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 TIỂU LUẬN
TRÍ TUỆ NHÂN TẠO

ĐỀ TÀI: XÂY DỰNG WEBSITE HỖ TRỢ TRA CỨU KIẾN THỨC TOÁN RỜI RẠC

Giảng viên hướng dẫn : NGUYỄN ĐÌNH HIỂN

Sinh viên thực hiện : NGUYỄN MAI NGHIÊM

NGUYỄN SÔNG NGÂN

TRẦN THỊ THANH NGÂN

TRẦN ANH DUY

ĐINH XUÂN GIANG

Lớp : CÔNG NGHỆ THÔNG TIN

Khóa : 59

Tp. Hồ Chí Minh, tháng 1 năm 2022

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 TIỂU LUẬN
TRÍ TUỆ NHÂN TẠO

ĐỀ TÀI: XÂY DỰNG WEBSITE HỖ TRỢ TRA CỨU KIẾN THỨC TOÁN RỜI RẠC

Giảng viên hướng dẫn : NGUYỄN ĐÌNH HIỂN

Sinh viên thực hiện : NGUYỄN MAI NGHIÊM

NGUYỄN SÔNG NGÂN

TRẦN THỊ THANH NGÂN

TRẦN ANH DUY

ĐINH XUÂN GIANG

Lớp : CÔNG NGHỆ THÔNG TIN

Khóa : 59

Tp. Hồ Chí Minh, tháng 1 năm 2022

TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI
PHÂN HIỆU TẠI THÀNH PHỐ HỒ CHÍ MINH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

THIẾT KẾ TỔNG QUAN ĐỀ TÀI

Mã sinh viên : 5951071063 — **Họ tên : Nguyễn Mai Nghiêm**
5951071062 — **Nguyễn Sông Ngân**
5951071061 — **Trần Thị Thanh Ngân**
5951071010 — **Trần Anh Duy**
5951071019 — **Đinh Xuân Giang**

Khóa : 59 — **Lớp : Công Nghệ Thông Tin**

1. **Tên đề tài :** Xây dựng Website hỗ trợ tra cứu toán rời rạc

2. **Mục đích, yêu cầu :**

   a) **Mục đích :**
   Tìm hiểu , nghiên cứu về các thuật toán đã học để tìm ra thuật toán tốt cho việc mã hóa và tìm kiếm dữ liệu từ đó ứng dụng vào chương trình thực tế.

   b) **Yêu cầu :**
   - Yêu cầu kiến thức :
     + Có kiến thức cơ bản về Python .
     + Có kiến thức cơ bản về các thuật toán và mã hóa .
   - Yêu cầu chức năng :
     + Đào tạo được 1 mô hình máy học có độ chính xác tương đối từ đó vận dụng phối hợp để đưa vào website hỗ trợ cho việc tra cứu kết quả.
   - Yêu cầu phi chức năng :
     + Kết quả đúng , giao diện ưa nhìn , thân thiện người dùng , dễ dàng thao tác.

3. **Nội dung và phạm vi đề tài**

   a) **Nội dung :**

   - Tìm hiểu ngôn ngữ lập trình Python.

- Tìm hiểu môi trường lập trình Jupyter Notebook.
- Tìm hiểu về các thuật toán cơ bản và cách mã hóa.
- Thực hiện các mã hóa dữ liệu và áp dụng các thuật toán để cho ra 1 mô hình tương đối chính xác.
- Xây dựng 1 website cơ bản.
- Ứng dụng model đã train vào website để tìm kiếm kết quả.

b) Phạm vi đề tài :

Mọi người có nhu cầu tìm hiểu và thực hiện về môn trí tuệ nhân tạo , áp dụng trí tuệ nhân tạo vào website để tìm kiếm kết quả.

4. Công nghệ, công cụ và ngôn ngữ lập trình

a) Công nghệ : Python.
b) Công cụ : Jupyter Notebook (Anacoda 3).
c) Ngôn ngữ lập trình : Python.

5. Các kết quả chính dự kiến sẽ đạt được và ứng dụng

- Hiển thị được giao diện cơ bản với các chương và tài liệu môn toán rời rạc.
- Ứng dụng model đã train áp dụng vào tìm kiếm kết quả mong muốn.
- Website cho ra kết quả đúng
- Giao diện ưa nhìn, đẹp mắt , dễ sử dụng.

6. Giảng viên và cán bộ hướng dẫn

Họ tên : Nguyễn Đình Hiển
Đơn vị công tác : Bộ môn CNTT Trường Đại học GTVT phân hiệu tại TP.HCM
Điện thoại : 0918735299
Email : nguyendinhhien.khm@st.utc2.edu.vn

Ngày tháng 1 năm 2022

Giảng viên hướng dẫn

LỜI CẢM ƠN

Lời đầu tiên em xin gửi lời cảm ơn chân thành đến quý thầy, cô giáo trong **Bộ môn Công nghệ thông tin – Phân hiệu Trường Đại học Giao thông vận tải.** Những người đã truyền dạy, đã trang bị cho em kho tàng kiến thức về bầu trời công nghệ thông tin rộng lớn.

Ở đây, em không chỉ học được kiến thức về sách vở mà em còn học được các bài học, kỷ năng sống trước khi tạm biệt mái trường đại học thân yêu này và tiến ra biển đời mênh mông rộng lớn. Đặc biệt, em xin gửi lời cảm ơn chân thành và sâu sắc đến thầy **Nguyễn Đình Hiển** , người đã đồng hành cùng em trong suốt quá trình làm đồ án.

Trong quá trình học tập và tìm hiểu em đã nổ lực rất nhiều với mong muốn hoàn thành đồ án một cách tốt nhất, nhưng đời người sẽ có những thiếu sót không thể tránh khỏi, và với những người chưa chững chạc và trưởng thành như em thì sai lầm là không thể không mắc phải. Em mong thầy, cô bộ môn có thể thông cảm và cho em những ý kiến, đóng góp để em có thể hoàn thành đồ án của mình một cách trọn vẹn nhất trước khi rời xa ngôi trường thân yêu này.

Sau cùng, em xin kính chúc Quý Thầy Cô trong **Bộ môn Công nghệ thông tin** lời chúc sức khoẻ, luôn hạnh phúc và thành công hơn nữa trong công việc cũng như trong cuộc sống.

Em xin chân thành cảm ơn!

LỜI MỞ ĐẦU

Ngành Công nghệ thông tin là một ngành khoa học đang trên đà phát triển mạnh và ứng dụng rộng rãi trên nhiều lĩnh vực. Cùng với xu hướng phát triển của các phương tiện truyền thông như Báo, Radio... thì việc sử dụng Internet ngày càng phổ biến.

Cổng thông tin điện tử trên Internet ra đời cùng với việc Internet đang nhanh chóng lan rộng khắp toàn cầu, nó sẽ trở thành công cụ chủ yếu và đặc lực cho việc trao đổi, tìm kiếm thông tin trên phạm vi toàn thế giới. Bây giờ thì hầu như bất cứ nhu cầu nào của bạn cũng đều có thể được đáp ứng ngay tức khắc. Với một máy tính cá nhân có kết nối mạng, bạn có thể lướt trên các Website tin tức, các trang báo điện tử, thoải mái tìm kiếm các thông tin mình cần ngay tại chỗ.

Để đáp ứng với việc học tập và tra cứu thông tin thì website tra cứu là một nhu cầu tất yếu. Do đó, chúng em đã xây dựng ứng dụng Website tra cứu, hỗ trợ học tập dành cho các bạn học ngành công nghệ thông tin , đặc biệt là các bạn có học môn Toán Rời Rạc. Đề tài là Website tra cứu kiến thức về Toán rời rạc.

NHẬN XÉT CỦA GIẢNG VIÊNG HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp. Hồ Chí Minh, ngày ……. tháng …… năm ……

Giảng viên hướng dẫn

Mục lục

Mục lục hình ảnh

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1 Trí Tuệ Nhân Tạo (AI)

Trí tuệ nhân tạo (Artifical Intelligence - AI) là một nhánh của khoa học máy tính (comupter science) liên quan đến việc làm cho máy tính có những khả năng của trí tuệ con người người, tiêu biểu như các khả năng "suy nghĩ", "hiểu ngôn ngữ", và biết "học tập".

AI là ngành nghiên cứu về các hành xử thông minh (intelligent behaviour) bao gồm : thu thập, lưu trữ tri thức, suy luận, hoạt động và kỹ năng.

Nền tảng của AI : AI được ứng dụng trong rất nhiều nền tảng như Triết học , Toán học, Tâm lý học, Ngôn ngữ học , Công nghệ máy tính, Điều khiển học, Kinh tế học,…

Kết luận : AI là xu thế phát triển tất yếu của công nghiệp hiện nay, đặc biệt là các lĩnh vực về Công nghệ tri thức, Máy học. Nghiên cứu, xác định các vấn đề trọng tâm để từ đó đề xuất các phương pháp cải tiến, thay đổi dựa trên các công nghệ AI

1.2 Ngôn ngữ lập trình Python

Python là ngôn ngữ lập trình hướng đối tượng, cấp cao, mạnh mẽ, được tạo ra bởi Guido van Rossum. Nó dễ dàng để tìm hiểu và đang nổi lên như một trong những ngôn ngữ lập trình nhập môn tốt nhất cho người lần đầu tiếp xúc với ngôn ngữ lập trình. Python hoàn toàn tạo kiểu động và sử dụng cơ chế cấp phát bộ nhớ tự động. Python có cấu trúc dữ liệu cấp cao mạnh mẽ và cách tiếp cận đơn giản nhưng hiệu quả đối với lập trình hướng đối tượng. Cú pháp lệnh của Python là điểm cộng vô cùng lớn vì sự rõ ràng, dễ hiểu và cách gõ linh động làm cho nó nhanh chóng trở thành một ngôn ngữ lý tưởng để viết script và phát triển ứng dụng trong nhiều lĩnh vực, ở hầu hết các nền tảng.

1.3 Thuật toán phân lớp Random Forest.

1.3.1 Giới thiệu

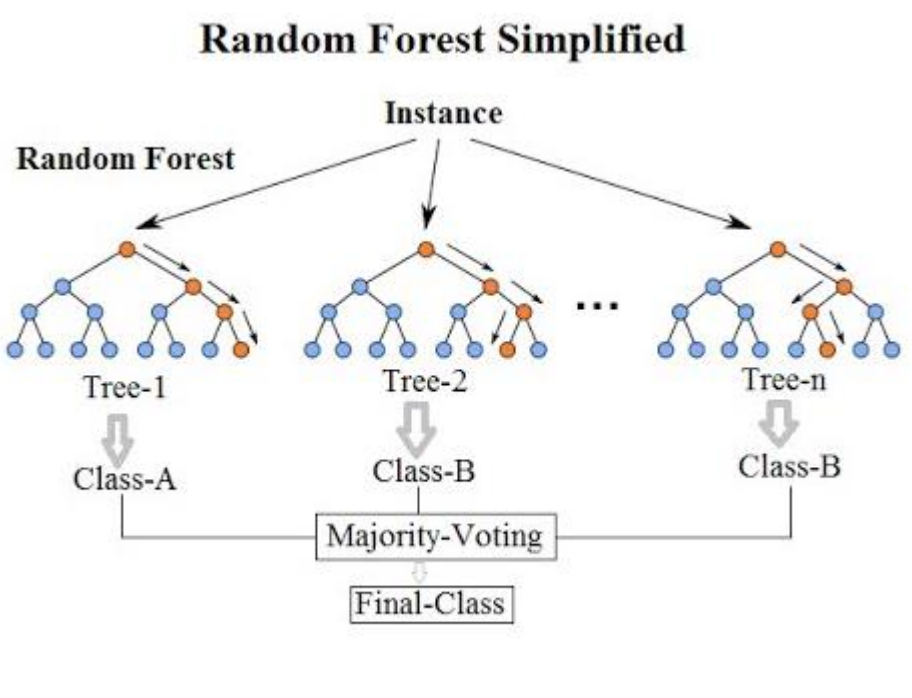
Random Forest là một tập hợp mô hình (ensemble) rất hiệu quả cho các bài toán phân loại vì nó huy động cùng lúc hàng trăm mô hình nhỏ hơn bên trong với quy luật khác nhau để đưa ra quyết định cuối cùng. Mỗi mô hình con có thể mạnh yếu khác nhau, nhưng theo nguyên tắc "wisdom of the crowd", ta sẽ có cơ hội phân loại chính xác hơn so với khi sử dụng bất kì một mô hình đơn lẻ nào.

Như tên gọi của nó, Random Forest (RF) dựa trên cơ sở:

- Random = Tính ngẫu nhiên;
- Forest = Nhiều cây quyết định (decision tree).

Đơn vị của RF là thuật toán cây quyết định, với số lượng hàng trăm. Mỗi cây quyết định được tạo ra một cách ngẫu nhiên từ việc: Tái chọn mẫu (bootstrap, random sampling) và chỉ dùng một phần nhỏ tập biến ngẫu nhiên (random features) từ toàn bộ các biến trong dữ liệu. Ở trạng thái sau cùng, mô hình RF thường hoạt động rất chính xác, nhưng đổi lại, ta không thể nào hiểu được cơ chế hoạt động bên trong mô hình vì cấu trúc quá phức tạp. RF do đó là một trong số những mô hình hộp đen (black box).

Trong quá khứ, chúng ta thường chấp nhận đánh đổi tính tường minh để đạt được tính chính xác. Từ mô hình Random Forest, chúng ta chỉ có thể làm một số khảo sát hạn chế, bao gồm vai trò tương đối của các biến (features) và vẽ các biểu đồ 2 chiều thể hiện ranh giới các vùng phân loại.



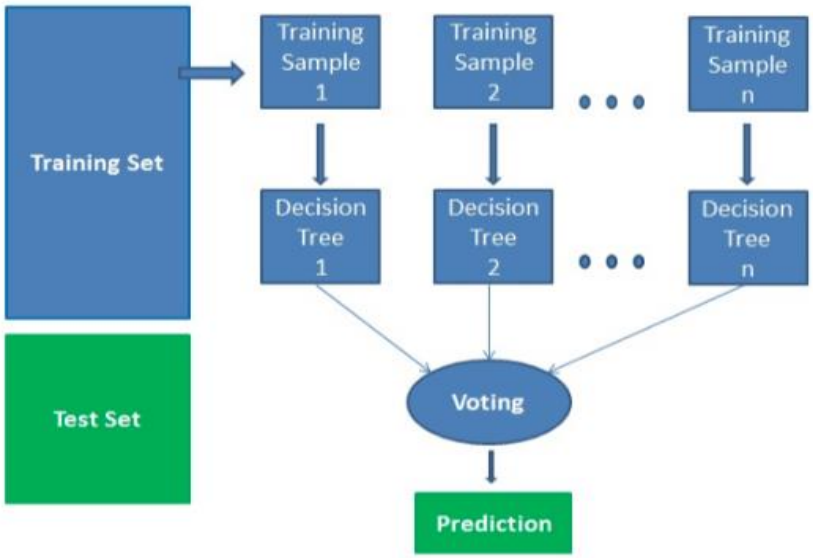
Hình 1 : Sơ đồ hoạt động của Random Forest

1.3.2 Cách hoạt động

Random Forest hoạt động theo 4 bước:

1. Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho.
2. Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây.

3. Hãy bỏ phiếu cho mỗi kết quả dự đoán.
4. Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.



Hình 2 : Các mô hình trong RandomForest

1.3.3 Ưu điểm

Random Forest được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình này. Nó không bị vấn đề overfitting. Lý do chính là nó mất trung bình của tất cả các dự đoán, trong đó hủy bỏ những thành kiến.

Thuật toán có thể được sử dụng trong cả hai vấn đề phân loại và hồi quy.

Random Forest cũng có thể xử lý các giá trị còn thiếu. Có hai cách để xử lý các giá trị này: sử dụng các giá trị trung bình để thay thế các biến liên tục và tính toán mức trung bình gần kề của các giá trị bị thiếu. Bạn có thể nhận được tầm quan trọng của tính năng tương đối, giúp chọn các tính năng đóng góp nhiều nhất cho trình phân loại.

1.3.4 Nhược điểm

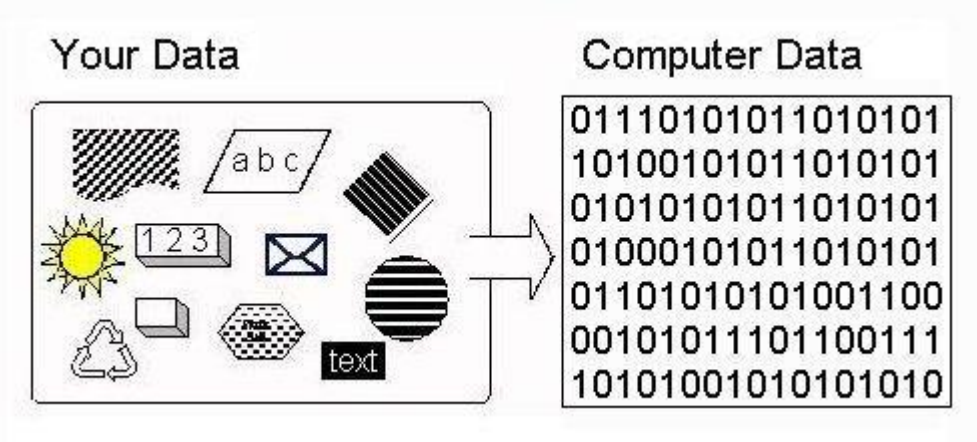
Random Forest chậm tạo dự đoán bởi vì nó có nhiều cây quyết định. Bất cứ khi nào nó đưa ra dự đoán, tất cả các cây trong rừng phải đưa ra dự đoán cho cùng một đầu vào cho trước và sau đó thực hiện bỏ phiếu trên đó.

Toàn bộ quá trình này tốn thời gian.

Mô hình khó hiểu hơn so với cây quyết định, nơi bạn có thể dễ dàng đưa ra quyết định bằng cách đi theo đường dẫn trong cây.

1.4 Categorical Data Encoding.

Trong các mô hình ML, chúng ta thường được yêu cầu chuyển đổi các tính năng văn bản phân loại thành biểu diễn số của nó. Hai cách phổ biến nhất để làm điều này là sử dụng Label Encoder hoặc One-hot Encoder. Tuy nhiên, hầu hết những người mới nghiên cứu về ML sẽ không quen thuộc với tác động của việc lựa chọn mã hóa có trên mô hình của họ, độ chính xác của mô hình có thể thay đổi theo số lượng lớn bằng cách sử dụng mã hóa phù hợp ở đúng kịch bản.



Hình 3 : Mã hóa dữ liệu thường sang dữ liệu máy tính

1.4.1 Categorical Data là gì?

Categorical data là dạng dữ liệu có tính chất phân biệt các giá trị thành hữu hạn các nhóm. Chúng cũng thường được gọi là các phân lớp hoặc các nhãn của các thuộc tính. Các giá trị rời rạc này có thể là văn bản hoặc các số tự nhiên (hoặc có thể là dữ liệu phi cấu trúc như hình ảnh). Có hai dạng chính của categorical data là danh nghĩa (**nominal**) và thứ tự (**ordinal**).

Vì nhóm em sẽ làm việc trên categorical variables nên dưới đây là một vài ví dụ thể hiện. Categorical variables thường được biểu diễn dưới dạng ‘string’ hoặc ‘categories’ và có số lượng hữu hạn.

Ví dụ:

- Thành phố mà một người sống: Delhi, Mumbai, Ahmedabad, Bangalore, etc.
- Bằng cấp cao nhất mà một người sở hữu: High school, Diploma, Bachelors, Masters, PhD.
- Điểm số của một học sinh, sinh viên: A+, A, B+, B, B- etc.

Trong các ví dụ trên, các biến chỉ có các giá trị xác định có thể có. Hơn nữa, chúng ta có thể thấy có hai loại dữ liệu phân loại:

- Ordinal Data: Các danh mục có thứ tự vốn có .
- Nominal Data: Các danh mục không có thứ tự vốn có.

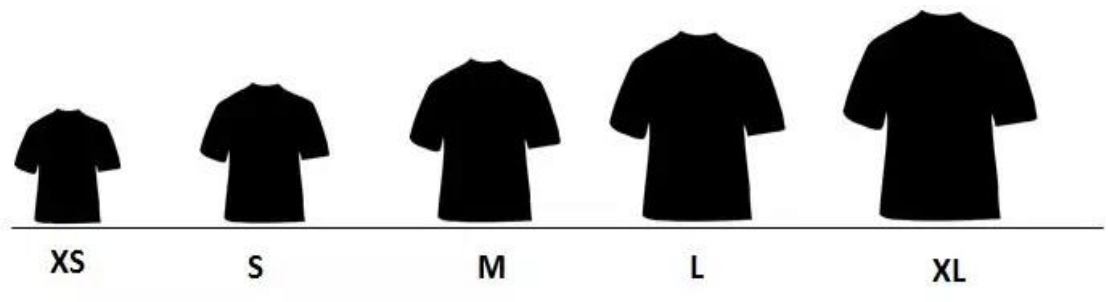
Trong khi encoding Ordinal Data, chúng ta nên giữ lại thông tin liên quan đến thứ tự mà category được cung cấp. Giống như trong ví dụ trên, bằng cấp cao nhất mà một người sở hữu cung cấp thông tin quan trọng về trình độ của anh ta. Bằng cấp là một đặc điểm quan trọng để quyết định một người có phù hợp với vị trí đăng tuyển hay không.

Đối với khi encoding Nominal Data, chúng ta phải xem xét sự hiện diện hay vắng mặt của một đối tượng địa lý và không cần quan tâm đến thứ tự.Ví dụ: thành phố mà một người sống. Đối với dữ liệu, điều quan trọng là phải giữ lại nơi một người sống. Nó là tương đồng nếu một người sống ở Delhi hoặc Bangalore.

1.4.2 Label Encoding (Ordinal Encoding)

Chúng ta sử dụng categorical data encoding này khi đối tượng phân loại là có thứ tự. Trong trường hợp này, việc quan trọng là duy trì thứ tự. Do đó encoding phải phản ánh trình tự. Trong Label encoding, mỗi label được convert thành một giá trị số nguyên từ 0,1,2,3,…...

Ví dụ hình dưới đây cho việc phân loại kích cỡ áo. Thứ tự giữa các nhóm được thể hiện rất rõ ràng trong trường hợp này như khi nói về "size" áo sơ mi thì S < M < L



Hình 4 : Ví dụ Size áo cho LabelEncoder

Hoặc chúng ta sẽ tạo LabelEndcoder cho tập dữ liệu đại diện cho trình độ học vấn và xem kết quả.

	Degree
0	High school
1	Masters
2	Diploma
3	Bachelors
4	Bachelors
5	Masters
6	Phd
7	High school
8	High school

LabelEndcoder()

	Degree
0	1
1	4
2	2
3	3
4	3
5	4
6	5
7	1
8	1

Hình 5 : Ví dụ LabelEndcoder

1.4.3 One-hot Encoding

Chúng ta sử dụng categorical data encoding này khi các tính năng là nominal (không có bất kỳ thứ tự nào). Cách truyền thống nhất để đưa dữ liệu này về dạng số là mã hóa one-hot. Trong cách mã hóa này, một "từ điển" cần được xây dựng chứa tất cả các giá trị khả dĩ của từng dữ liệu hạng mục. Sau đó mỗi giá trị hạng mục sẽ được mã hóa bằng một vector nhị phân với toàn bộ các phần tử bằng 0 trừ một phần tử bằng 1 tương ứng với vị trí của giá trị hạng mục đó trong từ điển.

Ví dụ như ta có 1 từ điển Animal ["Dog", "Cat", "Sheep", "Horse", "Lion"] tương ứng với mã hóa one-hot sẽ được kết quả như sau :

Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

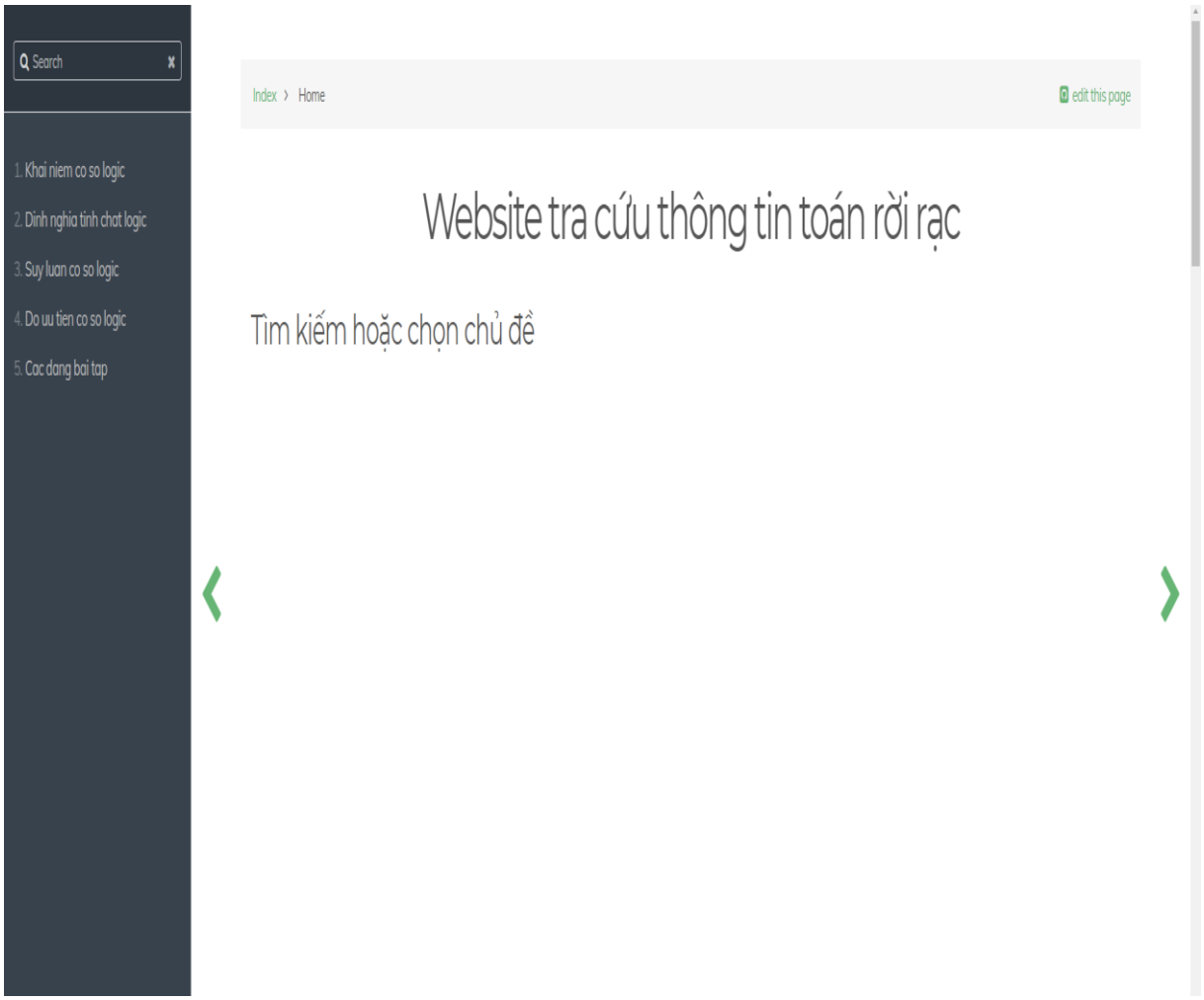
One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

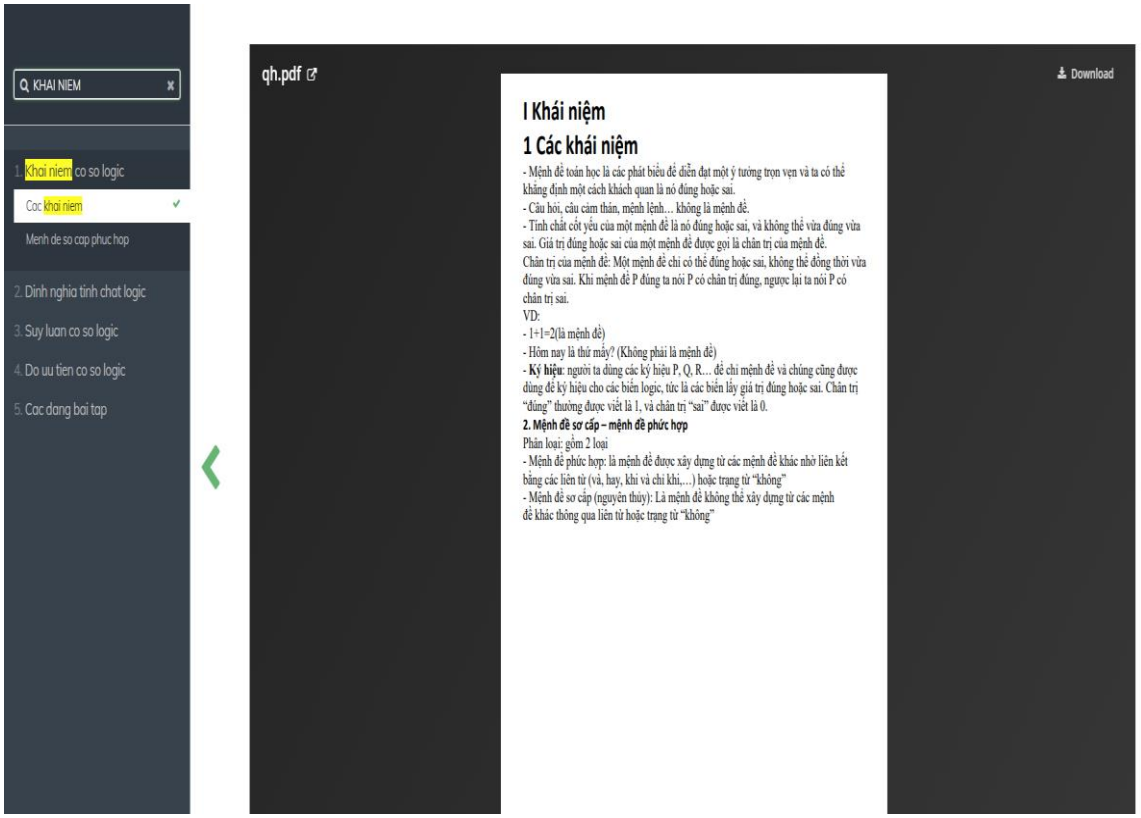
Hình 6 : Ví dụ mã hóa OneHotEndcoding

CHƯƠNG 2: XÂY DỰNG WEBSITE TRA CỨU KIẾN THỨC TOÁN RỜI RẠC

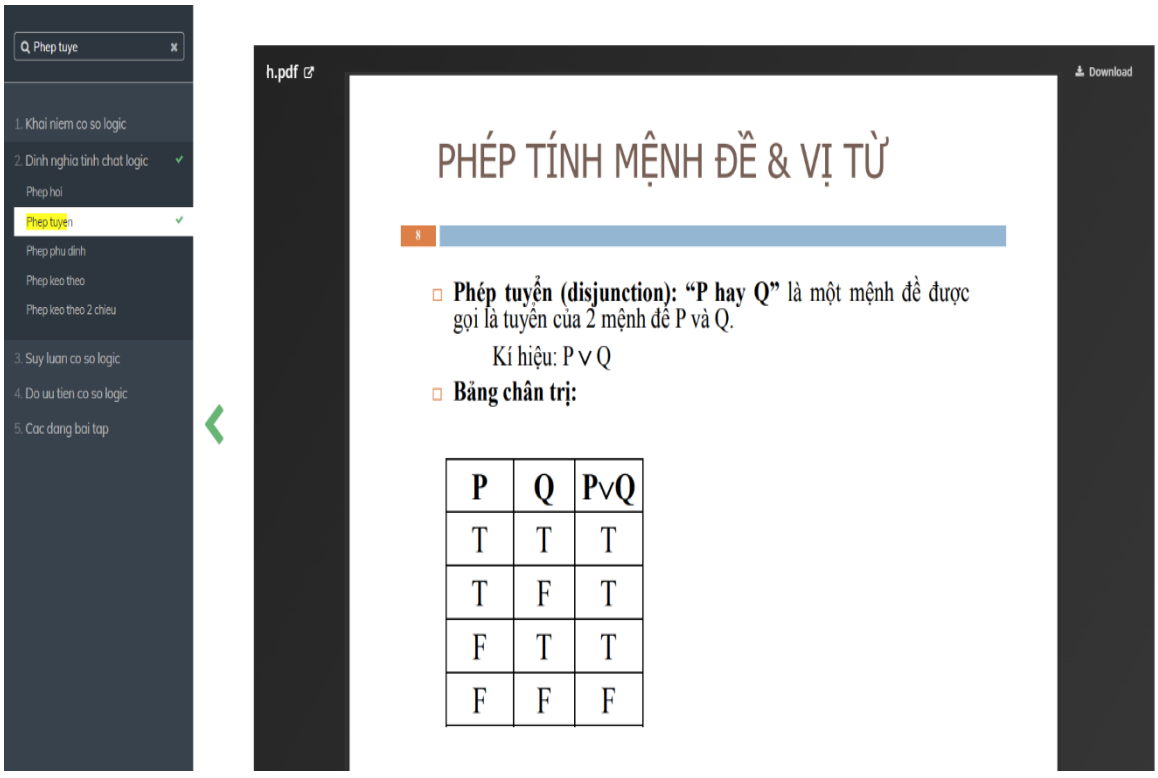
Giao diện website đã xây dựng được :



Hình 7 : Giao diện Website tra cứu Toán rời rạc



Hình 8 : Kết quả tìm kiếm



Hình 9 : Kết quả tìm kiếm

CHƯƠNG 3: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

3.1 Kết quả đạt được :

- Hiểu biết được rất nhiều kiến từ mô hình máy học như các thuật toán , thu thập và xử lý dữ liệu , cách train 1 mô hình máy , xử lý cho kết quả cải thiện hơn , khả năng làm việc nhóm hiệu quả hơn ,……
- Website có giao diện ưa nhìn , dễ sử dụng , có chức năng tìm kiếm với dữ liệu liên quan đến Toán rời rạc.
- Mô hình cho được kết quả tương đối chính xác.

3.2 Hướng phát triển

- Bổ sung thêm nhiều dữ liệu và chi tiết hơn để nhiều mọi người có thể tìm được nhiều thông tin theo mong muốn.
- Thiết kế thêm cho website đẹp hơn.
- Thêm nhiều dữ liệu tìm kiếm giúp cho mô hình đạt được độ chính xác cao hơn.

TÀI LIỆU THAM KHẢO

[1]. https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/

[2]. https://docs.microsoft.com/en-us/aspnet/core/blazor/security/webassembly/standalone-with-authentication-library?view=aspnetcore-6.0&tabs=visual-studio

[3]. Mã hóa one-hot — Machine Learning cho dữ liệu dạng bảng (machinelearningcoban.com)

[4]. Feature Engineering (Phần 3): Feature engineering với dữ liệu dạng phân loại (Categorical Data) (viblo.asia)

[5]. Chi tiết bài học Tiền xử lý dữ liệu trong lĩnh vực học máy (Phần 3) (vimentor.com)

DANH SÁCH PHÂN CÔNG KHỐI LƯỢNG THỰC HIỆN ĐỐI TƯỢNG ĐỀ TÀI

MSSV	Họ tên	Nhiệm vụ	Hoàn thành
5951071063	Nguyễn Mai Nghiêm	Tìm tài liệu + tổng hợp báo cáo , powerpoint + xây dựng model	100%
5951071061	Trần Thị Thanh Ngân	Tìm tài liệu + Xử lý dữ liệu và làm sạch dữ liệu + Làm báo cáo + Làm powerpoint	100%
5951071062	Nguyễn Sông Ngân	Tìm tài liệu + Xử lý dữ liệu và làm sạch dữ liệu + Làm báo cáo + Làm powerpoint	100%
5951071010	Trần Anh Duy	Tìm tài liệu + Xây dựng Website	100%
5951071019	Đinh Xuân Giang	Tìm tài liệu + Làm báo cáo + Làm powerpoint + hỗ trợ xây dựng model	100%