TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

BÁO CÁO

KHAI KHOÁNG DỮ LIỆU

MÃ HP: CT312

NỘI DUNG

PHÂN LOẠI ĐỘNG VẬT (TẬP DỮ LIỆU ZOO)

Giáo viên hướng dẫn:
T.S Lưu Tiến Đạo

Sinh viên thực hiện:
Trần Nhựt Linh – B1609828
Đinh Thành Công – B1609811
Ngành: **Khoa học máy tính**
Khóa: **42**

Cần Thơ, 07/2020

NHẬN XÉT CỦA GIẢNG VIÊN

LỜI CẢM ƠN

Chúng em xin chân thành cảm ơn Khoa CNTT&TT, Trường Đại học Cần Thơ đã tạo điều kiện thuận lợi cho chúng em học tập và thực hiện đề bài tập nhóm này.

Chúng em xin chân thành cảm ơn giảng viên, TS. Lưu Tiến Đạo đã giúp đỡ và hướng dẫn chúng em tận tình trong suốt thời gian làm đề tài cáo báo nhóm, tạo cho chúng em những tiền đề, những kiến thức để tiếp cận, phân tích và giải quyết vấn đề. Nhờ đó mà chúng em hoàn thành đề tài báo cáo nhóm được tốt hơn. Chúng em cũng xin cảm ơn bạn bè, anh chị đã tận tình chỉ bảo, giúp đỡ chúng em trong quá trình hoàn thành đề tài báo cáo nhóm, giúp chúng em hiểu thêm về những kiến thức mới.

Nhưng do kiến thức và kĩ năng còn hạn chế, phương pháp tiếp cận, phân tích, giải quyết vấn đề còn chưa được thấu đáo, chi tiết nên không tránh khỏi những thiếu sót trong quá trình nghiên cứu và trình bày. Rất kính mong được sự đóng góp ý kiến của các thầy, cô giảng viên để đề tài báo cáo nhóm được hoàn chỉnh hơn.

Một lần nữa chúng em xin cảm ơn và xin chúc tất cả quý Thầy Cô giảng viên được dồi dào sức khỏe và thành đạt hơn trong sự nghiệp!

Cần Thơ, ngày 03 tháng 07 năm 2020

Người viết	Người viết
(Kí và ghi rõ họ tên)	(Kí và ghi rõ họ tên)
Đinh Thành Công	Trần Nhựt Linh

MỤC LỤC

TÓM TẮT

Trong bài viết này, chúng tôi xây sử dụng thuật toán máy học Support Vector Machine(SVM), K-Nearest Neighbors (KNN) và Cây quyết định (Decision Tree) để xây dựng giao diện phân loại động vật dựa trên tập dữ liệu Zoo với 101 loài động vật thuộc 7 lớp: Mammal, Bird, Fish, Amphibian, Invertebrate, Replite, Insert. Kết quả sau khi đánh giá mô hình cho thấy thuật toán SVM cho độ chính xác cao hơn so với hai thuật toán còn lại. Cây quyết định sinh ra được các luật giúp phân loại động vật một cách trực quan. Xây dựng giao diện website phân loại động vật để tiết kiệm thời gian phân loại động vật.

PHẦN GIỚI THIỆU

1. Đặt vấn đề

Trong cuộc sống, chúng ta có thể thấy có rất nhiều loài động vật với những đặc điểm khác nhau như: có chân, sinh con, có cánh, sống trên cạn hay sống dưới nước,... Nhưng mỗi loài động vật đều thuộc một loại lớp động vật riêng trong 7 lớp: Mammal (Động vật có vú), Bird (Chim), Fish (Động vật dưới nước), Amphibian (Lưỡng cư), Invertebrate (Động vật không xương sống), Replite (Bò sát), Insert (Côn trùng). Tuy nhiên, có một số loài chúng ta khi nhìn qua thì có thể biết chúng thuộc lớp động vật nào nhưng cũng có một số loài mới xuất hiện hoặc chúng ta biết tên của chúng nhưng không biết được chúng thuộc loại lớp động vật nào.

Xuất phát từ các yêu cầu thực tế trên, đang rất cần có những nghiên cứu về vấn đề này. Chính vì vậy chúng tôi đã áp dụng thuật toán phân lớp dựa trên mô hình SVM, KNN và cây quyết định cho bài toán phân loại động vật với mong muốn phần nào áp dụng vào thực tế.

2. Giải quyết vấn đề

Nhóm sử dụng tập dữ liệu Zoo.aff thu thập được qua việc khảo sát nhiều con vật khác nhau trong vườn thú Mapperley Nottingham bởi Richard Forsyth cùng các cộng sự ngày 2/15/1990. Nhóm tiến hành phân tích dữ liệu, tìm hiểu từng thuộc tính của dữ liệu và sự tương quan giữa các thuộc tính của các loài. Sau đó nhóm tiến hành phân loại động vật động qua thuật toán SVM, KNN và cây quyết định.

3. Mục tiêu bài tập nhóm

Giúp các thành viên trong nhóm củng cố những kiến thức đã học trên lớp qua sự giảng dạy của giảng viên bộ môn vào nghiên cứu và tìm ra các phương pháp ứng dụng vào bài tập nhóm.

4. Đối tượng và phạm vi nghiên cứu

a) Đối tượng nghiên cứu

- Dữ liệu, các thuật toán, giải thuật SVM, KNN và cây quyết định.
- Ngôn ngữ lập trình Python.

b) Phạm vi nghiên cứu

Bài tập nhóm nghiên cứu tập trung vào phạm vi những kiến thức đã học trong lớp, và những kiến thức sinh viên tự tìm hiểu trên mạng hay nguồn tài liệu sách giáo khoa trong thư viện của trường.

5. Phương pháp nghiên cứu

- Tìm hiểu lý thuyết về SVM, KNN, cây quyết định, các tập trainning, testing, tìm hiểu giải thuật và code.
- Nghiên cứu qua tài liệu học trên lớp và Internet.
- Chọn lọc và thảo luận ý kiến giữa các thành viên về bài toán.
- Cài đặt các gói thư viện liên quan hỗ trợ cho ngôn ngữ Python.
- Thiết kế và cài đặt bài toán.
- Tạo giao diện web dự đoán cho phù hợp.
- Tổng kết đưa ra kết quả đánh giá và nhận xét.

6. Kết quả đạt được

- Xây dựng thành công bài toán phân loại động vật.
- Đáp ứng được các yêu cầu đề ra.

7. Bố cục bài báo cáo

Phần giới thiệu

Giới thiệu tổng quát về bài tập nhóm.

Phần nội dung

Chương 1 : Giới thiệu bài toán: phân loại động vật.

Chương 2 : Thiết kế và cài đặt.

Chương 3 : Kết quả thực nghiệm.

Phần kết luận

Trình bày kết quả đạt được và hướng phát triển hệ thống.

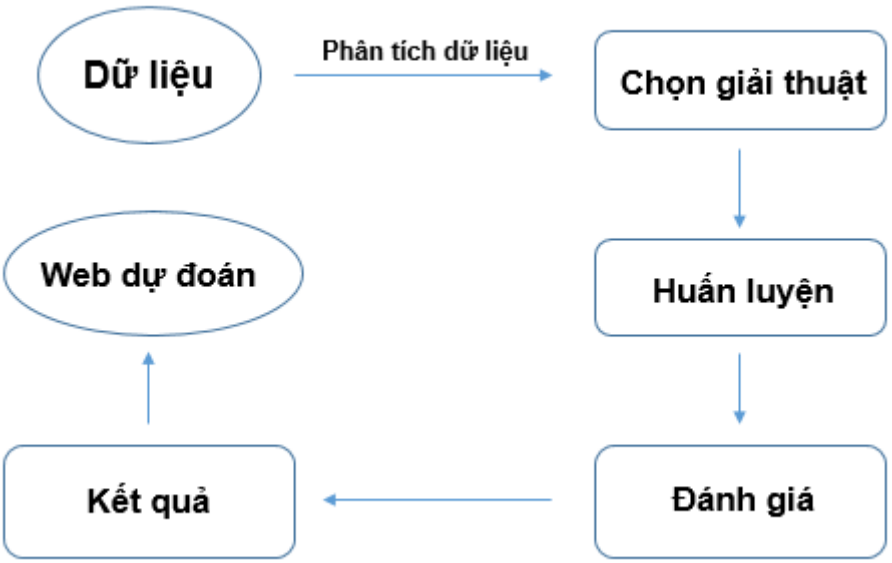
PHẦN NỘI DUNG

CHƯƠNG 1: GIỚI THIỆU BÀI TOÁN PHÂN LOẠI ĐỘNG VẬT

1. Giới thiệu bài toán:

Xây dựng hệ thống phân loại động vật bằng thuật toán SVM, cây quyết định và KNN được thực hiện: Sử dụng tập dữ liệu để huấn luyện (trainning) bao gồm tất cả các thuộc tính tạo ra được mô hình. Sau khi có được mô hình tiến hành kiểm tra (testing) lại mô hình xem độ chính xác có cao không bằng nghi thức hold-out và k-fold. Sử dụng mô hình tốt nhất để dự đoán.

2. Mô tả bài toán



Hình 1 Sơ đồ mô tả bài toán

Bài toán phân loại động vật bao gồm các giai đoạn: chọn giải thuật để huấn luyện – huấn luyện – đánh giá – kết quả – xây dựng ứng dụng.

- Lựa chọn giải thuật: lựa chọn và sử dụng 3 giải thuật để huấn luyện sao, phù hợp với yêu cầu và mục đích sử dụng: SVM, KNN và cây quyết định.
- Huấn luyện:
  - SVM: sử dụng kernel RBF để huận luyện.
  - KNN: sử dụng mô hình với k = 3.

- Cây quyết định: xây dựng mô hình dựa trên chỉ số Entropy.

- Đánh giá: Sử dụng cả 2 phương thức đánh giá là k-fold và hold-out cho cả 3 giải thuật:
  - Đối với nghi thức kiểm tra hold-out: chia tập test với 30% dữ liệu và được kiểm tra 20 lần và lấy giá trị trung bình để nâng cao sự tin cậy cho chỉ số đánh giá.
  - Cao nhất:
    - SVM: 100%
    - KNN: 100%
    - Cây quyết định: 96%
  - Trung bình:
    - SVM: 90%
    - KNN: 89%
    - Cây quyết định: 88%
  - Đối với nghi thức kiểm tra chéo k-fold: sử dụng kiểm tra với k=5. Chia dữ liệu ra thành năm phần rồi lần lượt kiểm tra qua từng k.
  - Accuracy:
    - SVM: 95%
    - KNN: 94%
    - Cây quyết định: 88%
  - Recall:
    - SVM: 95%
    - KNN: 94%
    - Cây quyết định: 87%
  - F1-Score:
    - SVM: 93%
    - KNN: 93%
    - Cây quyết định: 86%
- Kết quả: SVM là giải thuật cho kết quả cao nhất tuy nhiên việc hiểu và phân loại động vật theo cách trực quan thì SVM không thể đáp ứng. Do đó sử dụng cây quyết định, xây dựng bộ luật để tìm ra các loại động vật dựa vào 1 số thuộc tính quan trọng.
  - Bộ luật:
    - R1: Nếu các động vật mà có lông vũ thì nó thuộc *lớp chim*.
    - R2: Nếu các động vật không có lông vũ và có sữa thì nó thuộc *lớp động vật có vú*.
    - R3: Nếu các động vật không có lông vũ , không có sữa , có xương sống và có vây thì thuộc lớp *động vật dưới nước*.

- R4: Nếu các động vật không có lông vũ , không có sữa , có xương sống , không có vây và nó sống dưới nước thì nó thuộc *lớp lưỡng cư*.
- R5: Nếu các động vật không có lông vũ , không có sữa , có xương sống , không có vây và nó không sống dưới nước thì nó thuộc *lớp bò sát*.
- R6: Nếu các động vật không có lông vũ , không có sữa ,không có xương sống và sống trên cạn thì nó thuộc *lớp côn trùng*.
- R7: Nếu các động vật không có lông vũ , không có sữa ,không có xương sống và không sống trên cạn thì nó thuộc *lớp động vật không có xương sống*.

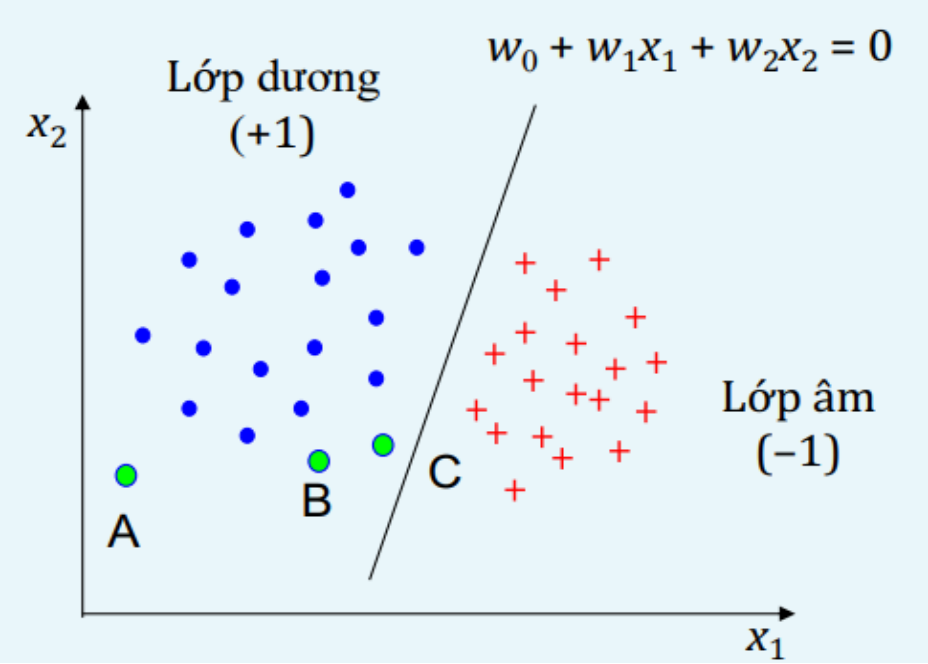
- Website dự đoán: Sau khi có được mô hình huấn luyện, sử dụng mô hình để ứng dụng, hướng tới một webside cho người dùng có thể dự đoán dựa vào bộ luật đang có, hoặc có thể tiến hành dự đoán nhờ vào việc nhập vào các thuộc tính của động vật muốn dự đoán.

CHƯƠNG 2: THIẾT KẾ VÀ CÀI ĐẶT

1. Cơ sở lý thuyết

a) Tìm hiểu về thuật toán

- SVM
  - Máy học véc-tơ hỗ trợ (SVM) được đề xuất bởi Vapnik từ năm 1995 là mô hình học hiệu quả và phổ biến cho vấn đề phân lớp, hồi quy tuyến tính và phi tuyến.
  - Giải thuật được phát triển mạnh vào những năm 1990.
  - Là công cụ hữu hiệu và phổ biến của lãnh vực máy học, nhận dạng và khai mỏ dữ liệu.
  - Áp dụng thành công trong: nhận dạng mặt người, phân loại văn bản, phân loại bệnh ung thư, ...
  - SVM cho bài toán phân lớp được biết đến như bài toán phân lớp nhị phân.



Hình 2 Máy học véc-tơ cho bài toán phân lớp

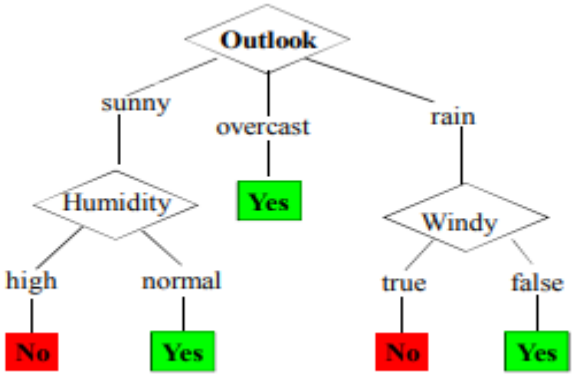
- KNN
  - K-nearest neighbor là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning.
  - KNN là thuật toán đi tìm đầu ra của một điểm dữ liệu mới bằng cách *chỉ* dựa trên thông tin của K điểm dữ liệu trong training set gần nó nhất (K-lân cận), *không quan tâm đến*

việc có một vài điểm dữ liệu trong những điểm gần nhất này là nhiễu.

- 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.
- 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 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.

➢ Cây quyết định

- Kết quả sinh ra dễ dịch (if … then …).
- Khá đơn giản, nhanh, hiệu quả được sử dụng nhiều.
- Liên tục trong nhiều năm qua, cây quyết định được bình chọn là giải thuật được sử dụng nhiều nhất và thành công nhất.
- Giải quyết các vấn đề của phân loại, hồi quy.
- Làm việc cho dữ liệu số và kiểu liệt kê
- Được ứng dụng thành công trong hầu hết các lĩnh vực về phân tích dữ liệu, phân loại text, spam.



Hình 3 Mô hình cây quyết định

- Nút trong : được tích hợp với điều kiện để kiểm tra rẽ nhánh.
- Nút lá : được gán nhãn tương ứng với lớp của dữ liệu.
- Một nhánh : trình bày cho dữ liệu thỏa mãn điều kiện kiểm tra.
- Ở mỗi nút, 1 thuộc tính được chọn để phân hoạch dữ liệu học sao cho tách rời các lớp tốt nhất có thể.
- Một luật quyết định có dạng IF-THEN được tạo ra từ việc thực hiện AND trên các điều kiện theo đường dẫn từ nút gốc đến nút lá..
- Dữ liệu mới đến được phân loại bằng cách duyệt từ nút gốc của cây cho đến khi đụng đến nút lá, từ đó rút ra lớp của đối tượng cần xét.

b) Thiết kế hệ thống

Phân loại động vật dựa trên tập dữ liệu Zoo được viết trên ngôn ngữ Python 3.6.6. Đây là một trong những ngôn ngữ phổ biến và sử dụng nhiều nhất hiện nay.

Chương trình website đự đoán:

- Input: Là 17 thuộc tính của động vật.
- Output: Là kết quả của loại động vật đó thuộc 1 trong 7 lớp động vật.

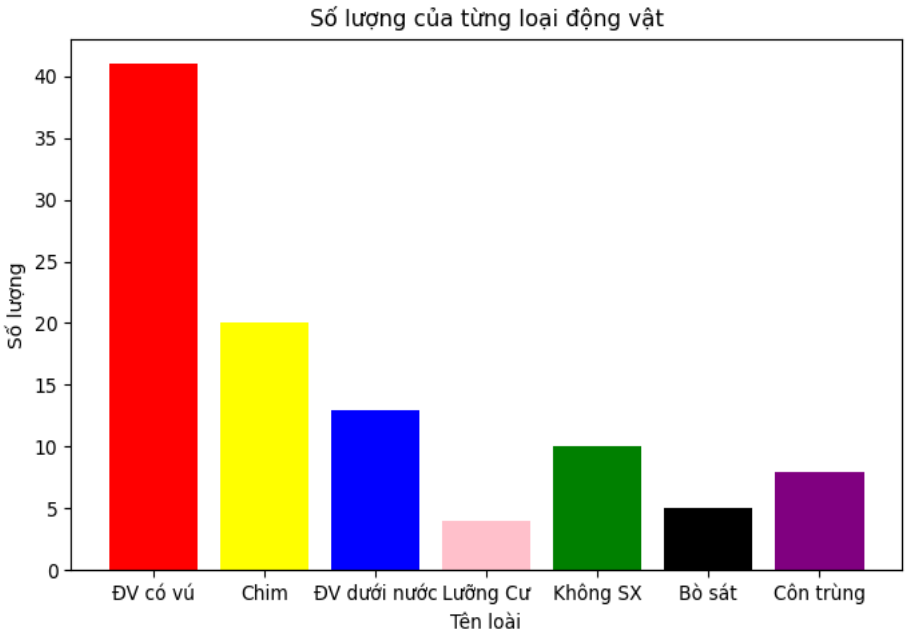
2. Phân tích dữ liệu

- Dữ liệu Zoo gồm 18 thuộc tính.
- Dữ liệu không có phần tử rỗng.
- Bảng miêu tả tên, kiểu dữ liệu, các giá trị của từng thuộc tính:

STT	Tên thuộc tính	Kiểu dữ liệu	Các giá trị
1	Animal name/ Tên con vật		Unique for each instance
2	Hair/ Lông	Nominal	True, false
3	Feathers/ Lông vũ	Nominal	True, false
4	Eggs/ Đẻ trứng	Nominal	True, false
5	Milk/ Sữa	Nominal	True, false
6	Airborne/ Sống trên cạn	Nominal	True, false
7	Aquatic/ Sống dưới nước	Nominal	True, false
8	Predator/ Ăn thịt	Nominal	True, false

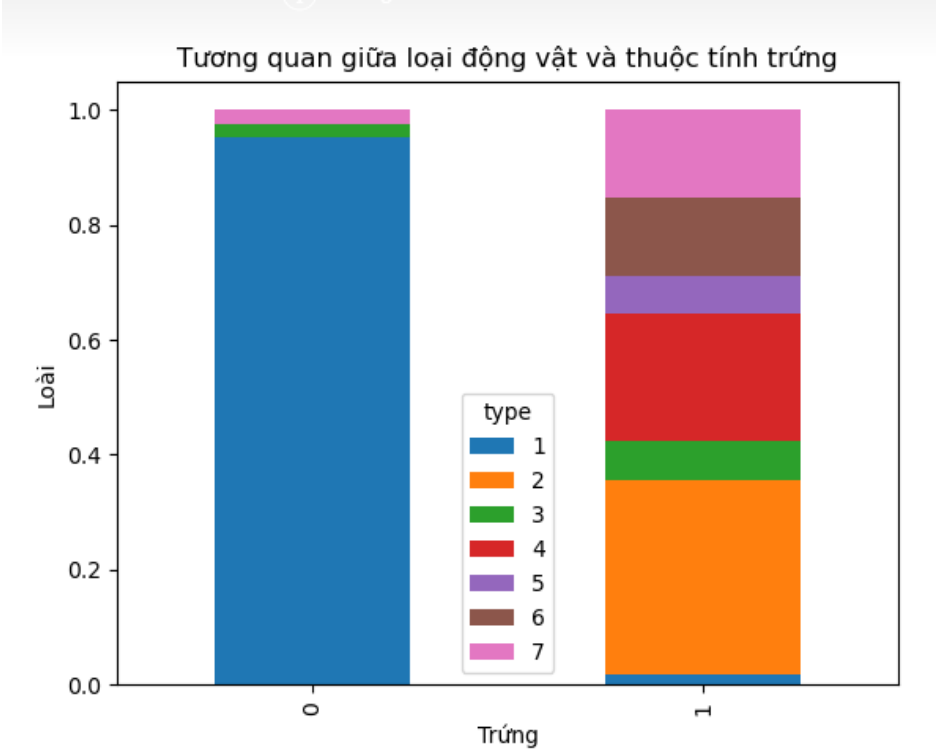
9	Toothed/ Có răng	Nominal	True, false
10	Backbone / Có xương sống	Nominal	True, false
11	Breathes / Hô hấp	Nominal	True, false
12	Venomuos / Có nộc độc	Nominal	True, false
13	Fins / Có cánh	Nominal	True, false
14	Legs / Số chân	Nominal	Numeric (set of value: {0,2,4,6,8})
15	Tail / Đuôi	Nominal	True, false
16	Domestic / Sống bầy đàn	Nominal	True, false
17	Catsize / Catsize	Nominal	True, false
18	Type / Loại lớp	Nominal	Numeric (interger values in range [1,7])

- Type (loại lớp) gồm 7 lớp với 101 động vật:
  - Mammal / Động vật có vú: 41
  - Bird / Chim: 20
  - Fish / ĐV dưới nước: 13
  - Amphibian / Lưỡng cư: 4
  - Invertebrate / ĐV không xương sống: 10
  - Reptile / Bò sát: 5
  - Insert / Côn trùng: 8

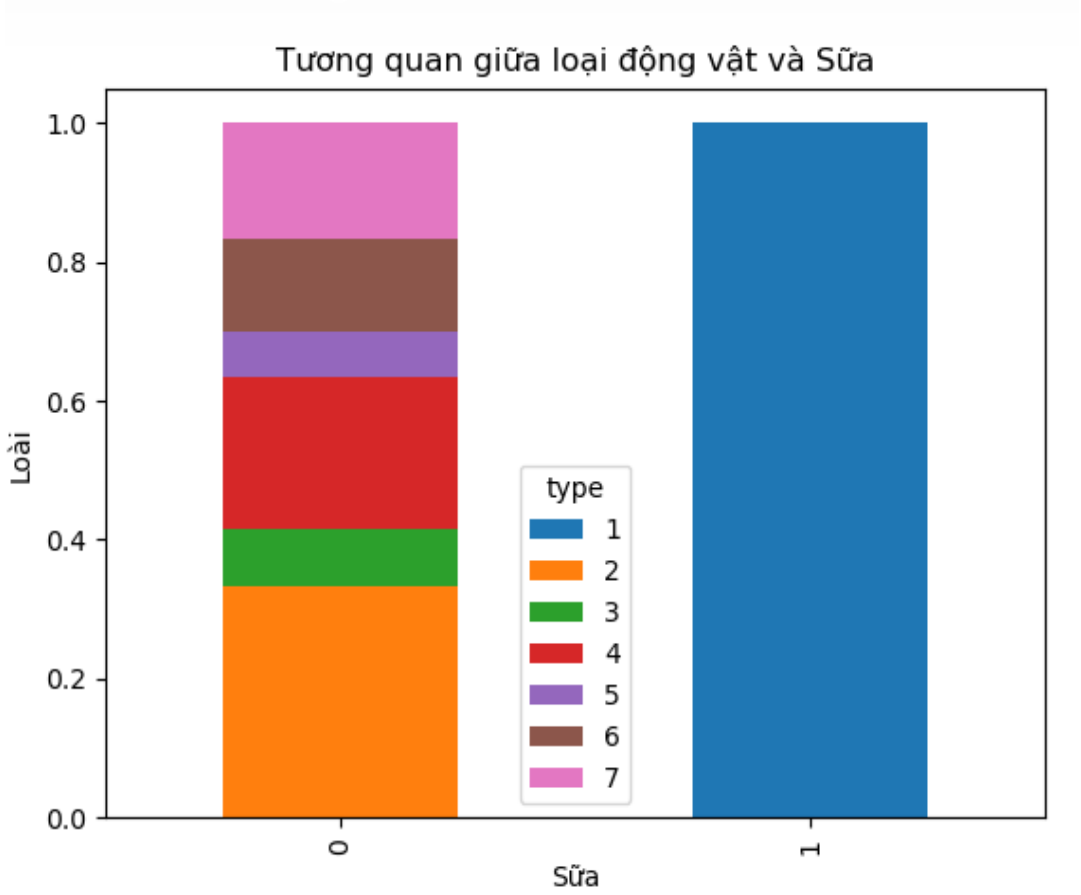


Hình 4 Biểu đồ phân lớp động vật

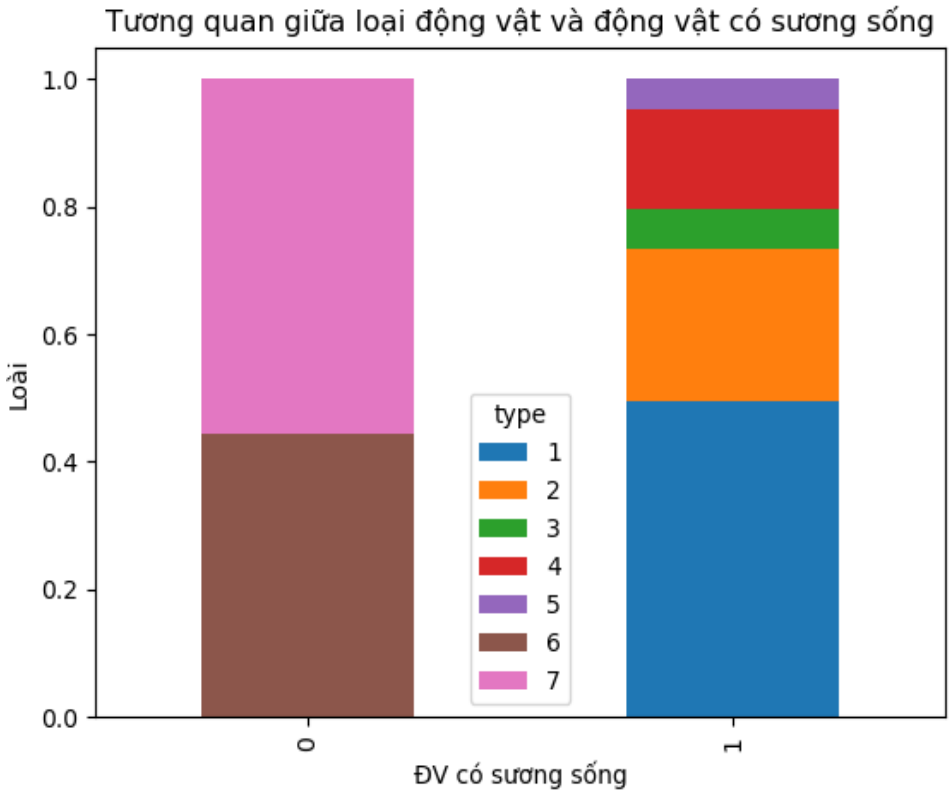
Độ tương quan giữa các thuộc tính với nhãn:



Hình 5 Độ tương quan giữa thuộc tính tính với loài động vật



Hình 6 Độ tương quan giữa thuộc tính sữa với loài động vật



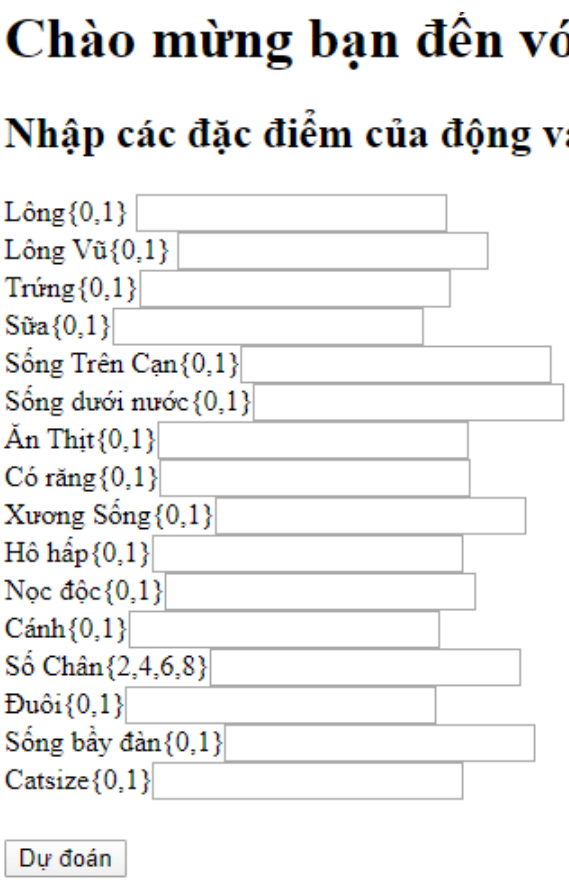
Hình 7 Độ tương quan giữa thuộc tính xương sống với loài động vật

3. Giao diện

Sử dụng Framework Flask để xây dựng ứng dụng website phân loại động vật.

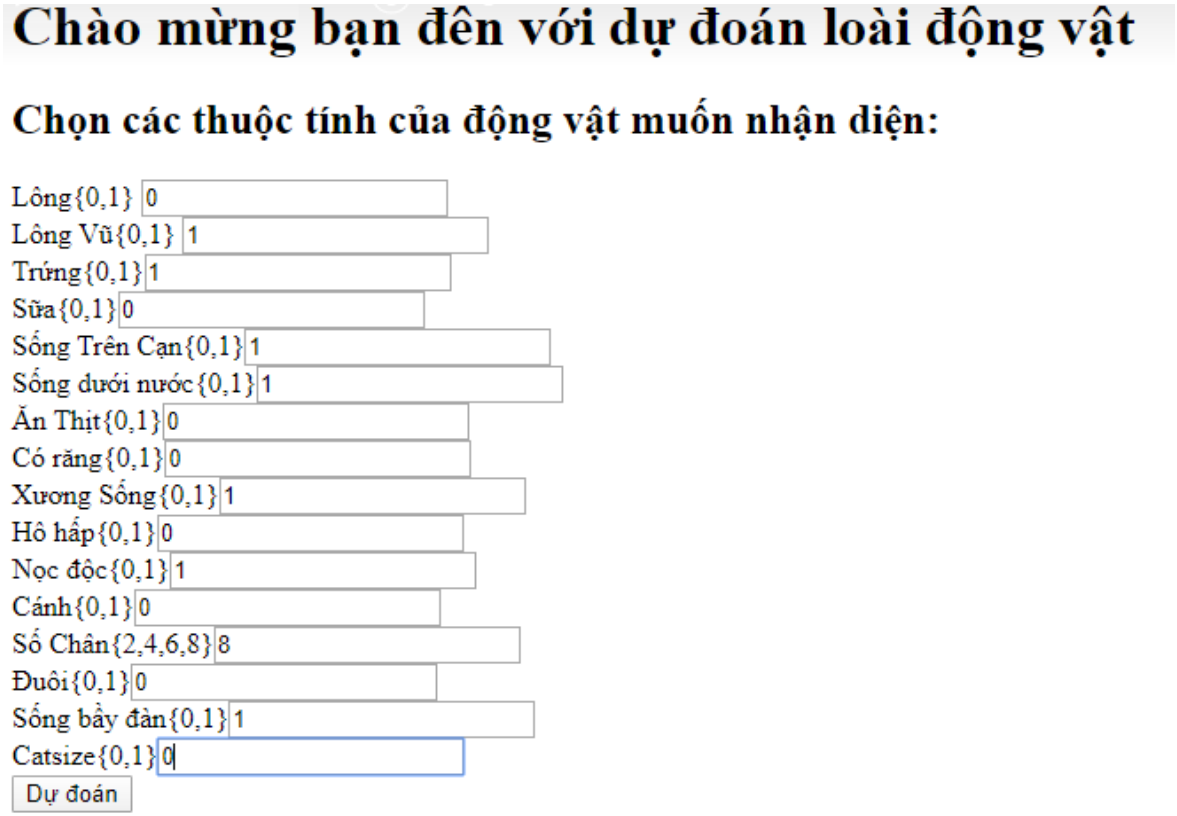
CHƯƠNG 3: KẾT QUẢ THỰC NGHIỆM

- Giao diện:



Hình 8. Demo kết quả

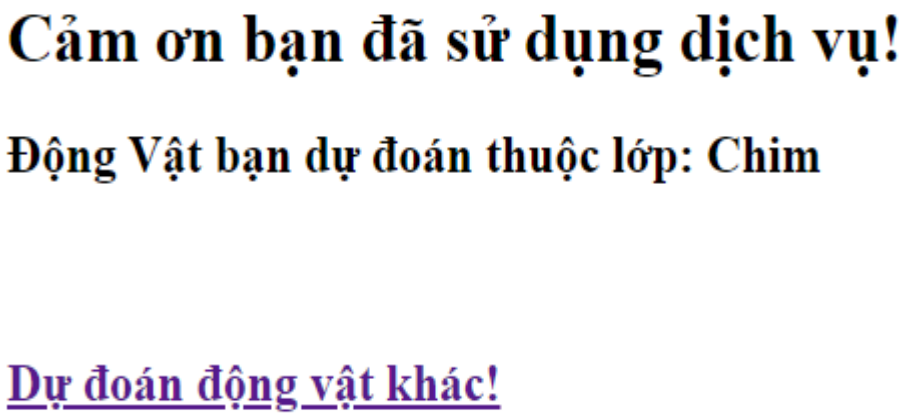
- Nhập các đặc điểm của động vật:



Hình 9. Demo Nhập các thuộc tính của động vật

Chọn các thuộc tính cần dự đoán để làm dữ liệu đầu vào và nhấn chọn vào nút dự đoán để kiểm tra động vật cần dự đoán sẽ thuộc lớp động vật nào.

- Kết quả dự đoán:



Hình 10. Kết quả khi dự đoán

PHẦN KẾT LUẬN

1. Kết quả đạt được

- Xây dựng được giao diện website phân loại động vật với độ chính xác cao.
- Xây dựng được bộ luật phân loại động vật.

2. Hạn chế

- Dữ liệu không đạt được kết quả tối ưu (100%) do có dữ liệu lỗi:
  - Frog: Có 2 loài động vật trùng tên frog và có thuộc tính bên trong khác.
  - Girl: phần lớn các động vật có sương sống đều có "đuôi" (tail) trong khi "girl" không có.
- Giao diện website phân loại còn đơn giản.

3. Hướng phát triển

- Xây dựng app phân loại động vật trên thiết bị di động.
- Xây dựng giao diện website phân loại động vật hiện đại hơn.

TÀI LIỆU THAM KHẢO

1. Google: https://machinelearningcoban.com/2017/01/04/kmeans2/
2. Sile máy học vector hỗ trợ - SVM của thầy Phạm Nguyên Khang
3. Google: https://machinelearningcoban.com/2017/04/09/smv/
4. Google: https://1upnote.me/post/2018/11/ds-ml-svm-mnist/
5. Google: https://github.com/pallets/flask
6. W3School: https://www.w3schools.com/
7. Wiki: https://vi.wikipedia.org/wiki/Wikipedia
8. Google: https://machinelearningcoban.com/2017/01/08/knn/
9. Slide bài báo khoa học của thầy Lưu Tiến Đạo.