giảm độ phức tạp tính toán.

**Indexing:** Sử dụng cấu trúc dữ liệu index như k-d trees hoặc Ball trees để tối ưu việc tìm kiếm: giúp giảm độ phức tạp của việc tìm kiếm và cải thiện hiệu suất của thuật toán.

**Parallelization:** Sử dụng tính toán song song để xử lý nhiều điểm dữ liệu cùng một lúc, giúp tăng tốc độ tính toán đáng kể, nhất là trên các hệ thống mạnh.

**Lazy Learning:** Tránh việc huấn luyện mô hình k-NN trước khi có dữ liệu cụ thể cần phân loại. Thuật toán k-NN thuộc loại "lazy learning", nghĩa là không cần quá trình huấn luyện một mô hình, điều này giúp tiết kiệm thời gian.

**Sampling Techniques:** Sử dụng kỹ thuật lấy mẫu (sampling) như Random Projection hoặc LSH (Locality-Sensitive Hashing) để giảm số lượng điểm dữ liệu cần xem xét mà vẫn đảm bảo tính chính xác của thuật toán.

**Optimize Distance Calculation:** Tối ưu hóa quá trình tính toán khoảng cách giữa các điểm dữ liệu bằng cách sử dụng các thư viện tối ưu hóa hoặc cải tiến thuật toán tính khoảng cách.

Kết hợp các chiến lược trên có thể giúp tăng tốc độ và hiệu suất cho thuật toán k-NN trong việc phân loại và dự báo trên các tập dữ liệu lớn.

## Kết luận chương 1:

Trong chương 1, chúng ta đã tìm hiểu về các kiến thức cơ bản của thuật toán K-nearest neighbor, các loại tính khoảng cách trong không gian vector, Chúng ta đã cũng đã xem xét các ưu - nhược điểm của thuật toán này.

Từ đó, chúng ta có cái nhìn tổng quan về cách mà thuật toán K-nearest neighbor hoạt động. Việc hiểu rõ về cơ sở lý thuyết của thuật toán K-nearest neighbor và các yếu tố liên quan đã cung cấp cho chúng ta nền tảng quan trọng để tiếp tục khám phá và nghiên cứu sâu hơn về ứng dụng của thuật toán này trong các bài toán dự đoán và phân loại trong lĩnh vực máy học và khoa học dữ liệu.

# XÂY DỰNG DỮ LIỆU ĐỒ THỊ VÀ BÀI TOÁN DỰ ĐOÁN

## Xây dựng dữ liệu đồ thị từ tập dữ liệu

Thực hiện thu thập dữ liệu (thông tin cá nhân, thông tin học vấn, dự định học tập,…), xác nhận xếp loại tốt nghiệp, từ đó ta có tập dữ liệu càn thiết.

Từ tập dữ liệu thu được thông tin cần thiết của các sinh viên như điểm GPA cũng như số học kỳ của điểm GPA đó.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **MSV** | **HỌ TÊN** | **GPA** | **HỌC KỲ** | **KẾT QUẢ** |
| 1 | 2020607678 | Đậu Xuân Long | 3.74 | 8 | xuất sắc |
| 2 | 2020605521 | Nguyễn Minh Đức | 3.16 | 4 | xuất sắc |
| 3 | 2020603331 | Nguyễn Văn Tiến | 3.61 | 4 | xuất sắc |
| 4 | 2020602509 | Nguyễn Thị Quỳnh | 3.78 | 3 | xuất sắc |
| 5 | 2020603396 | Nguyễn Thị Minh | 3.79 | 4 | xuất sắc |
| 6 | 2020606450 | Chu Minh Quyết | 2.32 | 6 | trung bình |
| 7 | 2020607485 | Nguyễn Tài Huy | 2.54 | 8 | khá |
| 8 | 2020607939 | Giàng A Chúng | 2.61 | 8 | khá |
| 9 | 2020607007 | Lê Thế Hùng | 2.61 | 8 | khá |
| 10 | 2020606987 | Nguyễn Gia Thắng | 2.61 | 8 | khá |
| 11 | 2020607847 | Nguyễn Thanh Long | 2.63 | 8 | khá |
| 12 | 2020608311 | Phạm Đình Thành | 2.63 | 8 | khá |
| 13 | 2020607575 | Hoàng Công Đức | 2.65 | 8 | khá |
| 14 | 2020606615 | Phan Mạnh Tuân | 2.66 | 8 | khá |
| 15 | 2020608237 | Trương Mạnh Quang | 2.67 | 8 | khá |
| 16 | 2020608387 | Nguyễn Đức Dũng | 2.7 | 8 | khá |
| 17 | 2020608155 | Đặng Hoàng Gia | 2.72 | 8 | khá |
| 18 | 2020608202 | Chu Văn Trường | 2.75 | 8 | khá |
| 19 | 2020606998 | Đoàn Văn Giang | 2.77 | 8 | khá |
| 20 | 2020607489 | Lê Trung Nguyên | 2.78 | 8 | khá |
| 21 | 2020607404 | Nguyễn Văn Đức | 2.79 | 8 | khá |
| 22 | 2020607285 | Đặng Bảo Chung | 2.8 | 8 | khá |
| 23 | 2020600419 | Nguyễn Quang Ninh | 2.8 | 7 | khá |
| 24 | 2020606739 | Nguyễn Ngô An | 2.81 | 8 | khá |
| 25 | 2020607612 | Đinh Đắc Đại | 2.81 | 8 | khá |
| 26 | 2020607540 | Phạm Quang Định | 2.81 | 8 | khá |
| 27 | 2020607726 | Bùi Anh Tuấn | 2.86 | 8 | khá |
| 28 | 2020606101 | Ngô Văn Triết | 2.89 | 8 | khá |
| 29 | 2020606958 | Nguyễn Đức Kiên | 2.89 | 8 | khá |
| 30 | 2020606891 | Cao Đức Hiếu | 2.92 | 8 | khá |
| 31 | 2020607784 | Nguyễn Xuân Nam | 2.96 | 8 | khá |
| 32 | 2020606764 | Nguyễn Minh Quân | 2.98 | 8 | khá |
| 33 | 2020604485 | Vũ Quang Nam | 3.01 | 8 | khá |
| 34 | 2020607212 | Nguyễn Văn Luận | 3.03 | 8 | khá |
| 35 | 2020607814 | Phạm Quyết Thắng | 3.04 | 8 | khá |
| 36 | 2020608245 | Tạ Văn Tú | 3.04 | 8 | khá |
| 37 | 2020607776 | Mai Thế Đoàn | 3.05 | 8 | khá |
| 38 | 2020606901 | Trần Trung Hiếu | 3.05 | 8 | khá |
| 39 | 2020607943 | Nguyễn Văn Huy | 3.07 | 8 | khá |
| 40 | 2020607338 | Tống Duy Anh | 3.08 | 8 | khá |
| 41 | 2020607346 | Nguyễn Văn Bình | 3.08 | 8 | khá |
| 42 | 2020604455 | Nguyễn Văn Dẫn | 3.1 | 8 | khá |
| 43 | 2020607981 | Lại Hoàng Tùng | 3.1 | 8 | khá |
| 44 | 2020607692 | Phan Xuân An | 3.12 | 8 | khá |
| 45 | 2020607524 | Nguyễn Hoàng Hiếu | 3.17 | 8 | khá |
| 46 | 2020600478 | Lục Cao Bắc | 1.5 | 1 | khá |
| 47 | 2020600620 | Nguyễn Phùng An | 1.75 | 1 | khá |
| 48 | 2020600962 | Phạm Trung Hiếu | 2.13 | 1 | khá |
| 49 | 2020601048 | Tống Mỹ Linh | 2.13 | 1 | khá |
| 50 | 2020601384 | Lê Tuấn Kiệt | 2.25 | 1 | khá |
| 51 | 2020600816 | Vũ Văn Mạnh | 2.25 | 1 | khá |
| 52 | 2020604542 | Trần Quang Huy | 2.27 | 4 | khá |
| 53 | 2020605139 | Lưu Đặng Trung Cheng | 2.33 | 4 | khá |
| 54 | 2020605418 | Cao Xuân Minh | 2.37 | 6 | khá |
| 55 | 2020600813 | Vương Quốc Tuấn | 2.38 | 1 | khá |
| 56 | 2020600686 | Nguyễn Đăng Phi | 2.5 | 1 | khá |
| 57 | 2020600187 | Lê Xuân Sỹ | 2.5 | 1 | khá |
| 58 | 2020605216 | Võ Văn Đức | 2.56 | 4 | khá |
| 59 | 2020605970 | Nguyễn Nam Lâm | 2.56 | 6 | khá |
| 60 | 2020604246 | Đoàn Hữu Thành | 2.57 | 4 | khá |
| 61 | 2020606594 | Hoàng Chí Chung | 2.57 | 6 | khá |
| 62 | 2020602657 | Nguyễn Văn Trúc | 2.59 | 3 | khá |
| 63 | 2020605883 | Nguyễn Đức Quân | 2.6 | 6 | khá |
| 64 | 2020605332 | Nguyễn Duy Sáng | 2.61 | 6 | khá |
| 65 | 2020605815 | Mai Thế Trí Hiếu | 2.61 | 6 | khá |
| 66 | 2020601564 | Trần Thị Thu Trang | 2.62 | 3 | khá |
| 67 | 2020600079 | Phạm Văn Cương | 2.63 | 1 | khá |
| 68 | 2020600758 | Đỗ Minh Đức | 2.63 | 1 | khá |
| 69 | 2020600157 | Vũ Quang Tùng | 2.63 | 1 | khá |
| 70 | 2020604881 | Bùi Tuấn Dương | 2.64 | 4 | khá |
| 71 | 2020606278 | Hồ Việt Anh | 2.65 | 6 | khá |
| 72 | 2020603110 | Phạm Quốc Hoàng | 2.66 | 4 | khá |
| 73 | 2020602028 | Phạm Văn Đạt | 2.67 | 3 | khá |
| 74 | 2020606004 | Ma Thị Lợi | 2.67 | 6 | khá |
| 75 | 2020606158 | Lương Tuấn Anh | 2.68 | 6 | khá |
| 76 | 2020602410 | Trần Thành Công | 2.7 | 3 | khá |
| 77 | 2020606145 | Kiều Văn Anh | 2.72 | 6 | khá |
| 78 | 2020606692 | Phùng Thế Khánh | 2.72 | 6 | khá |
| 79 | 2020605358 | Vũ Huy Công | 2.74 | 4 | khá |
| 80 | 2020601000 | Kiều Đức Anh | 2.75 | 1 | khá |
| 81 | 2020600796 | Nguyễn Duy Hùng | 2.75 | 1 | khá |
| 82 | 2020600154 | Cao Đức Thái Nhật | 2.75 | 1 | khá |
| 83 | 2020600432 | Lê Văn Phúc | 2.75 | 1 | khá |
| 84 | 2020600209 | Nguyễn Bỉnh Trường Sơn | 2.75 | 1 | khá |
| 85 | 2020603701 | Nguyễn Văn Phương | 2.75 | 4 | khá |
| 86 | 2020605090 | Châu Tiến Hưng | 2.75 | 4 | khá |
| 87 | 2020605337 | Hoàng Xuân Thắng | 2.75 | 6 | khá |
| 88 | 2020603011 | Nguyễn Việt Khánh | 2.75 | 6 | khá |
| 89 | 2020604764 | Nguyễn Văn Dũng | 2.77 | 4 | khá |
| 90 | 2020602616 | Phạm Ngọc Minh | 2.8 | 3 | khá |
| 91 | 2020606261 | Nguyễn Tuấn Anh | 2.8 | 6 | khá |
| 92 | 2020606391 | Nguyễn Văn Linh | 2.8 | 6 | khá |
| 93 | 2020606297 | Trần Ngọc Huy | 2.81 | 6 | khá |
| 94 | 2020602741 | Trần Tiến Đạt | 2.82 | 3 | khá |
| 95 | 2020602396 | Phạm Văn Đồng | 2.82 | 3 | khá |
| 96 | 2020605070 | Nguyễn Ngọc Đạt | 2.82 | 4 | khá |
| 97 | 2020605939 | Nguyễn Văn Hưng | 2.83 | 6 | khá |
| 98 | 2020605876 | Mai Trung Lâm | 2.83 | 6 | khá |
| 99 | 2020605913 | Ngô Khắc Lập | 2.83 | 6 | khá |
| 100 | 2020603930 | Đoàn Trung Đức | 2.84 | 3 | khá |
| 101 | 2020605563 | Dương Thanh Nam | 2.85 | 6 | khá |
| 102 | 2020605574 | Lê Minh Quang | 2.85 | 6 | khá |
| 103 | 2020604672 | Trần Khánh Toàn | 2.87 | 6 | khá |
| 104 | 2020601046 | Lương Trung Nghĩa | 2.88 | 1 | khá |
| 105 | 2020600779 | Lưu Văn Việt | 2.88 | 1 | khá |
| 106 | 2020601854 | Hoàng Công Hùng | 2.88 | 3 | khá |
| 107 | 2020603086 | Đặng Hữu Thắng | 2.88 | 4 | khá |
| 108 | 2020605567 | Nghiêm Đình Hồng | 2.88 | 4 | khá |
| 109 | 2020606337 | Nguyễn Chí Dũng | 2.88 | 6 | khá |
| 110 | 2020603007 | Nguyễn Tiến Dũng | 2.89 | 6 | khá |
| 111 | 2020602825 | Lại Minh Tiến | 2.9 | 3 | khá |
| 112 | 2020603802 | Nguyễn Hoàng Lâm | 2.91 | 4 | khá |
| 113 | 2020605849 | Phạm Quý Dương | 2.91 | 6 | khá |
| 114 | 2020604110 | Nguyễn Bá Khang | 2.92 | 4 | khá |
| 115 | 2020601821 | Lê Quang Linh | 2.93 | 3 | khá |
| 116 | 2020603398 | Đào Trung Hiếu | 2.93 | 3 | khá |
| 117 | 2020605502 | Trần Quang Tú | 2.93 | 6 | khá |
| 118 | 2020606695 | Nguyễn Văn Linh | 2.93 | 6 | khá |
| 119 | 2020605647 | Bùi Thị Minh Nguyệt | 2.93 | 6 | khá |
| 120 | 2020605740 | Hoàng Văn Ninh | 2.93 | 6 | khá |
| 121 | 2020606587 | Lý Văn Khải | 2.94 | 6 | khá |
| 122 | 2020604849 | Vũ Thị Quỳnh | 2.95 | 6 | khá |
| 123 | 2020606542 | Trần Hoàng Đức | 2.95 | 6 | khá |
| 124 | 2020602266 | Vũ Minh Đức | 2.96 | 3 | khá |
| 125 | 2020603080 | Trần Tuấn Hùng | 2.97 | 4 | khá |
| 126 | 2020606646 | Đinh Ngọc Hải | 2.97 | 6 | khá |
| 127 | 2020602617 | Nguyễn Văn Toàn | 2.99 | 3 | khá |
| 128 | 2020600461 | Lê Quang Huy | 3 | 1 | khá |
| 129 | 2020600976 | Trịnh Dương Nhật | 3 | 1 | khá |
| 130 | 2020602544 | Hoàng Minh Hiếu | 3 | 3 | khá |
| 131 | 2020602004 | Nguyễn Duy Đức | 3.01 | 3 | khá |
| 132 | 2020602694 | Vũ Ngọc Nghĩa | 3.01 | 3 | khá |
| 133 | 2020604619 | Vũ Đình An | 3.02 | 4 | khá |
| 134 | 2020602504 | Phạm Minh Hoàng | 3.03 | 3 | khá |
| 135 | 2020603133 | Nguyễn Quang Trường | 3.03 | 4 | khá |
| 136 | 2020604175 | Nguyễn Minh Thông | 3.04 | 4 | khá |
| 137 | 2020606688 | Nguyễn Thị Hương Lan | 3.04 | 6 | khá |
| 138 | 2020602536 | Nguyễn Tùng Dương | 3.06 | 3 | khá |
| 139 | 2020602847 | Nguyễn Văn Long | 3.06 | 3 | khá |
| 140 | 2020605607 | Bùi Việt Hoàng | 3.06 | 4 | khá |
| 141 | 2020605860 | Nguyễn Tuấn Đạt | 3.06 | 6 | khá |
| 142 | 2020604195 | Nguyễn Hải Triều | 3.07 | 4 | khá |
| 143 | 2020602342 | Đoàn Văn Quân | 3.09 | 3 | khá |
| 144 | 2020601885 | Phạm Minh Đức | 3.1 | 3 | khá |
| 145 | 2020604260 | Phạm Mạnh Hùng | 3.11 | 4 | khá |
| 146 | 2020601808 | Trần Xuân Tùng | 3.12 | 3 | khá |
| 147 | 2020600735 | Nguyễn Quỳnh Anh | 3.13 | 1 | khá |
| 148 | 2020600678 | Nguyễn Kim Tường | 3.13 | 1 | khá |
| 149 | 2020603405 | Trần Hùng Cường | 3.13 | 3 | khá |
| 150 | 2020603178 | Nguyễn Công Thành | 3.13 | 4 | khá |
| 151 | 2020604129 | Bùi Văn Đức | 3.16 | 3 | khá |
| 152 | 2020602889 | Nguyễn Văn Hải | 3.17 | 3 | khá |
| 153 | 2020602156 | Lê Ngọc Sơn | 3.18 | 3 | khá |
| 154 | 2020604531 | Vương Tú Linh | 3.18 | 4 | khá |
| 155 | 2020604119 | Nguyễn Hoài Nam | 3.19 | 4 | khá |
| 156 | 2020604037 | Nguyễn Hữu Thịnh | 3.22 | 4 | khá |
| 157 | 2020601337 | Ma Thị Nhị Hằng | 3.25 | 1 | khá |
| 158 | 2020600653 | Tạ Đức Nghĩa | 3.25 | 1 | khá |
| 159 | 2020600793 | Nguyễn Văn Quân | 3.25 | 1 | khá |
| 160 | 2020601249 | Nguyễn Quang Sáng | 3.25 | 1 | khá |
| 161 | 2020605339 | Đặng Đình Điềm | 3.33 | 4 | khá |
| 162 | 2020600207 | Phạm Hưng Khang | 3.38 | 1 | khá |
| 163 | 2020600156 | Nguyễn Văn Thịnh | 3.38 | 1 | khá |
| 164 | 2020602942 | Đặng Tùng Khánh | 3.41 | 3 | khá |
| 165 | 2020601657 | Nguyễn Thành Nam | 3.41 | 3 | khá |
| 166 | 2020600810 | Nguyễn Văn Quang | 3.5 | 1 | khá |
| 167 | 2020606772 | Nguyễn Thị Lan | 3.22 | 8 | giỏi |
| 168 | 2020605777 | Trần Thị Thư | 3.23 | 8 | giỏi |
| 169 | 2020607189 | Nguyễn Duy Hoàng | 3.23 | 8 | giỏi |
| 170 | 2020606981 | Phạm Thanh Nhàn | 3.24 | 8 | giỏi |
| 171 | 2020606214 | Nguyễn Việt Tiến | 3.28 | 8 | giỏi |
| 172 | 2020607528 | Nguyễn Danh Khánh | 3.29 | 8 | giỏi |
| 173 | 2020606766 | Đinh Văn Huy | 3.35 | 8 | giỏi |
| 174 | 2020608104 | Nguyễn Trung Thành | 3.37 | 8 | giỏi |
| 175 | 2020608127 | Nguyễn Hoàng Minh | 3.42 | 8 | giỏi |
| 176 | 2020606909 | Nguyễn Thị Thu Thủy | 3.43 | 8 | giỏi |
| 177 | 2020607923 | Nguyễn Quang Linh | 3.49 | 8 | giỏi |
| 178 | 2020608019 | Hoàng Thị Sao Mai | 3.52 | 8 | giỏi |
| 179 | 2020600685 | Nguyễn Quốc Anh | 2 | 1 | giỏi |
| 180 | 2020601159 | Trần Thanh Phong | 2.5 | 1 | giỏi |
| 181 | 2020600265 | Hoàng Quốc Chung | 2.75 | 1 | giỏi |
| 182 | 2020600215 | Phạm Vũ Anh Đức | 2.75 | 1 | giỏi |
| 183 | 2020601227 | Nguyễn Thành Lâm | 2.88 | 1 | giỏi |
| 184 | 2020600528 | Nguyễn Hoàng Long | 2.88 | 1 | giỏi |
| 185 | 2020602996 | Nguyễn Quang Anh | 2.88 | 1 | giỏi |
| 186 | 2020604419 | Trần Giang Lân | 2.98 | 4 | giỏi |
| 187 | 2020601133 | Bùi Hoàng Đức Anh | 3 | 1 | giỏi |
| 188 | 2020600274 | Dương Ngọc Anh | 3 | 1 | giỏi |
| 189 | 2020600227 | Đỗ Xuân Chiến | 3 | 1 | giỏi |
| 190 | 2020600872 | Diệp Thị Linh | 3 | 1 | giỏi |
| 191 | 2020602363 | Phạm Thanh Tùng | 3.06 | 3 | giỏi |
| 192 | 2020603098 | Trần Thị Thủy | 3.08 | 4 | giỏi |
| 193 | 2020602212 | Đặng Thị Thu Hà | 3.09 | 3 | giỏi |
| 194 | 2020603508 | Đặng Đức Long | 3.1 | 4 | giỏi |
| 195 | 2020605224 | Nguyễn Văn Thọ | 3.1 | 6 | giỏi |
| 196 | 2020601412 | Lê Văn Đức Đạt | 3.11 | 3 | giỏi |
| 197 | 2020605888 | Dương Quốc Khánh | 3.11 | 6 | giỏi |
| 198 | 2020600115 | Lê Việt Anh | 3.13 | 1 | giỏi |
| 199 | 2020600452 | Đặng Thanh Hải | 3.13 | 1 | giỏi |
| 200 | 2020600635 | Ngô Mạnh Hùng | 3.13 | 1 | giỏi |
| 201 | 2020602247 | Nguyễn Ngọc Thạch | 3.13 | 3 | giỏi |
| 202 | 2020605689 | Lê Khả Hiếu | 3.14 | 6 | giỏi |
| 203 | 2020605318 | Đặng Trường Khánh | 3.15 | 4 | giỏi |
| 204 | 2020605080 | Bùi Văn Thành | 3.17 | 6 | giỏi |
| 205 | 2020605486 | Trịnh Tiến Đạt | 3.19 | 4 | giỏi |
| 206 | 2020603917 | Đàm Mạnh Đào | 3.2 | 3 | giỏi |
| 207 | 2020604382 | Nguyễn Vũ Tiến Độ | 3.2 | 3 | giỏi |
| 208 | 2020603205 | Nguyễn Thế Hoàng | 3.2 | 4 | giỏi |
| 209 | 2020602271 | Phạm Ánh Dương | 3.21 | 3 | giỏi |
| 210 | 2020603752 | Nguyễn Thành Đạt | 3.21 | 3 | giỏi |
| 211 | 2020606268 | Phạm Thị Hạnh | 3.21 | 6 | giỏi |
| 212 | 2020602021 | Bùi Văn Thuận | 3.22 | 3 | giỏi |
| 213 | 2020603814 | Nguyễn Thị Thương | 3.22 | 4 | giỏi |
| 214 | 2020602995 | Lê Thị Huyền | 3.24 | 3 | giỏi |
| 215 | 2020600968 | Nguyễn Viết Quang Tùng | 3.25 | 1 | giỏi |
| 216 | 2020606049 | Phan Kim Sinh | 3.27 | 6 | giỏi |
| 217 | 2020605591 | Nguyễn Văn Anh | 3.29 | 4 | giỏi |
| 218 | 2020604967 | Vũ Đình Dũng | 3.29 | 4 | giỏi |
| 219 | 2020601414 | Phạm Thị Phương Thảo | 3.3 | 3 | giỏi |
| 220 | 2020606267 | Vũ Thị Huê | 3.3 | 6 | giỏi |
| 221 | 2020600026 | Tạ Đức Mạnh | 3.3 | 6 | giỏi |
| 222 | 2020605423 | Nguyễn Công Mạnh | 3.31 | 4 | giỏi |
| 223 | 2020602059 | Trịnh Tiến Đại | 3.33 | 3 | giỏi |
| 224 | 2020604313 | Nguyễn Thị Minh Ngọc | 3.33 | 4 | giỏi |
| 225 | 2020605642 | Trần Huy Long | 3.33 | 4 | giỏi |
| 226 | 2020606266 | Nguyễn Thị Khánh Ly | 3.33 | 6 | giỏi |
| 227 | 2020602706 | Lê Văn Hiếu | 3.36 | 3 | giỏi |
| 228 | 2020605544 | Ngô Triệu Vi | 3.36 | 6 | giỏi |
| 229 | 2020604232 | Nguyễn Đăng Tạ Khôi | 3.37 | 4 | giỏi |
| 230 | 2020605125 | Lê Huỳnh Đức | 3.37 | 4 | giỏi |
| 231 | 2020600770 | Ngô Tất Sơn | 3.38 | 1 | giỏi |
| 232 | 2020603568 | Lê Đức Lượng | 3.38 | 4 | giỏi |
| 233 | 2020604805 | Nguyễn Thị Xuân | 3.39 | 6 | giỏi |
| 234 | 2020602253 | Phạm Xuân Trường | 3.41 | 3 | giỏi |
| 235 | 2020605812 | Đặng Xuân Hoàng | 3.42 | 6 | giỏi |
| 236 | 2020602562 | Nguyễn Hải Dương | 3.43 | 3 | giỏi |
| 237 | 2020604595 | Dư Đăng Quang | 3.45 | 6 | giỏi |
| 238 | 2020605518 | Trương Thành Hưng | 3.47 | 4 | giỏi |
| 239 | 2020603876 | Ngô Đình Luân | 3.48 | 4 | giỏi |
| 240 | 2020603386 | Trần Hải Nam | 3.48 | 4 | giỏi |
| 241 | 2020600388 | Chử Thị Cẩm Ly | 3.5 | 1 | giỏi |
| 242 | 2020602356 | Nguyễn Huy Ngọ | 3.5 | 3 | giỏi |
| 243 | 2020604803 | Nguyễn Văn Việt | 3.5 | 6 | giỏi |
| 244 | 2020604137 | Trương Thị Lan Anh | 3.52 | 3 | giỏi |
| 245 | 2020605433 | Nguyễn Minh Đức | 3.52 | 4 | giỏi |
| 246 | 2020604411 | Nguyễn Mạnh Cường | 3.54 | 3 | giỏi |
| 247 | 2020602050 | Lê Thị Thu Trà | 3.61 | 3 | giỏi |
| 248 | 2020600285 | Nguyễn Văn Sơn | 3.63 | 1 | giỏi |
| 249 | 2020600875 | Nguyễn Văn Thắng | 3.63 | 1 | giỏi |
| 250 | 2020603502 | Nguyễn Văn Dũng | 3.72 | 3 | giỏi |
| 251 | 2020603329 | Đoàn Thị Thúy | 3.74 | 4 | giỏi |
| 252 | 2020600886 | Nguyễn Văn Việt | 3.75 | 1 | giỏi |

*Bảng 1: Tập dữ liệu thông tin sinh viên thu thập được*

Dữ liệu vào đã được đồ thị hóa bằng cách loại bỏ đi các cột thuộc tính mà thuật toán k-NN không cần đến là một file csv có 4 cột với cột đầu tiên là chỉ số, 2 cột giữa là thông số của thuộc tính và cột cuối cùng là kết quả tốt nghiệp.

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **GPA** | **HOCKY** | **KETQUA** |
| 1 | 3.74 | 8 | xuất sắc |
| 2 | 3.16 | 4 | xuất sắc |
| 3 | 3.61 | 4 | xuất sắc |
| 4 | 3.78 | 3 | xuất sắc |
| 5 | 3.79 | 4 | xuất sắc |
| 6 | 2.32 | 6 | trung bình |
| 7 | 2.54 | 8 | khá |
| 8 | 2.61 | 8 | khá |
| 9 | 2.61 | 8 | khá |
| 10 | 2.61 | 8 | khá |
| 11 | 2.63 | 8 | khá |
| 12 | 2.63 | 8 | khá |
| 13 | 2.65 | 8 | khá |
| 14 | 2.66 | 8 | khá |
| 15 | 2.67 | 8 | khá |
| 16 | 2.7 | 8 | khá |
| 17 | 2.72 | 8 | khá |
| 18 | 2.75 | 8 | khá |
| 19 | 2.77 | 8 | khá |
| 20 | 2.78 | 8 | khá |
| 21 | 2.79 | 8 | khá |
| 22 | 2.8 | 8 | khá |
| 23 | 2.8 | 7 | khá |
| 24 | 2.81 | 8 | khá |
| 25 | 2.81 | 8 | khá |
| 26 | 2.81 | 8 | khá |
| 27 | 2.86 | 8 | khá |
| 28 | 2.89 | 8 | khá |
| 29 | 2.89 | 8 | khá |
| 30 | 2.92 | 8 | khá |
| 31 | 2.96 | 8 | khá |
| 32 | 2.98 | 8 | khá |
| 33 | 3.01 | 8 | khá |
| 34 | 3.03 | 8 | khá |
| 35 | 3.04 | 8 | khá |
| 36 | 3.04 | 8 | khá |
| 37 | 3.05 | 8 | khá |
| 38 | 3.05 | 8 | khá |
| 39 | 3.07 | 8 | khá |
| 40 | 3.08 | 8 | khá |
| 41 | 3.08 | 8 | khá |
| 42 | 3.1 | 8 | khá |
| 43 | 3.1 | 8 | khá |
| 44 | 3.12 | 8 | khá |
| 45 | 3.17 | 8 | khá |
| 46 | 1.5 | 1 | khá |
| 47 | 1.75 | 1 | khá |
| 48 | 2.13 | 1 | khá |
| 49 | 2.13 | 1 | khá |
| 50 | 2.25 | 1 | khá |
| 51 | 2.25 | 1 | khá |
| 52 | 2.27 | 4 | khá |
| 53 | 2.33 | 4 | khá |
| 54 | 2.37 | 6 | khá |
| 55 | 2.38 | 1 | khá |
| 56 | 2.5 | 1 | khá |
| 57 | 2.5 | 1 | khá |
| 58 | 2.56 | 4 | khá |
| 59 | 2.56 | 6 | khá |
| 60 | 2.57 | 4 | khá |
| 61 | 2.57 | 6 | khá |
| 62 | 2.59 | 3 | khá |
| 63 | 2.6 | 6 | khá |
| 64 | 2.61 | 6 | khá |
| 65 | 2.61 | 6 | khá |
| 66 | 2.62 | 3 | khá |
| 67 | 2.63 | 1 | khá |
| 68 | 2.63 | 1 | khá |
| 69 | 2.63 | 1 | khá |
| 70 | 2.64 | 4 | khá |
| 71 | 2.65 | 6 | khá |
| 72 | 2.66 | 4 | khá |
| 73 | 2.67 | 3 | khá |
| 74 | 2.67 | 6 | khá |
| 75 | 2.68 | 6 | khá |
| 76 | 2.7 | 3 | khá |
| 77 | 2.72 | 6 | khá |
| 78 | 2.72 | 6 | khá |
| 79 | 2.74 | 4 | khá |
| 80 | 2.75 | 1 | khá |
| 81 | 2.75 | 1 | khá |
| 82 | 2.75 | 1 | khá |
| 83 | 2.75 | 1 | khá |
| 84 | 2.75 | 1 | khá |
| 85 | 2.75 | 4 | khá |
| 86 | 2.75 | 4 | khá |
| 87 | 2.75 | 6 | khá |
| 88 | 2.75 | 6 | khá |
| 89 | 2.77 | 4 | khá |
| 90 | 2.8 | 3 | khá |
| 91 | 2.8 | 6 | khá |
| 92 | 2.8 | 6 | khá |
| 93 | 2.81 | 6 | khá |
| 94 | 2.82 | 3 | khá |
| 95 | 2.82 | 3 | khá |
| 96 | 2.82 | 4 | khá |
| 97 | 2.83 | 6 | khá |
| 98 | 2.83 | 6 | khá |
| 99 | 2.83 | 6 | khá |
| 100 | 2.84 | 3 | khá |
| 101 | 2.85 | 6 | khá |
| 102 | 2.85 | 6 | khá |
| 103 | 2.87 | 6 | khá |
| 104 | 2.88 | 1 | khá |
| 105 | 2.88 | 1 | khá |
| 106 | 2.88 | 3 | khá |
| 107 | 2.88 | 4 | khá |
| 108 | 2.88 | 4 | khá |
| 109 | 2.88 | 6 | khá |
| 110 | 2.89 | 6 | khá |
| 111 | 2.9 | 3 | khá |
| 112 | 2.91 | 4 | khá |
| 113 | 2.91 | 6 | khá |
| 114 | 2.92 | 4 | khá |
| 115 | 2.93 | 3 | khá |
| 116 | 2.93 | 3 | khá |
| 117 | 2.93 | 6 | khá |
| 118 | 2.93 | 6 | khá |
| 119 | 2.93 | 6 | khá |
| 120 | 2.93 | 6 | khá |
| 121 | 2.94 | 6 | khá |
| 122 | 2.95 | 6 | khá |
| 123 | 2.95 | 6 | khá |
| 124 | 2.96 | 3 | khá |
| 125 | 2.97 | 4 | khá |
| 126 | 2.97 | 6 | khá |
| 127 | 2.99 | 3 | khá |
| 128 | 3 | 1 | khá |
| 129 | 3 | 1 | khá |
| 130 | 3 | 3 | khá |
| 131 | 3.01 | 3 | khá |
| 132 | 3.01 | 3 | khá |
| 133 | 3.02 | 4 | khá |
| 134 | 3.03 | 3 | khá |
| 135 | 3.03 | 4 | khá |
| 136 | 3.04 | 4 | khá |
| 137 | 3.04 | 6 | khá |
| 138 | 3.06 | 3 | khá |
| 139 | 3.06 | 3 | khá |
| 140 | 3.06 | 4 | khá |
| 141 | 3.06 | 6 | khá |
| 142 | 3.07 | 4 | khá |
| 143 | 3.09 | 3 | khá |
| 144 | 3.1 | 3 | khá |
| 145 | 3.11 | 4 | khá |
| 146 | 3.12 | 3 | khá |
| 147 | 3.13 | 1 | khá |
| 148 | 3.13 | 1 | khá |
| 149 | 3.13 | 3 | khá |
| 150 | 3.13 | 4 | khá |
| 151 | 3.16 | 3 | khá |
| 152 | 3.17 | 3 | khá |
| 153 | 3.18 | 3 | khá |
| 154 | 3.18 | 4 | khá |
| 155 | 3.19 | 4 | khá |
| 156 | 3.22 | 4 | khá |
| 157 | 3.25 | 1 | khá |
| 158 | 3.25 | 1 | khá |
| 159 | 3.25 | 1 | khá |
| 160 | 3.25 | 1 | khá |
| 161 | 3.33 | 4 | khá |
| 162 | 3.38 | 1 | khá |
| 163 | 3.38 | 1 | khá |
| 164 | 3.41 | 3 | khá |
| 165 | 3.41 | 3 | khá |
| 166 | 3.5 | 1 | khá |
| 167 | 3.22 | 8 | giỏi |
| 168 | 3.23 | 8 | giỏi |
| 169 | 3.23 | 8 | giỏi |
| 170 | 3.24 | 8 | giỏi |
| 171 | 3.28 | 8 | giỏi |
| 172 | 3.29 | 8 | giỏi |
| 173 | 3.35 | 8 | giỏi |
| 174 | 3.37 | 8 | giỏi |
| 175 | 3.42 | 8 | giỏi |
| 176 | 3.43 | 8 | giỏi |
| 177 | 3.49 | 8 | giỏi |
| 178 | 3.52 | 8 | giỏi |
| 179 | 2 | 1 | giỏi |
| 180 | 2.5 | 1 | giỏi |
| 181 | 2.75 | 1 | giỏi |
| 182 | 2.75 | 1 | giỏi |
| 183 | 2.88 | 1 | giỏi |
| 184 | 2.88 | 1 | giỏi |
| 185 | 2.88 | 1 | giỏi |
| 186 | 2.98 | 4 | giỏi |
| 187 | 3 | 1 | giỏi |
| 188 | 3 | 1 | giỏi |
| 189 | 3 | 1 | giỏi |
| 190 | 3 | 1 | giỏi |
| 191 | 3.06 | 3 | giỏi |
| 192 | 3.08 | 4 | giỏi |
| 193 | 3.09 | 3 | giỏi |
| 194 | 3.1 | 4 | giỏi |
| 195 | 3.1 | 6 | giỏi |
| 196 | 3.11 | 3 | giỏi |
| 197 | 3.11 | 6 | giỏi |
| 198 | 3.13 | 1 | giỏi |
| 199 | 3.13 | 1 | giỏi |
| 200 | 3.13 | 1 | giỏi |
| 201 | 3.13 | 3 | giỏi |
| 202 | 3.14 | 6 | giỏi |
| 203 | 3.15 | 4 | giỏi |
| 204 | 3.17 | 6 | giỏi |
| 205 | 3.19 | 4 | giỏi |
| 206 | 3.2 | 3 | giỏi |
| 207 | 3.2 | 3 | giỏi |
| 208 | 3.2 | 4 | giỏi |
| 209 | 3.21 | 3 | giỏi |
| 210 | 3.21 | 3 | giỏi |
| 211 | 3.21 | 6 | giỏi |
| 212 | 3.22 | 3 | giỏi |
| 213 | 3.22 | 4 | giỏi |
| 214 | 3.24 | 3 | giỏi |
| 215 | 3.25 | 1 | giỏi |
| 216 | 3.27 | 6 | giỏi |
| 217 | 3.29 | 4 | giỏi |
| 218 | 3.29 | 4 | giỏi |
| 219 | 3.3 | 3 | giỏi |
| 220 | 3.3 | 6 | giỏi |
| 221 | 3.3 | 6 | giỏi |
| 222 | 3.31 | 4 | giỏi |
| 223 | 3.33 | 3 | giỏi |
| 224 | 3.33 | 4 | giỏi |
| 225 | 3.33 | 4 | giỏi |
| 226 | 3.33 | 6 | giỏi |
| 227 | 3.36 | 3 | giỏi |
| 228 | 3.36 | 6 | giỏi |
| 229 | 3.37 | 4 | giỏi |
| 230 | 3.37 | 4 | giỏi |
| 231 | 3.38 | 1 | giỏi |
| 232 | 3.38 | 4 | giỏi |
| 233 | 3.39 | 6 | giỏi |
| 234 | 3.41 | 3 | giỏi |
| 235 | 3.42 | 6 | giỏi |
| 236 | 3.43 | 3 | giỏi |
| 237 | 3.45 | 6 | giỏi |
| 238 | 3.47 | 4 | giỏi |
| 239 | 3.48 | 4 | giỏi |
| 240 | 3.48 | 4 | giỏi |
| 241 | 3.5 | 1 | giỏi |
| 242 | 3.5 | 3 | giỏi |
| 243 | 3.5 | 6 | giỏi |
| 244 | 3.52 | 3 | giỏi |
| 245 | 3.52 | 4 | giỏi |
| 246 | 3.54 | 3 | giỏi |
| 247 | 3.61 | 3 | giỏi |
| 248 | 3.63 | 1 | giỏi |
| 249 | 3.63 | 1 | giỏi |
| 250 | 3.72 | 3 | giỏi |
| 251 | 3.74 | 4 | giỏi |
| 252 | 3.75 | 1 | giỏi |

*Bảng 2: Dữ liệu đầu vào*

## Áp dụng thuật toán k-NN vào dữ liệu đồ thị

Bảng dữ liệu sau khi được đồ thị hóa sẽ là dữ liệu đồ thị, tại đây mỗi sinh viên sẽ là 1 điểm dữ liệu đồ thị được biểu thị bằng 2 thông số là GPA và Học kỳ. Dữ liệu đồ thị sẽ được phân chia thành 2 phần là tập dữ liệu tập huấn và tập dữ liệu kiểm thử. Đáp án hay đầu ra là 1 trong 4 kết quả xuất sắc, giỏi, khá, trung bình. Để có được đáp án, thuật toán k-NN sử dụng phương pháp tính khoảng cánh giữa điểm dữ liệu cần test với tất các các điểm dữ liệu trong training set. Sau cùng thuật toán sử dụng phương pháp major voting để cho ra kết quả dự đoán.

## Bài Toán Dự Đoán Trên Dữ Liệu Đồ Thị

Trainingset

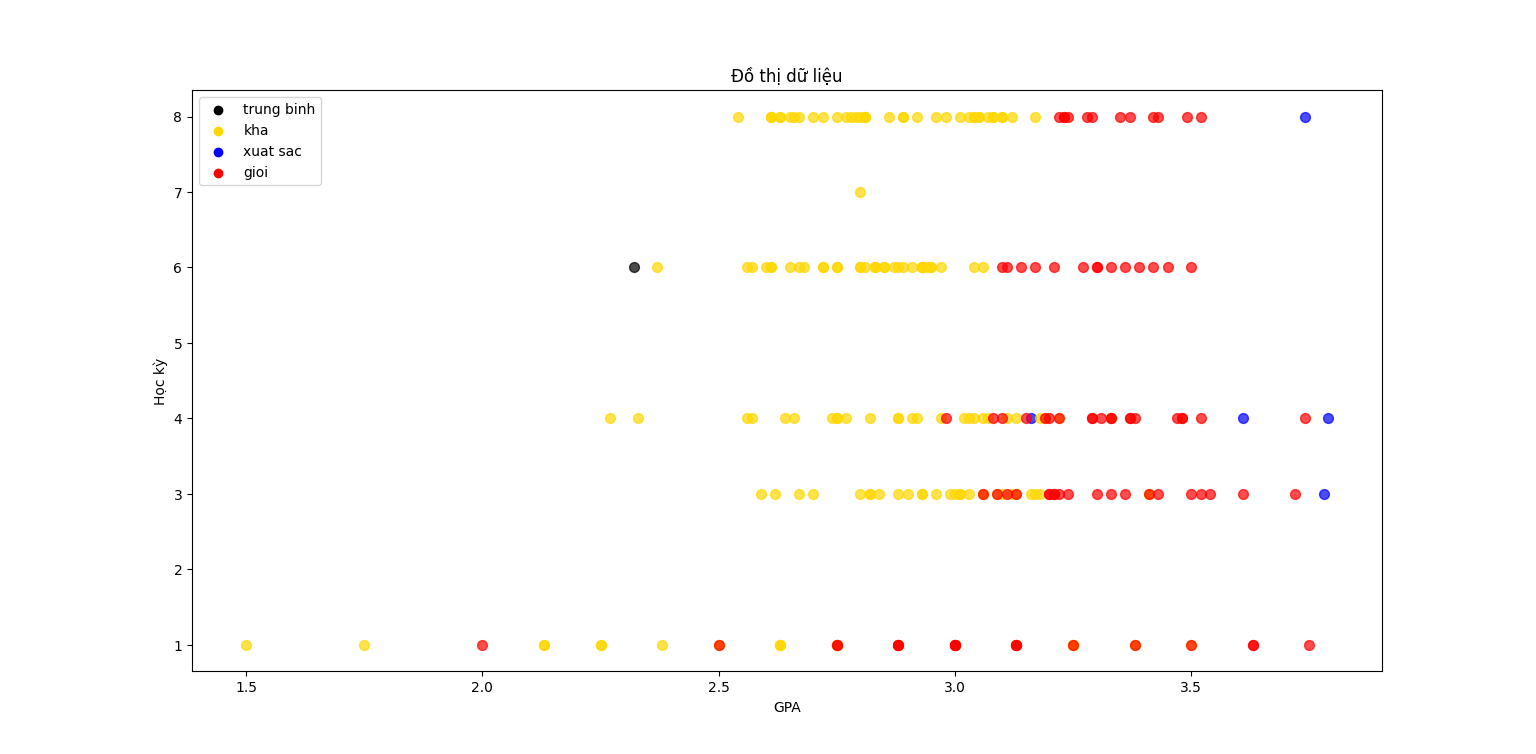
Dữ liệu

Tập các kết quả dự đoán

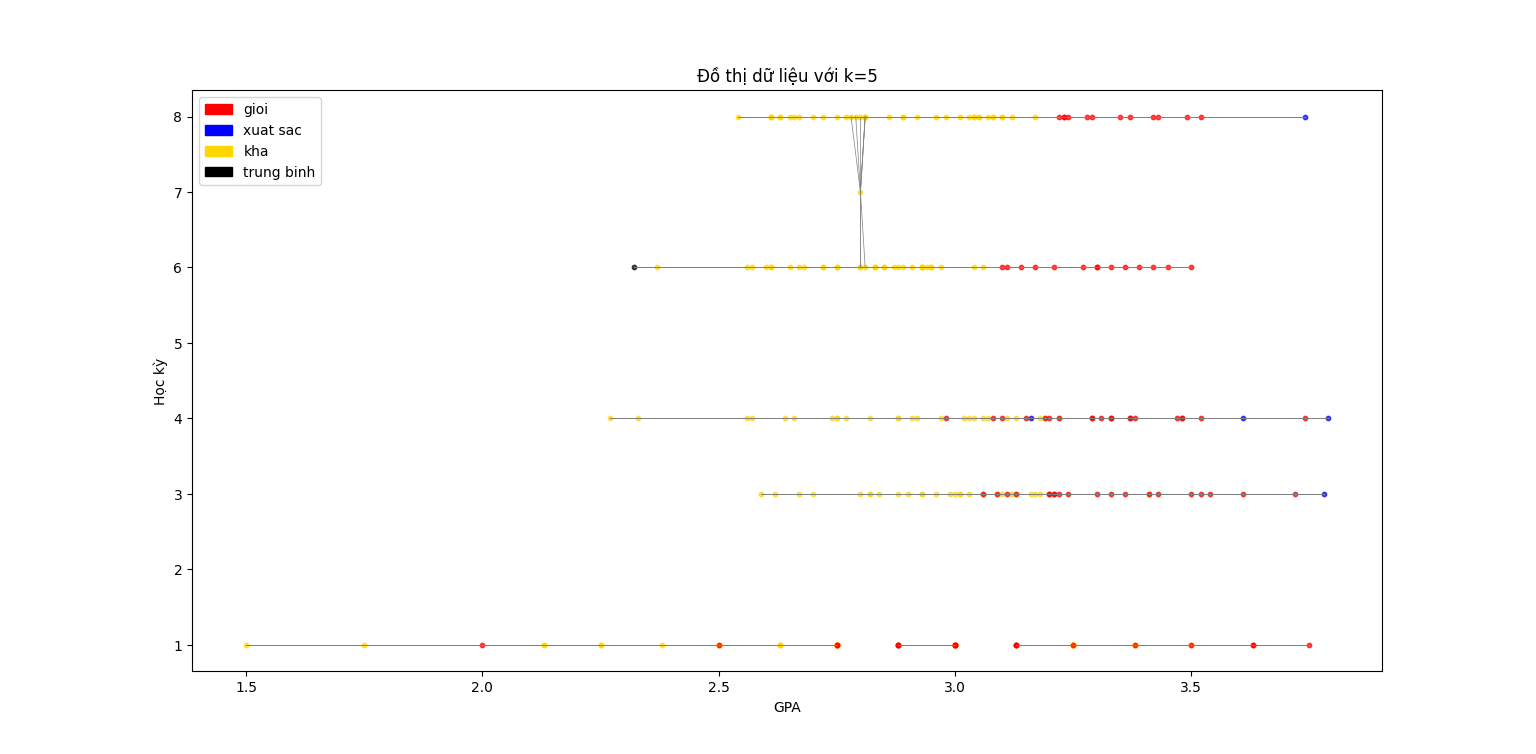
K-nearest neighbor

Testset

*Hình 4. Mô hình bài toán k-NN*



*Hình 4: Dữ liệu đồ thị*

**

*Hình 5: Đồ thị dữ liệu trong bài toán k-NN với k=9*

## Kết luận chương 2:

Trong chương 2, chúng ta đã tiến hành xây dựng dữ liệu đồ thị từ tập dữ liệu ban đầu, áp dụng thuật toán K-Nearest Neighbors (k-NN) vào dữ liệu đồ thị và đưa ra bài toán dự đoán trên dữ liệu đồ thị cùng với mô hình bài toán k-NN.

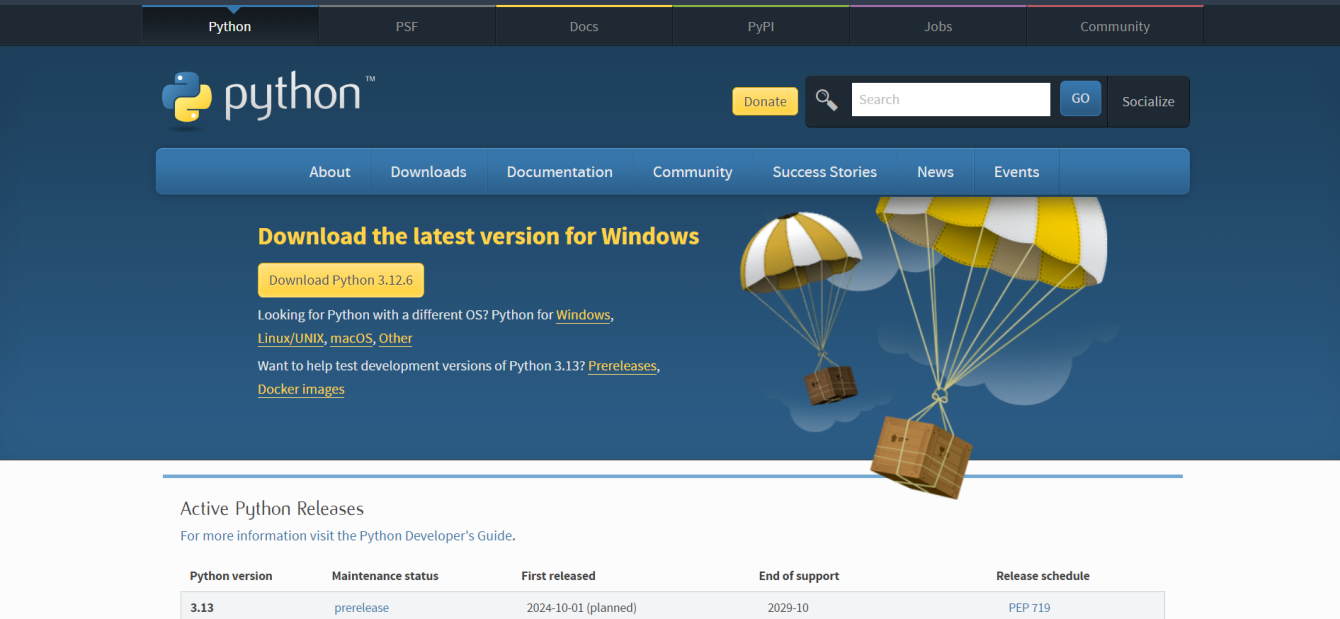
# THỰC NGHIỆM

## Môi trường triển khai

Chương trình được xây dựng bằng ngôn ngữ vì vậy cần thiết lập môi trường để chương trình có thể hoạt động.

Tải python tại đây: <https://www.python.org/downloads/>

Xuất hiện trang như hình dưới: Nhấp chuột vào download Python 3.12.6.

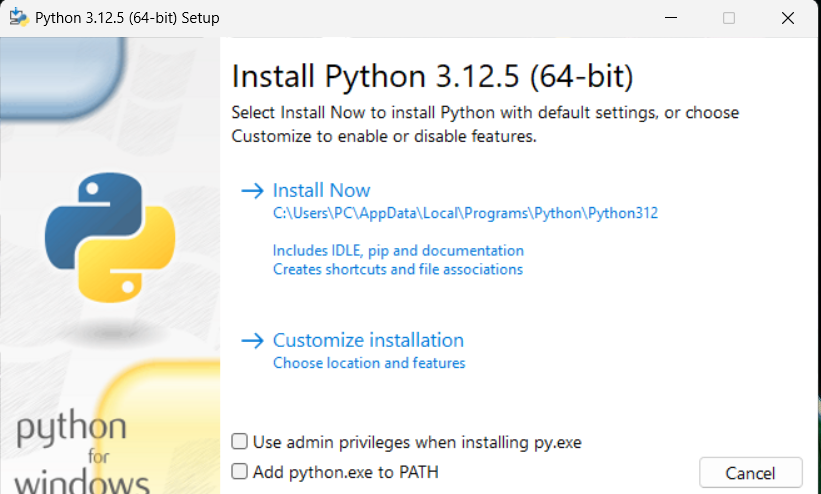


Lúc đó xuất hiện trang có chứa các phiên bản của Python, tải về phiên bản 3.12.6

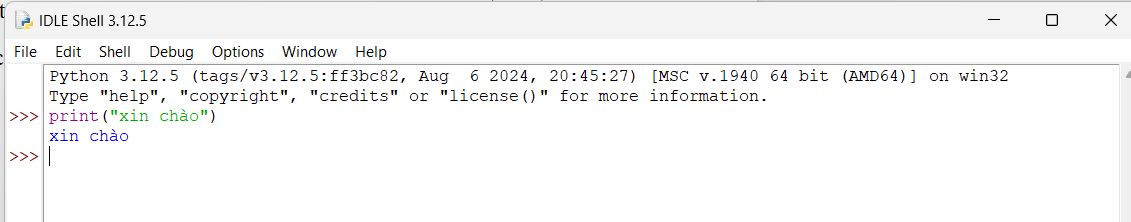
Khi tải về hoàn tất ta bắt đầu tiến hành cài đặt:

Nhấp đúp vào tệp vừa tải về đẻ cài đặt. Tại đây có hai lựa chọn:

* Install now: Mặc định cài python vào ổ C, cài sẵn IDLE, pip và tài liệu…
* Customize installation: cho phép chọn cài đặt và tính năng cần thiết.



Khi đã cài đặt xong, mở IDLE để chạy thử một đoạn code như sau: print(“xin chào”):

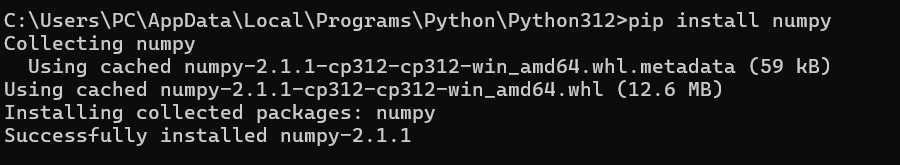


Như vậy đã hoàn thành xong việc cài đặt python, việc còn lại của chúng ta là cài đặt những thư viện cần thiết để hỗ trợ cho bài toán.

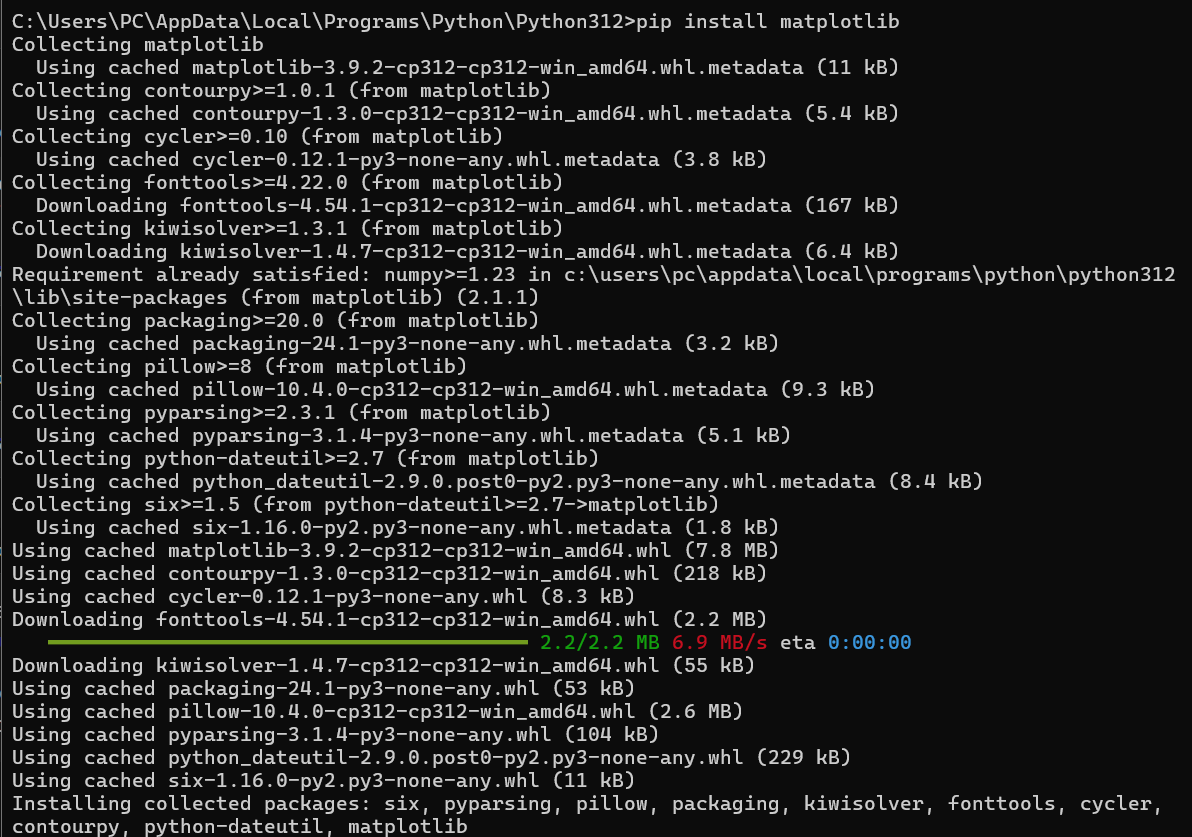
Ở đây, bài toán chúng ta cần các thư viện sau: numpy, matplotlib, sklearn.

Sử dụng cmd : Nếu pip chưa có sẵn gói thì ta phải cài đặt cho nó. Sau khi cài đặt thành công gói pip ta tiếp tục cài đặt thư viện numpy, matplotlib, sklearn.

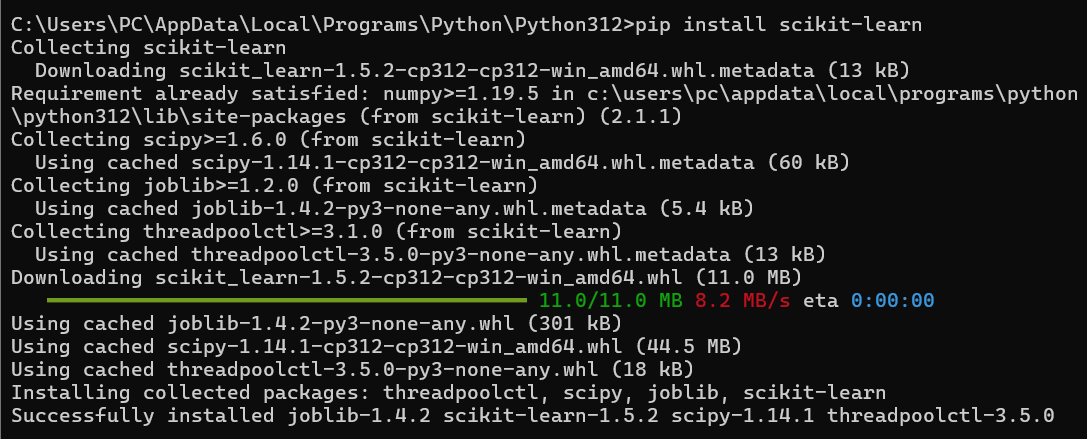
- Cài đặt thư viện numpy:



- Cài đặt thư viện matplotlib:



- Cài dặt thư viện sklearn:



Như vậy là đã hoàn thành cài đặt các thư viện hỗ trợ cho python cũng như bài toán.

## Quy trình triển khai

Bước 1: Thực hiện cài đặt các môi trường cần thiết cho việc triển khái, ví dụ như tải các IDE cần thiết, môi trường hỗ trợ ngôn ngữ Python, các thư viện hỗ trợ.

Bước 2: Tập dữ liệu đã được đồ thị hóa vào file csv

Bước 3: thiết lập chương trình.

Bước 4: chạy thực nghiệm, tổng hợp kết quả, so sánh, đánh giá.

## Thiết lập thực nghiệm

khái báo các thư viện cần thiết: math, numpy,…

import csv

import numpy as np

import math

import matplotlib.pyplot as plt

from sklearn import neighbors, datasets

Load dữ liệu, tách training set và test set: Dữ liệu được lưu trong file csv và được lưu vào data, sau đó tiến hành xóa các cột/hàng không cần thiết, tiến hành trộn các điểm dữ liệu và chia data vào trainset và testset.

def loadData(path):

    f = open(path, "r")

    data = csv.reader(f) #csv format

    data = np.array(list(data))# covert to matrix

    data = np.delete(data, 0, 0)# delete header

    data = np.delete(data, 0, 1) # delete index

    np.random.shuffle(data) # shuffle data

    f.close()

    trainSet = data[:220] #training data from 1->220

    testSet = data[220:]# the others is testing data

    return trainSet, testSet

Tính khoảng cách giữa các điểm đồ thị: Khoảng cách được tính bằng công thức Euclidean với 2 thông số đầu vào đối với mỗi điểm dữ liệu:

def calcDistancs(pointA, pointB, numOfFeature=2):

    tmp = 0

    for i in range(numOfFeature):

        tmp += (float(pointA[i]) - float(pointB[i])) \*\* 2

    return math.sqrt(tmp)

k-NN dự đoán: hàm KNearestNeighbor gọi hàm calDistancs để tính khoảng cách và sắp xếp các khoảng cách theo thứ tự tăng dần, sau đó hàm findMostOccus sẽ tiến hành đếm các nhãn của K điểm gần nhất, nhãn xuất hiện nhiều nhất sẽ là câu trả lời cho điểm dữ liệu đang test:

def kNearestNeighbor(trainSet, point, k):

    distances = []

    for item in trainSet:

        distances.append({

            "label": item[-1],

            "value": calcDistancs(item, point)

        })

    distances.sort(key=lambda x: x["value"])

    labels = [item["label"] for item in distances]

return labels[:k]

def findMostOccur(arr):

    labels = set(arr) # set label

    ans = ""

    maxOccur = 0

    for label in labels:

        num = arr.count(label)

        if num > maxOccur:

            maxOccur = num

            ans = label

    return ans

Xét trường hợp K=9, tức là với mỗi điểm dữ liệu test ta xét 5 điểm dữ liệu training gần nhất và lấy nhãn cả điểm đó để dự đoán cho điểm dữ liệu test

if \_\_name\_\_ == "\_\_main\_\_":

    trainSet, testSet = loadData("./dulieu.csv")

    numOfRightAnwser = 0

    for item in testSet:

        k-NN = kNearestNeighbor(trainSet, item, 9)

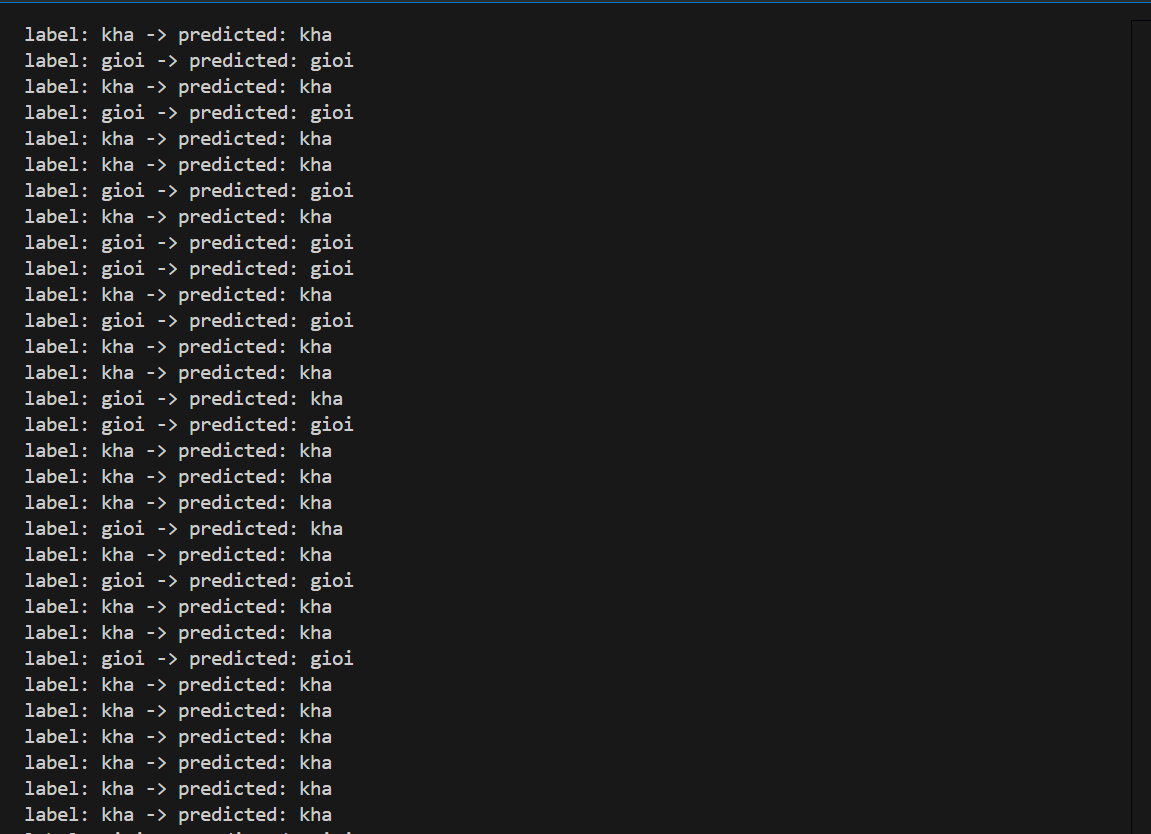
        answer = findMostOccur(k-NN)

        numOfRightAnwser += item[-1] == answer

        print("label: {} -> predicted: {}".format(item[-1], answer))

    print("Accuracy", numOfRightAnwser/len(testSet))

Kết quả nhận được:



Kết quả cho thấy nhãn dự đoán gần giống với label thật của dữ liệu test, chỉ có 2 trong số 31 điểm được hiển thị có kết quả sai lệch.

## So sánh và đánh giá

* Đánh giá:

Với k = 1:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Lần test** | 1 | 2 | 3 | 4 | 5 | **Trung bình** |
| **Hiệu suất** | 68.75% | 71.875% | 75% | 75% | 81.25% | 74,375% |

Với k= 3:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Lần test** | 1 | 2 | 3 | 4 | 5 | **Trung bình** |
| **Hiệu suất** | 84.375% | 75% | 81.25% | 68.75% | 81.25% | 78,125% |

Với k = 5:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Lần test** | 1 | 2 | 3 | 4 | 5 | **Trung bình** |
| **Hiệu suất** | 81.25% | 93.75% | 84.375% | 78.125% | 68.75% | 81.25% |

Với k = 7:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Lần test** | 1 | 2 | 3 | 4 | 5 | **Trung bình** |
| **Hiệu suất** | 87.5% | 78.125% | 84.375% | 78.125% | 90.625% | 83.75% |

Với k = 9:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Lần test** | 1 | 2 | 3 | 4 | 5 | **Trung bình** |
| **Hiệu suất** | 87.5% | 81.25% | 87.5% | 87.5% | 93.75% | 87.5% |

Với k =11:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Lần test** | 1 | 2 | 3 | 4 | 5 | **Trung bình** |
| **Hiệu suất** | 84.375% | 81.25% | 87.5% | 84.375% | 81.25% | 83.75% |

- Từ các lần thử nghiệm với các kịch bản hệ số k lần lượt là 1, 3, 5, 7, 9, 11. Ta nhận thấy với K = 9, mô hình dự đoán có hiệu suất trung bình cao nhất với 87.5%. Tiến hành thử nghiệm mở rộng thêm 5 lần nữa đối trường hợp k = 9:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Lần test** | 1 | 2 | 3 | 4 | 5 | 6 |
| **Hiệu suất** | 87.5% | 81.25% | 87.5% | 87.5% | 93.75% | 78.125% |
| **Lần test** | 7 | 8 | 9 | 10 | **Trung bình** | |
| **Hiệu suất** | 87.5% | 81.25% | 90.625% | 75% | 85% | |

- Hiệu suất trung bình của mô hình là 85%, cho thấy khả năng dự đoán trung bình của mô hình trong các lần test.

* So sánh:

- GCN là một thuật toán học sâu được sử dụng cho phân loại dữ liệu đồ thị.

- GCN thường được sử dụng hiệu quả trong việc xử lý dữ liệu đồ thị có cấu trúc phức tạp.

- So với mô hình dự đoán hiện tại, GCN có thể cung cấp kết quả tốt hơn đối với dữ liệu có tính chất đồ thị.

- Trong trường hợp dữ liệu có cấu trúc đồ thị và yêu cầu xử lý đồ thị phức tạp, việc sử dụng GCN hoặc các phương pháp học sâu khác có thể cung cấp hiệu suất cao hơn và phù hợp hơn.

# KẾT LUẬN

* **Kết quả đạt được:**

Đề tài đã hệ thống hóa được các kiến thức cơ bản về thuật toán k-NN: Từ khái niệm, nguyên lý hoạt động đến mô hình bài toán của k-NN. Đồng thời cũng liệt kê được các ứng dụng của k-NN trong thực tế đời sống.

Đề tài đã xây dựng và triển khai thành công mô hình k-NN giúp dự đoán kết quả tốt nghiệp của sinh viên. Kết quả cho thấy khả năng dự báo khi áp dụng thuật toán k-NN. Đặc biệt hữu ích trong các lĩnh vực như chứng khoán, tiền tệ.

* **Kết quả chưa đạt được:**

Mặc dù đề tài đã đạt được nhiều kết quả khả quan, nhưng vẫn tồn tại một số hạn chế: Độ chính xác của mô hình còn hạn chế: Kết quả của mô hình chưa thực sự đạt được độ chính xác tối ưu.

* **Hướng phát triển:**

Cải thiện độ chính xác của mô hình: Tiếp tục nghiên cứu và thử nghiệm các kỹ thuật tối ưu hóa cho mô hình k-NN để cải thiện độ chính xác của kết quả.

Tăng cường khả năng xử lý dữ liệu lớn:Nghiên cứu các phương pháp tăng tốc thuật toán

Mở rộng ứng dụng của k-NN: Nghiên cứu và phát triển thêm các ứng dụng khác của thuật toán k-NN trong các lĩnh vực như sinh học, y tế, tài chính,…

# TÀI LIỆU THAM KHẢO

[1] Nguyễn Văn Núi, Phân lớp dữ liệu hoa iris sử dụng các thuật toán naive bayes, random forest và knn, 2019

[2] Phạm Ngọc Hoài, Phan Thị Thanh Huyền, Lê Nguyễn, Nguyễn Thu Hiền, Trần Thành Thái, Lương Lê Lâm, Đánh giá khả năng dự báo mặn trên sông Hàm Luông của thuật toán k-Nearest Neighbors, Tạp chí khoa học và công nghệ thủy lợi số 74, 2022

[3] Vũ Hữu Tiệp, Machine Learning cơ bản, NXB Khoa Học và Kỹ Thuật, 2018.

[4] Nguyễn Phương Nga, Giáo trình trí tuệ nhân tạo, NXB Thông Kê, 2021.

[5] [https://machinelearningcoban.com/2017/01/08/k-NN/](https://machinelearningcoban.com/2017/01/08/knn/)

[6] <https://vi.wikipedia.org/wiki/Hoc_máy>

[7] https://codelearn.io/sharing/thuat-toan-k-nearest-neighbors-k-NN

# PHỤ LỤC

Báo cáo bản mềm, mã nguồn và các tài liệu liên quan khác:

https://github.com/Ninh216/DoAnTotNghiep/tree/master/OneDrive/Desktop/DoAn