XÂY DỰNG MÔ HÌNH PHÂN LOẠI CÁC LOÀI THÚ DỰA TRÊN TẬP DỮ LIỆU ZOO DATA SET TRÊN WEBSITE UCI

Nguyễn Hoàng Khánh Vân
Khoa CNTT-TT, Trường Đại học Cần Thơ
Ngành Khoa Học Máy Tính
vanb1509903@student.ctu.edu.vn

TÓM TẮT— *Trong bài viết này, chúng tôi trình bày tiếp cận xây dựng mô hình phân loại các loài thú dựa trên tập dữ liệu zoo data set trên website UCI. Mô hình cây quyết định, máy học véc-tơ hỗ trợ và rừng ngẫu nhiên là các mô hình được sử dụng phổ trong dự báo do tính chính xác của chúng. Sau đó dùng phương pháp gom nhóm KMeans Clustering đê gom nhóm các loài thú có cùng đặc điểm để quản lý và chăm sóc. Kết quả thu được mô hình để có thể phân loại được các loài thú trong sở thú dễ dàng với độ chính xác khá cao.*

Từ khóa— *Dự báo mật số rầy nâu, máy học véc-tơ hỗ trợ, rừng ngẫu nhiên, Apache Spark.*

I. GIỚI THIỆU

Vườn bách thú, thường gọi là vườn thú hay sở thú hay còn gọi là thảo cầm viên là một nơi mà nhiều loài động vật khác nhau được lưu giữ để mọi người có thể xem và theo dõi hoạt động của chúng.Vườn thú hiện đại không chỉ để cho mục đích giải trí, mà còn dùng cho các hoạt động giáo dục, nghiên cứu, và việc bảo tồn và bảo vệ động vật. Nhiều vườn thú là các trung tâm có chức năng bảo tồn động vật quý hiếm đang ở trong nguy cơ tuyệt chủng. Những vườn thú hiện đại cũng muốn cung cấp cho các động vật một đời sống tự nhiên, để chúng có sức khỏe và có một đời sống bình thường như trong tự nhiên cũng như để cho quan khách có thể nhìn thấy các loài động vật tương tự như trong môi trường tự nhiên thay vì trong một vườn thú. Trong quá khứ, và thậm chí cả ngày nay, có rất nhiều vườn động vật không có các điều kiện như các loại vườn động vật hiện đại. Có các loài động vật được nuôi nhốt trong điều kiện rất tệ như: nuôi giữ trong lồng nhỏ khiến chúng buồn chán và bị bệnh.Đối với một số loài nếu không được chăm sóc đúng cách sẽ không thể phát triển bình thường được nếu chỉ chăm sóc chúng giống như tất cả loài động vật khác.

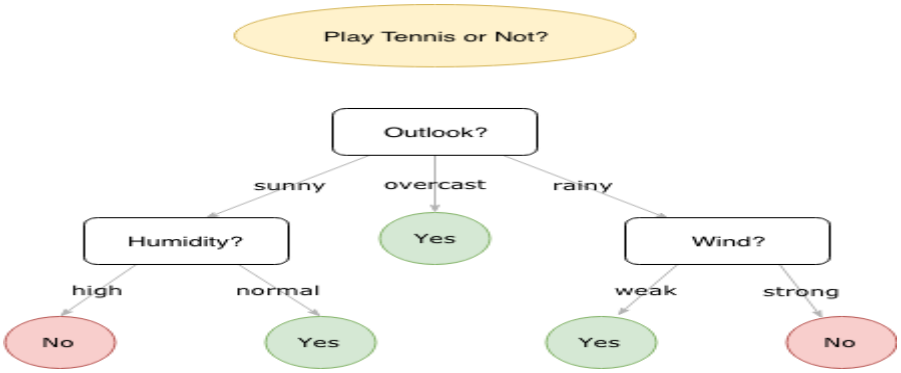
Chính vì lý do trên, xây dựng mô hình phục vụ việc phân loại các loài thú rất cần thiết. Mục tiêu chính là giúp cho các sở thú dễ dàng phân loại và chăm sóc cũng như xây dựng nơi ở thích hợp với từng loại thú khác nhau, giúp chúng phát triển bình thường như khi ở thế giới tự nhiên

Trong phạm vi của nghiên cứu này, trình bày kết quả thu được từ việc áp dụng các mô hình máy học vào tập dữ liệu zoo data set trên website UCI và phân loại các loài thú có trong đó. Dữ liệu sẽ được lấy trực tiếp từ website UCI, thực hiện các thao tác tiền xử lý và làm sạch dữ liệu. Bước tiếp theo thực hiện xây dựng mô hình phi tuyến, cây quyết định (Decision Tree [Ross Quinlan, 1986]), máy học véc-tơ hỗ trợ (Support Vector Machines – SVM [Vapnik, 1995]), rừng ngẫu nhiên (Random Forests – RF [Breiman, 01]), để phân loại các loài thú. Dùng phương pháp gom nhóm KMeans Clustering và Hierarchical Clustering đê gom nhóm các loài thú có cùng đặc điểm để quản lý và chăm sóc.

Phần còn lại của bài viết được tổ chức như sau: phần 2 trình bày tóm tắt về các mô hình phân loại thú; phần 3 trình bày phương pháp gom nhóm; kết quả thực nghiệm được trình bày trong phần 4 trước khi kết luận và hướng phát triển được trình bày trong phần 5.

II. CÁC MÔ HÌNH DỰ BÁO

A. *Cây quyết định*



Hình 1. Cây quyết định

Cây quyết định là một đồ thị của các quyết định và các hậu quả có thể của nó (bao gồm rủi ro và hao phí tài nguyên). *Cây quyết định* được sử dụng để xây dựng một kế hoạch nhằm đạt được mục tiêu mong muốn

Các cây quyết định được dùng để hỗ trợ quá trình ra quyết định. Cây quyết định là một dạng đặc biệt của cấu trúc cây. Mô hình hồi quy tuyến tính có dạng:

$$gini(T)=1-\sum_{j=1}^{n} P_j^2 \quad (1)$$

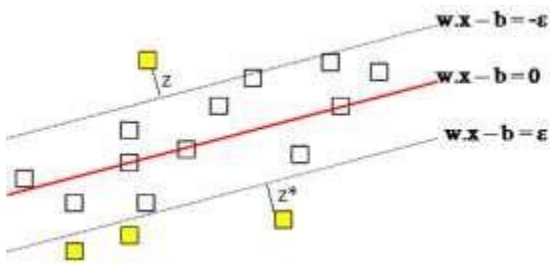
với P_j^2 là xác suất của lớp j trong T

Gini (T) là nhỏ nhất nếu những lớp trong T bị lệch: $gini_{split}(T) = \frac{N1}{N}(gini(T1) + \frac{N2}{N} gini(T2)$ (2)

Thuộc tính có $gini_{split}$ (T) nhỏ nhất sẽ được chọn

B. *Máy học véc-tơ hỗ trợ*

Thuật toán SVM ban đầu được tìm ra bởi Vladimir N. Vapnik và dạng chuẩn hiện nay sử dụng lề mềm được tìm ra bởi Vapnik và Corinna Cortes năm 1995.



Hình 2. Máy học véc-tơ hỗ trợ cho vấn đề hồi quy

Giải thuật SVM tìm siêu phẳng tối ưu (xác định bởi véc-tơ pháp tuyến *w* và độ lệch của siêu phẳng *b*), đi qua tất cả các phần tử dữ liệu với độ lệch chuẩn là ε (dựa trên 2 siêu phẳng hỗ trợ, $w.x - b = \varepsilon$ và $w.x - b = -\varepsilon$). Những phần tử nằm phía ngoài siêu phẳng hỗ trợ được coi như lỗi. Khoảng cách lỗi được biểu diễn bởi $z_i \geq 0$ (với x_i nằm phía trong của 2 siêu phẳng hỗ trợ của nó thì khoảng cách lỗi tương ứng $z_i = 0$, còn ngược lại thì $z_i > 0$ là khoảng cách từ điểm x_i đến siêu phẳng hỗ trợ tương ứng của nó). Huấn luyện máy học SVM cho xử lý vấn đề hồi quy dẫn đến việc giải bài toán quy hoạch toàn phương (4) như sau:

$$\min \Psi(w, b, z^*, z) = (1/2)\,||w||^2 + c \sum_{i=1}^{m} (z_i^* + z_i)$$

$$s.t. \quad (4)$$

$$w.x_i - b - y_i - z_i^* \leq \varepsilon$$

$$w.x_i - b - y_i + z_i \geq -\varepsilon$$

z_i^* , $z_i \geq 0$ *(i=1, 2, …, m)*

với hằng c > 0 được sử dụng để chỉnh độ rộng lề và lỗi.

Giải bài toán quy hoạch toàn phương (4) sẽ thu được siêu phẳng hồi quy (*w, b*) của SVM. Dự báo cho phần tử mới đến *x* dựa trên siêu phẳng (*w, b*) được tính theo công thức (5):

$$predict(x) = (w.x - b) \quad (5)$$

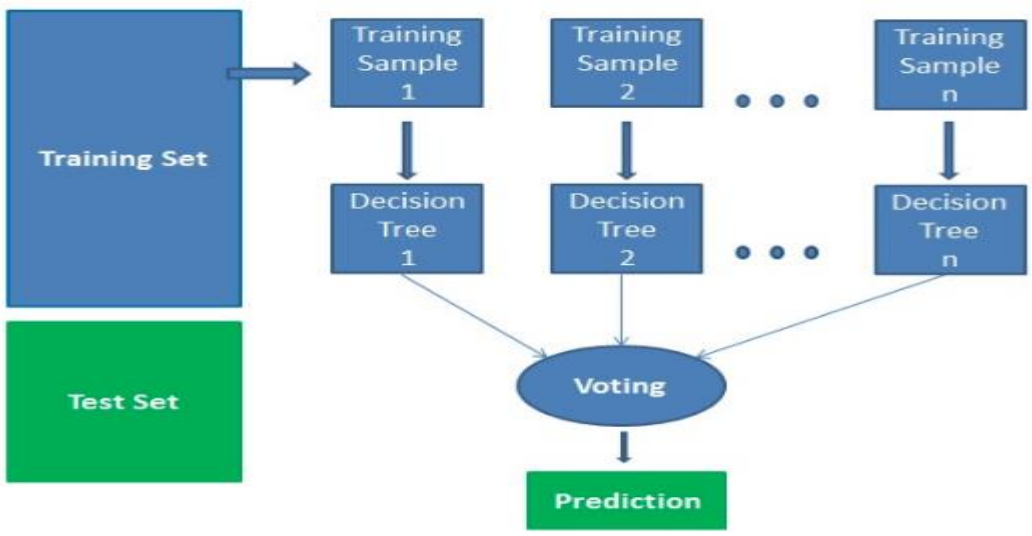
Máy học SVM có thể sử dụng các hàm nhân khác nhau để giải quyết lớp các bài toán phân lớp phi tuyến [Cristianini & Shawe-Taylor, 00]. Để xử lý các vấn đề phân lớp phi tuyến, không cần bất kỳ thay đổi nào hơn từ giải thuật mà chỉ cần thay thế hàm nhân tuyến tính trong công thức bằng các hàm nhân khác. Có 2 hàm nhân phi tuyến phổ biến là:

- Hàm đa thức bậc d: $K(x_i, x_j) = (x_i \cdot x_j + 1)^d$ (6)
- Hàm cơ sở bán kính (Radial Basic Function – RBF): $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$ (7)

Mô hình máy học SVM cho kết quả cao, ổn định, chịu đựng nhiễu tốt và phù hợp với các bài toán phân lớp, hồi quy. Nhiều ứng thành công của SVM đã được công bố trong nhiều lĩnh vực như nhận dạng ảnh, phân loại văn bản và sinh-tin học .

C. *Rừng ngẫu nhiên*

Rừng ngẫu nhiên là một thuật toán học có giám sát. Như tên gọi của nó, rừng ngẫu nhiên dùng các cây để làm nền tảng. Rừng ngẫu nhiên là tập hợp của các Decision Tree mà trong đó mỗi cây sẽ được chọn dựa vào thuật toán ngẫu nhiên



Hình 3. Mô hình rừng ngẫu nhiên

Nó hoạt động theo bốn bước:

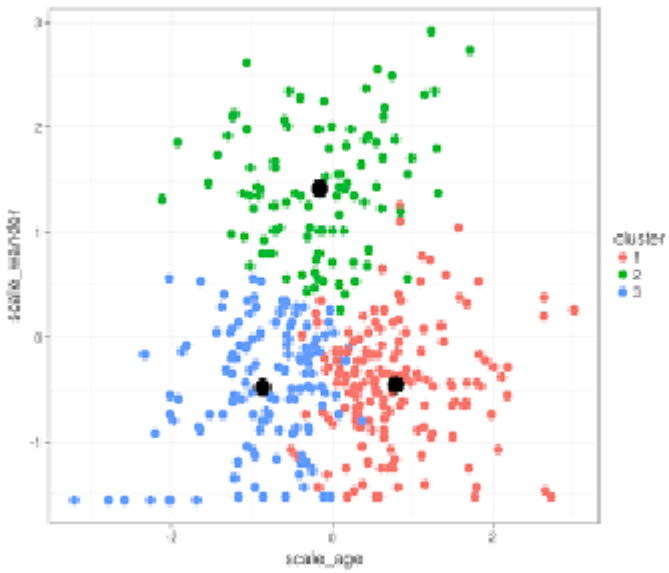
1. Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho.
2. Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây.
3. Hãy bỏ phiếu cho mỗi kết quả dự đoán.
4. Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.

Random forests được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình này, Random forests 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

III. GOM NHÓM DỮ LIỆU

Gom nhóm của dữ liệu thường không có nhiều thông tin sẵn có như lớp (nhãn) , gom nhóm : mô hình gom cụm dữ liệu (không có nhãn) sao cho các dữ liệu cùng nhóm có các tính chất tương tự nhau và dữ liệu của 2 nhóm khác nhau sẽ có các tính chất khác nhau, có nhiều nhóm giải thuật khác nhau : hierarchical clustering, partitioning, density-based, model-based. Được sử dụng nhiều : K-Means, Dendrogram, SOM, EM và được ứng dụng thành công trong hầu hết các lãnh vực tìm kiếm thông tin, phân tích dữ liệu.

Gom nhóm thường dựa trên cơ sở khoảng cách, nên chuẩn hóa dữ liệu, khoảng cách được tính theo từng kiểu của dữ liệu : số, nhị phân, loại, kiểu symbol (interval, histogram, taxonomy



Hình 4. Gom nhóm dữ liệu

Giải thuật K-Means: khởi động ngẫu nhiên K tâm (center) của K clusters. Mỗi phần tử được gán cho tâm gần nhất với phần tử dựa vào khoảng cách (khoảng cách Euclid). Cập nhật lại các tâm của K clusters, mỗi tâm là giá trị trung bình (mean) của các phần tử trong cluster của nó. Lặp lại bước 2,3 cho đến khi hội tụ

IV. BIỂU DIỄN DỮ LIỆU

Một sở thú muốn tìm hiểu các đặc tính cũng như ngoại hình của từng loài thú có trong sở thú đó (cụ thể có thể là lông hay tóc,loại,…) đối với các loài thú khác nhau thuộc các loại khác nhau như trên cạn hay dưới nước. Tập dữ liệu zoo dataset trên website UCI đã cho chúng ta 17 đặc trưng trong số rất nhiều đặc trưng của rất nhiều loài động vật trên thế giới. Nối kết bảng class.csv với zoo.csv theo hai thuộc tính class_type và Class_Number tạo ra một tập dữ liệu mới bao gồm các thuộc tính trong zoo.csv và Class_Number, Number_Of_Animal_Species_In_Class, Class_Type, Animal_Names để dễ dàng phân loại các loài thú theo lớp**.** Việc xây dựng mô hình phân loại này sẽ giúp cho việc chăm sóc dễ dàng hơn đối với các loài động vật có trong sở thú. Chi tiết dữ liệu sẽ được biểu diễn dưới bảng sau:

Thuộc tính	Định nghĩa	Giá trị
animal_name	Tên động vật	Aardvark, antelope,…
hair	Bộ lông	0,1
feathers	Lông vũ	0,1
eggs	Trứng	0,1

milk	Sữa	0,1
airborne	Sống trên cạn	0,1
aquatic	Sống dưới nước	0,1
predator	Ăn thịt	0,1
toothed	Răng	0,1
backbone	Xương sống	0,1
breathes	Hơi thở	0,1
venomous	Nọc độc	0,1
fins	Vây	0,1
legs	Chân	0,2,4,5,6,8
tail	Đuôi	0,1
domestic	Có gia đình	0,1
catsize	catsize	0,1
class_type	Lớp	1-7
Class_Number	Số thứ tự lớp	1-7
Number_Of_Animal_Species_In_Class	Số lượng trong lớp	41,13,…
Class_Type	Lớp	Mammal,Fish,…
Animal_Names	Tên động vật	aardvark, antelope, bear, boar, buffalo, calf,…

v. KẾT QUẢ THỰC NGHIỆM

Để tiến hành đánh giá độ chính xác trong phân loại các loài thú trong tập dữ liệu zoo dataset trên website UCI các giải thuật Cây quyết định, Rừng ngẫu nhiên và Máy học vectơ hỗ trợ đã được cài bằng ngôn ngữ trong ngôn ngữ Python có sử dụng gói thư viện Scikit-learn .Thư viện Scikit-learn cung cấp các giải thuật để xây dựng mô hình hồi quy tuyến tính (LM [Hastie et al., 01]), *k* láng giềng (*k*NN [Fix & Hodges, 52]), máy học véc-tơ hỗ trợ cho hồi quy (SVR), rừng ngẫu nhiên (RF).

Sử dụng Google Colab để viết và thực thi mô hình trong trình duyệt.

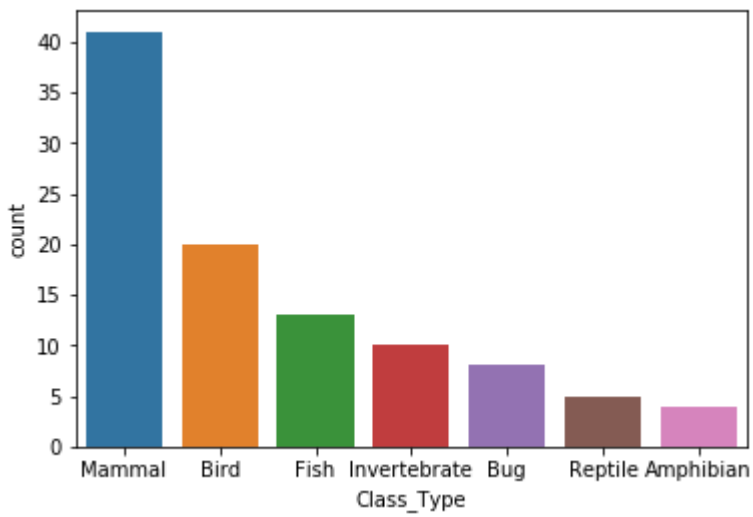
A. *Chuẩn bị tập dữ liệu*

Dữ liệu được lấy trực tiếp từ website UCI. Sau khi tiền xử lý, loại bỏ các thuộc tính không dùng trong dự báo như: animal_name. Thu được 17 thuộc tính, trong đó có 14 thuộc tính kiểu Boolean và 2 thuộc tính kiểu Numeric. Các thuộc tính bao gồm:

1. hair (0,1)
2. feathers (0,1)
3. eggs (0,1)
4. milk (0,1)
5. airborne (0,1)
6. aquatic (0,1)
7. predator (0,1)
8. toothed (0,1)
9. backbone (0,1)
10. breathes (0,1)
11. venomous (0,1)
12. fins (0,1)
13. legs (0,2,4,5,6,8)
14. tail (0,1)
15. domestic (0,1)
16. catsize(0,1)
17. class_type(1-7)

B. *Xây dựng mô hình*

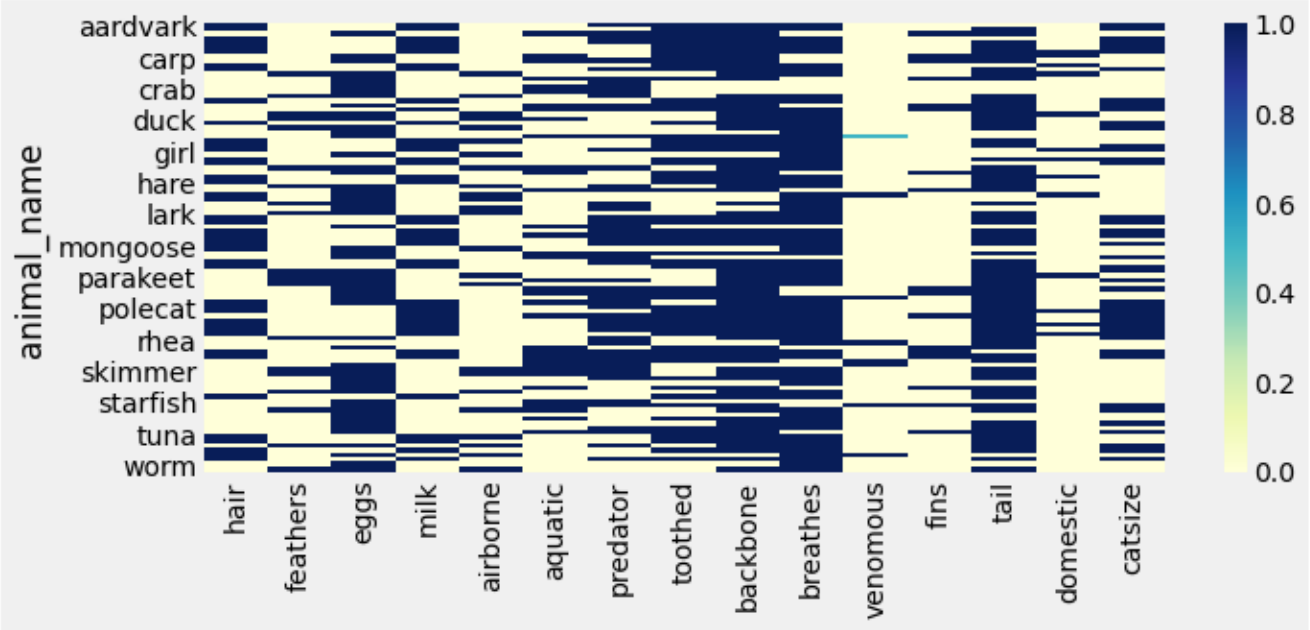
Vẽ biểu đồ thể hiện sự phân bố của các loài động vật trong các lớp.



Hình 6. Biểu đồ thể hiện số lượng các loài động vật trong lớp

Qua hình 6 có thể thấy các loài thú trong tập dữ liệu sẽ gồm có 7 lớp và số lượng sẽ dao động từ 5 đến hơn 40 con trong một lớp chủ yếu là các loại thú thuộc lớp Mammal, và ở các lớp khác số lượng không chênh lệch quá nhiều.

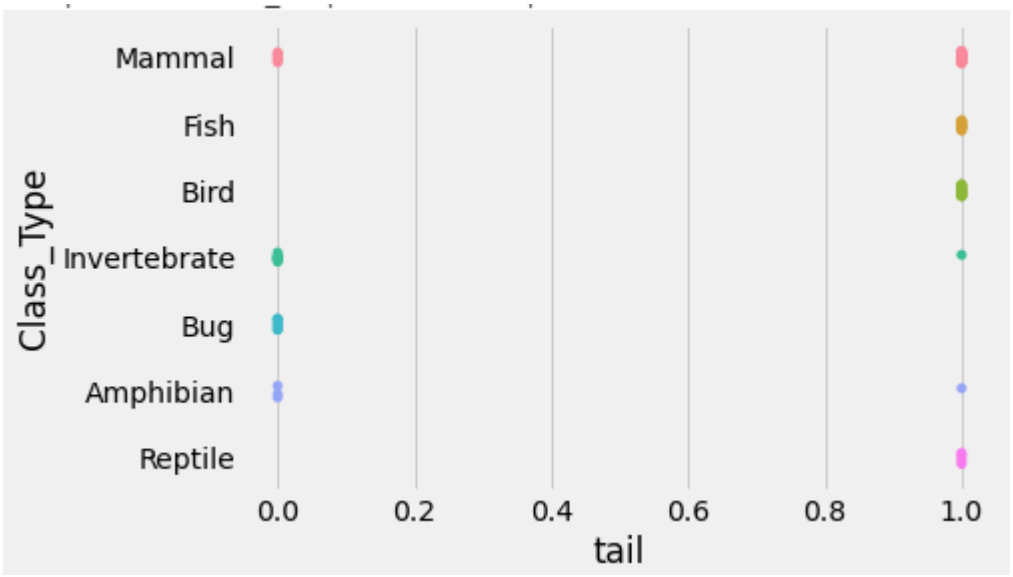
Xem dữ liệu qua heat map:



Hình 7. Biểu đồ thể hiện thuộc tính của các loài động vật

Trong hình 7, có thể thấy các loài động vật trong sở thú dự trên tập dữ liệu chủ yếu là các loài động vật có xương sống và phân bố ở cả trên cạn lẫn dưới nước với các thuộc tính chủ yếu như lông, tóc, hơi thở, rang,… Trong số đó chỉ có một vài loài có nộc độc, có vây, lông vũ và có gia đình.

Tìm hiểu sự khác nhau giữa các thuộc tính trong từng lớp: Sử dụng thuộc tính “legs” để xem sự khác nhau giữa các lớp với thuộc tính.



Hình 8. Biểu đồ thể hiện sự khác biệt thuộc tính “tail” giữa các lớp

Xây dựng mô hình phân loại các loài thú sử dụng tập dữ liệu có được để phân loại các loài thú từ 16 thuộc tính.Chia tập dữ liệu ngẫu nhiên thành hai tập dữ liệu con training set và test set theo tỷ lệ 70/30. Sau đó Xây dựng mô hình với giải thuật Cây quyết định(Decision Tree), Rừng ngẫu nhiên (Random Forest) và Máy học vectơ hỗ trợ (Support Vector Machines) để dự đoán độ chính xác.

Sau khi đưa dữ liệu đã xử lý vào các mô hình độ chính xác là:

	Mô hình	Độ chính xác
1	Cây quyết định	0.96
2	Rừng ngẫu nhiên	0.96
3	Máy học vectơ hỗ trợ	0.90

Hình 5. Độ chính xác của giải thuật

Gom nhóm dữ liệu sử dụng giải thuật kmeans
Dùng hai thuộc tính “hair” và “tail” để gom nhóm

Hình 8. Gom nhóm với giải thuật K-Means

Dựa vào hình 8 có thấy thấy dựa theo hai thuộc tính “hair” và “tail” có thể phân các loài động vật thành 2 nhóm:

- Nhóm 1 là các loài động vật không có lông và không có tóc hoặc các loài động vật có tóc nhưng không có lông
- Nhóm 2 là các loài động vật có lông và không có tóc hoặc các loài động vật có tóc và có lông

VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong bài viết vừa trình bày về xây dựng mô hình phân loại các loài thú với các mô hình cây quyết định, máy học véc-tơ hỗ trợ và rừng ngẫu nhiên để dự báo độ chính xác. Đồng thời dùng giải thuật K-Means để gom nhóm một vài thuộc tính trong tập dữ liệu.

Cơ bản đã xây dựng được mô hình để phân loại các loài thú để nhằm hướng đến phát triển trong tương lai xây dựng thêm nhiều mô hình để có thể phân loại thêm các đối tượng khác trong đời sống.

TÀI LIỆU THAM KHẢO

[1] https://machinelearningcoban.com/

[2] https://stackoverflow.com/

[3] https://www.kaggle.com/