**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI TP. HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN TỐT NGHIỆP**

**ĐỀ TÀI: NGHIÊN CỨU THUẬT TOÁN K-NEAREST NEIGHBORS VÀ ỨNG DỤNG VÀO BÀI TOÁN DỰ ĐOÁN LY HÔN**

Giảng viên hướng dẫn: TRẦN PHONG NHÃ

Sinh viên thực hiện: TRẦN XUÂN LÂM

Lớp: CÔNG NGHỆ THÔNG TIN

Khoá:57

TP. Hồ Chí Minh, tháng 08 năm 2020

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**PHÂN HIỆU TẠI TP. HỒ CHÍ MINH**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN TỐT NGHIỆP**

**ĐỀ TÀI: NGHIÊN CỨU THUẬT TOÁN K-NEAREST NEIGHBORS VÀ ỨNG DỤNG VÀO BÀI TOÁN DỰ ĐOÁN LY HÔN**

Giảng viên hướng dẫn: TRẦN PHONG NHÃ

Sinh viên thực hiện: TRẦN XUÂN LÂM

Lớp: CÔNG NGHỆ THÔNG TIN

Khoá:57

TP. Hồ Chí Minh, tháng 08 năm 2020

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI **CỘNG HÒA XÃ HỘI CHỦ NGHIÃ VIỆT NAM**

**PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH** Độc lập – Tự do – Hạnh phúc

# NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP

BỘ MÔN: **CÔNG NGHỆ THÔNG TIN**

-------\*\*\*-------

**Mã sinh viên:** 5751071021 **Họ tên SV:** Trần Xuân Lâm

**Khóa:** 57 **Lớp:** Công Nghệ Thông Tin

1. **Tên đề tài.**

***“NGHIÊN CỨU THUẬT TOÁN K-NEAREST NEIGHBORS VÀ ỨNG DỤNG VÀO BÀI TOÁN DỰ ĐOÁN LY HÔN”.***

1. **Mục đích, yêu cầu.**
   1. **Mục đích.**

* Xây dựng một công cụ dự đoán, dự báo nhằm phục vụ cho các cá nhân, tổ chức, doanh nghiệp có nhu cầu trong việc xác định tình trạng hôn nhân.
  1. **Yêu cầu.**
* Tìm hiểu về Machine Learning.
* Nghiên cứu về thuật toán K-Nearest Neighbors.
* Xây dựng, thiết kế, đánh giá thuật toán dựa trên tập dữ liệu mẫu.
* Ứng dụng thuật toán vào bài toán dự đoán ly hôn.

1. **Nội dung và phạm vi đề tài.**
   1. **Nội dung đề tài.**

* Tổng quan tình hình nghiên cứu liên quan đến đề tài
* Giới thiệu về thuật toán K-Nearest Neighbors.
* Nghiên cứu, phân tích, đánh giá thuật toán K-Nearest Neighbors.
* Ứng dụng thuật toán vào bài toán thực tế, cụ thể là bài toán dự đoán ly hôn.
  1. **Phạm vi đề tài.**
* Nghiên cứu thuật toán K-Nearest Neighbors và ứng dụng thuật toán vào dự đoán ly hôn.

1. **Công nghệ, công cụ và ngôn ngữ lập trình.**
   1. **Công nghệ:** HTML, CSS, Javascript, Bootstrap.
   2. **Công cụ:**
   3. **Ngôn ngữ lập trình:** Sublime Text phiên bản 3.2.2.
2. **Các kết quả chính dự kiến sẽ đạt được và ứng dụng**

* Hoàn chỉnh cuốn báo cáo đề tài.
* Khái quát được tổng quan về Machine Learning.
* Nắm được thuật toán K-Nearest Neighbors và có thể áp dụng được thuật toán cho bất kỳ bài toán nào liên quan.
* Nắm được các ưu, nhược điểm của thuật toán, phương pháp tối ưu cho thuật toán.
* Xây dựng được website demo dự doán ly hôn.

1. **Giáo viên và cán bộ hướng dẫn**

Họ tên: TRẦN PHONG NHÃ

Đơn vị công tác: Bộ môn Công Nghệ Thông Tin – Trường Đại học Giao thông Vận tải phân hiệu tại TP HCM

Điện thoại: 0906761014 Email: tpnha@utc2.edu.vn

|  |  |
| --- | --- |
| **Ngày 15 tháng 08 năm 2020**  **BM Công Nghệ Thông Tin** | **Đã giao nhiệm vụ TKTN**  **Giáo viên hướng dẫn** |
|  | **Trần Phong Nhã** |

Đã nhận nhiệm vụ TKTN

Sinh viên: Trần Xuân Lâm Ký tên:

Điện thoại: 0363131199 Email: trxlamit@gmail.com

# LỜI CẢM ƠN

Lời nói đầu tiên, em xin kính gửi lời cảm ơn chân thành nhất tới Quý thầy cô trong Bộ môn Công Nghệ Thông Tin, cũng như Ban Giám Hiệu Trường Đại học Giao thông Vận tải phân hiệu tại Thành phố Hồ Chí Minh, đã cho phép em thực hiện đề tài tốt nghiệp: ***“Nghiên cứu thuật toán K-Nearest Neighbors và Ứng dụng vào bài toán dự đoán ly hôn”***.

Để hoàn thành nhiệm vụ được giao này, ngoài sự nỗ lực học hỏi không ngừng của bản thân còn có sự hướng dẫn tận tình của thầy giáo **Trần Phong Nhã**, người đã hướng dẫn cho em những hướng đi, truyền đạt cho em những kiến thức, kỹ năng để em có thể hoàn thành đề tài tốt nghiệp này.

Mặc dù đã cố gắng hết sức để hoàn thành đề tài, nhưng chắc chắn rằng sẽ khó tránh khỏi những thiếu sót. Em rất mong nhận được những sự đánh giá, góp ý của Quý thầy cô để em có thể rút ra cho mình những bài học, kinh nghiệm quý báu.

Sau cùng, em cũng không biết nói gì hơn ngoài kính chúc Quý thầy cô trong Bộ môn Công Nghệ Thông Tin và đặc biệt là thầy giáo **Trần Phong Nhã** thật dồi dào sức khỏe và ngày càng gặt hái được nhiều thành công hơn nữa trong cuộc sống cũng như trong sự nghiệp giảng dạy của mình.

Em xin chân thành cảm ơn!

***TP. Hồ Chí Minh, ngày 15 tháng 08 năm 2020***

**Sinh viên thực hiện**

**Trần Xuân Lâm**

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

|  |
| --- |
| ***TP. Hồ Chí Minh, ngày 15 tháng 08 năm 2020***  **Giáo viên hướng dẫn**  **Trần Phong Nhã** |

MỤC LỤC

[NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP 16](#_Toc47827348)

[LỜI CẢM ƠN 18](#_Toc47827349)

[DANH MỤC THUẬT NGỮ 23](#_Toc47827350)

[DANH MỤC BẢNG BIỂU 24](#_Toc47827351)

[DANH MỤC HÌNH ẢNH 25](#_Toc47827352)

[CHƯƠNG 1. TỔNG QUAN. 1](#_Toc47827353)

[1.1. Tính cấp thiết của đề tài. 1](#_Toc47827354)

[1.2. Mục tiêu nghiên cứu. 1](#_Toc47827355)

[1.3. Đối tượng và phạm vi nghiên cứu. 1](#_Toc47827356)

[1.4. Phương pháp nghiên cứu. 2](#_Toc47827357)

[1.5. Ý nghĩa khoa học. 2](#_Toc47827358)

[1.6. Các ứng dụng thực tiễn. 2](#_Toc47827359)

[**1.6.1.** **Dự báo thời tiết.** 2](#_Toc47827360)

[**1.6.2.** **Dự báo chứng khoán.** 3](#_Toc47827361)

[**1.6.3.** **Dự báo trong y tế.** 4](#_Toc47827362)

[**1.6.4.** **Dự báo trong khoa học - môi trường.** 5](#_Toc47827363)

[1.7. Cấu trúc nội dung báo cáo. 5](#_Toc47827364)

[1.8. Kết luận chương 1. 5](#_Toc47827365)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT. 6](#_Toc47827366)

[2.1. Machine Learning. 6](#_Toc47827367)

[**2.1.1.** **Giới thiệu về Machine Learning.** 6](#_Toc47827368)

[**2.1.2.** **Phân nhóm các thuật toán Machine Learning.** 9](#_Toc47827369)

[**2.1.3.** **Các bước thực hiện Machine Learning.** 12](#_Toc47827370)

[**2.1.4.** **Các ứng dụng.** 15](#_Toc47827371)

[2.2. Python. 16](#_Toc47827372)

[**2.2.1.** **Giới thiệu về Python.** 16](#_Toc47827373)

[**2.2.2.** **Đặc điểm của Python.** 16](#_Toc47827374)

[**2.2.3.** **Một số thư viện liên quan.** 17](#_Toc47827375)

[2.3. HTML. 17](#_Toc47827376)

[**2.3.1.** **Giới thiệu về HTML.** 17](#_Toc47827377)

[**2.3.2.** **Vai trò của HTML.** 18](#_Toc47827378)

[2.4. CSS. 18](#_Toc47827379)

[**2.4.1.** **Giới thiệu về CSS.** 18](#_Toc47827380)

[**2.4.2.** **Ưu điểm của CSS.** 19](#_Toc47827381)

[2.5. JavaScript. 19](#_Toc47827382)

[**2.5.1.** **Giới thiệu về JavaScript.** 19](#_Toc47827383)

[**2.5.2.** **Ưu điểm của JavaScript.** 19](#_Toc47827384)

[**2.5.3.** **Nhược điểm của JavaScript.** 20](#_Toc47827385)

[2.6. Bootstrap. 20](#_Toc47827386)

[**2.6.1.** **Giới thiệu về Bootstrap.** 20](#_Toc47827387)

[**2.6.2.** **Lý do chọn Bootstrap.** 20](#_Toc47827388)

[**2.6.3.** **Cấu trúc và tính năng của Bootstrap.** 21](#_Toc47827389)

[2.7. Kết luận chương 2. 22](#_Toc47827390)

[CHƯƠNG 3. PHÂN TÍCH – ĐÁNH GIÁ. 23](#_Toc47827391)

[3.1. Sơ lược về K-Nearest Neighbors. 23](#_Toc47827392)

[**3.1.1.** **Giới thiệu về K-Nearest Neighbors.** 23](#_Toc47827393)

[**3.1.2.** **Khoảng cách trong không gian vector**. 25](#_Toc47827394)

[3.2. Phân tích ý tưởng. 25](#_Toc47827395)

[3.3. Ưu và nhược điểm của K-Nearest Neighbors. 26](#_Toc47827396)

[3.4. Ví dụ minh hoạ. 26](#_Toc47827397)

[3.5. Kết luận chương 3. 28](#_Toc47827398)

[CHƯƠNG 4. ỨNG DỤNG THUẬT TOÁN K-NEAREST NEIGHBORS VÀO BÀI TOÁN DỰ ĐOÁN LY HÔN. 29](#_Toc47827399)

[4.1. Phát biểu bài toán. 29](#_Toc47827400)

[4.2. Tập dữ liệu sử dụng. 29](#_Toc47827401)

[4.3. Mô hình hệ thống. 32](#_Toc47827402)

[4.4. Phương pháp tiếp cận. 32](#_Toc47827403)

[4.5. Xây dựng thuật toán. 32](#_Toc47827404)

[**4.5.1.** **Thiết kế.** 32](#_Toc47827405)

[**4.5.2.** **Thực thi thuật toán.** 35](#_Toc47827406)

[4.6. Xây dựng website demo. 35](#_Toc47827407)

[**4.6.1.** **Sơ đồ chức năng.** 35](#_Toc47827408)

[**4.6.2.** **Giao diện website.** 35](#_Toc47827409)

[4.7. Kết luận chương 4. 37](#_Toc47827410)

[CHƯƠNG 5. KẾT QUẢ VÀ KIẾN NGHỊ 38](#_Toc47827411)

[5.1. Kết quả đạt được. 38](#_Toc47827412)

[5.2. Kiến nghị. 38](#_Toc47827413)

[TÀI LIỆU THAM KHẢO 39](#_Toc47827414)

# DANH MỤC THUẬT NGỮ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **TỪ VIẾT TẮT** | **TÊN TIẾNG ANH** | **Ý NGHĨA TIẾNG VIỆT** | **GHI CHÚ** |
| 1 | CNTT | Information Technology | Công Nghệ Thông Tin |  |
| 2 | SPAM | Stupid Pointless Annoying Messages | Thư rác |  |
| 3 | AI | Artificial Intelligence | Trí tuệ nhân tạo |  |
| 4 | ML | Machine Learning | Học máy |  |
| 5 | DL | Deep Learning | Học sâu |  |
| 6 | CAD | Computer-Aided Design | Thiết kế dưới sự hỗ trợ của máy tính |  |
| 7 | GUI | Graphical User Interface | Giao diện đồ họa người dùng |  |
| 8 | Numpy | Nummerical Python |  |  |
| 9 | MVC | Model-View-Control |  |  |
| 10 | DSF | Django Software Foundation | Tổ chức phần mềm Django |  |
| 12 | [CMS](https://wagtail.io/) | Content Management System | Hệ thống quản lý nội dung |  |
| 13 | MVT | Model – View – Template |  |  |
| 14 | CLI | Command Line Interface | Giao diện dòng lệnh |  |
| 15 | URL | Uniform Resource Locator | Định vị tài nguyện thống nhất |  |
| 16 | SLR | Simple Linear Regression | Hồi quy tuyến tính đơn giản |  |
| 17 | OLS | Ordinary Least Square | Trung bình bình phương tối thiểu |  |

# DANH MỤC BẢNG BIỂU

[Bảng 3. 1 Khái niệm tương ứng người-máy 24](#_Toc47827658)

# DANH MỤC HÌNH ẢNH

[Hình 1. 1. Một ví dụ về dự báo thời tiết. 3](#_Toc535564883)

[Hình 1. 2. Một ví dụ về dự báo thị trường chứng khoán. 3](#_Toc535564884)

[Hình 1. 3. Hình ảnh quét não có thể dự đoán tự tử. 4](#_Toc535564885)

[Hình 1. 4. Một trạm quan trắc dự báo về môi trường không khí. 5](#_Toc535564886)

[Hình 2. 1. Các nhánh của Machine Learning trong Trí tuệ nhân tạo. 6](#_Toc47775174)

[Hình 2. 2. Mối quan hệ giữa AI, Machine Learning và Deep Learning. 7](#_Toc47775175)

[Hình 2. 3. Lựa chọn thuật toán phù hợp trong Machine Learning. 14](#_Toc47775176)

[Hình 3. 1 Bản đồ 1NN 25](#_Toc47827424)

[Hình 4. 1 Câu lệnh import các thư viện cần dùng. 32](#_Toc47827464)

[Hình 4. 2. Nhập dữ liệu từ file xlsx. 32](#_Toc47827465)

[Hình 4. 3 Lấy các nhãn và 54 đặc điểm. 32](#_Toc47827466)

[Hình 4. 4. Hàm huấn luyện model. 33](#_Toc47827467)

[Hình 4. 5 Đánh giá độ chính xác thuật toán. 33](#_Toc47827468)

[Hình 4. 6. Kết quả khi thực thi thuật toán. 35](#_Toc47827469)

[Hình 4. 7 Giao diện webdemo 35](#_Toc47827470)

[Hình 4. 8. Giao diện khi nhập liệu. 36](#_Toc47827471)

[Hình 4. 9 Giao diện dự đoán “không ly hôn” 36](#_Toc47827472)

[Hình 4. 10 Giao diện dự đoán “Ly hôn” 37](#_Toc47827473)

# CHƯƠNG 1. TỔNG QUAN.

* 1. **Tính cấp thiết của đề tài.**

Trong những năm gần đây, sự phát triển mạnh mẽ của CNTT đã làm cho khả năng thu thập và lưu trữ thông tin của các hệ thống thông tin tăng nhanh một cách chóng mặt. Bên cạnh đó, việc tin học hóa một cách ồ ạt và nhanh chóng các hoạt động sản xuất, kinh doanh cũng như nhiều lĩnh vực hoạt động khác đã tạo ra cho chúng ta một lượng dữ liệu lưu trữ khổng lồ. Sự bùng nổ này đã dẫn tới một yêu cầu cấp thiết là cần có những kỹ thuật và công cụ mới để tự động chuyển đổi lượng dữ liệu khổng lồ kia thành các tri thức có ích. Từ đó, các kỹ thuật Machine Learning (ML) đã trở thành một lĩnh vực vô cùng nóng bỏng của nền CNTT thế giới hiện nay nói chung và Việt Nam nói riêng.

ML đang ngày càng được áp dụng một cách rộng rãi trong nhiều lĩnh vực khác nhau trong đời sống con người như: marketing, tài chính - ngân hàng, bảo hiểm, khoa học, y tế, an ninh… Rất nhiều tổ chức và công ty lớn trên thế giới đã áp dụng các kỹ thuật ML vào các hoạt động sản xuất kinh doanh của mình và thu được những lợi ích to lớn.

Các tổ chức, doanh nghiệp, tập đoàn bất động sản cũng là một trong số đó. Vì cơ bản thị trường bất động sản là một trong những thị trường quan trọng của nền kinh tế vì liên quan trực tiếp đến một lượng tài sản rất lớn về quy mô, tính chất cũng như giá trị nhiều mặt trong nền kinh tế quốc dân. Và thị trường nhà ở là bộ phận quan trọng nhất của thị trường bất động sản.

Vì tất cả lý do trên em quyết định thực hiện đề tài: ***"Nghiên cứu thuật toán K-Nearest Neighbors trong Machine Learning và Ứng dụng vào bài toán dự đoán ly hôn”*** để làm đề tài đồ án tốt nghiệp.

## **Mục tiêu nghiên cứu.**

Sau khi đã xác định được đề tài, em đặt mục tiêu nghiên cứu các vấn đề sau:

* Nghiên cứu về Machine Learning.
* Nghiên cứu thuật toán K-Nearest Neighbors.
* Sử dụng thuật toán đó để áp dụng vào trong dự đoán ly hôn.

## **Đối tượng và phạm vi nghiên cứu.**

* Đối tượng nghiên cứu: ML và thuật toán K-Nearest Neighbors.
* Phạm vi nghiên cứu của đề tài tập trung vào thuật toán K-Nearest Neighbors.

## **Phương pháp nghiên cứu.**

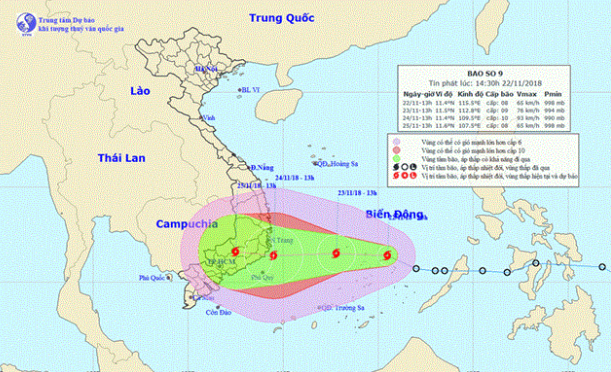
Chủ động tiến hành thu thập và phân tích các tài liệu, thông tin liên quan đến đề tài. Từ đó, lựa chọn phương hướng giải quyết vấn đề, tìm hiểu thuật toán và ứng dụng của thuật toán.

* 1. **Ý nghĩa khoa học.**

Việc xây dựng được các hệ thống dự báo, dự đoán tốt giúp ích rất nhiều cho các cá nhân, tổ chức, doanh nghiệp.

* Đối với các cá nhân, những dự báo chính xác sẽ giúp cải thiện đời sống vật chất cũng như tinh thần, giúp con người có cái nhìn đa chiều hơn về sự vật, hiện tượng đó.
* Đối với các doanh nghiệp, nếu công tác dự báo được thực hiện một cách nghiêm túc, chính xác sẽ tạo điều kiện nâng cao khả năng cạnh tranh trên thị trường, đồng thời giảm thiểu những rủi ro, tổn thất không mong muốn.
* Đối với các tổ chức xã hội, những dự báo chính xác sẽ là căn cứ để các nhà lãnh đạo có thể đưa ra các chính sách phát triển kinh tế, văn hoá, xã hội phù hợp nhằm đưa nền kinh tế ngày càng phát triển.
  1. **Các ứng dụng thực tiễn.**
     1. **Dự báo thời tiết.**

Nhắc đến lĩnh vực ứng dụng các mô hình dự báo thì chắc chắn thời tiết sẽ là một trong những lĩnh vực đầu tiên được nhắc tới. Đây là một ngành khoa học rất quan trọng và thiết yếu bởi vì nó cung cấp thông tin nhằm bảo vệ cuộc sống con người, giảm thiểu tối đa rủi ro về tài sản, an toàn, tính mạng của con người,... Công tác dự báo thời tiết được thực hiện bằng cách thu thập số liệu về trạng thái hiện tại của bầu khí quyển và áp dụng những hiểu biết về khoa học để tiên đoán sự thay đổi của khí quyển.



*Hình 1. 1. Một ví dụ về dự báo thời tiết.[[1]](#footnote-1)*

* + 1. **Dự báo chứng khoán.**

Thị trường chứng khoán là một trong những thị trường biến động bậc nhất từ trước nay. Đây là một lĩnh vực vô cùng đa dạng và phong phú, nhưng cũng rất phức tạp. Và hiện nay, với sự phát triển của khoa học công nghệ, việc dự đoán những biến động về giá chứng khoán không còn là vấn đề lớn nữa.



*Hình 1. 2. Một ví dụ về dự báo thị trường chứng khoán.[[2]](#footnote-2)*

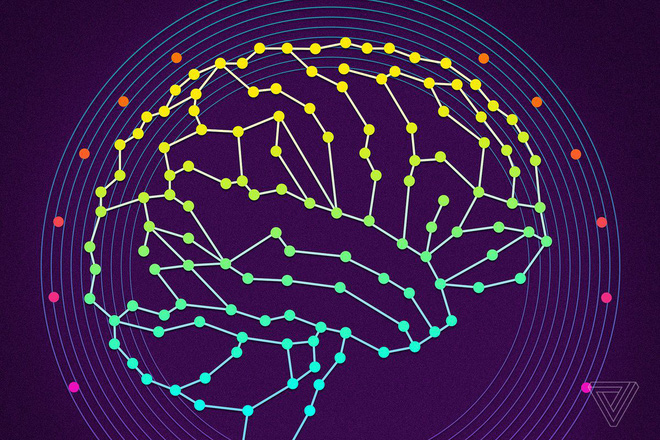
* + 1. **Dự báo trong y tế.**

Thực tế, có khoảng gần 1 triệu người trên thế giới tử vong do tự sát mỗi năm, dự đoán hành vi xu hướng này vẫn còn gặp rất nhiều khó khăn, đặc biệt là khi con người không cảm thấy thoải mái để chia sẻ vấn đề này.3

Theo đó, nghiên cứu được thực hiện bởi các nhà khoa học tại Đại học Carnegie Mellon. Các tác giả nghiên cứu đã tạo ra một chương tình máy tính nhằm chẩn đoán ý định tự tử dự trên các cuộc nghiên cứu về não bộ. Mặc dù được thực hiện với quy mô nhỏ, nhưng kết quả này có thể được sử dụng để chẩn đoán tình trạng bệnh nhân tâm thần.3

Nghiên cứu này được công bố trên tạp chí Truyền thông tự nhiên, các nhà khoa học đã quan sát hoạt động của não bộ ở 37 người được chia làm 2 nhóm. Trong đó họ phân ra một nhóm có ý định tự tử và một nhóm không.3

Sau đó, họ đưa dữ liệu này vào một thuật toán để phân tích những người có ý định tự sát với độ chính xác 90%. Thuật toán này cũng dự đoán thành công 94% những ai đã từng cố tự kết liễu đời mình.3



*Hình 1. 3. Hình ảnh quét não có thể dự đoán tự tử.[[3]](#footnote-3)*

* + 1. **Dự báo trong khoa học - môi trường.**

Các phân tích, thống kê, khảo sát về các sinh vật sinh sống trong các sông, suối, ao, hồ là các cơ sở cho phép đưa ra các dự đoán về môi trường mà các khu dân cư, khu công nghiệp, nhà máy trong phạm vi xung quanh đó tác động tới hệ sinh thái dưới nước.



*Hình 1. 4. Một trạm quan trắc dự báo về môi trường không khí.[[4]](#footnote-4)*

## **Cấu trúc nội dung báo cáo.**

Đồ án được chia thành 5 chương như sau:

Chương 1: Tổng quan.

Chương 2: Cơ sở lý thuyết.

Chương 3: Phân tích – Đánh giá.

Chương 4: Ứng dụng thuật toán K-Nearest Neighbors vào bài toán dự đoán ly hôn.

Chương 5: Kết quả và kiến nghị.

* 1. **Kết luận chương 1.**

Trong chương này, em đã giới thiệu khái quát tổng quan những vấn đề mà đề tài nghiên cứu.

# CHƯƠNG 2. CƠ SỞ LÝ THUYẾT.

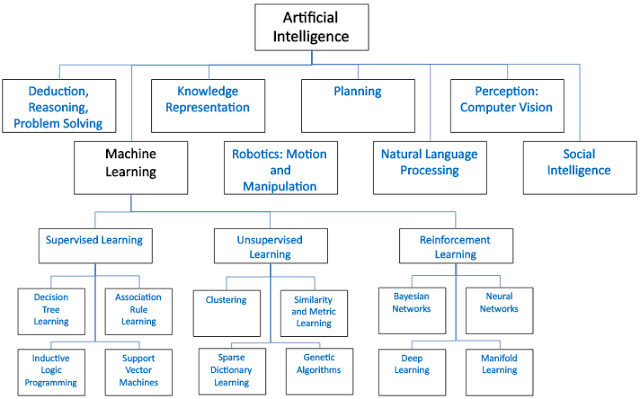
## **Machine Learning.**

* + 1. **Giới thiệu về Machine Learning.**

Những năm gần đây, AI – Trí tuệ nhân tạo nổi lên như một bằng chứng của cuộc cách mạng công nghiệp 4.0. Càng ngày, AI càng len lỏi vào mọi lĩnh vực trong đời sống mà có thể chúng ta không nhận ra.

Xe tự lái của Google và Tesla, hệ thống tự nhận diện 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 trí tuệ nhân tạo.

Và Machine Learning chính là một tập con trong trí tuệ nhân tạo. Nó là một lĩnh vực nhỏ trong ngành khoa học máy tính, có khả năng tự học hỏi dựa tên dữ liệu được đưa vào mà không cần phải lập trình cụ thể.



*Hình 2. 1. Các nhánh của Machine Learning trong Trí tuệ nhân tạo.[[5]](#footnote-5)*

Thực tế gần đây, khi mà khả năng tính toán của các máy tính ngày càng được nâng lên một tầm cao mới, song song với đó là sự bùng nổ của BigData, ML đã tiến thêm một bước dài và một lĩnh vực mới được ra đời gọi là Deep Learning (DL).

DL đã giúp máy tính thực thi những việc tưởng chừng như không thể vào 10 năm trước như: phân loại cả ngàn vật thể khác nhau trong các bức ảnh, tự tạo chú thích cho ảnh, bắt chước giọng nói và chữ viết của con người, giao tiếp với con người, hay thậm chí cả sáng tác văn hay âm nhạc,…



*Hình 2. 2. Mối quan hệ giữa AI, Machine Learning và Deep Learning.[[6]](#footnote-6)*

1. **Định nghĩa Machine Learning.**

Machine Learning là một thuật toán có khả năng học tập từ dữ liệu, có nghĩa là chương trình máy tính sẽ học hỏi từ kinh nghiệm E từ các tác vụ T, với kết quả được đo bằng hiệu suất P. Nếu hiệu suất của nó áp dụng trên tác vụ T khi được đánh giá bởi P và cải thiện theo kinh nghiệm E.[1]

* Ví dụ 1: Giả sử như bạn muốn máy tính xác định một tin nhắn có phải là SPAM hay không? [1]
* Tác vụ T: Xác định 1 tin nhắn có phải SPAM hay không?
* Kinh nghiệm E: Xem lại những tin nhắn đánh dấu là SPAM xem có những đặc tính gì để có thể xác định nó là SPAM.
* Độ đo P: Là phần trăm số tin nhắn SPAM được phân loại đúng.
* Ví dụ 2: Chương trình nhận dạng số (số từ 0 -> 9) [1]
* T: Là nhận dạng được ảnh chứa ký tự số.
* E: Đặc trưng để phân loại ký tự số từ tập dữ liệu số cho trước.
* P: Độ chính xác của quá trình nhận dạng.

1. **Sự hữu ích của Machine Learning.**

Từ lâu đã có nhiều thuật toán ML nổi tiếng nhưng khả năng tự động áp dụng các phép tính phức tạp vào Big Data, lặp đi lặp lại với tốc độ nhanh hơn, chỉ mới phát triển gần đây.[2]

Các ứng dụng của ML đã trở nên quá quen thuộc như: [2]

* Xe tự lái, giảm thiểu tai nạn của Google? Chính là bản chất của ML. [2]
* Các ưu đãi Recommendation Online như của Amazon & Netflix? Ứng dụng của Machine Learning trong cuộc sống hằng ngày. [2]
* Muốn biết người dùng nói gì về bạn trên Twitter? ML kết hợp với sự sáng tạo của quy tắc ngôn ngữ. [2]
* Nhận diện lừa đảo? Một trong những nhu cầu sử dụng hiển nhiên ngày nay. [2]

1. **Đối tượng sử dụng**

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 sâu sắc từ nguồn dữ liệu này, sẽ giúp các tổ chức vận hành hiệu quả hơn hoặc tạo được lợi thế cạnh tranh so với các đối thủ. [2]

* Các dịch vụ tài chính

Ngân hàng và những doanh nghiệp hoạt động trong lĩnh vực tài chính sử dụng công nghệ ML với 2 mục đích chính: xác định insights trong dữ liệu và ngăn chặn lừa đảo. Insights sẽ biết được các cơ hội đầu tư hoặc thông báo đến nhà đầu tư thời điểm giao dịch hợp lý. Khai phá dữ liệu cũng có thể tìm được những khách hàng đang có hồ sơ rủi ro cao hoặc sử dụng giám sát mạng để chỉ rõ những tín hiệu lừa đảo. [2]

* Chính phủ

Các tổ chức chính phủ hoạt động về an ninh cộng đồng hoặc tiện ích xã hội sở hữu rất nhiều nguồn dữ liệu có thể khai thác insights. Ví dụ, khi phân tích dữ liệu cảm biến, chính phủ sẽ tăng mức độ hiệu quả của dịch vụ và tiết kiệm chi phí. ML còn hỗ trợ phát hiện gian lận và giảm thiểu khả năng trộm cắp danh tính. [2]

* Chăm sóc sức khỏe

ML là 1 xu hướng phát triển nhanh chóng trong ngành chăm sóc sức khỏe, nhờ vào sự ra đời của các thiết bị và máy cảm ứng đeo được sử dụng dữ liệu để đánh giá tình hình sức khỏe của bệnh nhân trong thời gian thực. Công nghệ ML còn giúp các chuyên gia y tế xác định những xu hướng hoặc tín hiệu để cải thiện khả năng điều trị, chẩn đoán bệnh. [2]

* Marketing và sales

Dựa trên hành vi mua hàng trước đây, các trang website sử dụng ML phân tích lịch sử mua hàng, từ đó giới thiệu những vật dụng mà bạn có thể sẽ quan tâm và yêu thích. Khả năng tiếp nhận dữ liệu, phân tích và sử dụng những dữ liệu đó để cá nhân hóa trải nghiệm mua sắm hoặc thực hiện chiến dịch Marketing chính là tương tai của ngành bán lẻ. [2]

* Dầu khí

Tìm kiếm những nguồn nguyên liệu mới. Phân tích các mỏ dầu dưới đất. Dự đoán tình trạng thất bại của bộ cảm biến lọc dầu. Sắp xếp các kênh phân phối để đạt hiệu quả và tiết kiệm chi phí. Có thể nói, số lượng các trường hợp sử dụng ML trong ngành công nghiệp này cực kì lớn và vẫn ngày càng mở rộng. [2]

* Vận tải

Phân tích dữ liệu để xác định mô hình và các xu hướng là trọng tâm trong ngành vận tải vì đây là ngành phụ thuộc vào khả năng tận dụng hiệu quả trên mỗi tuyến đường và dự đoán các vấn đề tiềm tàng để gia tăng lợi nhuận. Các chức năng phân tích dữ liệu và modeling của ML đóng vai trò quan trọng với các doanh nghiệp vận chuyển, vận tải công cộng và các tổ chức vận chuyển khác. [2]

### **Phân nhóm các thuật toán Machine Learning.**

Các thuật toán ML thường được chia làm 4 nhóm: [3]

- Supervise learning.

- Unsupervised learning.

- Semi-supervised lerning.

- Reinforcement learning.

* 1. **Supervised Learning (Học có giám sát).**

Supervised learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới dựa trên các cặp (*input, outcome*) đã biết từ trước. Cặp dữ liệu này còn được gọi là (*data, label*), tức (*dữ liệu, nhãn*). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning. [3]

Một cách toán học, Supervised learning là khi chúng ra có một tập hợp biến đầu vào X={x1,x2,…,xN} và một tập hợp nhãn tương ứng Y={y1,y2,…,yN} trong đó xi,yi là các vector. Các cặp dữ liệu biết trước (xi,yi)∈X×Y được gọi là tập *training data* (dữ liệu huấn luyện). Từ tập traing data này, chúng ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập X sang một phần tử (xấp xỉ) tương ứng của tập Y: yi≈f(xi),  ∀i=1,2,…,N Mục đích là xấp xỉ hàm số f thật tốt để khi có một dữ liệu xx mới, chúng ta có thể tính được nhãn tương ứng của nó y=f(x). [3]

**Ví dụ:** trong nhận dạng chữ viết tay, ta có ảnh của hàng nghìn ví dụ của mỗi chữ số được viết bởi nhiều người khác nhau. Chúng ta đưa các bức ảnh này vào trong một thuật toán và chỉ cho nó biết mỗi bức ảnh tương ứng với chữ số nào. Sau khi thuật toán tạo ra một mô hình, tức một hàm số mà đầu vào là một bức ảnh và đầu ra là một chữ số, khi nhận được một bức ảnh mới mà mô hình chưa nhìn thấy bao giờ, nó sẽ dự đoán bức ảnh đó chứa chữ số nào. [3]

Ví dụ này khá giống với cách học của con người khi còn nhỏ. Ta đưa bảng chữ cái cho một đứa trẻ và chỉ cho chúng đây là chữ A, đây là chữ B. Sau một vài lần được dạy thì trẻ có thể nhận biết được đâu là chữ A, đâu là chữ B trong một cuốn sách mà chúng chưa nhìn thấy bao giờ. [3]

Thuật toán supervised learning còn được tiếp tục chia nhỏ ra thành hai loại chính:

**Classification (Phân loại)**

Một bài toán được gọi là classification nếu các label của input data được chia thành một số hữu hạn nhóm. [3]

**Ví dụ**: Gmail xác định xem một email có phải là spam hay không; các hãng tín dụng xác định xem một khách hàng có khả năng thanh toán nợ hay không. Ba ví dụ phía trên được chia vào loại này. [3]

**Regression (Hồi quy)**

Nếu label không được chia thành các nhóm mà là một giá trị thực cụ thể. [3]

**Ví dụ:** một căn nhà rộng x m2, có y phòng ngủ và cách trung tâm thành phố z km sẽ có giá là bao nhiêu? [3]

* 1. **Unsupervised Learning (Học không giám sát).**

Trong thuật toán này, chúng ta không biết được outcome hay nhãn mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm hoặc giảm số chiều của dữ liệu để thuận tiện trong việc lưu trữ và tính toán. [3]

Một cách toán học, Unsupervised learning là khi chúng ta chỉ có dữ liệu vào X mà không biết nhãn Y tương ứng. [3]

Những thuật toán loại này được gọi là Unsupervised learning vì không giống như Supervised learning, chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào. Giống như khi ta học, không có thầy cô giáo nào chỉ cho ta biết đó là chữ A hay chữ B. Cụm không giám sát được đặt tên theo nghĩa này. [3]

Các bài toán Unsupervised learning được tiếp tục chia nhỏ thành hai loại:

**Clustering (phân nhóm)**

Một bài toán phân nhóm toàn bộ dữ liệu XX thành các nhóm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi nhóm.

Ví dụ: phân nhóm khách hàng dựa trên hành vi mua hàng. Điều này cũng giống như việc ta đưa cho một đứa trẻ rất nhiều mảnh ghép với các hình thù và màu sắc khác nhau, ví dụ tam giác, vuông, tròn với màu xanh và đỏ, sau đó yêu cầu trẻ phân chúng thành từng nhóm. Mặc dù không cho trẻ biết mảnh nào tương ứng với hình nào hoặc màu nào, nhiều khả năng chúng vẫn có thể phân loại các mảnh ghép theo màu hoặc hình dạng. [3]

**Association (Suy diễn hội)**

Là bài toán khi chúng ta muốn khám phá ra một quy luật dựa trên nhiều dữ liệu cho trước.

Ví dụ: những khách hàng nam mua quần áo thường có xu hướng mua thêm đồng hồ hoặc thắt lưng; những khán giả xem phim Spider Man thường có xu hướng xem thêm phim Bat Man, dựa vào đó tạo ra một hệ thống gợi ý khách hàng, thúc đẩy nhu cầu mua sắm. [3]

* 1. **Semi-Supervised Learning (Học bán giám sát).**

Các bài toán khi chúng ta có một lượng lớn dữ liệu XX nhưng chỉ một phần trong chúng được gán nhãn được gọi là Semi-Supervised Learning. Những bài toán thuộc nhóm này nằm giữa hai nhóm được nêu bên trên. [3]

Một ví dụ điển hình của nhóm này là chỉ có một phần ảnh hoặc văn bản được gán nhãn (ví dụ bức ảnh về người, động vật hoặc các văn bản khoa học, chính trị) và phần lớn các bức ảnh, văn bản khác chưa được gán nhãn được thu thập từ internet. Thực tế cho thấy rất nhiều các bài toán ML thuộc vào nhóm này vì việc thu thập dữ liệu có nhãn tốn rất nhiều thời gian và có chi phí cao. Rất nhiều loại dữ liệu thậm chí cần phải có chuyên gia mới gán nhãn được. Ngược lại, dữ liệu chưa có nhãn có thể được thu thập với chi phí thấp từ internet. [3]

* 1. **Reinforcement Learning (Học Củng Cố)**

Reinforcement learning là các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao. Hiện tại, Reinforcement Learning chủ yếu được áp dụng vào Lý Thuyết Trò Chơi, các thuật toán cần xác định nước đi tiếp theo để đạt được điểm số cao nhất. [3]

**Ví dụ:** [AlphaGo gần đây nổi tiếng với việc chơi cờ vây thắng cả con người](https://gogameguru.com/tag/deepmind-alphago-lee-sedol/). [Cờ vây được xem là có độ phức tạp cực kỳ cao](https://www.tastehit.com/blog/google-deepmind-alphago-how-it-works/) với tổng số nước đi là xấp xỉ 1076110761, so với cờ vua là 1012010120 và tổng số nguyên tử trong toàn vũ trụ là khoảng 10801080.

Vì vậy, thuật toán phải chọn ra 1 nước đi tối ưu trong số hàng nhiều tỉ tỉ lựa chọn, và tất nhiên, không thể áp dụng thuật toán tương tự như IBM Deep. [3]

Về cơ bản, AlphaGo bao gồm các thuật toán thuộc cả Supervised learning và Reinforcement Learning. Trong phần Supervised Learning, dữ liệu từ các ván cờ do con người chơi với nhau được đưa vào để huấn luyện. Tuy nhiên, mục đích cuối cùng của AlphaGo không phải là chơi như con người mà phải thậm chí thắng cả con người. [3]

Vì vậy, sau khi học xong các ván cờ của con người, AlphaGo tự chơi với chính nó với hàng triệu ván chơi để tìm ra các nước đi mới tối ưu hơn. Thuật toán trong phần tự chơi này được xếp vào loại Reinforcement learning. [3]

* + 1. **Các bước thực hiện Machine Learning.**

Thực hiện Machine Learning bao gồm các bước như sau:

**Thu thập và chuẩn bị dữ liệu**:

Yếu tố ban đầu cần thiết để thực hiện ML là cần có dữ liệu. Dữ liệu có thể được lấy từ nhiều nguồn khác nhau, có thể ít, có thể nhiều, có thể sạch, có thể nhiều dữ liệu lỗi…  
Sau khi thu thập, dữ liệu cần được làm sạch.

**Chọn thành phần**

Với mỗi tập dữ liệu có thể có rất nhiều thành phần, nhưng không phải thành phần nào cũng liên quan tới bài toán mà ta cần giải, việc lựa chọn thành phần (loại bỏ các thành phần không cần thiết) giúp cho việc học của ta trở nên nhanh và hiệu quả hơn. Tuy nhiên, việc lựa chọn đòi hỏi sự thấu hiểu về dữ liệu và bài toán, chủ yếu làm bằng tay và sức người.  
Sau khi chọn được thành phần, nhiều khi ta quay lại bước 1, tiến hành loại bỏ các dữ liệu không liên quan để thu nhỏ tập dữ liệu.

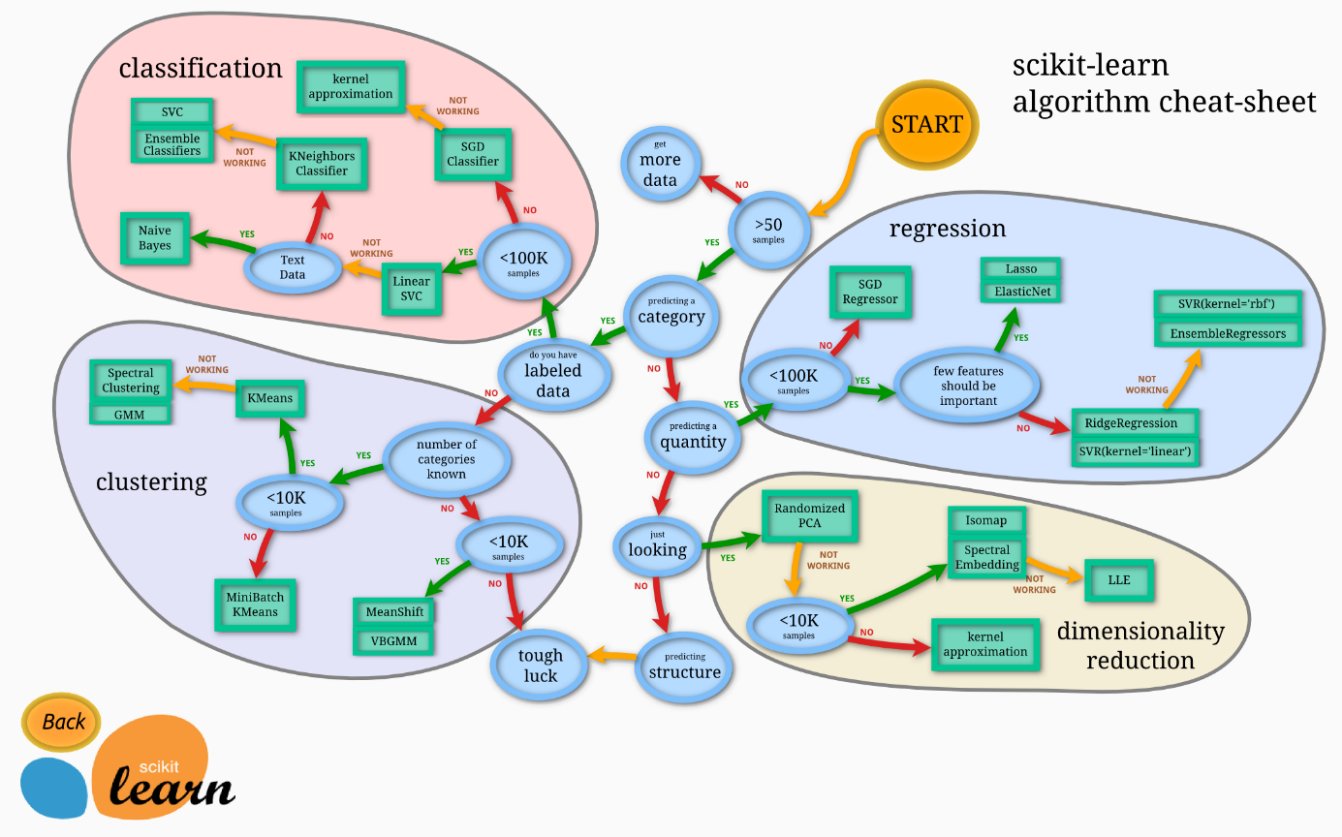
Việc thu nhỏ tập dữ liệu cũng khiến cho việc học của ta tốn ít thời gian hơn, tuy nhiên dữ liệu ít quá cũng khiến việc học đạt độ chính xác không cao, cần cân đối giữa các yếu tố này.

**Chuẩn hoá dữ liệu**

Nhiều khi dữ liệu của từng thành phần có định dạng, kích thước khác biệt lớn, ví dụ thành phần 1 có dữ liệu trong khoảng [0, 1], thành phần 2 có dữ liệu trong khoảng [-1000, 1000], thành phần 3 có dữ liệu là hình ảnh… Để tăng tốc độ và hiệu quả của việc học, ta cần chuẩn hoá dữ liệu, đưa dữ liệu của tất cả các thành phần về cùng một định dạng (số hoá), và cùng khoảng biến thiên.

**Chọn thuật toán**

Tuỳ vào dữ liệu, bài toán mà ta lựa chọn thuật toán ML tương ứng. Đôi khi lựa chọn thuật toán cũng cần dựa trên kinh nghiệm. Một số gợi ý cho việc lựa chọn thuật toán là tham khảo của Scikit-learn.



*Hình 2. 3. Lựa chọn thuật toán phù hợp trong Machine Learning.[[7]](#footnote-7)*

**Chọn parameter cho thuật toán**

Tuỳ mỗi thuật toán mà có nhiều các cách cài đặt khác nhau. Hơn thế nữa các tham số tính toán cũng quyết định không nhỏ tới kết quả tính toán. Vì thế khi sau khi chọn thuật toán thì việc chọn parameter phù hợp với dữ liệu cũng khá quan trọng. Việc chọn Parameter chủ yếu dựa trên kinh nghiệm, nhiều khi có thể sử dụng các thư viện support.

**Huấn luyện và Đánh giá**

Sau khi chọn được thuật toán (một hoặc nhiều thuật toán) và parameter tương ứng ta cho dữ liệu vào train. Tiến hành cross-validation để điều chỉnh model.

Sau khi train được model, ta đưa data test vào kiểm tra và đánh giá độ chính xác của model vừa train được.

**Phân chia dữ liệu**

Thông thường với tập data cho trước ta thường chia làm 3 tập để sử dụng với mục đích khác nhau:

* **Tập huấn luyện** (Training set): dùng để huấn luyện model.
* **Tập kiểm chứng** (Validation set): dùng để đánh giá, điều chỉnh model. Ví dụ như ta dùng tập huấn luyện cho nhiều thuật toán khác nhau rồi dùng tập kiểm chứng để chọn thuật toán phù hợp nhất. Hoặc tìm parameter phù hợp nhất cho một thuật toán cụ thể.
* **Tập kiểm tra** (Test set): dùng để đánh giá model sau khi huấn luyện. Mô hình đã được đánh giá bằng tập test phải là mô hình cuối cùng, không được thay đổi nữa.

Việc phân chia tỉ lệ giữa 3 tập này được khuyên là 60% - 20% - 20%.

Nếu model cuối cùng thu được sau khi huấn luyện và kiểm chứng cần phải được đánh giá bằng tập kiểm tra. Nếu kết quả không tốt, cần thực hiện huấn luyện lại từ đầu.

* + 1. **Các ứng dụng.**

Có rất nhiều ứng dụng mang tính thực tế cao của máy học mà khó có thể kể hết được. Những ứng dụng dưới đây là những ứng dụng phổ biến và được chọn lọc theo góc nhìn cá nhân.

* Nhận diện và phát hiện khuôn mặt: Nhận diện và phát hiện khuôn mặt là ứng dụng khá thú vị của máy học và được áp dụng khá nhiều vào đời sống. Tiêu biểu là tính năng phát hiện khuôn mặt ở máy chụp ảnh. Ứng dụng được phát triển thêm thành phát hiện chớp mắt, phát hiện cười....
* Xe tự lái: Xe tự lái mặc dù phát triển từ đầu thập niên 90 nhưng cho tới nay vẫn còn là vấn đề được nhiều người quan tâm. Các hãng lớn như Google, NVDIA đang nỗ lực để tạo ra một cỗ máy có thể hoàn toàn tự động lái xe và giảm thiểu tai nạn cho con người.
* Phân lớp ảnh: Tìm kiếm ảnh trên Google hiện rất quen thuộc với nhiều người. Ứng dụng của phân lớp ảnh giúp người dùng sử dụng ảnh làm từ khóa tìm kiếm thay thế cho việc tìm kiếm truyền thống trên Google. Bạn upload một ảnh lên và Google sẽ giúp bạn tìm kiếm những thông tin liên quan đến bức ảnh đó.
* Nhận dạng giọng nói: Các trợ lý ảo như Siri, Conrtana hay Google Now là ví dụ điển hình cho nhận dạng giọng nói. Một ví dụ khác nữa là tính năng dịch thuật trực tuyến của Youtube. Với ứng dụng của DL, khả năng dịch thuật chín xác ngôn ngữ từ các video Youtube đang ngày một phát triển vượt bậc.
* Anti-virus: Có thể nhiều người không nghĩ rằng các phần mềm diệt virus lại áp dụng máy học. Tuy nhiên, áp dụng máy học vào để phân tích và dự đoán xu hướng các loại virus sẽ giúp ích rất nhiều trong việc bảo vệ dữ liệu máy tính.
  1. **Python.**
     1. **Giới thiệu về Python.**

Python là một ngôn ngữ lập trình bậc cao, thông dịch, hướng đối tượng, đa mục đích và cũng là một ngôn ngữ lập trình động.

Cú pháp của Python là khá dễ dàng để học và ngôn ngữ này cũng mạnh mẽ và linh hoạt không kém các ngôn ngữ khác trong việc phát triển các ứng dụng. Python hỗ trợ mẫu đa lập trình, bao gồm lập trình hướng đối tượng, lập trình hàm và mệnh lệnh hoặc là các phong cách lập trình theo thủ tục.

Python không chỉ làm việc trên lĩnh vực đặc biệt như lập trình web, và đó là tại sao ngôn ngữ này là đa mục đích bởi vì nó có thể được sử dụng với web, enterprise, 3D CAD, …

Với Python, việc phát triển ứng dụng và debug trở nên nhanh hơn bởi vì không cần đến bước biên dịch và chu trình edit-test-debug của Python là rất nhanh.

* + 1. **Đặc điểm của Python.**

Dưới đây là một số đặc điểm chính của Python:

* Dễ dàng để sử dụng: Python là một ngôn ngữ bậc cao rất dễ dàng để sử dụng. Python có một số lượng từ khóa ít hơn, cấu trúc của Python đơn giản hơn và cú pháp của Python được định nghĩa khá rõ ràng, … Tất cả các điều này là Python thực sự trở thành một ngôn ngữ thân thiện với lập trình viên.
* Bạn có thể đọc code của Python khá dễ dàng. Phần code của Python được định nghĩa khá rõ ràng và rành mạch.
* Python có một thư viện chuẩn khá rộng lớn. Thư viện này dễ dàng tương thích và tích hợp với UNIX, Windows, và Macintosh.
* Python là một ngôn ngữ thông dịch. Trình thông dịch thực thi code theo từng dòng, điều này giúp cho quá trình debug trở nên dễ dàng hơn và đây cũng là yếu tố khá quan trọng giúp Python thu hút được nhiều người học và trở nên khá phổ biến.
* Python cũng là một ngôn ngữ lập trình hướng đối tượng. Ngoài ra, Python còn hỗ trợ các phương thức lập trình theo hàm và theo cấu trúc.

Ngoài các đặc điểm trên, Python còn khá nhiều đặc điểm khác như hỗ trợ lập trình GUI, mã nguồn mở, có thể tích hợp với các ngôn ngữ lập trình khác, …

* + 1. **Một số thư viện liên quan.**
* Numpy là một thư viện không thể thiếu khi chúng ta xây dựng các ứng dụng Máy học trên Python. Numpy cung cấp các đối tượng và phương thức để làm việc với mảng nhiều chiều và các phép toán đại số tuyến tính. Trong numpy, chiều của mảng gọi là axes, trong khi số chiều gọi là rank.
* Pandas là một trong những thư viện được dùng rộng rãi nhất trong Python cùng với Numpy. Pandas cung cấp nhiều đối tượng và phương thức cho các cấu trúc dữ liệu. Pandas là thư viện không thể thiếu cho chúng ta trong suốt quá trình xử lý dữ liệu, từ chuyển đổi hay ánh xạ dữ liệu thô sang dạng dữ liệu mà chúng ta mong muốn, nhằm có thể phân tích dễ dàng hơn.
* Scikit-learn  là một thư viện mã nguồn mở dành cho học máy - một ngành trong trí tuệ nhân tạo, rất mạnh mẽ và thông dụng với cộng đồng Python, được thiết kế trên nền NumPy và SciPy. Scikit-learn chứa hầu hết các thuật toán machine learning hiện đại nhất, đi kèm với documentations, luôn được cập nhật.
  1. **HTML.**
     1. **Giới thiệu về HTML.**

HTML là chữ viết tắt của cụm từ HyperText Markup Language) (dịch là Ngôn ngữ đánh dấu siêu văn bản) được sử dụng để tạo một trang web, trên một website có thể sẽ chứa nhiều trang và mỗi trang được quy ra là một tài liệu HTML (thi thoảng mình sẽ ghi là một tập tin HTML). Cha đẻ của HTML là Tim Berners-Lee, cũng là người khai sinh ra World Wide Web và chủ tịch của World Wide Web Consortium (W3C – tổ chức thiết lập ra các chuẩn trên môi trường Internet).

Một tài liệu HTML được hình thành bởi các phần tử HTML (HTML Elements) được quy định bằng các cặp thẻ (tag), các cặp thẻ này được bao bọc bởi một dấu ngoặc ngọn (ví dụ <html>) và thường là sẽ được khai báo thành một cặp, bao gồm thẻ mở và thẻ đóng (ví <strong> dụ </strong> và ). Các văn bản muốn được đánh dấu bằng HTML sẽ được khai báo bên trong cặp thẻ (ví dụ <strong>Đây là chữ in đậm</strong>). Nhưng một số thẻ đặc biệt lại không có thẻ đóng và dữ liệu được khai báo sẽ nằm trong các thuộc tính (ví dụ như thẻ <img>).

Một tập tin HTML sẽ bao gồm các phần tử HTML và được lưu lại dưới đuôi mở rộng là .html hoặc .htm.

* + 1. **Vai trò của HTML.**

HTML là một ngôn ngữ đánh dấu siêu văn bản nên nó sẽ có vai trò xây dựng cấu trúc siêu văn bản trên một website, hoặc khai báo các tập tin kỹ thuật số (media) như hình ảnh, video, nhạc.

Điều đó không có nghĩa là chỉ sử dụng HTML để tạo ra một website mà HTML chỉ đóng một vai trò hình thành trên website.

HTML – Xây dựng cấu trúc và định dạng các siêu văn bản.

CSS – Định dạng các siêu văn bản dạng thô tạo ra từ HTML thành một bố cục website, có màu sắc, ảnh nền…

Javascript – Tạo ra các sự kiện tương tác với hành vi của người dùng (ví dụ nhấp vào ảnh trên nó sẽ có hiệu ứng phóng to).

Dễ hiểu hơn, bạn hãy nghĩ rằng nếu website là một cơ thể hoàn chỉnh thì HTML chính là bộ xương của cơ thể đó, nó như là một cái khung sườn vậy.

Như vậy, dù website thuộc thể loại nào, giao tiếp với ngôn ngữ lập trình nào để xử lý dữ liệu thì vẫn phải cần HTML để hiển thị nội dung ra cho người truy cập xem.

* 1. **CSS.**

### **Giới thiệu về CSS.**

CSS là chữ viết tắt của Cascading Style Sheets, nó chỉ đơn thuần là một dạng file text với phần tên mở rộng là .css. Trong Style Sheet này chứa những câu lệnh CSS. Mỗi một lệnh của CSS sẽ định dạng một phần nhất định của HTML ví dụ như: font của chữ, đường viền, màu nền, căn chỉnh hình ảnh…

Trước đây khi chưa có CSS, những người thiết kế web phải trộn lẫn giữa các thành phần trình bày và nội dung với nhau. Nhưng với sự xuất hiện của CSS, người ta có thể tách rời hoàn toàn phần trình bày và nội dung. Giúp cho phần code của trang web cũng gọn hơn và quan trọng hơn cả là dễ chỉnh sửa hơn.

CSS được phát triển bởi W3C (World Wide Web Consortium) vào năm 1996, vì một lý do đơn giản. HTML không được thiết kế để gắn tag để giúp định dạng trang web. Bạn chỉ có thể dùng nó để “đánh dấu” lên site.

Những tag như <font> được ra mắt trong HTML phiên bản 3.2, nó gây rất nhiều rắc rối cho lập trình viên. Vì website có nhiều font khác nhau, màu nền và phong cách khác nhau. Để viết lại code cho trang web là cả một quá trình dài, cực nhọc. Vì vậy, CSS được tạo bởi W3C là để giải quyết vấn đề này.

Mối tương quan giữa HTML và CSS rất mật thiết. HTML là ngôn ngữ markup (nền tảng của site) và CSS định hình phong cách (tất cả những gì tạo nên giao diện website), chúng là không thể tách rời.

CSS về lý thuyết không có cũng được, nhưng khi đó website sẽ chỉ là một trang chứa văn bản mà không có gì khác.

### **Ưu điểm của CSS.**

Sự khác biệt giữa site có CSS và không có CSS rất dễ nhận biết.

Trước khi sử dụng CSS, tất cả những phong cách của CSS cần được đính kèm vào trong HTML markup. Có nghĩa là bạn cần tách ra để xác định các thành phần như background, font colors, canh hàng…

CSS giúp định kiểu mọi thứ trên một file khác, bạn có thể tạo phong cách trước rồi sau đó tích hợp file CSS lên trên cùng của file HTML. Việc này giúp HTML markup rõ ràng và dễ quản lý hơn nhiều.

Tóm lại, với CSS bạn không cần lặp lại các mô tả cho từng thanh phần. Nó tiết kiệm thời gian, làm code ngắn lại để bạn có thể kiểm soát lỗi dễ dàng hơn.

* 1. **JavaScript.**

### **Giới thiệu về JavaScript.**

JavaScript là ngôn ngữ lập trình phổ biến nhất trên thế giới trong suốt 20 năm qua. Nó cũng là một trong số 3 ngôn ngữ chính của lập trình web:

* HTML: Giúp bạn thêm nội dung cho trang web.
* CSS: Định dạng thiết kế, bố cục, phong cách, canh lề của trang web.
* JavaScript: Cải thiện cách hoạt động của trang web.

### **Ưu điểm của JavaScript.**

JavaScript có rất nhiều ưu điểm khiến nó vượt trội hơn so với các đối thủ, đặc biệt trong các trường hợp thực tế. Sau đây chỉ là một số lợi ích của JavaScript:

Bạn không cần một compiler vì web browser có thể biên dịch nó bằng HTML.

* Nó dễ học hơn các ngôn ngữ lập trình khác.
* Lỗi dễ phát hiện hơn và vì vậy dễ sửa hơn.
* Nó có thể được gắn trên một số element của trang web hoặc event của trang web như là thông qua click chuột hoặc di chuột tới.
* JS hoạt động trên nhiều trình duyệt, nền tảng…
* Bạn có thể sử dụng JavaScript để kiểm tra input và giảm thiểu việc kiểm tra thủ công khi truy xuất qua cơ sở dữ liệu.
* Nó giúp website tương tác tốt hơn với khách truy cập.
* Nó nhanh hơn và nhẹ hơn các ngôn ngữ lập trình khác.

### **Nhược điểm của JavaScript.**

Mọi ngôn ngữ lập trình đều có các khuyết điểm. Một phần là vì ngôn ngữ đó khi phát triển đến một mức độ như JavaScript, nó cũng sẽ thu hút lượng lớn hacker, scammer, và những người có ác tâm luôn tìm kiếm những lỗ hổng và các lỗi bảo mật để lợi dụng nó. Một số khuyết điểm có thể kể đến là:

* Dễ bị khai thác.
* Có thể được dùng để thực thi mã độc trên máy tính của người dùng.
* Nhiều khi không được hỗ trợ trên mọi trình duyệt.
* Có thể bị triển khai khác nhau tùy từng thiết bị dẫn đến việc không đồng nhất.
  1. **Bootstrap.**

### **Giới thiệu về Bootstrap.**

Bootstrap là một framework bao gồm các HTML, CSS và JavaScript template dùng để phát triển website chuẩn responsive. Bootstrap cho phép quá trình thiết kế website diễn ra nhanh chóng và dễ dàng hơn dựa trên những thành tố cơ bản sẵn có như typography, forms, buttons, tables, grids, navigation, image carousels…

Bootstrap là một bộ sưu tập miễn phí của các mã nguồn mở và công cụ dùng để tạo ra một mẫu webiste hoàn chỉnh. Với các thuộc tính về giao diện được quy định sẵn như kích thước, màu sắc, độ cao, độ rộng…, các designer có thể sáng tạo nhiều sản phẩm mới mẻ nhưng vẫn tiết kiệm thời gian khi làm việc với framework này trong quá trình thiết kế giao diện website.

* + 1. **Lý do chọn Bootstrap.**

Bootstrap rất phổ biến và là một lựa chọn tối ưu trong thiết kế web.

Giữa muôn vàn ứng dụng thiết kế website hiện nay, Bootstrap vẫn có khả năng cạnh tranh cao là nhờ những đặc điểm nổi bật sau:

* Dễ dàng thao tác.

Cơ chế hoạt động của Bootstrap là dựa trên xu hướng mã nguồn mở HTML, CSS và Javascript. Người dùng cần trang bị kiến thức cơ bản 3 mã này mới có thể sử dụng Bootstrap hiệu quả. Bên cạnh đó, các mã nguồn này cũng có thể dễ dàng thay đổi và chỉnh sửa tùy ý.

* Tùy chỉnh dễ dàng.

Bootstrap được tạo ra từ các mã nguồn mở cho phép designer linh hoạt hơn. Giờ đây có thể lựa chọn những thuộc tính, phần tử phù hợp với dự án họ đang theo đuổi. CDN Boostrap còn giúp bạn tiết kiệm dung lượng vì không cần tải mã nguồn về máy.

* Chất lượng sản phẩm đầu ra hoàn hảo.

Bootstrap là sáng tạo của các lập trình viên giỏi trên khắp thế giới. Bootstrap đã được nghiên cứu và thử nghiệm trên các thiết bị. Được kiểm tra nhiều lần trước khi đưa vào sử dụng. Do đó, khi chọn Bootstrap, bạn có thể tin rằng mình sẽ tạo nên những sản phẩm với chất lượng tốt nhất.

* Độ tương thích cao.

Điểm cộng lớn nhất của Bootstrap là khả năng tương thích với mọi trình duyệt và nền tảng. Đây là một điều cực kì quan trọng và cần thiết trong trải nghiệm người dùng. Sử dụng Grid System cùng với hai bộ tiền xử lý Less và Sass, Bootstrap mặc định hỗ trợ Responsive và ưu tiên cho các giao diện trên thiết bị di động hơn. Bootstrap có khả năng tự động điều chỉnh kích thước trang website theo khung browser. Mục đích để phù hợp với màn hình của máy tính để bàn, tablet hay laptop.

* + 1. **Cấu trúc và tính năng của Bootstrap.**

Cấu trúc gọn nhẹ khiến chức năng của Bootstrap trở nên linh hoạt.

Bootstrap chứa các tập tin JavaScript, CSS và fonts đã được biên dịch và nén lại. Ngoài ra, Bootstrap được thiết kế dưới dạng các mô-đun. Do đó, dễ dàng tích hợp với hầu hết các mã nguồn mở như WordPress, Joomla, Magento, …Trong đó, Bootstrap mang đến nhiều chức năng nổi bật.

* Bootstrap cho phép người dùng truy cập vào thư viện “khổng lồ” các thành tố dùng để tạo nên giao diện của một website hoàn chỉnh như font, typography, form, table, grid…
* Bootstrap cho phép bạn tùy chỉnh framework của website trước khi tải xuống và sử dụng nó tại trang web của khung.
* Tái sử dụng các thành phần lặp đi lặp lại trên trang web.
* Bootstrap được tích hợp jQuery. Bạn chỉ cần khai báo chính xác các tính năng trong quá trình lập trình web của bạn.
* Định nghĩa glyphicons nhằm giảm thiểu việc sử dụng hình ảnh làm biểu tượng và tăng tốc độ tải trang.
  1. **Kết luận chương 2.**

Trong chương này, em đã giới thiệu về tổng quan Machine Learning, phân nhóm các loại thuật toán, cũng như trình bày các bước khi thực hiện một dự án với ML. Ngoài ra, em còn giới thiệu về các ngôn ngữ Python, HTML, CSS, JavaScript, Bootstrap và lý do tại sao lại lựa chọn chúng.

# CHƯƠNG 3. PHÂN TÍCH – ĐÁNH GIÁ.

* 1. **Sơ lược về K-Nearest Neighbors.**

Nếu như con người có kiểu học “nước đến chân mới nhảy”, thì trong Machine Learning cũng có một thuật toán như vậy.[6]

* + 1. **Giới thiệu về K-Nearest Neighbors.**

K-nearest neighbor là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning. Khi training, thuật toán này không học một điều gì từ dữ liệu training (đây cũng là lý do thuật toán này được xếp vào loại lazy learning), mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. K-nearest neighbor có thể áp dụng được vào cả hai loại của bài toán Supervised learning là Classification và Regression. KNN còn được gọi là một thuật toán Instance-based hay Memory-based learning.[6]

Để dễ hiểu hơn, hãy thử xem một câu chuyện vui sau:

Có một anh bạn chuẩn bị đến ngày thi cuối kỳ. Vì môn này được mở tài liệu khi thi nên anh ta không chịu ôn tập để hiểu ý nghĩa của từng bài học và mối liên hệ giữa các bài. Thay vào đó, anh thu thập tất cả các tài liệu trên lớp, bao gồm ghi chép bài giảng (lecture notes), các slides và bài tập về nhà + lời giải. Để cho chắc, anh ta ra thư viện và các quán Photocopy quanh trường mua hết tất cả các loại tài liệu liên quan (khá khen cho cậu này chịu khó tìm kiếm tài liệu). Cuối cùng, anh bạn của chúng ta thu thập được một chồng cao tài liệu để mang vào phòng thi.[6]

Vào ngày thi, anh tự tin mang chồng tài liệu vào phòng thi. Aha, đề này ít nhất mình phải được 8 điểm. Câu 1 giống hệt bài giảng trên lớp. Câu 2 giống hệt đề thi năm ngoái mà lời giải có trong tập tài liệu mua ở quán Photocopy. Câu 3 gần giống với bài tập về nhà. Câu 4 trắc nghiệm thậm chí cậu nhớ chính xác ba tài liệu có ghi đáp án. Câu cuối cùng, 1 câu khó nhưng anh đã từng nhìn thấy, chỉ là không nhớ ở đâu thôi.[6]

Kết quả cuối cùng, cậu ta được 4 điểm, vừa đủ điểm qua môn. Cậu làm chính xác câu 1 vì tìm được ngay trong tập ghi chú bài giảng. Câu 2 cũng tìm được đáp án nhưng lời giải của quán Photocopy sai! Câu ba thấy gần giống bài về nhà, chỉ khác mỗi một số thôi, cậu cho kết quả giống như thế luôn, vậy mà không được điểm nào. Câu 4 thì tìm được cả 3 tài liệu nhưng có hai trong đó cho đáp án A, cái còn lại cho B. Cậu chọn A và được điểm. Câu 5 thì không làm được dù còn tới 20 phút, vì tìm mãi chẳng thấy đáp án đâu - nhiều tài liệu quá cũng mệt!![6]

Thuật toán KNN khá giống với cách học/thi của anh bạn kém may mắn kia.

Có một vài khái niệm tương ứng người-máy như sau:

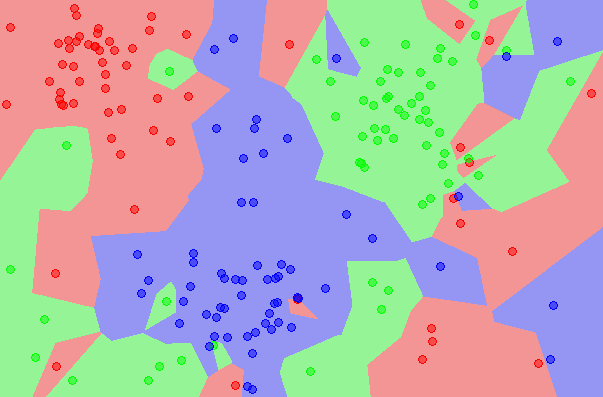
| **Ngôn ngữ người** | **Ngôn ngữ Máy Học** | **in Machine Learning** |
| --- | --- | --- |
| Câu hỏi | Điểm dữ liệu | Data point |
| Đáp án | Đầu ra, nhãn | Output, Label |
| Ôn thi | Huấn luyện | Training |
| Tập tài liệu mang vào phòng thi | Tập dữ liệu tập huấn | Training set |
| Đề thi | Tập dữ liểu kiểm thử | Test set |
| Câu hỏi trong đề thi | Dữ liệu kiểm thử | Test data point |
| Câu hỏi có đáp án sai | Nhiễu | Noise, Outlier |
| Câu hỏi gần giống | Điểm dữ liệu gần nhất | Nearest Neighbor |

*Bảng 3. 1 Khái niệm tương ứng người-máy*

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.[6]

Trong bài toán Regresssion, 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.[6]

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.[6]



*Hình 3. 1 Bản đồ 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.[6]

* + 1. **Khoảng cách trong không gian vector**.

Trong không gian một chiều, khoảng cách giữa hai điểm là trị tuyệt đối giữa hiệu giá trị của hai điểm đó. Trong không gian nhiều chiều, khoảng cách giữa hai điểm có thể được định nghĩa bằng nhiều hàm số khác nhau, trong đó độ dài đường thằng nổi hai điểm chỉ là một trường hợp đặc biệt trong đó.[6]

* 1. **Phân tích ý tưởng.**

Ý tưởng: [7]

Hãy cho tôi biết bạn của bạn là ai, tôi sẽ cho bạn biết bạn là người như thế nào.

Đó là một câu danh ngôn mà có lẽ ai cũng biết. Và có lẽ, phần lớn đều đồng ý. Đây là ý tưởng đằng sau thuật toán KNN. Để dự đoán mẫu dữ liệu mới xnew thuộc về lớp nào, ta dựa vào số k dữ liệu gần nó nhất. Ví dụ k=3, nghĩa là gần xnew có 3 điểm dữ liệu. Giả sử trong đó có 2 điểm dữ liệu thuộc về lớp B và 1 điểm dữ liệu thuộc về lớp A. Như vậy, ta sẽ gán xnew thuộc về lớp B do xnew có 2 điểm dữ liệu thuộc lớp B gần nó nhất.

Thuật toán:[7]

* Thuật toán xác định lớp cho mẫu mới xnew
* Tính khoảng cách giữa xnew và tất cả các mẫu trong tập huấn luyện
* Chọn k mẫu gần nhất với xnew trong tập huấn luyện
* Gán xnew vào lớp có nhiều mẫu nhất trong số k mẫu láng giềng đó (hoặc xnew nhận giá trị trung bình của k mẫu)
  1. **Ưu và nhược điểm của K-Nearest Neighbors.**

Ưu điểm của KNN:[6]

* 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 của KNN:[6]

* 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.
  1. **Ví dụ minh hoạ.**

Ta sẽ sử dụng thuật toán KNN của scikit-learn để phân lớp dữ liệu với tập huấn luyện là Iris. Tập dữ liệu hoa Iris được Ronald Fisher giới thiệu vào năm 1936. Tập dữ liệu này được dùng cho các bài toán phân lớp. Dữ liệu gồm 50 mẫu thu thập từ ba loại hoa Iris (Iris setosa, Iris virginica và Iris versicolor). Bốn thuộc tính cho mỗi mẫu gồm chiều dài và chiều rộng của đài hoa và cánh hoa được tính theo đơn vị centimet.

In [1]:

*# k-Nearest Neighbor*

**from** **sklearn** **import** datasets

**from** **sklearn** **import** metrics

**from** **sklearn.neighbors** **import** KNeighborsClassifier

*# load iris the datasets*

dataset = datasets.load\_iris()

*# 6 dòng dữ liệu đầu tiên của tập dữ liệu Iris*

dataset.data[0:6]

Out[1]:

array([[5.1, 3.5, 1.4, 0.2],

[4.9, 3. , 1.4, 0.2],

[4.7, 3.2, 1.3, 0.2],

[4.6, 3.1, 1.5, 0.2],

[5. , 3.6, 1.4, 0.2],

[5.4, 3.9, 1.7, 0.4]])

Tiếp theo, ta sẽ cài đặt chương trình minh hoạ cho thuật toán KNN.

In [2]:

*# fit a k-nearest neighbor model to the data*

model = KNeighborsClassifier()

model.fit(dataset.data, dataset.target)

print(model)

KNeighborsClassifier(algorithm='auto', leaf\_size=30, metric='minkowski',

metric\_params=None, n\_jobs=1, n\_neighbors=5, p=2,

weights='uniform')

In [3]:

*# make predictions*

expected = dataset.target

predicted = model.predict(dataset.data)

*# summarize the fit of the model*

print(metrics.classification\_report(expected, predicted))

precision recall f1-score support

0 1.00 1.00 1.00 50

1 0.96 0.94 0.95 50

2 0.94 0.96 0.95 50

avg / total 0.97 0.97 0.97 150

In [4]:

print(metrics.confusion\_matrix(expected, predicted))

[[50 0 0]

[ 0 47 3]

[ 0 2 48]]

Qua ví dụ trên, ta thấy thuật toán KNN ứng dụng khá tốt trên tập dữ liệu Iris. KNN cho độ chính xác (precision) trung bình là 97%, (recall) trung bình là 97%, và (f1-score) là 97%. Để cài đặt thuật toán, ta chỉ cần khai báo lớp KNeighborsClassifier và gọi hàm fit() ứng với mô hình phân lớp này.

* 1. **Kết luận chương 3.**

Trong chương này, em đã giới thiệu tổng quan về thuật toán K-Nearest Neighbors, cũng như nêu ra các hạn chế của thuật toán.

# CHƯƠNG 4. ỨNG DỤNG THUẬT TOÁN K-NEAREST NEIGHBORS VÀO BÀI TOÁN DỰ ĐOÁN LY HÔN.

## **Phát biểu bài toán.**

Những năm gần đây, tình trạng ly hôn, li dị ngày một gia tăng ở các gia đình, năm sau cao hơn năm trước, trong đó phần lớn là những người còn trẻ. Theo một nghiên cứu tỷ lệ ly hôn ở Việt Nam đang tăng nhanh và chiếm 31%-40%, hàng năm có hàng chục ngàn trẻ em phải chịu cảnh thiếu cha mẹ do ly hôn. Ly hôn gia tăng đã và đang để lại hệ lụy cho người trong cuộc, cho thế hệ trẻ và cho xã hội.

Có thể nói, ly hôn là sự lựa chọn của đàn ông đàn bà phụ nữ với nhau, nhưng hệ quả của nó có tác động đến tâm sinh lý của những đứa trẻ; để lại gánh nặng cho xã hội nếu như con cái của họ bị bỏ rơi, không được chăm sóc, nuôi dưỡng, giáo dục chu đáo; chúng sẽ thiếu đi sự chăm sóc, tình cảm của người cha hoặc người mẹ, thậm chí cả hai. Từ đó, sẽ ảnh hưởng tới quá trình phát triển nhân cách, dễ sa ngã vào những tệ nạn xã hội... Đây cũng là một trong những lý do mà trong những năm gần đây tình trạng tội phạm tuổi vị thành niên có xu hướng gia tăng.

Kết hôn rồi ly hôn, dễ đến rồi lại dễ đi. Lý do để ra đi thì nhiều lắm thế mà không thể tìm lấy một lý do để ở lại. Khoảnh khắc hai người cùng kí vào tờ đơn ly hôn cũng chính là lúc họ đã tước đi mái ấm, phá bỏ đi hạnh phúc gia đình của chính những đứa con của họ.

Do đó, dự doán là để có giải pháp tránh cho tình trạng ly hôn xảy ra hay còn để cải thiện mối quan hệ.

* 1. **Tập dữ liệu sử dụng.**

Tập dữ liệu lấy từ UCIMachinelearning và nó cung cấp cho bạn tất cả thông tin liên quan cần thiết cho việc dự đoán Ly hôn. Nó chứa 54 đặc điểm và trên cơ sở những đặc điểm này, chúng ta phải dự đoán rằng cặp đôi đã ly hôn hay chưa. Giá trị 1 đại diện cho Đã ly hôn và giá trị 0 đại diện cho không ly hôn. Các đặc điểm gồm:

1. Nếu một trong số chúng tôi xin lỗi khi cuộc thảo luận của chúng tôi xấu đi, cuộc thảo luận kết thúc.
2. Tôi biết chúng ta có thể bỏ qua sự khác biệt của mình, ngay cả khi đôi khi mọi thứ trở nên khó khăn.
3. Khi chúng tôi cần, chúng tôi có thể trao đổi với vợ / chồng của tôi ngay từ đầu và sửa nó.
4. Khi tôi tranh cãi với vợ/chồng tôi, cuối cùng tôi sẽ liên lạc với cô/anh ấy.
5. Thời gian tôi dành cho vợ/chồng thật đặc biệt đối với chúng tôi.
6. Chúng tôi gần như không bao giờ có thời gian ở nhà như là cặp đôi.
7. Chúng tôi giống như hai người xa lạ có chung môi trường ở nhà hơn là gia đình.
8. Tôi tận hưởng kỳ nghỉ của chúng tôi với vợ/chồng tôi.
9. Tôi thích đi du lịch cùng vợ/chồng.
10. Hầu hết các mục tiêu của chúng tôi là chung cho vợ / chồng tôi.
11. Tôi nghĩ rằng một ngày nào đó trong tương lai, khi nhìn lại, tôi thấy rằng vợ / chồng tôi và tôi đã hòa hợp với nhau.
12. Vợ/chồng tôi và tôi có những giá trị tương đồng nhau về quyền tự do cá nhân.
13. Vợ chồng tôi có những trò giải trí giống nhau.
14. Hầu hết các mục tiêu của chúng ta đối với mọi người (trẻ em, bạn bè, v.v.) đều giống nhau.
15. Ước mơ của chúng tôi được sống với vợ/chồng tôi rất giống nhau và hòa hợp.
16. Chúng tôi tương thích với vợ/chồng của tôi về tình yêu nên là gì.
17. Chúng tôi có cùng quan điểm về việc hạnh phúc trong cuộc sống của chúng tôi với vợ/chồng của tôi.
18. Vợ/chồng tôi và tôi có những suy nghĩ giống nhau về cuộc hôn nhân nên như thế nào.
19. Vợ/chồng tôi và tôi có những ý kiến tương tự nhau về vai trò nên có trong hôn nhân.
20. Vợ tôi và tôi có những giá trị tương tự về sự tin tưởng.
21. Tôi biết chính xác những gì vợ/chồng tôi thích.
22. Tôi biết vợ/chồng của tôi muốn được chăm sóc như thế nào khi cô ấy / anh ấy bị ốm.
23. Tôi biết món ăn yêu thích của vợ / chồng tôi.
24. Tôi có thể nói cho bạn biết loại căng thẳng mà vợ / chồng tôi đang phải đối mặt trong cuộc sống của cô ấy / anh ấy.
25. Tôi có kiến ​​thức về thế giới nội tâm của vợ / chồng tôi.
26. Tôi biết những lo lắng cơ bản của vợ / chồng tôi.
27. Tôi biết nguồn căng thẳng hiện tại của vợ / chồng tôi là gì.
28. Tôi biết hy vọng và mong muốn của vợ / chồng tôi.
29. Tôi biết vợ/chồng của mình rất tốt.
30. Tôi biết bạn bè của vợ / chồng tôi và các mối quan hệ xã hội của họ.
31. Tôi cảm thấy hung hăng khi cãi nhau với vợ / chồng.
32. Khi thảo luận với vợ/chồng của tôi, tôi thường sử dụng các thành ngữ như 'Bạn luôn luôn' hoặc 'Bạn không bao giờ'.
33. Tôi có thể sử dụng những tuyên bố tiêu cực về tính cách của vợ / chồng tôi trong các cuộc thảo luận của chúng tôi.
34. Tôi có thể sử dụng các biểu hiện xúc phạm trong các cuộc thảo luận của chúng tôi.
35. Tôi có thể xúc phạm vợ/chồng của mình trong các cuộc thảo luận của chúng tôi.
36. Tôi có thể bị sỉ nhục khi chúng tôi thảo luận.
37. Cuộc tranh cãi của tôi với vợ/chồng của tôi không được bình tĩnh.
38. Tôi ghét cách vợ / chồng của tôi mở một chủ đề.
39. Những trận đánh nhau thường xảy ra đột ngột.
40. Chúng tôi chỉ bắt đầu một cuộc chiến trước khi tôi biết chuyện gì đang xảy ra.
41. Khi tôi nói chuyện với vợ/chồng của mình về điều gì đó, sự bình tĩnh của tôi đột nhiên bị phá vỡ.
42. Khi tôi tranh cãi với vợ, nó chỉ nổ ra và tôi không nói một lời.
43. Tôi chủ yếu muốn làm dịu môi trường một chút.
44. Đôi khi tôi nghĩ rằng thật tốt cho tôi khi tôi rời khỏi nhà một thời gian.
45. Tôi thà im lặng hơn là cãi vã với vợ / chồng tôi.
46. Ngay cả khi tôi đúng trong cuộc tranh luận, tôi khát khao không làm mất lòng bên kia.
47. Khi tôi thảo luận với vợ/chồng của mình, tôi im lặng vì tôi sợ không thể kiềm chế cơn giận của mình.
48. Tôi cảm thấy đúng trong các cuộc thảo luận của chúng tôi.
49. Tôi không liên quan gì đến những gì tôi đã bị buộc tội.
50. Tôi thực sự không phải là người có tội về những gì tôi bị buộc tội.
51. Tôi không phải là người sai về các vấn đề ở nhà.
52. Tôi sẽ không ngần ngại nói với vợ / chồng của tôi về sự bất cập của cô ấy / anh ấy.
53. Khi trao đổi, tôi nhắc nhở cô/anh ấy về những vấn đề chưa thỏa đáng của vợ/chồng tôi.
54. Tôi không ngại nói với vợ / chồng mình về sự kém cỏi của anh ấy / cô ấy.

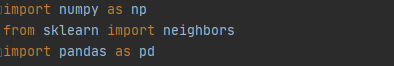
Bản thân tập dữ liệu chứa 170 hàng và 55 cột, tôi đã nhờ thêm 11 người làm khảo sát để tăng thêm số lượng dữ liệu.

## **Mô hình hệ thống.**

* 1. **Phương pháp tiếp cận.**
  2. **Xây dựng thuật toán.**
     1. **Thiết kế.**

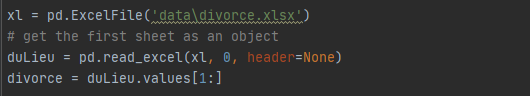
Để xây dựng thuật toán, ở đây ta sử dụng một số thư viện của Python như: numpy, pandas, sklearn,... đã được trình bày ở phần trên.

Đầu tiên, ta import các thư viện cần dùng:



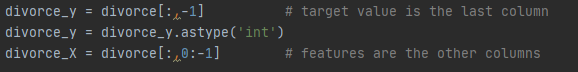
*Hình 4. 1 Câu lệnh import các thư viện cần dùng.*

Tiếp theo, chúng ta nhập dữ liệu vào:



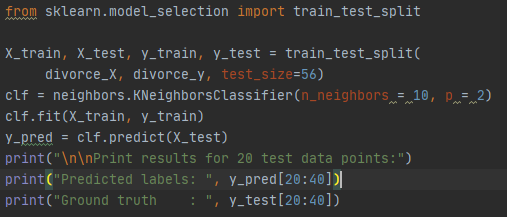
*Hình 4. 2. Nhập dữ liệu từ file xlsx.*

Và rồi, ta tách thuộc tính nhãn và 54 đặc điểm ra riêng:



*Hình 4. 3 Lấy các nhãn và 54 đặc điểm.*

Sau đó, chúng ta viết code huấn luyện model:



*Hình 4. 4. Hàm huấn luyện model.*

Mã trên chạy mô hình kNN trên tập dữ liệu huấn luyện gồm ***X\_train*** đại diện cho tập 54 thuộc tính của tình trạng hôn nhân và ***Y\_train*** đại diện cho kết quả ly hôn hay không ly hôn. Code trả về một giá trị đánh giá điểm của mô hình khi chạy trên tập kiểm tra.

Ý tưởng của chúng ta là sẽ phân chia tập dữ liệu mẫu thành hai tập con là tập dữ liệu huấn luyện và tập dữ liệu kiểm tra.

Đoạn code bên trên phân chia tập dữ liệu thành 3 phần tương ứng ***67%*** cho training và ***33%*** cho testing.

Cuối cùng, ta đánh giá độ chính xác của thuật toán:



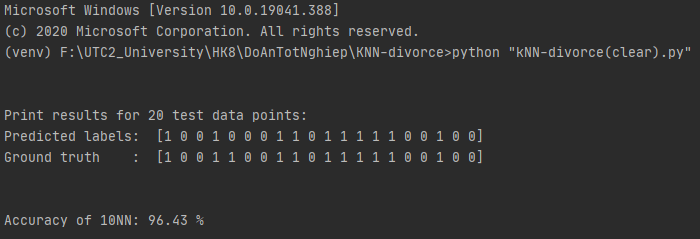
*Hình 4. 5 Đánh giá độ chính xác thuật toán.*

Source code đầy đủ:

import numpy as np  
from sklearn import neighbors  
import pandas as pd  
  
xl = pd.ExcelFile('data\divorce.xlsx')  
# get the first sheet as an object  
duLieu = pd.read\_excel(xl, 0, header=None)  
divorce = duLieu.values[1:]  
  
divorce\_y = divorce[:,-1] # target value is the last column  
divorce\_y = divorce\_y.astype('int')  
divorce\_X = divorce[:,0:-1] # features are the other columns  
  
from sklearn.model\_selection import train\_test\_split  
  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(  
 divorce\_X, divorce\_y, test\_size=56)  
clf = neighbors.KNeighborsClassifier(n\_neighbors = 10, p = 2)  
clf.fit(X\_train, y\_train)  
y\_pred = clf.predict(X\_test)  
print("\n\nPrint results for 20 test data points:")  
print("Predicted labels: ", y\_pred[20:40])  
print("Ground truth : ", y\_test[20:40])  
  
from sklearn.metrics import accuracy\_score  
print("\n\nAccuracy of 10NN: %.2f %%" %(100\*accuracy\_score(y\_test, y\_pred)))

* + 1. **Thực thi thuật toán.**

Kết quả sau khi chạy thuật toán:



*Hình 4. 6. Kết quả khi thực thi thuật toán.*

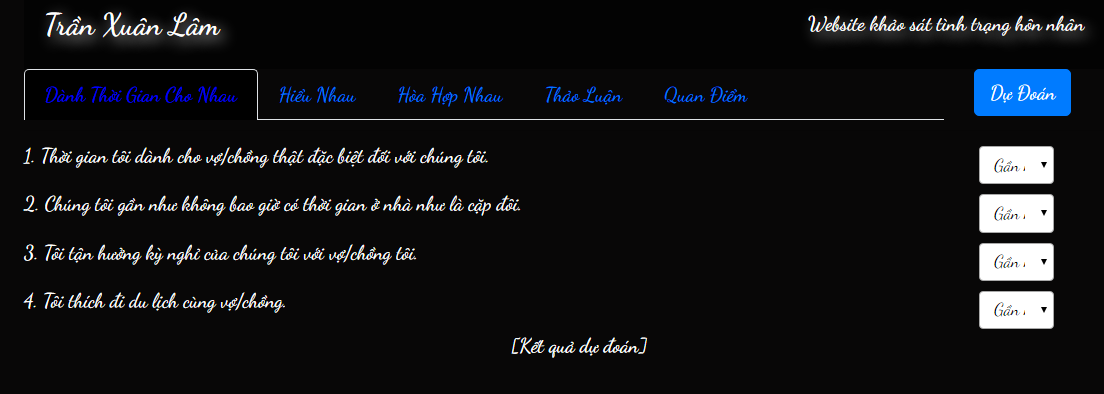
* 1. **Xây dựng website demo.**

Để minh họa cho thuật toán đã xây dựng, trong phần này em sẽ xây dựng một website demo và áp dụng thuật toán đó (được code lại bằng JavaScript) vào trong dự đoán ly hôn. Để làm được điều này em đã sử dụng ngôn ngữ HTML, CSS, JavaScript và một số thư viện mã nguồn mở như Bootstrap.

Sau khi tổng hợp và chia 54 đặc điểm trên thành 5 phần chính: Dành thời gian cho nhau, Hiểu nhau, Hòa hợp nhau, Thảo luận, Quan điểm.

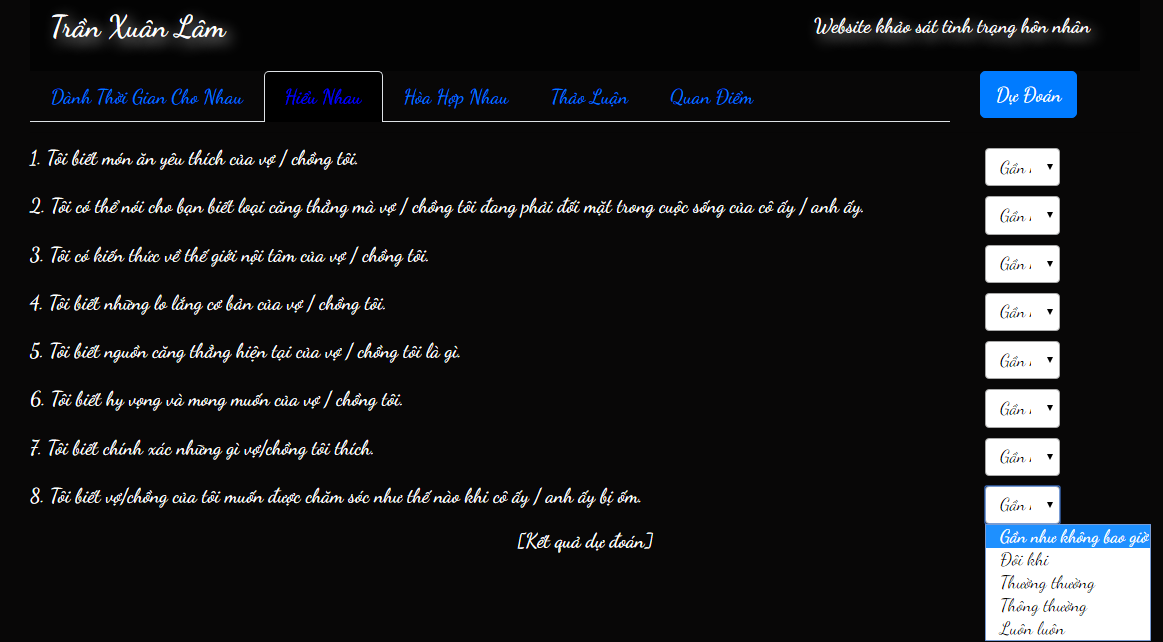
* + 1. **Sơ đồ chức năng.**
    2. **Giao diện website.**

Giao diện trang dự đoán ly hôn:



*Hình 4. 7 Giao diện webdemo*

Giao diện khi nhập dữ liệu và chọn dự đoán:



*Hình 4. 8. Giao diện khi nhập liệu.*

Giao diện dự đoán:



*Hình 4. 9 Giao diện dự đoán “không ly hôn”*



*Hình 4. 10 Giao diện dự đoán “Ly hôn”*

* 1. **Kết luận chương 4.**

Trong chương này, em đã trình bày cách ứng dụng thuật toán, để giải bài toán dự đoán ly hôn ở sử dụng mô hình k-nearest neighbors, dựa trên tập dữ liệu mẫu để tiến hành đào tạo, kiểm tra và đánh giá mô hình.

# CHƯƠNG 5. KẾT QUẢ VÀ KIẾN NGHỊ

## **5.1. Kết quả đạt được.**

- Nắm được các kiến thức nền tảng về Machine Learning – một lĩnh vực đầy tiềm năng trong thời kỳ cách mạng lần thứ 4 (cách mạng 4.0).

- Tìm hiểu được các thuật toán trong Machine Learning, cụ thể là thuật toán K-Nearest Neighbors.

- Nắm được cách xây dựng một mô hình K-Nearest Neighbors.

- Nắm được cách cài đặt thuật toán K-Nearest Neighbors.

- Nắm được ý nghĩa của thuật toán, từ đó có thể áp dụng vào các bài toán khác liên quan.

- Xậy dựng được thuật toán demo.

- Ứng dụng được thuật toán vào cài đặt web demo.

## **5.2. Kiến nghị.**

**-** Dự kiến trong thời gian tới em sẽ cố gắng tối ưu thuật toán hơn nữa, thiết kế ứng dụng thân thiện và dễ sử dụng nhất có thể.

- Bên cạnh đó, em cũng sẽ cố gắng tìm kiếm, trích xuất, sàng lọc ra những bộ dữ liệu tốt nhất, sát với thực tế ở quốc gia mình nhất để có thể đáp ứng nhu cầu của người tra cứu, tìm kiếm, sử dụng ở nước ta.

-Và dùng các dữ liệu từ 54 đặc điểm để có thể tư vấn cho người dùng nên cải thiện từng đặc điểm như thế nào để hôn nhân bền vững hơn.

- Ngoài ra, em sẽ tìm hiểu và nghiên cứu thêm các thuật toán khác để có thể áp dụng vào từng bài toán cụ thể khác nhau.

Vì thời gian thực hiện đề tài có hạn nên trong quá trình làm việc, nghiên cứu không thể tránh khỏi những thiếu sót, rất mong Hội đồng xem xét.

Em xin chân thành cảm ơn !

# TÀI LIỆU THAM KHẢO

1. Nguyễn Hoàng Huy, Phạm Văn Toàn, Hoàng Thị Thanh Giang, Tạp chí Khoa Học Nông nghiệp Việt Nam 2016, tập 14, số 9: 1441-1447, “*Xây dựng thuật toán hiệu quả cho định giá bất động sản quận Long Biên và tỉnh Montreal*”, [*https://www.vnua.edu.vn/*](https://www.vnua.edu.vn/)*.*
2. Tất tần tật về Machine Learning & ứng dụng trong những ngành công nghiệp lớn, [*https://techtalk.vn/tat-tan-tat-moi-kien-thuc-co-ban-ve-machine-learning.html*](https://techtalk.vn/tat-tan-tat-moi-kien-thuc-co-ban-ve-machine-learning.html)*.*
3. Vũ Hữu Tiệp’s Blog, “*Machine Learning cơ bản*”, [*https://machinelearningcoban.com*](https://machinelearningcoban.com)*.*
4. Cáp Hữu Quân’s Blog, [*https://caphuuquan.blogspot.com/2016/05/machine-learning-la-gi.html*](https://caphuuquan.blogspot.com/2016/05/machine-learning-la-gi.html).
5. Hồ Sỹ Hùng, *“Machine Learning là gì & tại sao nó lại quan trọng - Phần 1”,* [*https://techmaster.vn/posts/33836/machine-learning-la-gi-tai-sao-machine-learning-lai-quan-trong*](https://techmaster.vn/posts/33836/machine-learning-la-gi-tai-sao-machine-learning-lai-quan-trong)*.*
6. Vũ Hữu Tiệp’s Blog, “*Machine Learning cơ bản*”, <https://machinelearningcoban.com/2017/01/08/knn/>
7. Ông Xuân Hồng, <https://nbviewer.jupyter.org/github/ongxuanhong/hong_notebooks/blob/master/python/knn.ipynb>
8. Wikipedia, <https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm>

1. http://soha.vn/ap-thap-nhiet-doi-da-manh-len-thanh-bao-so-9-20181122155742767.htm [↑](#footnote-ref-1)
2. https://www.vndirect.com.vn/portal/cong-cu-phan-tich-chung-khoan/bieu-do-ky-thuat.shtml [↑](#footnote-ref-2)
3. http://soha.vn/phuong-phap-moi-co-the-du-doan-duoc-hanh-vi-tu-tu-chinh-xac-90-20171101115125237.htm [↑](#footnote-ref-3)
4. http://moitruong.com.vn/tin-tuc-su-kien/su-kien-moi-truong/ha-noi-dua-vao-hoat-dong-15-tram-quan-trac-moi-truong-18222.htm [↑](#footnote-ref-4)
5. https://caphuuquan.blogspot.com/2016/05/machine-learning-la-gi.html [↑](#footnote-ref-5)
6. https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/ [↑](#footnote-ref-6)
7. https://scikit-learn.org/stable/tutorial/machine\_learning\_map/index.html [↑](#footnote-ref-7)