BÁO CÁO ĐỒ ÁN CUỐI KỲ

Lớp: CS114.(K21+K21.KHTN)
Môn: MÁY HỌC

GV: PGS.TS Lê Đình Duy - THS. Phạm Nguyễn Trường An
Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

NHẬN DIỆN RẮN ĐỘC VÀ KHÔNG ĐỘC

Lê Võ Ngọc Anh - 18520452 - CS114.K21

Link Github: https://github.com/levongocanh/CS114.K21

Bùi Thanh Tuấn - 18520395 - CS114.K21

Link Github: https://github.com/18520395/CS114.K21

Tóm tắt

Tên đề tài: NHẬN DIỆN RẮN ĐỘC VÀ KHÔNG ĐỘC

Tóm tắt:

- Là một bài toán phân lớp
- Input: Hình ảnh một con rắn
- Output: Là rắn độc hay không độc

Kết quả đạt được:

- Xây dựng được nhiều model với model tốt nhất đạt độ chính xác 83%

Ảnh thành viên

Lê Võ Ngọc Anh

Bùi Thanh Tuấn

Bài toán

Mỗi năm có từ 80000 đến 140000 người trên thế giới thiệt mạng vì bị rắn độc cắn. Đây được xem là một trong những nguyên nhân mang đến chết chóc lớn nhất đối với loài người nhưng chưa được nhắc tới nhiều, cũng như chưa được đánh giá một cách đầy đủ. Theo thống kê, châu Á và châu Phi là hai khu vực có số người tử vong vì bị rắn độc cắn nhiều nhất với con số lần lượt là 57.000 - 100.000 người và 20.000 - 32.000 người, tiếp theo là Mỹ Latinh - Caribe (3.400 - 5.000), châu Đại Dương (200 - 500) và cuối cùng là châu Âu (30 - 130). Việt Nam là nơi cư ngụ của gần 200 loài rắn, trong đó 53 loài là rắn độc chủ yếu thuộc hai họ rắn lục và rắn hổ. Các vết cắn của rắn độc để lại những hậu quả rất nặng nề, các vết cắn của rắn không độc vẫn rất đáng lưu tâm. Mặc dù hầu hết các loài rắn không chủ động tấn công con người tuy nhiên việc trang bị kiến thức kỹ càng để tránh các tình huống nguy hiểm. Xuất phát từ vấn đề đó, đề tài phân loại rắn độc và không độc giúp tăng hiểu biết về các loài rắn để có được tâm lý vững vàng trong những tỉnh huống đối mặt với các loài rắn khác nhau dù độc hay không. Đề tài này chỉ là bài toán nhỏ và đơn giản trong vấn đề xử lý các vụ rắn cắn.

Mô tả dữ liệu

Là hình ảnh của các loài rắn phân làm hai lớp độc và không độc

Kích thước: 8890 ảnh

- Độc: 4808 ảnh
- Không độc: 4082 ảnh

Mô tả dữ liệu

Dữ liệu được lấy từ:

- Kaggle:

https://www.kaggle.com/sameeharahman/preprocessed-snake-images/data#

- Github:

https://github.com/arjun921/Indian-Snakes-Dataset

Mô tả dữ liệu

Từ Kaggle:

Là bộ dữ liệu được xây dựng lại tốt hơn từ một phần của bộ dữ liệu của cuộc thi AIcrowd Snake Species Identification Challenge bao gồm 5 loài rắn

Từ Github:

Là bộ dữ liệu được xây dựng trên các loài rắn phổ biến ở Ấn Độ được chi làm hai lớp độc và không độc (nhiều loài phân bố từ Ấn Độ đến Đông Nam Á)

Kết quả đạt được

```
15.08 Seconds to train Support Vector Machines...
Accuracy Support Vector Machines =  0.8147
              precision    recall  f1-score   support

         0.0       0.80      0.80      0.80       838
         1.0       0.82      0.83      0.83       943

    accuracy                           0.81      1781
   macro avg       0.81      0.81      0.81      1781
weighted avg       0.81      0.81      0.81      1781
```

Kết quả của SVM

```
13.33 Seconds to train Logistic Regression...
Accuracy Logistic Regression =  0.8338
              precision    recall  f1-score   support

         0.0       0.82      0.83      0.82       838
         1.0       0.85      0.84      0.84       943

    accuracy                           0.83      1781
   macro avg       0.83      0.83      0.83      1781
weighted avg       0.83      0.83      0.83      1781
```

Kết quả của Logistic Regression

Kết quả đạt được

36.87 Seconds to train Random Forest Classifier...
Accuracy Random Forest Classifier = 0.8046

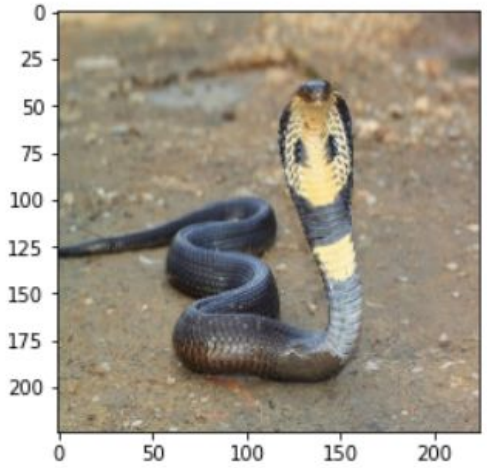
	precision	recall	f1-score	support
0.0	0.79	0.80	0.79	838
1.0	0.82	0.81	0.81	943
accuracy			0.80	1781
macro avg	0.80	0.80	0.80	1781
weighted avg	0.80	0.80	0.80	1781

Kết quả của Random Forest Classifier

Kết quả đạt được

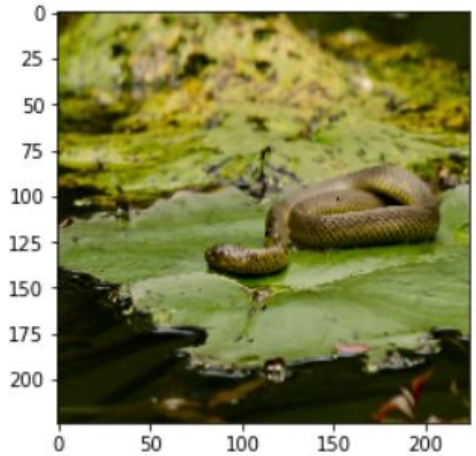
Một số kết quả dự đoán:

Image 1 :
Name of image: ho_mang_chua_doc.jpg



Support vector machine result: Có độc
Logistic Regression: Có độc
Random Forest Classifier: Có độc

Image 2 :
Name of image: ran_bongsung_khongdoc.jpg

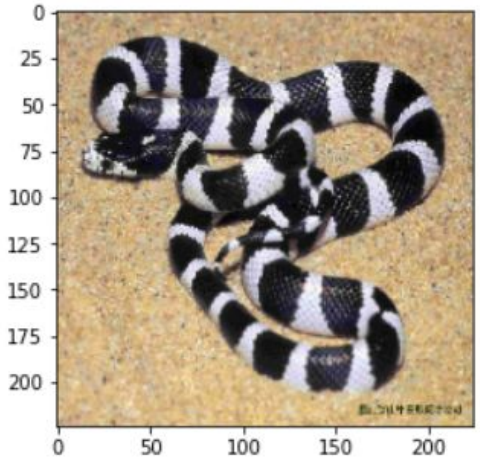


Support vector machine result: Không có độc
Logistic Regression: Không có độc
Random Forest Classifier: Không có độc

Kết quả đạt được

Một số kết quả dự đoán:

Image 3 :
Name of image: ran_cap_nia_doc.jpg



Support vector machine result: Không có độc
Logistic Regression: Không có độc
Random Forest Classifier: Có độc

Image 9 :
Name of image: ran_ho_trau_khongdoc.jpg

Support vector machine result: Không có độc
Logistic Regression: Không có độc
Random Forest Classifier: Có độc

Khó khăn

- Mô hình dự đoán không tốt với các loài rắn có nhiều đặc điểm ngoại hình giống nhau
- Mô hình dự đoán không tốt với những loài có đặc điểm ngoại hình giống với môi trường
- Dữ liệu ít và chưa đa dạng

Kết luận

Nhìn chung các mô hình cho ra tỷ lệ chính xác ổn và các kết quả thực nghiệm cũng cho kết quả tốt

Tuy nhiên do bộ dữ liệu còn nhiều hạn chế về số lượng lẫn chất lượng cùng với các ảnh thực nghiệm chưa quá đa dạng và mang tính kiểm tra kỹ nên không thể đánh giá mô hình là tốt

Vì lượng dữ liệu tương đối ít cộng với khó xác định con rắn trong các môi trường có màu sắc tương đồng nên các mô hình dự đoán không tốt với những trường hợp này

Hướng phát triển

- Sử dụng nhiều kỹ thuật trích xuất đặc trưng
- Sử dụng các thuật toán phân lớp ở mức sâu hơn
- Xây dựng dataset nhiều, đa dạng và dễ nhận biết hơn