BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CẦN THƠ

BÁO CÁO
KHAI KHOÁNG DỮ LIỆU

ĐỀ TÀI

PHÂN LOẠI NẤM
TẬP DỮ LIỆU MUSHROOM

Giảng viên hướng dẫn

Ts. Lưu Tiến Đạo

Sinh viên thực hiện

B1609830 – Lê Thanh Lương

B1611128 – Lâm Thanh Hòa

Cần Thơ – 2020

ĐÁNH GIÁ VÀ NHẬN XÉT

……

……

……

……

……

……

……

……

……

……

……

……

Giảng viên hướng dẫn

(ký và ghi rõ họ tên)

LỜI CẢM ƠN

Đầu tiên, chúng tôi xin bày tỏ lòng biết ơn sâu sắc tới giảng viên hướng dẫn là Ts. Lưu Tiến Đạo đã tận tâm, dành nhiều thời gian để hướng dẫn, định hướng phương pháp nghiên cứu đề tài khoa học này, đồng thời cung cấp nhiều tài liệu tham khảo và tạo điều kiện thuận lợi nhất trong suốt quá trình học tập nghiên cứu, để tôi có thể hoàn thành bài báo cáo khai khoáng dữ liệu này.

Cần Thơ, ngày tháng năm 2020

LỜI CAM ĐOAN

Bài báo cáo khai khoáng dữ liệu về đề tài "Phân loại nấm trên tập dữ liệu Mushroom" đánh dấu những thành quả, kiến thức tôi đã tiếp thu được trong suốt quá trình học tập rèn luyện tại trường. Tôi xin cam đoan bài báo cáo này được hoàn thành bằng quá trình học tập nghiên cứu của tôi dưới sự hướng dẫn của giảng viên Ts. Lưu Tiến Đạo.

Nội dung trong báo cáo này là của cá nhân tôi nghiên cứu tổng hợp và tham khảo các tài liệu trong thư viện và internet có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Tôi xin hoàn toàn chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Cần Thơ, ngày 01 tháng 07 năm 2020

MỤC LỤC

Mục lục

DANH SÁCH CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

KNN	K-nearest neighbor

DANH MỤC CÁC HÌNH VẼ

TÓM TẮT

Trong bài báo cáo này, chúng tôi thực hiện phân loại nấm trên tập dữ liệu Mushroom, sử sụng các giải thuật đánh giá mô hình như sau RandomForest, DesicionTree, LogisticRegression, để tìm mô hình đánh giá tốt nhất. Kết quả cho thấy giải thuật RandomForest và DecisionTree cho kết quả tốt nhất. Bên cạnh việc sử dụng các giải thuật như trên, chúng tôi còn xây dựng mô hình KNN, DecisionTree, GradientBoosting và giao diện phân loại nấm từ các thuộc tính.

GIỚI THIỆU

1. Đặt vấn đề

Trong đời sống hiện nay, nấm là một loại thực phẩm rất phổ biến được sử dụng hằng ngày trong các bữa ăn ở gia đình. Tuy nhiên ngoài những loại nấm ăn được mà chúng ta biết vẫn còn các loại nấm độc thường mọc hoang dại ở ven đường, trong những cách rừng,...nếu chẳng mai người ta sử dụng nấm độc thì có thể dẫn đến nguy kịch đến sức khỏe con người.

Với các lý do nêu trên, tôi chọn đề tài "Phân loại nấm trên tập dữ liệu Mushroom" để giúp mọi người hiểu rõ và xác định được những đặc điểm nào là của loài nấm ăn được và những đặc điểm nào của loài nấm có độc.

2. Lịch sử giải quyết vấn đề

Phân loại nấm trên tập dữ liệu Mushroom đã có nhiều cá nhân hay tổ chức xây dựng với nhiều ngôn ngữ lập trình khác nhau như python…và cả những phần mềm hỗ trợ phân tích dữ liệu như weka…

3. Mục tiêu đề tài

Mục tiêu của đề là nghiên cứu phân loại nấm thông qua thuộc tính của các loại nấm (màu sắc, hình dạnh thân, nón….) để đặc điểm nào là của loài nấm ăn được và đặc điểm nào là của nấm độc.

Xây dựng giao diện phân loại nấm dựa vào các thuộc tính.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: tập dữ liệu Mushroom, các giải thuật phân loại và đánh giá mô hình, ngôn ngữ lập trình Python.

Phạm vi nghiên cứu: đầu tiên nghiên cứu trên cơ sở lý thuyết để hiểu hơn về loài nấm, nghiên cứu các giải thuật đánh giá mô hình, và ngôn ngữ lập trình Python. Sau đó nghiên cứu xây dựng giao diện phân loại nấm thông qua các thuộc tính.

5. Phương pháp nghiên cứu

Các phương pháp nghiên cứu được áp dụng trong đề tài bao gồm:

- Phương pháp tổng hợp tài liệu: các tài liệu về nấm, các tài liệu Machine Learning…
- Phương pháp thực nghiệm: lập trình chương trình phân tích dữ liệu với tập dữ liệu Mushroom và xây dựng giao diện phân loại nấm đơn giãn thông qua các thuộc tính của nấm.

6. Kết quả đạt được

Phân tích được dữ liệu trên tập dữ liệu Mushrom, xây dựng được mô hình đánh giá với độ chính xác tương đối cao

Xây dựng giao diện phân loại nấm đơn giản để xác định nấm có độc và nấm không có độc

7. Bố cục

Phần Giới thiệu

1. Đặt vấn đề
2. Lịch sử giải quyết vấn đề
3. Mục tiêu đề tài
4. Đối tượng và phạm vi nghiên cứu
5. Phương pháp nghiên cứu
6. Kết quả nghiên cứu
7. Bố cục niên luận

Phần Nội dung

Chương I: Mô tả bài toán

1. Mô tả chi tiết bài toán
2. Các vấn đề có liên quan đến bài toán

Chương II: Thiết kế và cài đặt giải thuật

1. Thiết kế hệ thống

2. Thiết kế và cài đặt giải thuật

3. Giao diện

Phần Kết luận

Trình bày kết quả đạt được và khả năng phát triển của đề tài trong tương lai.

NỘI DUNG

CHƯƠNG I: MÔ TẢ BÀI TOÁN

I. Mô tả chi tiết bài toán

1. Giới thiệu chung về tập dữ liệu

Dataset gồm các mô tả về giải quyết trương ứng với 23 loài nấm mang trong họ Agraricus và Lepiota được nghiên cứu bởi The Audubon Society Field Guide về các loại nấm bắc Mĩ (1981). Bộ dữ liệu này nghiên cứu và cho kết luận về khả năng ăn được và độc hại của từng mẩu thử nấm

Có 8124 dữ liệu mẫu cùng với 23 thuộc tính. Trong đó Poisonous(p) có độc, edible(e) ăn được.

Số thứ tự	Tên thuộc tính (nấm)	Ý nghĩa
1	Classes	Phân loại (cột nhãn)
2	Cap-shape	Hình dạng mũ nấm
3	Cap-surface	Bề mặt mũ nấm
4	Cap-color	Màu mũ nấm
5	Bruises	Vết thâm
6	Odor	Mùi hương
7	Gill-attachment	Lá tia đính kèm
8	Gill-spacing	Mật độ lá tia
9	Gill-size	Kích cở lá tia
10	Gill-color	Màu lá tia
11	Stalk-shape	Hình dạng cuống
12	Stalk-root	Hình dạng cuống rễ
13	Stalk-surface-above-ring	Bề mặt cuống trên vòng
14	Stalk-surface-below-ring	Bề mặt cuống dưới vòng

15	Stalk-color-above-ring	Màu cuống trên vòng
16	Stalk-color-below-ring	Màu cuống dưới vòng
17	Veil-type	Loại mạng
18	Veil-color	Màu mạng
19	Ring-number	Số vòng
20	Spore-print-color	Màu bàu tử
21	Population	Mật độ
22	Habitat	Môi trường sống

2. Tiền xử lý dữ liệu:

- Gán nhãn cho dữ liệu trong cột Class với p (=1) và e (=0).

STT	Classes (gốc)	Class (sau tiền xử lý dữ liệu)
1	p	1
2	e	0
3	e	0
4	p	1
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
8121	e	0
8122	e	0
8123	e	0
8124	p	1
8125	e	0

- Loại bỏ các thuộc tính rỗng và dữ liệu đã xác định (veil-type và stalk-root)

stalk-root
?
?
?
?
?
?
b
?
?
?

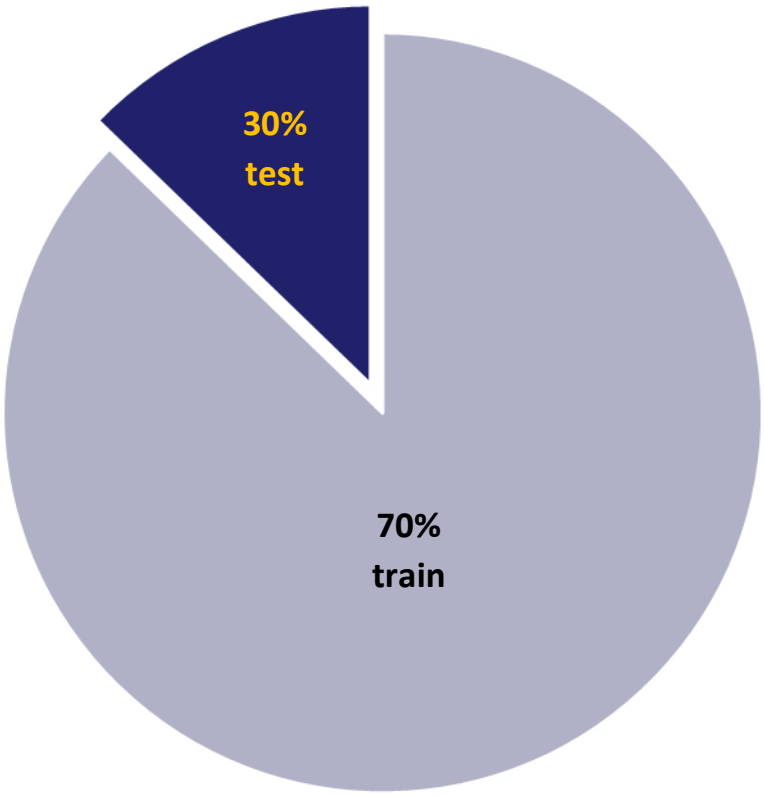
veil-type
p
p
p
p
p
p
p
p
p
p

- Biến đổi giá trị dữ liệu của tất cả các cột về kiểu số thực.

class	cap-shape	...	population	habitat
1	0.0	...	0.0	0.000000
0	0.0	...	0.2	0.166667
0	0.2	...	0.2	0.333333
1	0.0	...	0.0	0.000000
0	0.0	...	0.4	0.166667
...	...	...	...	...
0	0.8	...	1.0	1.000000
0	0.0	...	0.6	1.000000
0	0.6	...	1.0	1.000000
1	0.8	...	0.6	1.000000
0	0.0	...	1.0	1.000000

3. Mô hình đánh giá

Phân chia tập dữ liệu sử dụng 30% test và 70% dùng để train.



Hình 1: Phân chia dữ liệu theo hold-out

4. Kết quả

- Độ chính xác lần lượt như sau:
  - Rừng ngẫu nhiên (RandomForest): 100%
  - Cây quyết định (DecisionTree): 100%
  - Hồi quy tuyến tính (LogisticRegression): 97,29%

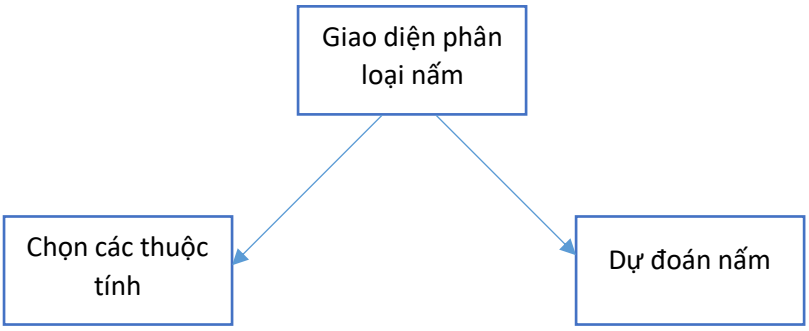
II. Các vấn đề liên quan đến bài toán

- Xây dựng các mô hình KNN, DecisionTree, GradientBoosting
- Đánh giá mô hình với các chỉ số Recall, Accuracy, F1-score
- Sử dụng các mô hình vừa xây dựng để phân loại nấm thông qua các thuộc tính

- Cài đặt flask để xây dựng giao diện phân loại nấm thông qua nền tảng website

CHƯƠNG II: THIẾT KẾ VÀ CÀI ĐẶT GIẢI THUẬT

1. Thiết kế hệ thống



2. Thiết kế và cài đặt giải thuật

a) Các giải thuật

- KNN
- DecisionTree
- GradientBoosting

⇨ Tạo model với KNN, DecisionTree, GradientBoosting, sử dụng model vừa xây dựng để phân loại nấm.

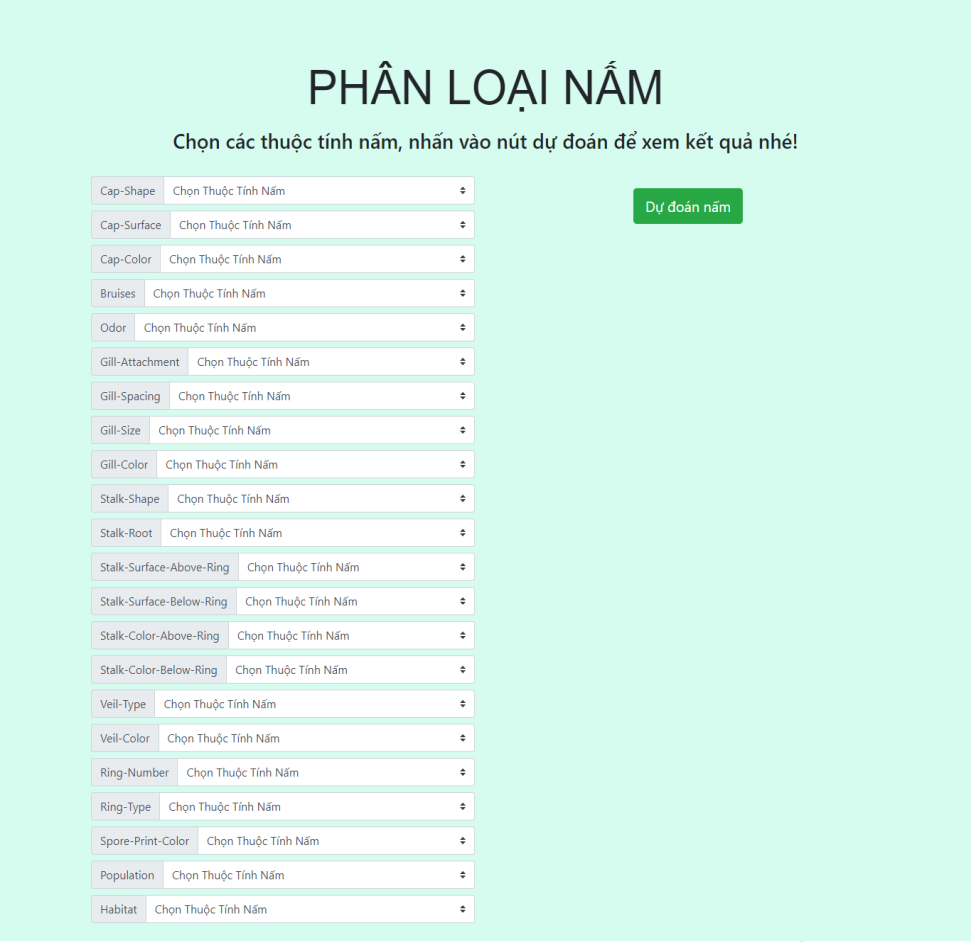
b) Microframework

- Sử dụng flask để chạy trên web xây dựng giao diện phân loại nấm.

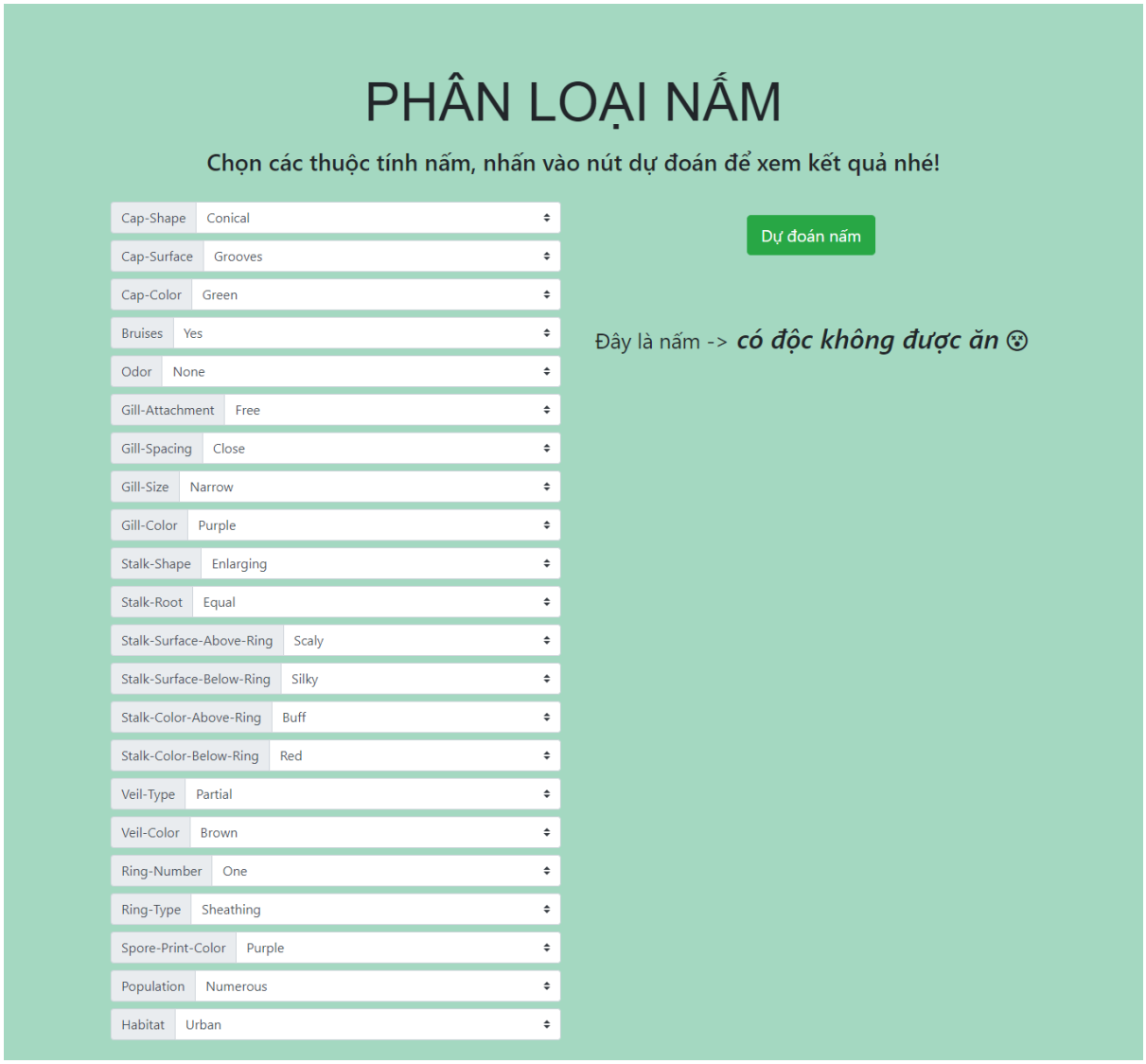
c) Xây dựng web

- Sử dụng html xây dựng cấu trúc web
- Sử dụng framework: bootstrap trang trí giao diện web
- Sử dụng các thư viện hỗ trợ jquery

3. Giao diện



Hình 2: Giao diện trang chủ phân loại nấm



Hình 3: Giao diện kết quả

KẾT LUẬN

1. Kết quả đạt được

- Rừng ngẫu nhiên(RandomForest) cho ra kết quả dự đoán đúng là 100%.
- Cây quyết định(DecisionTree) cho ra kết quả dự đoán đúng 100%.
- Hồi quy tuyến tính(LogisticRegression) cho ra kết quả thấp hơn chỉ với 97.29 %.
- Xây dựng giao diện web phân loại nấm sử dụng Flask và các giải thuật phân loại như KNN, DecisionTree, GradientBoosting

2. Hướng phát triển đề tài

- Sử dụng nhiều giải thuật để đánh gias tập dữ liệu.
- Cho mô hình đánh giá chạy từ 10 lần trở lên.

Tài liệu tham khảo

1. Giáo trình khai khoáng dữ liệu, TS. Đỗ Thanh Nghị
2. Giáo trình nguyên lý máy học, TS. Đỗ Thanh Nghị - TS. Phạm Nguyên Khang.