***Đề tài: Kĩ năng phân tích dữ liệu và ứng dụng.***

giới thiệu thành viên nhóm em gồm 9 anh em siêu nhân bla bla bla,...

**I. MỞ ĐẦU**

1. Lý do chọn đề tài

# Phân tích dữ liệu và những tương lai của ngành trong kỷ nguyên số

Cuộc cách mạng công nghiệp 4.0 đang mở ra một kỷ nguyên mới của thế giới. Đi kèm với đó là sự bùng nổ dữ liệu, nhu cầu lưu trữ và khai thác lớn chưa từng có. Hằng ngày, hằng giờ, thậm chí hằng giây mỗi người chúng ta đều tạo nên dữ liệu được ghi nhận lại. Chụp một bức ảnh số, đăng một nhận xét, mua một mặt hàng qua mạng, xem một trang web… tất cả những điều ấy đều là dữ liệu được ghi nhận lại. Năm 2020, mỗi giây, người dùng toàn thế giới tạo ra 100 TB dữ liệu, các công ty và các tập đoàn lớn đã tận dụng chúng và kiếm được hàng tỷ USD.

Cùng với [IoT](https://swinburne-vn.edu.vn/news/iot-la-gi/) (Internet of Things), AI (Artificial Intelligence – Trí tuệ nhân tạo), Blockchain (Chuỗi khối), Big Data (Dữ liệu lớn) là một trong bốn nền tảng quan trọng của cuộc cách mạng công nghệ 4.0. Mà trong đó Big Data được coi là cốt lõi của sự phát triển trong cuộc cách mạng này.

Dữ liệu còn đóng vai trò vô cùng quan trong đối với các doanh nghiệp, tổ chức sẽ sử dụng nguồn dữ liệu phân tích để tìm ra câu trả lời cho các câu hỏi: giảm chi phí, giảm thời gian, phát triển sản phẩm mới và dịch vụ tối ưu, ra quyết định thông minh. Đặc biệt, ở các lĩnh vực như ngân hàng đầu tư, bảo hiểm, du lịch, quốc phòng, hàng không vũ trụ và y học,... dữ liệu cũng là phần không thể thiếu.

Chính vì tầm quan trọng xuất hiện trong mọi lĩnh vực mà dữ liệu đã được coi là nguồn tài nguyên quý giá nhất trong kỉ nguyên số. Để có thể tận dụng được nguồn tài nguyên vô giá này thì kĩ năng phân tích dữ liệu là điều không thể thiếu. Chúng ta sẽ biết cách khai thác dữ liệu, hiểu rõ hơn vai trò và ứng dụng trong thực tế.

B. Ý nghĩa

* Hiểu được ý nghĩa các đại lượng đặc trưng của dữ liệu, Nắm được các phương pháp thu thập dữ liệu.
* Biết cách sử dụng được phần mềm để biểu diễn dữ liệu, tính toán được các số liệu đặc trưng.
* Hình thành khả năng phân tích dữ liệu cơ bản, xử lí vấn đề, tư duy logic. Biết được tính ứng dụng trong thực tiễn.

**II. Nội dung**

**Phần 1: Giới thiệu về bộ dữ liệu**

* Trong đề tài này chúng em thực hiện những phân tích trên bộ dữ liệu iris.
* Bộ dữ liệu Iris chứa bốn đặc điểm (chiều dài và chiều rộng của các lá đài và cánh hoa) của 50 mẫu của ba loài Iris (Iris setosa, Iris virginica và Iris versicolor).
* Bộ dữ liệu iris có kích thước không quá lớn, phù hợp để thực hiện những phân tích cơ bản khi mới làm quen với dữ liệu.
* Tập dữ liệu Iris được sử dụng với mục đích phân loại 3 loài hoa dựa vào các đặc điểm cho trước. Đồng thời phân loại dữ liệu cũng chính là một ứng dụng quan trọng trong phân tích dữ liệu, giúp phân biệt các đối tượng dữ liệu lớn vì có thể nhận diện nhiều đối tượng nhờ các đặc điểm thuộc tính đặc trưng.

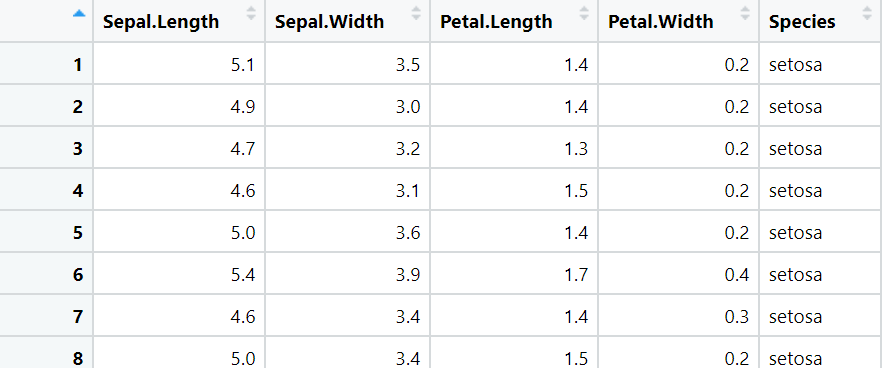
**Phần 2: Khám phá dữ liệu**

**2.1. Tổng quan về dữ liệu**

**\*Phần mềm được sử dụng RStudio**

Trước hết chúng ta nên có được cái nhìn bao quát chung về tổng thể tập dữ liệu

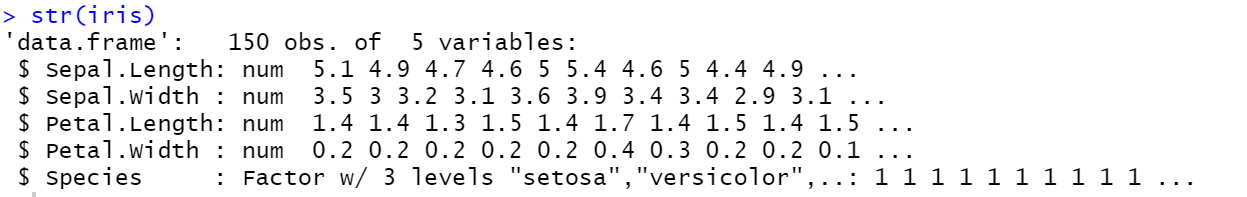




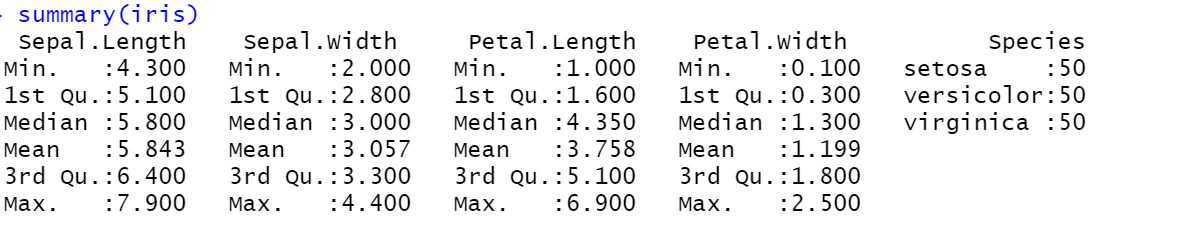
Quan sát bảng trên ta có thể thấy được tập dữ liệu iris được chia thành các hàng và cột. Trong đó 5 cột lần lượt thể hiện các thuộc tính: Chiều dài đài hoa ( Sepal.Length), Chiều rộng đài hoa ( Sepal.Width ), Chiều dài cánh hoa ( Petal.Length ), Chiều rộng cánh hoa ( Petal.Width ), Loài ( Species )

Thông thường chúng ta cần làm sạch dữ liệu trước khi đi đến những phân tích chi tiết hơn bằng cách loại bỏ những dữ liệu bị lỗi. Vì tập dữ liệu iris là tập dữ liệu chuẩn nên chúng ta có thể bỏ qua bước này.

Tuy nhiên, để hiểu được tập dữ liệu 1 cách rõ ràng hơn chúng ta nên biết về các kiểu dữ liệu có trong bảng trên cũng như 1 bản tóm tắt có thể giúp ta có nhiều thông tin hơn về tập dữ liệu này:



Ta có thể thấy rõ ràng hơn, tập dữ liệu này chứa 150 hàng, các số liệu về kích thước có định dạng số, có tổng cộng 3 loài hoa.



Tiếp theo là 1 bản tóm tắt thông tin đặc trưng của mỗi cột. Từ đây đối với cột Sepal.Length ta có thể biết được: giá trị nhỏ nhất là 4,3 ; giá trị lớn nhất là 7,9 ; giá trị trung vị là 5,8 ; giá trị trung bình là 5,843 ;.... Tương tự cho 3 cột kích thước còn lại.

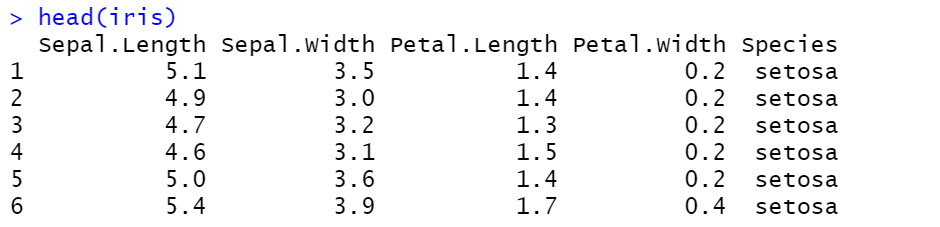
ở cột Species, ta có thể thấy mỗi loài đều có số lượng là 50.

Ngoài ra ta có thể sử dụng lệnh dim() nếu chỉ đơn giản muốn biết về kích thước hàng và cột của tập dữ liệu.



150 hàng - 5 cột

Chúng ta có thể chỉ lấy ra 1 vài dòng đầu của tập dữ liệu thay vì tất cả bằng cách sử dụng lệnh head(). ( Mặc định là 6 dòng đầu tiên )



**2.2 Truy xuất và tính toán với dữ liệu**

Sau khi đã tìm hiểu tổng quan về dữ liệu, chúng ta sẽ học cách truy xuất và lấy ra các dữ liệu phục vụ mục đích tính toán cũng như thao tác trên dữ liệu.

VD: iris[ 3, 4 ] # Lấy ra phần tử ở hàng thứ 3 , cột thứ 4.

iris[3, ] # Lấy ra các phần tử nằm ở hàng thứ 3.

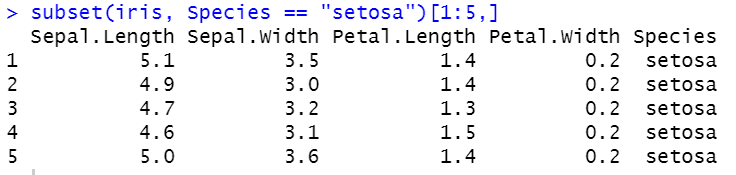
iris[ , 4 ] # Lấy ra các phần tử nằm ở cột thứ 4.

iris[ , (1:4) ] # Lấy ra các phần tử nằm ở cột thứ 1 đến cột thứ 4.

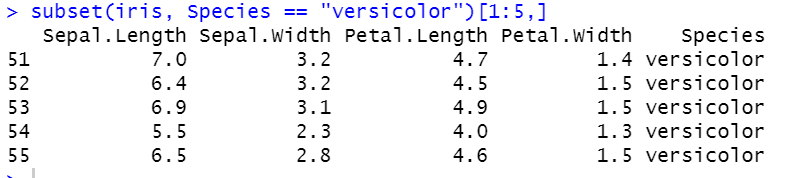
iris$Species # Lấy ra cột Species

Ngoài cách truy xuất như ở trên ta có thể sử dụng các hàm có sẵn.

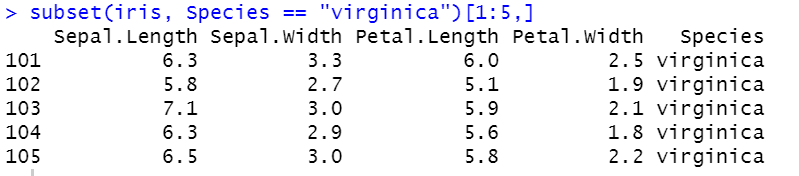
Lấy ra 5 hàng đầu tiên có Species là “ setosa ”:



Lấy ra 5 hàng đầu tiên có Species là “ versicolor ”:

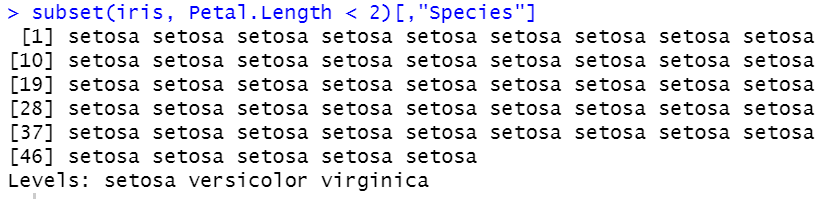


Lấy ra 5 hàng đầu tiên có Species là “ virginica ” :



Nhìn nhanh 3 bảng kết quả trên, chúng ta có thể thấy Petal.Length của loài “setosa” khá ngắn và ngắn hơn so với 2 loài còn lại. Liệu điều này có đúng cho cả tập dữ liệu ?

Vậy thì hãy cùng xem, bây giờ chúng ta sẽ thử truy xuất những phần tử ở cột “Species” mà có Petal.Length < 2 :



Từ bảng kết quả trên, ta có thể thấy chúng ta vừa đưa ra 1 nhận xét chính xác. Đây cũng chính là 1 bước đầu tiên để có thể hiểu rõ hơn về tập dữ liệu chúng ta đang làm cũng như xây dựng các mô hình tiếp theo

.

Tiếp theo sẽ là 1 số câu lệnh dùng để tính toán các đại lượng đặc trưng:

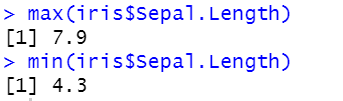
Hàm tính trung bình: mean()



Hàm tính trung vị: median()



Hàm tính giá trị lớn nhất, nhỏ nhất: max() , min()



Bên cạnh đó là các hàm tính phương sai, độ lệch chuẩn: var() , sd(),...

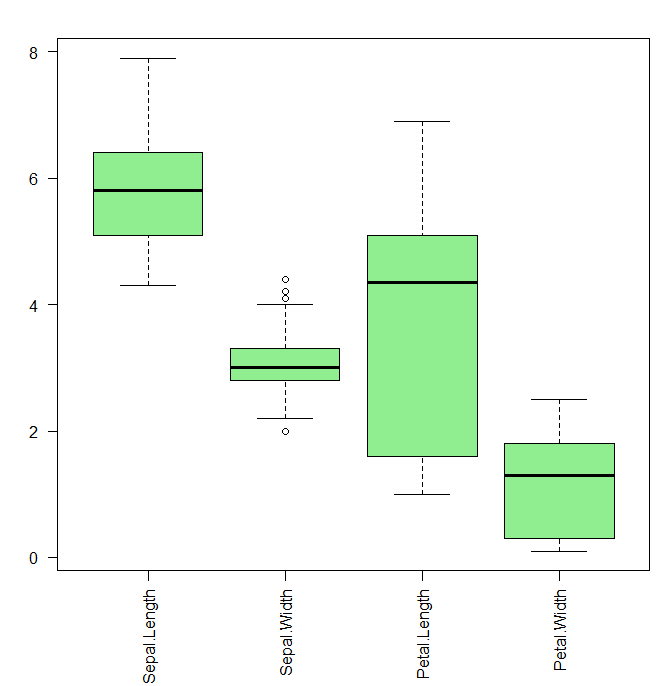
**Phần 3. Trực quan hóa dữ liệu**

Sau khi đã có cái nhìn tổng quan về tập dữ liệu cũng như 1 số phân tích đơn giản, những con số đã cho chúng ta biết nhiều điều, nhưng để có thể thực hiện phân tích dữ liệu một cách dễ hiểu, trình bày gọn gàng hơn đồng thời nhanh chóng và hiệu quả thì việc thực hiện trực quan hóa dữ liệu là điều vô cùng quan trọng.Cụ thể hơn, chúng ta sẽ biểu diễn và phân tích các dữ liệu thông qua hình ảnh trên biểu đồ.

1. Boxplot

Biểu đồ boxplot là hay box-and-whisker plot là biểu đồ giúp biểu diễn một số đại lượng quan trọng như giá trị nhỏ nhất (min), giá trị lớn nhất (max), tứ phân vị (quartile), khoảng biến thiên tứ phân vị (interquartile range) một cách trực quan và dễ hiểu.

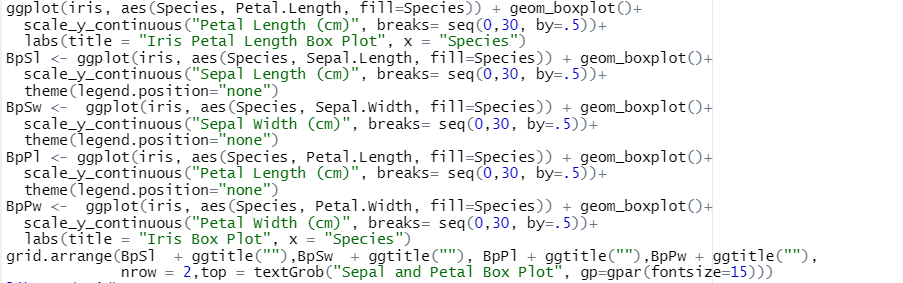


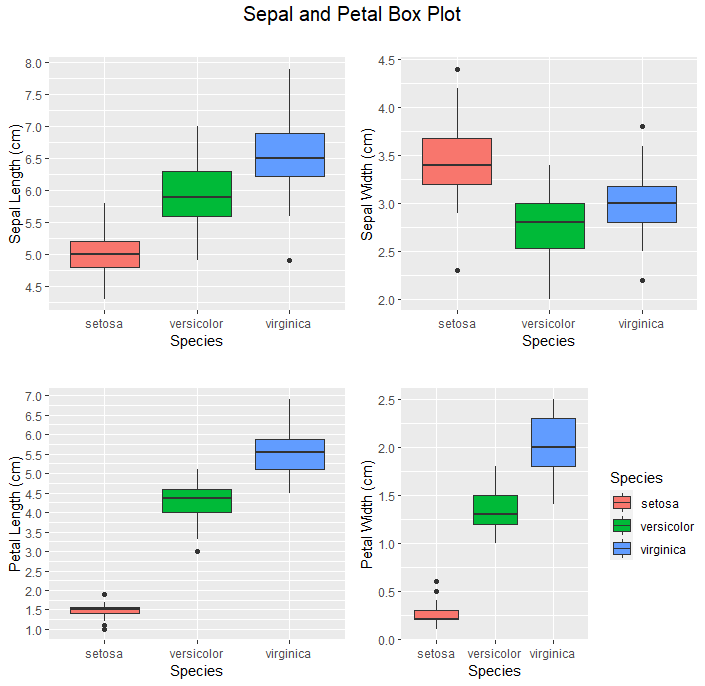


Từ biểu đồ boxplot trên, ta có thể rút ra một số nhận xét tổng quát về các thuộc tính chung của 3 loài hoa:

* Chiều dài đài hoa phân bố trong khoảng từ 4 đến 8
* Chiều rộng đài hoa phân bố trong khoảng từ 2 đến 4.
* Chiều dài cánh hoa phân bố trong khoảng từ 1 đến 7.
* Chiều rộng cánh hoa phân bố trong khoảng từ 0 đến 2
* Và dễ quan sát thấy dữ liệu về tỉ lệ chiều dài cánh hoa giữa các loài có sự dao động rất lớn điều này cho thấy các chiều dài về cánh hoa phân bố không đồng đều ở mỗi loài khác nhau. Vậy nên tỉ lệ về chiều dài cánh hoa có thể dùng để phân biệt giữa các loài nó mang lại tỉ lệ chính xác rất cao. Trái lại thì tỉ lệ về chiều rộng của đài hoa lại dao động hẹp và xuất hiện những điểm bất thường nên tỉ lệ này sẽ không hữu ích nếu đưa ra để phân biệt các loài.

Ta thấy, biểu đồ trên chỉ cung cấp những ước tính sơ bộ về sự phân bố cho từng thuộc tính. Nhưng để thấy được sự phân bố rõ ràng hơn của các thuộc tính theo từng loài, ta có thể lập biểu đồ boxplot cho từng loài hoa như sau:





Nhận xét về 4 biểu đồ boxplot

-Từ 4 biểu đồ hộp trên ta thấy chiều dài cánh hoa, chiều rộng cánh hoa, chiều dài đài hoa của loài Setosa hầu hết đều ngắn hơn so với các loài còn lại.

(chỉ có chiều rộng đài hoa là vượt hơn so với hai loài còn lại)

-Đối với loài Versicolor thì dễ thấy các chỉ số luôn dao động ở mức trung bình so với hai loài còn lại.

-Đối với loài Virginica ta thấy sự vượt trội về cả chiều dài đài hoa, chiều dài cánh hoa, chiều rộng cánh hoa.

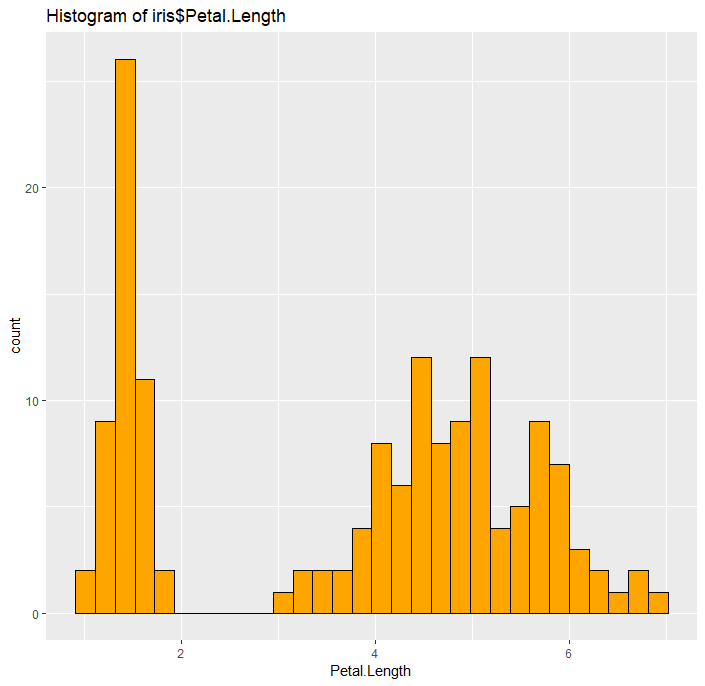
Đặc biệt chúng ta có thể thấy rằng dữ liệu về tỉ lệ chiều dài cánh hoa của loài Setosa tập trung ở mức rất thấp với dao động ở phạm vi hẹp.

Khác hẳn so với chiều dài cánh hoa của Setosa thì tỉ lệ chiều rộng cánh hoa của loài Virginica có mức tập trung ở mức cao với dao động ở phạm vi rất lớn cho thấy mức độ không đồng đều của các cánh hoa ở loài này.

1. Histogram

Ngoài ra, biểu đồ Histogram cũng rất hữu ích cho việc chỉ ra sự phân bố. Dữ liệu được biểu diễn bằng các cột trên biểu đồ và có độ cao khác nhau tùy thuộc vào tần suất phạm vi dữ liệu cụ thể xảy ra.



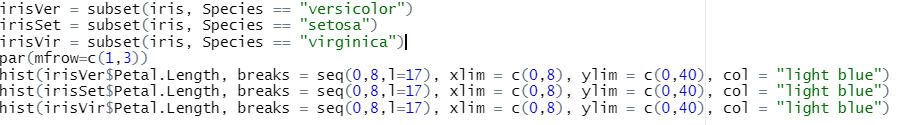


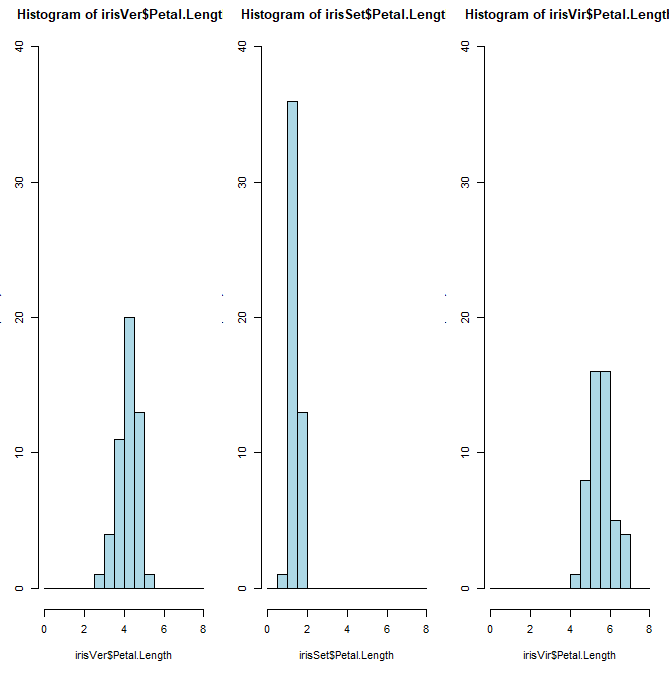
Từ biểu đồ trên chúng ta nhận thấy rằng chiều dài của cánh hoa tập trung nhiều nhất ở khoảng giá trị 1,5(cm).

Tuy nhiên mức độ tập trung của các chiều dài cánh hoa lại ở khoảng 4-6(cm).

Và với hình dạng biểu đồ được chia thành 2 đảo nhỏ như trên hình có thể cho thấy rằng các tệp dữ liệu của loài có sự trộn lẫn giữa 3 nhóm loài có giá trị phân bố khác nhau về chiều dài cánh hoa.

Cụ thể hơn cho từng loài, ta có các đồ thị sau:



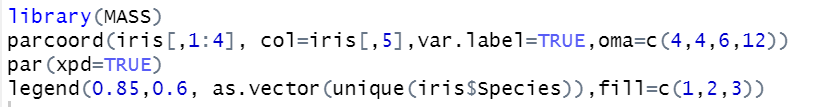


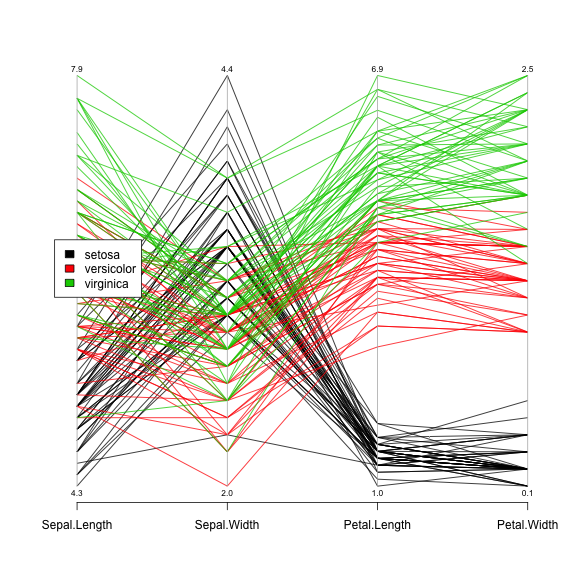
Những biểu đồ trên cho thấy rằng sự phân bố các giá trị cho Petal.Length là khác nhau đối với mỗi loài.

Có thể thấy rằng chiều dài cánh hoa của 3 loài Setosa, Versicolor, Virginica phân bố trong khoảng 0-8(cm) hầu như rất khác biệt nhau.

1. **parallel coordinate plot**

Một cách để vẽ biểu đồ giá trị của khung dữ liệu để xem các mối tương quan và giá trị nói chung là bằng biểu đồ parallel coordinate plot.

****



Từ biểu đồ trên, ta thấy được sự tương quan về chiều rộng của cánh hoa giữa 3 loài: loài hoa virginica có chiều rộng cánh hoa lớn nhất, bé hơn là loài hoa versicolor và bé nhất là loài hoa setosa.

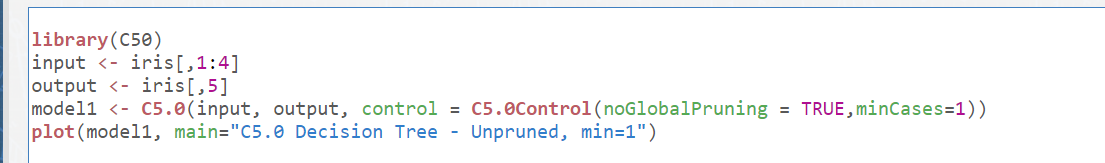
**Phần 4: Mô hình phân loài hoa và ứng dụng trong thực tế**

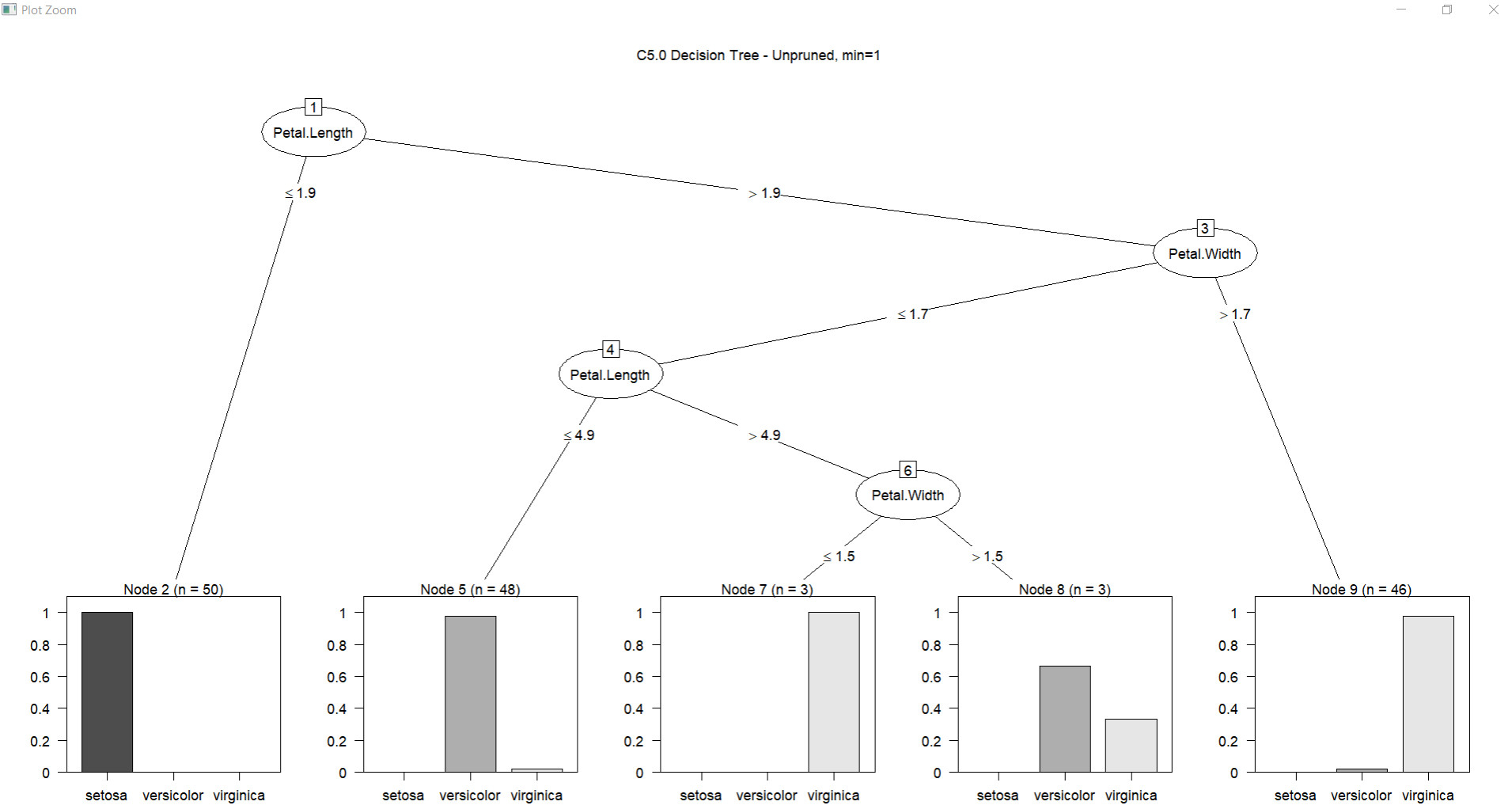
Ta đã biết việc phân tích và hiểu dữ liệu là rất quan trọng. Sau khi đã phân tích được dữ liệu, đưa ra các nhận xét thì chúng ta có thể ứng dụng vào những mô hình, tùy vào mục đích của người sử dụng.

Cụ thể hơn chúng em sẽ trình bày về *mô hình* dự đoán loài từ chiều rộng và chiều dài của cánh hoa và đài hoa. Một mô hình học máy đơn giản nhưng có thể giúp ta làm quen được một số thuật ngữ mới, hình dung được vai trò, tính ứng dụng của dữ liệu trong thực tiễn.

Một trong những mô hình dự đoán dễ hiểu và dễ thực hiện đó là cây quyết định, có thể được tạo nên bằng cách sử dụng package C5.0 ( Package C5.0 là thuật toán học máy được xây dựng sẵn có trong R ).

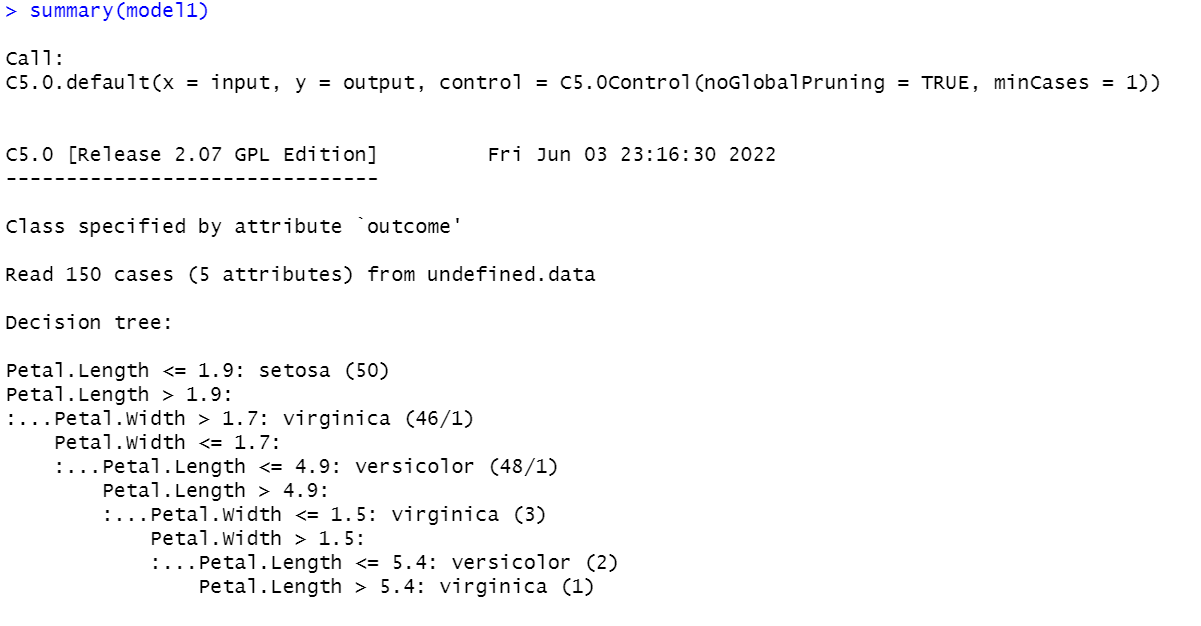
Bây giờ chúng ta tạo ra 1 cây quyết định để dự đoán loài. Kèm với đó là 1 bản tóm tắt về mô hình chúng ta vừa xây dựng.

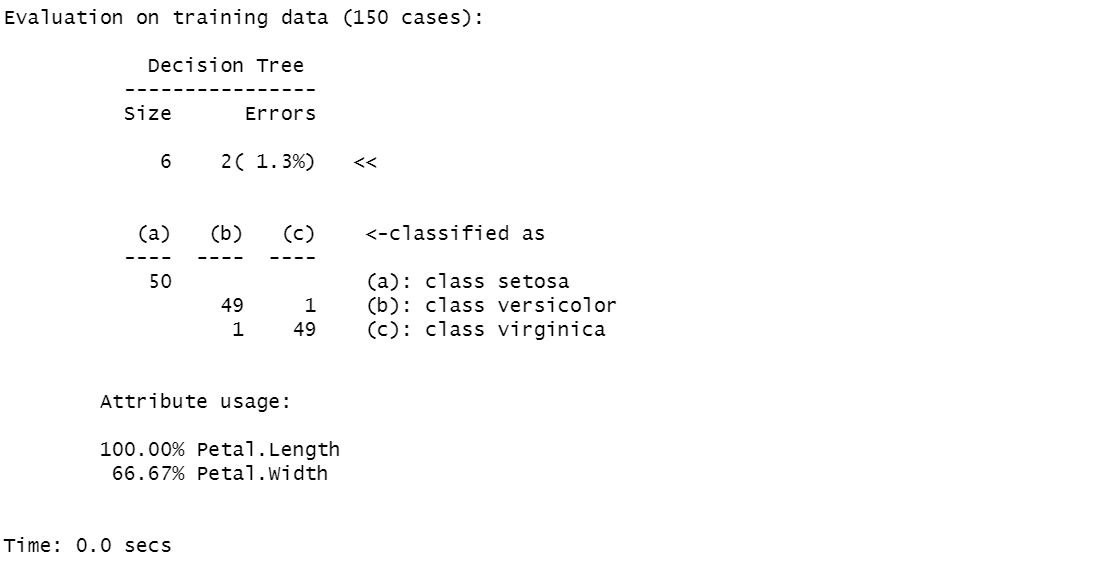




Từ cấu trúc cây phía trên ta có thể thấy từ dữ liệu đầu vào các thuộc tính được sử dụng sau đó tạo ra các ranh giới phân chia thành các nhóm nhỏ dần và đi đến dự đoán kết quả về loài.

Để có thể đánh giá mô hình này một cách tốt hơn ta nên kết hợp thêm với một bản tóm tắt về mô hình chúng ta vừa xây dựng.

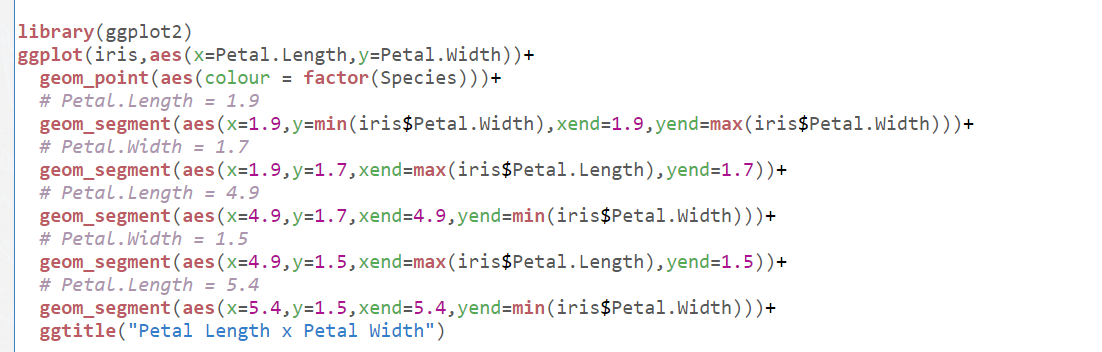


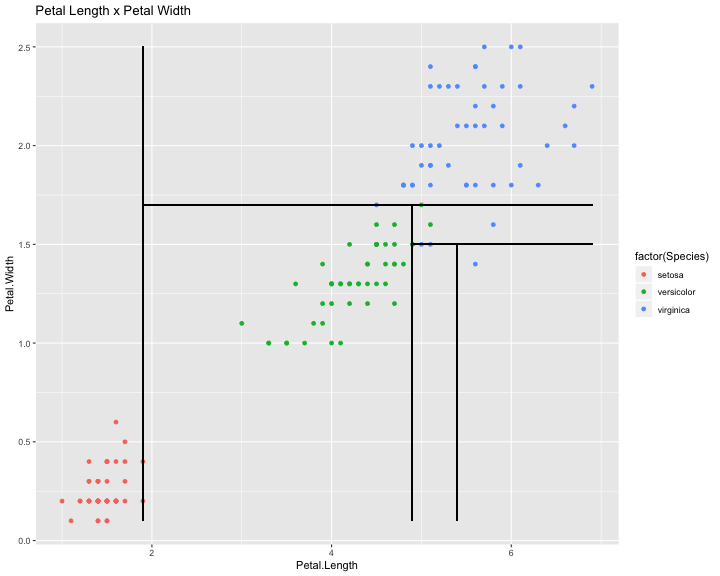


Có thể thấy được sự phân chia ở cây được thể hiện như sau. Đầu tiên Pental.Length được sử dụng nếu nhỏ hơn hoặc bằng 1,9 thì đưa đến kết quả chứa 50 loài setosa. nếu lớn hơn 1.9 thì cây tiếp tục phân nhánh dựa trên thuộc tính Pental.Width. Nếu lớn hơn 1.7 thì kết luận có 46 mẫu virginica và 1 mẫu không phải virginica, ngược lại nhỏ hơn hoặc bằng 1.7 cây tiếp tục phân nhánh dựa Pental.Length nếu nhỏ hơn hoặc bằng 4.9 thì có 48 mẫu là versicolor và 1 mẫu khác ,....

Đánh giá kết quả phân loại sau khi sử dụng cây quyết định thì có thể thấy tổng thể có 2 trường hợp bị phân loại sai. 50 trường hợp setosa được phân loại chính xác.Với versicolor thì 1 trường hợp bị phân loại sai thành virginica còn lại 49 trường hợp chính xác. Với Virginica thì có 1 trường hợp bị phân loại sai thành versicolor còn lại 49 trường hợp chính xác. Đồng thời 100% thuộc tính Pental.Length và 66,67% thuộc tính Pental.Width được sử dụng.

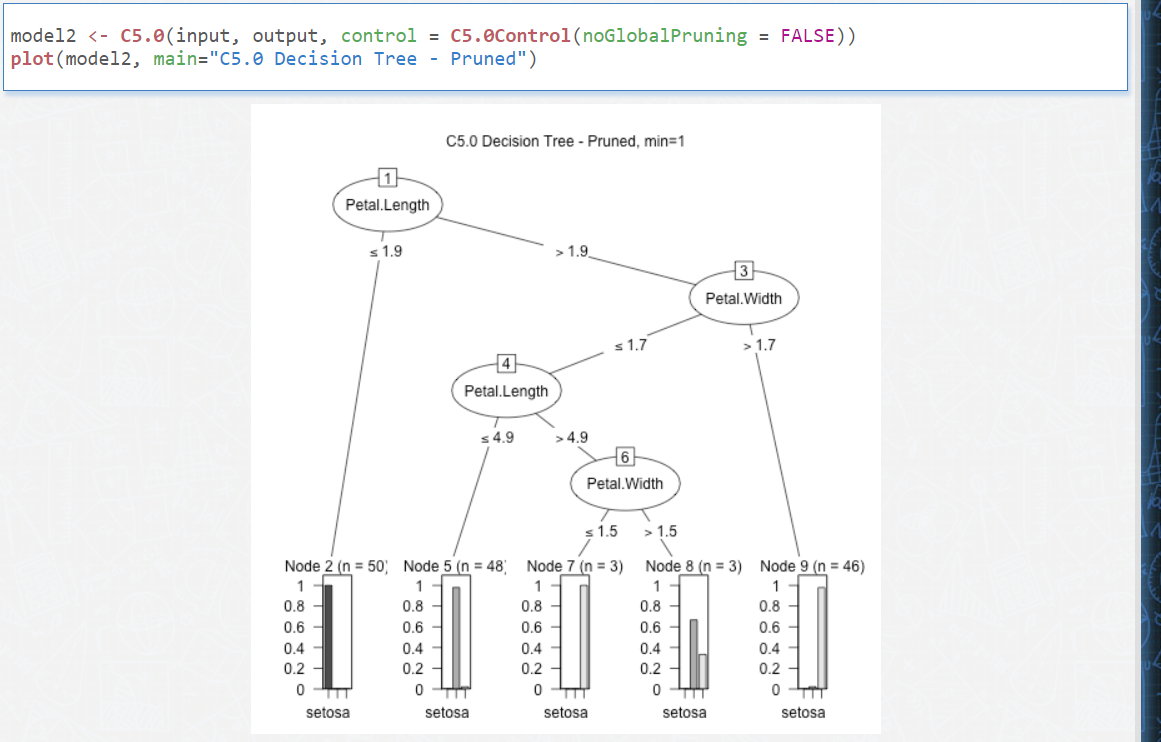
Để có thể thấy rõ hơn về sự phân bộ của mỗi loài ứng với 2 thuộc tính được sử dụng ta có thể phân chia các miền theo danh giới trên biểu đồ như sau:

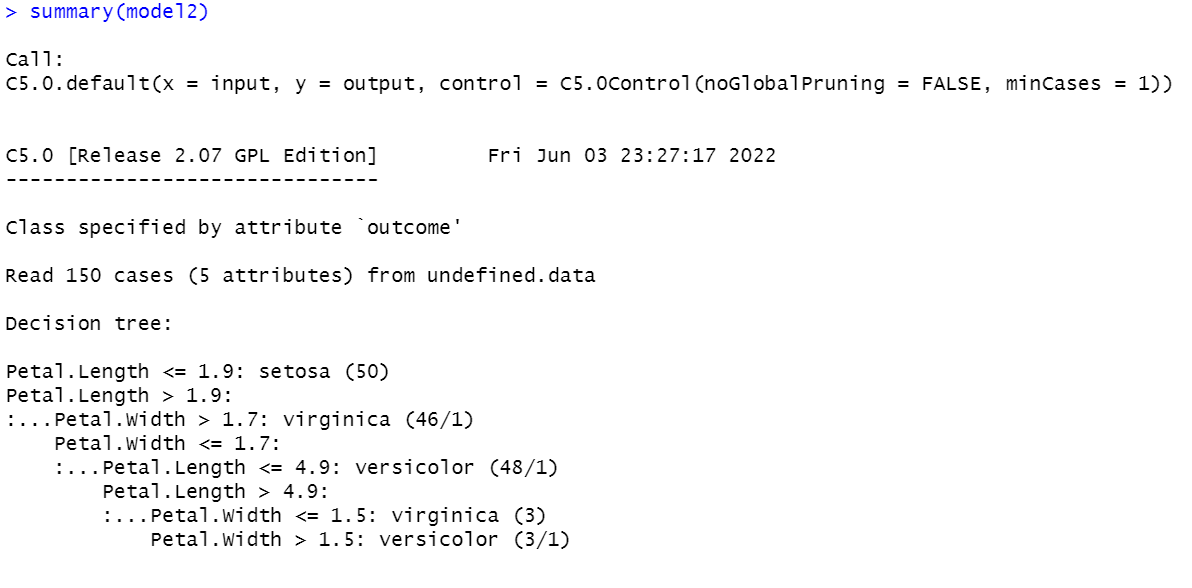


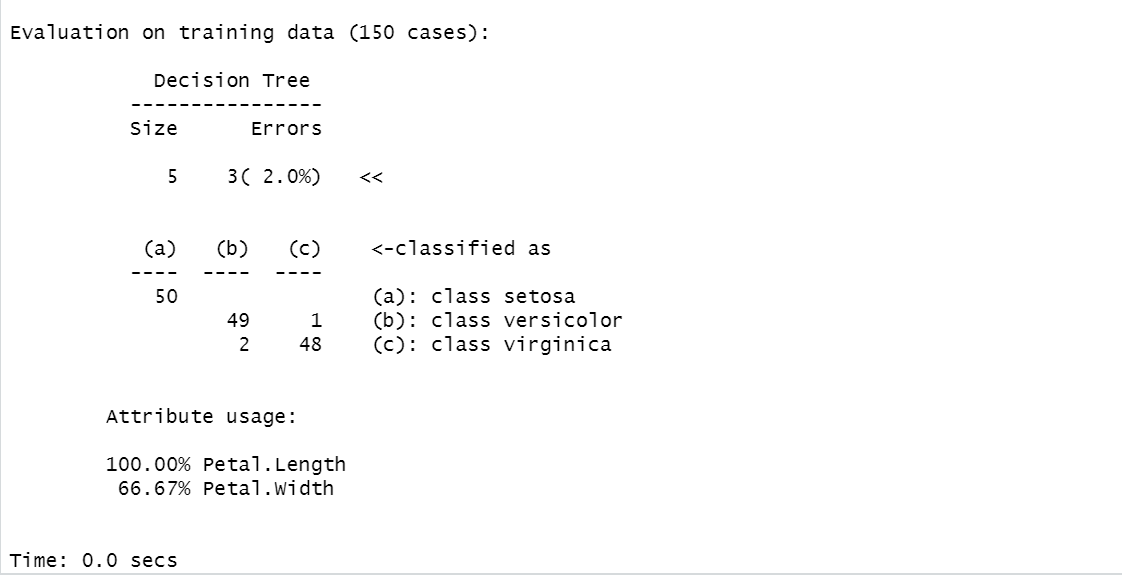


Từ đồ thị có thể thấy rõ ràng về sự phân bố của các loài theo các ranh giới có được từ mô hình chúng ta vừa xây dựng.

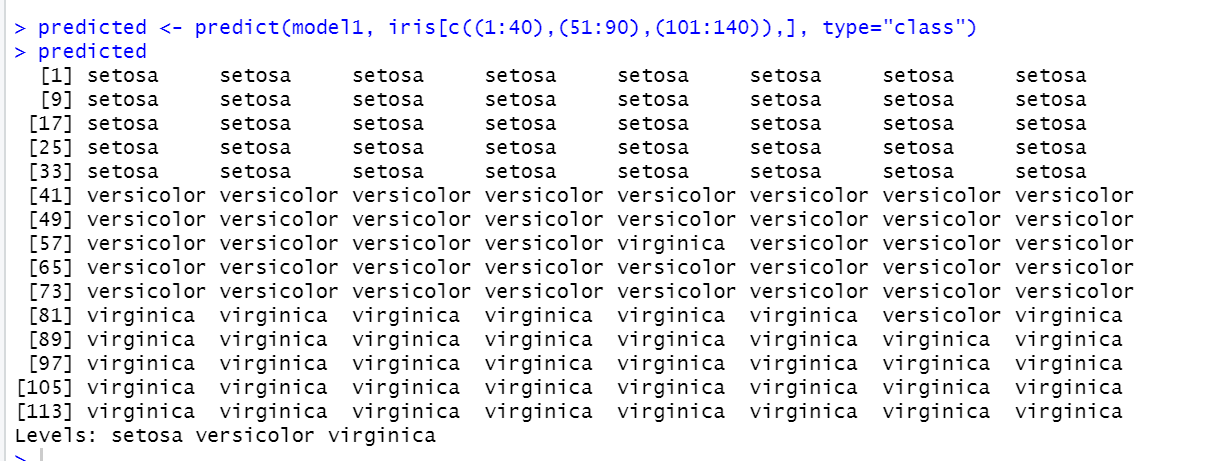
Ngoài ra chúng ta cũng có thể tạo ra 1 mô hình khác tương tự nhưng đơn giản hơn:



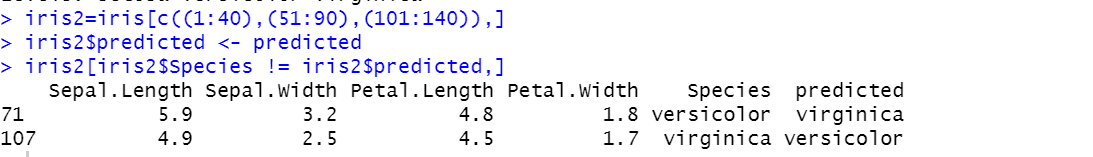




Sau khi đã xây dựng được mô hình chúng ta sẽ thực hiện kiểm thử bằng cách lấy ra phần phần dữ liệu và thử chạy qua mô hình :



Đây là kết quả mà mô hình của chúng ta dự đoán được vậy hãy xem có bao nhiêu trường hợp dự đoán khác với kết quả gốc.



Có 2 trường hợp kết quả dự đoán bị sai lệch. Từ đây có thể nhận xét mô hình này cho ta tỷ lệ dự đoán chính xác khá lớn.

Ta có thể thấy được rằng việc xử lí, phân tích được dữ liệu là công đoạn rất quan trọng và là đầu tiên để chúng ta có thể đưa ra các dự đoán, triển khai những mô hình lớn khác. Chúng có rất nhiều ứng dụng trong thực tế là tương lai phát triển của nhân loại. Tiêu biểu như mô hình cảnh báo rủi ro, tự động phân loại có thể đưa đến các thông tin phù hợp với từng người dùng mà các mạng xã hội lớn đều đang sử dụng, nhận diện hình ảnh, trí tuệ nhân tạo, IoT….

**III. KẾT LUẬN**

Tầm quan trọng của dữ liệu không nằm ở lượng dữ liệu mà chúng ta có, nó nằm ở việc chúng ta làm gì với những dữ liệu đó. Kĩ năng Phân tích dữ liệu giúp ta có thể thập, khai thác, quản lý và xử lý bộ dữ liệu , từ đó đưa ra các nhận định, dự đoán xu hướng hoạt động của tương lai.

Hiện nay các hệ thống phân tích dữ liệu tự động đang được đưa vào sử dụng. Tuy nhiên, nó vẫn chưa thể đáp ứng hoàn toàn nhu cầu của người sử dụng. Theo các nghiên cứu, 80% lượng công việc không thể tự động hóa; 20% còn lại có thể thực hiện bằng máy nhưng hiệu quả chưa cao. Hơn nữa, máy học tự động chỉ có thể giải quyết được những vấn đề đơn giản. Các vấn đề phức tạp hơn cần đến tư duy của con người mới có thể giải quyết được. Do đó, Phân tích dữ liệu sẽ luôn có một vị trí đặc biệt quan trọng ngay cả khi công nghệ phát triển trong tương lai.

**IV.LỜI CẢM ƠN**

Đầu tiên, nhóm em xin gửi lời cảm ơn chân thành đến cô Nguyễn Thị Phương Mai - người đã trực tiếp hướng dẫn cũng như cung cấp tài liệu, thông tin khoa học cần thiết cho bài luận này. Trong quá trình học tập, chúng em đã nhận được sự quan tâm, giúp đỡ rất tận tình của cô trong giờ thực hành cũng như giờ lý thuyết. Và những kiến thức đó đã được chúng em vận dụng vào mô hình bài toán phân tích để hoàn thành bài tiểu luận này.

Mặc dù đã có nhiều cố gắng trong quá trình thực hiện đề tài, nhưng khó tránh khỏi những thiếu sót.Chúng em rất mong nhận được những góp ý từ thầy cô và các bạn để bài tiểu luận của nhóm em được hoàn thiện nhất.

Kính chúc cô sức khỏe và thành công trên con đường sự nghiệp giảng dạy của mình!