**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

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

****

**ĐỒ ÁN TỐT NGHIỆP**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**TÊN ĐỀ TÀI ĐỒ ÁN TỐT NGHIỆP**

**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 ĐỒ ÁN TỐT NGHIỆP**

**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 muốn bày tỏ lòng biết ơn sâu sắc đến các thầy cô Trường Đại học Công Nghiệp Hà Nội, đặc biệt là các thầy, cô khoa Công nghệ thông tin đã tận tình dạy dỗ và hướng dẫn em trong suốt quãng thời gian học tập ở trường. Sự quan tâm, chỉ bảo và kiến thức chuyên môn 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 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 2](#_Toc3566)

[DANH MỤC CÁC BẢNG 3](#_Toc1309)

[DANH MỤC CÁC HÌNH 4](#_Toc23546)

[MỞ ĐẦU 5](#_Toc10321)

[1. Giới thiệu 5](#_Toc32248)

[2. Lý do chọn đề tài 6](#_Toc1204)

[3. Tổng quan nghiên cứu 7](#_Toc29274)

[4. Nội dung nghiên cứu 7](#_Toc6407)

[5. Mục tiêu của đề tài 7](#_Toc13974)

[6. Cấu trúc đồ án 7](#_Toc18387)

[Chương 1: CƠ SỞ LÝ THUYẾT 9](#_Toc14932)

[1.1 Khái niệm và nguyên lý hoạt động của thuật toán K-nearest neighbor 9](#_Toc1114)

[1.1.1 Định nghĩa 9](#_Toc17689)

[1.1.2 Nguyên lý hoạt động 9](#_Toc23576)

[1.2 Khoảng cách trong không gian vector 10](#_Toc10842)

[1.3 Tăng tốc cho thuật toán 12](#_Toc26885)

[1.4 Ưu điểm, nhược điểm của thuật toán 13](#_Toc18204)

[Chương 2: XÂY DỰNG DỮ LIỆU ĐỒ THỊ VÀ BÀI TOÁN DỰ ĐOÁN 15](#_Toc23361)

[2.1 Xây dựng dữ liệu đồ thị từ tập dữ liệu 15](#_Toc23940)

[2.2 Áp dụng thuật toán K-NN vào dữ liệu đồ thị 33](#_Toc31957)

[2.3 Bài Toán Dự Đoán Trên Dữ Liệu Đồ Thị 33](#_Toc17930)

*[Hình 3. Mô hình bài toán k-nn](#_Toc8906)* [34](#_Toc8906)

[Chương 3: TRIỂN KHAI VÀ THỬ NGHIỆM 35](#_Toc25067)

[3.1 Môi trường triển khai 35](#_Toc26910)

[3.2 Quy trình triển khai 36](#_Toc30004)

[3.3 Thiết lập thử nghiệm và đánh giá 36](#_Toc23952)

[3.4 So sánh và đánh giá 39](#_Toc27028)

[KẾT LUẬN 40](#_Toc30069)

[TÀI LIỆU THAM KHẢO 41](#_Toc1143)

# **DANH MỤC CÁC TỪ VIẾT TẮT**

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

# **DANH MỤC CÁC BẢNG**

|  |  |  |
| --- | --- | --- |
| STT | Tên bảng | Nội dung |
| 1 | Bảng 1 | Tập dữ liệu thu thập 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 giữa AI, ML và Deep Learning |
| 2 | Hình 2 | Bản đồ của 1NN |
| 3 | Hình 3 | Mô hình bài toán |

# MỞ ĐẦU

## Giới thiệu

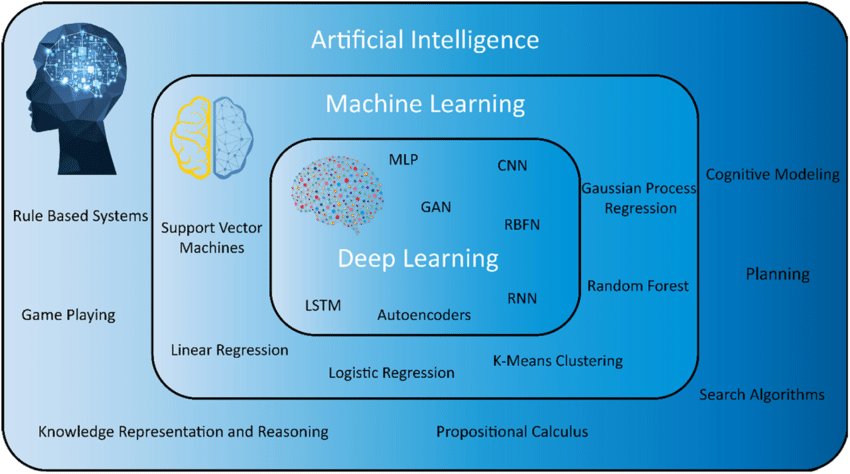
Những năm gần đây. AI nổi lên như một bằng chứng của cuộc cách mạng công nghiệp lần thứ tư. Tri tuệ nhân tạo có thể được định nghĩa như một ngành của khoa học máy tính liên quan đến việc tự động hóa các hành vi thông minh. Trí tuệ nhân tạo là một bộ phận của khoa học máy tính và do đó nó phải được đặt trên những nguyên lý lý thuyết vững chắc, có khả năng ứng dụng được của lĩnh vực này. Ở thời điểm hiện tại, thuật ngữ này thường dùng để nói đến các máy tính có mục đích không nhất định và ngành khoa học nghiên cứu về các lý thuyết và các ứng dụng của trí tuệ nhân tạo.

Theo đà phát triển của công nghệ, ứng dụng trí tuệ nhân tạo luôn là xu hướng công nghệ tương lai mà các hãng công nghệ trên toàn thế giới đua nhau sáng tạo, nó là nền tảng cốt lõi của cuốc cách mạng công nghệ 4.0.

ML (Machine Learning) là một lĩnh vực của trí tuệ nhân tạo, được sinh 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 đó.

Hầu hết mọi 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 dạng thời gian thực - sẽ giúp các tổ chức vận hành hiệu quả hơn hoặc tạo lợi thế cạnh tranh so với các đối thủ.

Các ứng dụng của ML đã quá quen thuộc với con người: xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trên Facebook, hệ thống gợi ý sản phẩm của Amazon, hệ thống gợi ý phim của Netflix..., chỉ là một vài trong vô vàn những ứng dụng của trí tuệ nhân tạo và cụ thể là ML.



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

Xu hướng phát triển công nghệ thông tin ngày càng tăng, song song với nó lượng dữ liệu được sinh ra cũng ngày một lớn. Vì vậy nhu cầu để xử lý dữ liệu cũng lớn hơn, ML đang góp phần giải quyết vấn đề này. Một trong những thuật toán thường dùng trong những thuật toán thường dùng trong ML là thuật toán K-nearest neighbor.

Ứng dụng của thuật toán này được sử dụng rất nhiều và 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" là sự kết hợp đầy hứa hẹn giữa tính linh hoạt của thuật toán k-NN và tiềm năng mở rộng trong lĩnh vực dự báo. Thuật toán k-NN không chỉ phổ biến mà còn mang tính ứng dụng cao trong nhiều lĩnh vực khác nhau. Áp dụng k-NN vào việc dự báo kết quả tốt nghiệp của sinh viên trên dữ liệu đồ thị có thể cung cấp thông tin quan trọng và chính xác hơn.

## Tổng quan nghiên cứu

Đề tài tập trung vào việc áp dụng thuật toán k-NN (k-Nearest Neighbors) trong bài toán dự báo trên dữ liệu đồ thị. Nghiên cứu chủ yếu là triển khai 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 thử 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.

## Nội dung nghiên cứu

Tìm hiểu thuật toán K-NN.

Xây dựng bộ dữ liệu.

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

## Mục tiêu của đề tài

Triển khai 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.

## 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 khai và thử nghiệm
* Kết luận

# CƠ SỞ LÝ THUYẾT

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

### Định nghĩa

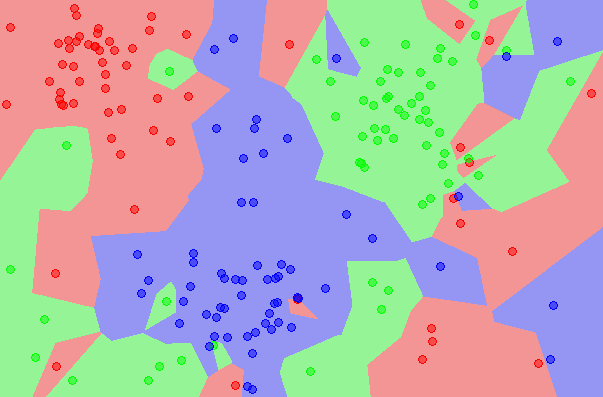
k-NN là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần phân lớp và tất cả các đối tượng trong dữ liệu huấn luyện.

### Nguyên lý hoạt động

Với KNN, trong bài toán Classification, label của một điểm dữ liệu mới (hay kết quả của câu hỏi trong bài thi) được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set. Label của một test data có thể được quyết định bằng major voting (bầu chọn theo số phiếu) giữa các điểm gần nhất, hoặc nó có thể được suy ra bằng cách đánh trọng số khác nhau cho mỗi trong các điểm gần nhất đó rồi suy ra label.

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

Một cách ngắn gọn, KNN 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 KNN trong classification với K = 1.

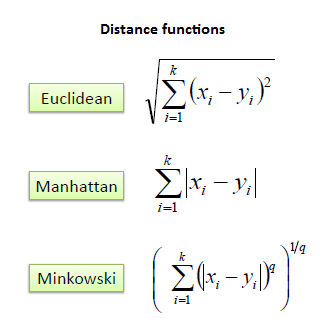


*Hình 2: Bản đồ của 1NN*

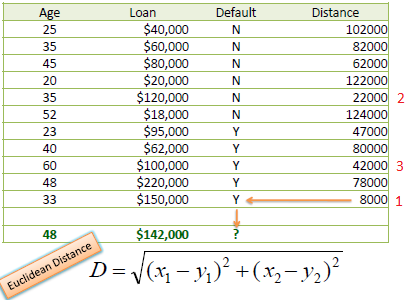
Ví dụ trên đây là bài toán Classification với 3 classes: Đỏ, Lam, Lục. Mỗi điểm dữ liệu mới (test data point) sẽ được gán label theo màu của điểm mà nó thuộc về. Trong hình này, có một vài vùng nhỏ xem lẫn vào các vùng lớn hơn khác màu. Ví dụ có một điểm màu Lục ở gần góc 11 giờ nằm giữa hai vùng lớn 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 sẽ có nhiều khả năng cho kết quả không chính xác.

## Khoảng cách trong không gian vector

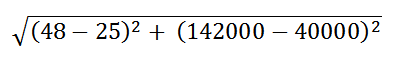
Thuật toán KNN cho rằng những dữ liệu tương tự nhau sẽ tồn tại gần nhau trong một không gian, từ đó công việc của chúng ta là sẽ tìm k điểm gần với dữ liệu cần kiểm tra nhất. Việc tìm khoảng cách giữa 2 điểm củng có nhiều công thức có thể sử dụng, tùy trường hợp mà chúng ta lựa chọn cho phù hợp. Đây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



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

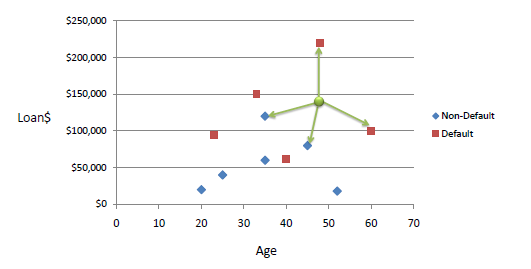


Dữ liệu cần phân loại của chúng ta là {age: 48, loan: 142000}. Đây dữ liệu 2 chiều và chúng ta cần dự đoán người này thuộc nguy cơ vở nợ hay không. Chúng ta sẽ dùng một cách khá phổ biến để tính khoảng cách là Euclidean. Ví dụ ở hàng đầu tiên khoảng cách sẽ được tính:



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



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.

## Tăng tốc cho thuật toán

Để tăng tốc độ cho thuật toán k-Nearest Neighbors (k-NN), có thể áp dụng một số kỹ 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.

## Ưu điểm, 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.

- Dễ sử dụng và cài đặt.

* Nhược điểm:

- KNN rất nhạy cảm với nhiễu khi K nhỏ.

- 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.

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

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

*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 đồ thị được biểu thị bằng 2 thông số là GPA và Học kỳ. Điểm dữ liệu hay câu hỏi sinh viên tốt nghiệp loại nào. 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 xuat sac, gioi, kha, trung binh. 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

Trong phần này, chúng ta sẽ tách 250 dữ liệu trong csv dataset ra thành hai phần, gọi là training set và test set. Thuật toán sẽ dựa vào thông tin ở training set để dự đoán xem mỗi dữ liệu trong test set tương ứng với kết quả tốt nghiệp nào. Dữ liệu được chuẩn đoán này sẽ được đối chiếu với kết quả tốt nghiệp thật của mỗi dữ liệu trong test set để đánh giá hiệu quả của K-NN.

Trainingset

Dữ liệu

Tập các kết quả dự đoán

K-nearest neighbor

Testset

*Hình 3. Mô hình bài toán k-nn*

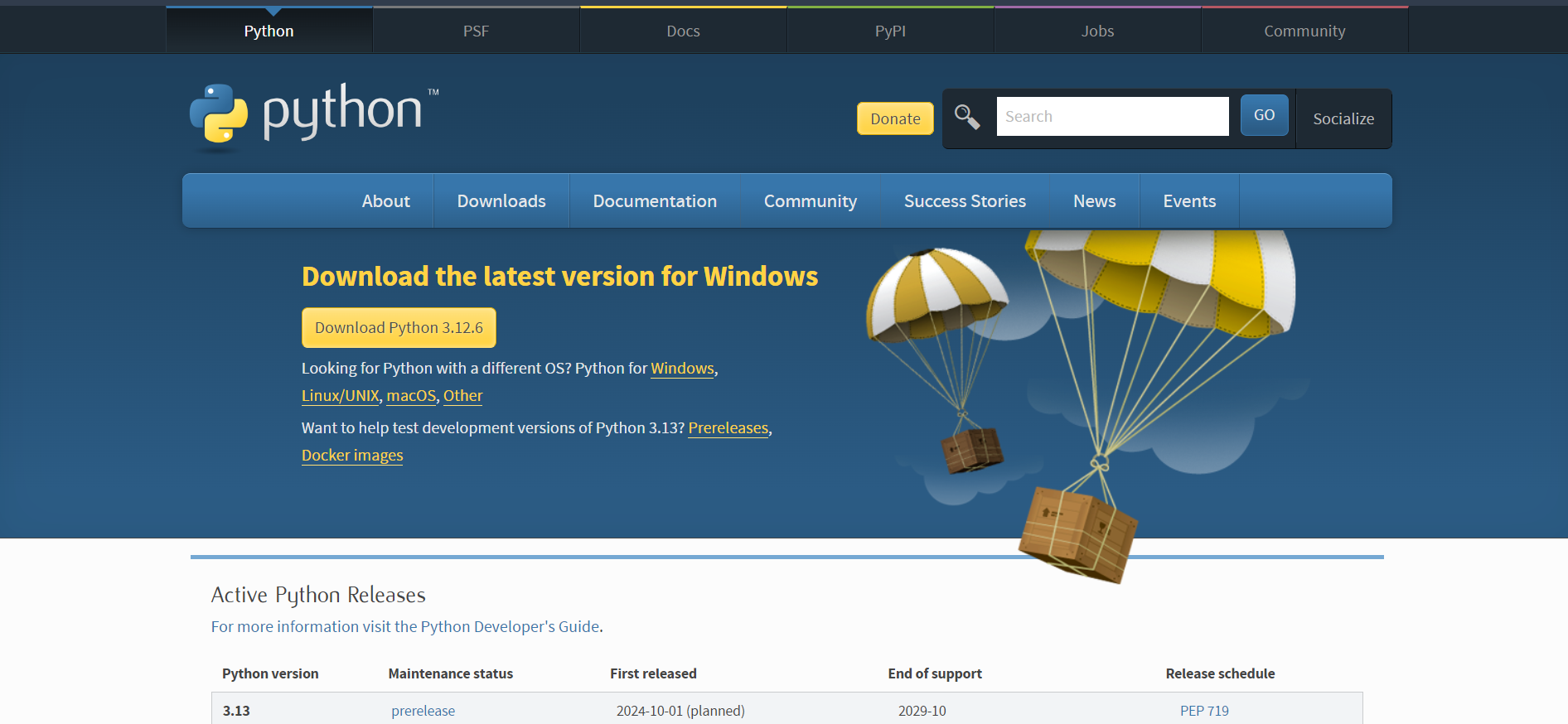
# TRIỂN KHAI VÀ THỬ NGHIỆM

## 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.



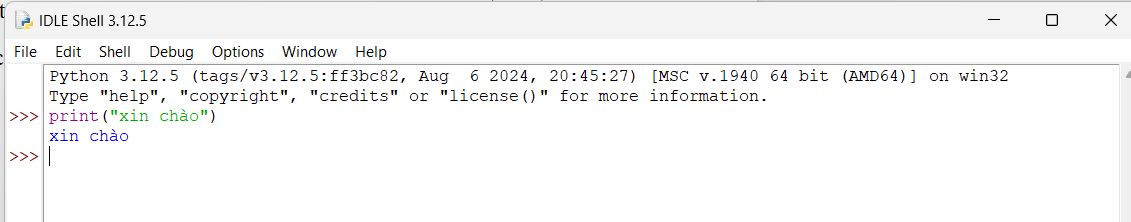
Lúc đó xuất hiện trang có chứa các phiên bản của Python, tải về phiên bản 3.12.6

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.

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.

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 khai, ví dụ như tải các IDE cần thiết, 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 thử nghiệm và đánh giá

Khai báo các thư viện cần thiết:

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:

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:

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)

KNN dự đoán:

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 hớp K=10, tức là với mỗi điểm dữ liệu test ta xét 10 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:

        knn = kNearestNeighbor(trainSet, item, 5)

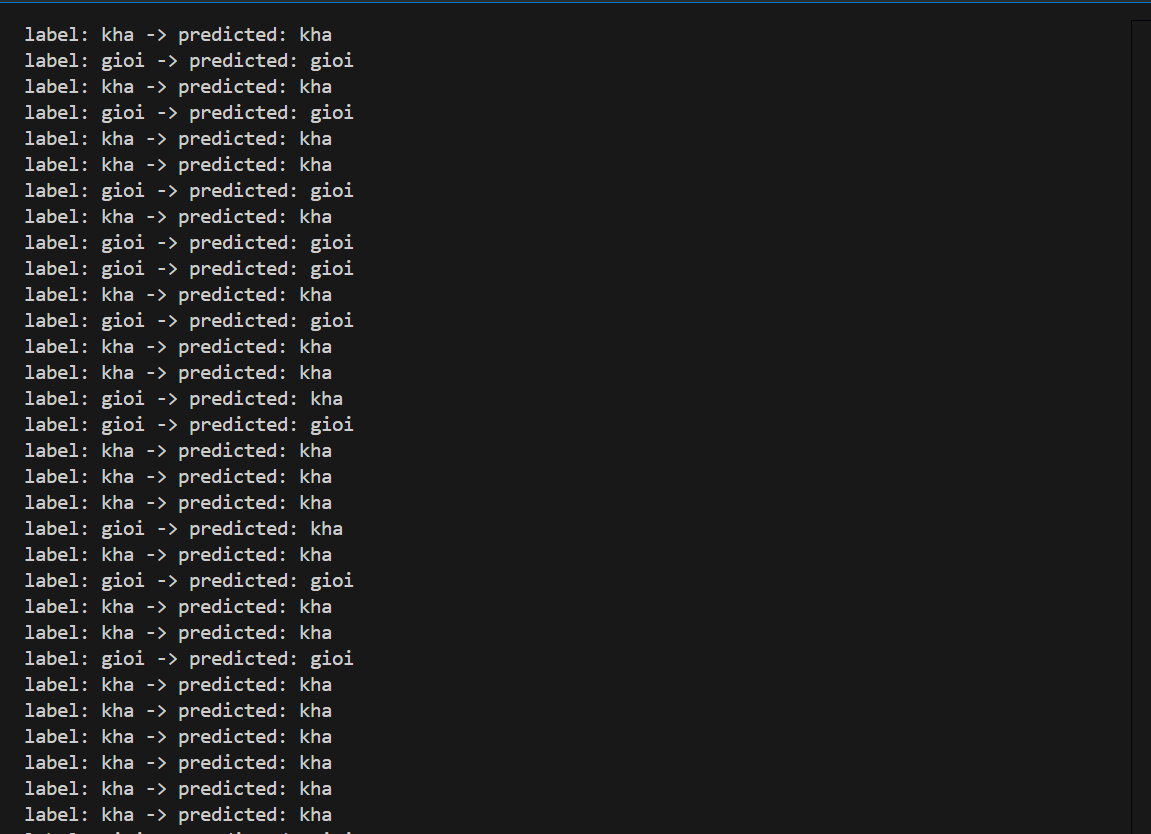
        answer = findMostOccur(knn)

        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:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Lần test | 1 | 2 | 3 | 4 | 5 | 6 |
| Hiệu xuất | 78.125% | 84.375% | 81.25% | 71.875% | 90.625% | 75% |
| Lần test | 7 | 8 | 9 | 10 | Trung bình | |
| Hiệu xuất | 75% | 75% | 78.125% | 84.375% | 79.375% | |

- Với K = 10 cho ra hiệu xuất dự đoán chính xác giao động trong khoảng 70% đến 90%

# KẾT LUẬN

Trong quá trình thực hiện đồ án tốt nghiệp, em đã cố gắng để tìm hiểu và học hỏi nhưng vì khả năng còn giới hạn không tránh khỏi những sai sót, nên có thể chưa giải quyết được tất cả những vấn đề, đặt ra. Em rất mong nhận được sự thông cảm của quý thầy cô và các bạn. Em xin chân thành cảm ơn.

Những kết quả đạt được:

Sự hiểu biết cơ bản về thuật toán K.

Làm quen với bài toán dự đoán kết quả tốt nghiệp.

Từ những gì đã làm có được sự hiểu biết nhất định về AI, ứng dụng của nó vào đời sống công nghệ hiện đại.

Làm quen với ngôn ngữ lập trình Python

Hạn chế:

Thuật toán phụ thuộc nhiều vào hệ số K.

Kết quả đưa ra có sự thay đổi (vì các điểm xét lấy ngẫu nhiên).

Hiệu xuất của bài toán chỉ đạt mức tốt(70-90%), chưa đạt được đến mức cao ( trên 90% hoặc tiệm cận 100%)

Chưa thự sự hiểu hết về bài toán.

# TÀI LIỆU THAM KHẢO

[1] <https://machinelearningcoban.com/2017/01/08/knn/>

[2] <https://en.wikipedia.org/wiki/Iris_flower_data_set>

[3] [https://machinelearningcoban.com/](https://machinelearningcoban.com/math/" \l "norm0)

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

[5] https://codelearn.io/sharing/thuat-toan-k-nearest-neighbors-knn