TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

KHOA CÔNG NGHỆ THÔNG TIN

TRỰC QUAN HÓA DỮ LIỆU – 21KHDL

BÁO CÁO LAB 3

PHÂN TÍCH CÁC YẾU TỐ GIÚP PHÂN LOẠI HOA IRIS (DIÊN VĨ)

DANH SÁCH THÀNH VIÊN

Họ và tên	MSSV	Mức độ đóng góp	Tỉ lệ thực hiện
Võ Duy Anh	21127221	100%	30%
Nguyễn Mậu Gia Bảo	21127583	100%	30%
Vũ Minh Phát	21127739	100%	40%

GIẢNG VIÊN HƯỚNG DẪN: Bùi Tiến Lên

Lê Ngọc Thành

Lê Nguyễn Nhựt Trường

Thành phố Hồ Chí Minh, ngày 20 tháng 04 năm 2024

Mục lục

1. THÔNG TIN CHUNG

1.1. Thông tin nhóm, mức độ đóng góp và tỉ lệ thực hiện của mỗi thành viên

Họ và tên	MSSV	Mức độ đóng góp	Tỉ lệ thực hiện
Võ Duy Anh	21127221	100%	30%
Nguyễn Mậu Gia Bảo	21127583	100%	30%
Vũ Minh Phát	21127739	100%	40%

1.2. Các câu hỏi chưa làm được

Cả nhóm đã hoàn thành toàn bộ công việc và trả lời mọi câu hỏi được giao.

2. GIỚI THIỆU VỀ BỘ DỮ LIỆU IRIS

Theo thông tin mô tả từ Kaggle, ta biết được:

- Bộ dữ liệu Iris được sử dụng trong bài báo kinh điển năm 1936 của R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems", và ta cũng có thể tìm thấy bộ dữ liệu này tại trang UCI Machine Learning Repository.
- Bộ dữ liệu này bao gồm ba loài hoa diên vĩ (hoa Iris) với 50 mẫu mỗi loài cùng một số thuộc tính về mỗi loài hoa. Trong đó, một loài hoa có thể phân tách tuyến tính với hai loài còn lại, nhưng hai loài hoa còn lại không thể phân tách tuyến tính với nhau.

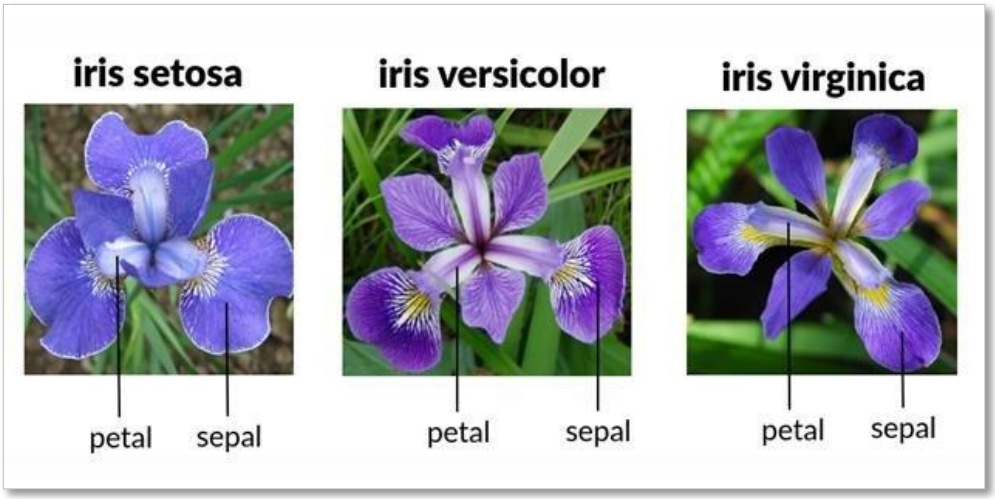
3. TÌM HIỂU VỀ BỘ DỮ LIỆU GỐC

3.1. Đếm số lượng dòng và số lượng cột của bộ dữ liệu gốc

- Bộ dữ liệu gốc có 150 dòng và 6 cột.

3.2. Ý nghĩa của mỗi dòng trong bộ dữ liệu gốc

- Bộ dữ liệu Iris là một trong những bộ dữ liệu nổi tiếng nhất trong lĩnh vực học máy và thống kê. Mỗi dòng trong số 150 dòng của bộ dữ liệu gốc sẽ cho biết tên gọi của loài hoa được lấy mẫu cùng với các đặc trưng như chiều dài và chiều rộng của cánh hoa (petal) và lá đài (sepal) của mẫu tương ứng. Hơn thế nữa, mỗi mẫu dữ liệu sẽ mang thông tin về một trong ba loài hoa thuộc họ diên vĩ là: "Iris setosa", "Iris versicolor" và "Iris virginica".



Hình 3.2: Hình ảnh ba loài hoa Iris trong bộ dữ liệu.

- Quan sát bảng dữ liệu, ta thấy có vẻ như không có dòng nào "lạc loài" (hay bất thường). Đây là một dấu hiệu rất tốt chứng tỏ dữ liệu của ta đủ chất lượng để tiến hành các bước phân tích tiếp theo.

3.3. Bảng mô tả về các cột trong bộ dữ liệu gốc

STT	Tên cột	Mô tả
1	Id	Mã định danh (riêng biệt) của mỗi mẫu dữ liệu.
2	SepalLengthCm	Chiều dài lá đài của mẫu dữ liệu (đơn vị: cm).
3	SepalWidthCm	Chiều rộng lá đài của mẫu dữ liệu (đơn vị: cm).
4	PetalLengthCm	Chiều dài cánh hoa của mẫu dữ liệu (đơn vị: cm).
5	PetalWidthCm	Chiều rộng cánh hoa của mẫu dữ liệu (đơn vị: cm).
6	Species	Tên của loài hoa mà mẫu dữ liệu thuộc về. Thuộc tính này chỉ nhận một trong ba giá trị: "Iris setosa", "Iris versicolor" và "Iris virginica".

Bảng 3.3: Bảng mô tả các cột trong bộ dữ liệu gốc.

3.4. Mục tiêu cần đạt được trong bài tập này

Chúng ta sẽ thực hiện quy trình phân tích khám phá dữ liệu (EDA) lên bộ dữ liệu Iris để khám phá ra các thuộc tính hữu ích có thể giúp ta giải quyết bài toán phân lớp các loài hoa khác nhau. Việc tìm được các thuộc tính như vậy sẽ giúp ta dễ dàng dự đoán giá trị nhãn của bất kỳ mẫu dữ liệu mới nào trong tương lai.

4. TIỀN XỬ LÝ DỮ LIỆU THÔ

4.1. Phân tích tỷ lệ trùng lặp (duplicate) và xử lý các dòng bị trùng lặp (nếu cần)

Ta sử dụng phương thức "duplicated()" của đối tượng DataFrame để kiểm tra xem có dòng nào xuất hiện nhiều hơn một lần hay không. Sau quá trình phân tích, ta rút ra nhận xét:

- Xem qua bảng dữ liệu của file CSV và quan sát kết quả phân tích thì có vẻ như không có một bông hoa nào xuất hiện nhiều hơn một lần trong bộ dữ liệu. Nghĩa là bộ dữ liệu thô không có dòng nào bị trùng lặp, và ta có thể chuyển sang các bước phân tích tiếp theo.

4.2. Phân tích tỷ lệ thiếu giá trị ở mỗi cột (missing rate) và xử lý các cột có giá trị thiếu (nếu cần)

Ta sử dụng phương thức "isnull()" của đối tượng DataFrame để kiểm tra xem có ô nào trong bảng dữ liệu bị thiếu giá trị hay không. Từ đó ta suy ra tỷ lệ thiếu giá trị ở mỗi cột.

STT	Tên cột	Số lượng giá trị bị thiếu	Tỷ lệ thiếu giá trị (%)
1	Id	0	0.0
2	SepalLengthCm	0	0.0
3	SepalWidthCm	0	0.0
4	PetalLengthCm	0	0.0
5	PetalWidthCm	0	0.0
6	Species	0	0.0

Bảng 4.2: Bảng mô tả tỷ lệ thiếu giá trị ở mỗi cột.

Nhận xét:

- Từ kết quả phân tích dữ liệu, ta thấy rằng không có cột nào trong bộ dữ liệu thô bị thiếu giá trị. Đây là một điều rất đáng mừng, cho thấy quá trình thu thập dữ liệu được tiến hành rất tốt. Như vậy, ta không cần phải xử lý gì thêm ở bước này và có thể chuyển sang các bước phân tích tiếp theo.

4.3. Loại bỏ các cột không có nhiều ý nghĩa cho việc phân tích dữ liệu theo mục tiêu đã đề ra

Theo bảng mô tả về các cột mà ta đã viết ở trên, cột "Id" cho biết mã định danh của mỗi mẫu dữ liệu, và mỗi mẫu dữ liệu sẽ có một mã định danh riêng. Đây là thông tin không có nhiều ý nghĩa trong việc giúp ta hiểu rõ hơn các đặc điểm nổi bật của mỗi loài hoa diên vĩ, để từ đó hỗ trợ cho bài toán phân lớp các loài hoa diên vĩ khác nhau.

Như vậy, ta quyết định sẽ loại bỏ cột "Id" khỏi bộ dữ liệu và chỉ giữ lại các đặc trưng quan trọng (là các cột còn lại).

Bộ dữ liệu sau khi loại bỏ cột “Id” có 150 dòng và 5 cột.

4.4. Phân tích kiểu dữ liệu của mỗi cột và xử lý các cột có kiểu dữ liệu chưa phù hợp (nếu cần)

Thay vì sử dụng thuộc tính "dtypes" của đối tượng DataFrame, ta có thể phân tích dữ liệu trong mỗi cột để xác định kiểu dữ liệu của cột tương ứng.

STT	Tên cột	Kiểu dữ liệu
1	SepalLengthCm	float64
2	SepalWidthCm	float64
3	PetalLengthCm	float64
4	PetalWidthCm	float64
5	Species	str

Bảng 4.4: Bảng mô tả kiểu dữ liệu của mỗi cột.

Nhận xét:

- Thông qua kết quả phân tích, ta thấy kiểu dữ liệu của các cột đều phù hợp với mô tả ban đầu của chúng:
  - Bốn cột "SepalLengthCm", "SepalWidthCm", "PetalLengthCm", "PetalWidthCm" đều có kiểu dữ liệu số "float64" cho biết các thông tin về kích thước của cánh hoa và lá đài ứng với mỗi bông hoa.
  - Cột "Species" có kiểu dữ liệu "str" cho biết tên của loài hoa mà mẫu dữ liệu thuộc về.
- Như vậy, kiểu dữ liệu của mỗi cột đều đã phù hợp để có thể thực hiện các bước phân tích tiếp theo.

4.5. Phân tích phân bố của các giá trị trong mỗi cột có kiểu dữ liệu dạng số (numerical) và xử lý các cột có giá trị bất thường

Với mỗi cột có kiểu dữ liệu dạng số, ta sẽ tính:

- Tỷ lệ thiếu giá trị (từ 0 đến 100).
- Giá trị tối thiểu.
- Giá trị tứ phân vị thứ nhất.
- Giá trị tứ phân vị thứ hai (giá trị trung vị).
- Giá trị tứ phân vị thứ ba.
- Giá trị tối đa.

Tên cột Chỉ số thống kê	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
Tỷ lệ thiếu giá trị (%)	0.0	0.0	0.0	0.0
Giá trị tối thiểu	4.3	2.0	1.0	0.1
Tứ phân vị thứ nhất (Q1)	5.1	2.8	1.6	0.3
Giá trị trung vị (Q2)	5.8	3.0	4.4	1.3
Tứ phân vị thứ ba (Q3)	6.4	3.3	5.1	1.8
Giá trị tối đa	7.9	4.4	6.9	2.5

Bảng 4.5: Phân bố của các giá trị trong mỗi cột có kiểu dữ liệu dạng số.

Nhận xét:

- Trong bộ dữ liệu mà ta đang xem xét, nhóm thuộc tính số bao gồm bốn cột là: "SepalLengthCm", "SepalWidthCm", "PetalLengthCm" và "PetalWidthCm". Sau khi quan sát bảng mô tả cho các cột có kiểu dữ liệu dạng số, ta thấy không có cột nào bị thiếu giá trị và có vẻ như dữ liệu cũng không có gì bất thường.

4.6. Phân tích phân bố của các giá trị trong mỗi cột có kiểu dữ liệu dạng phân loại (categorical) và xử lý các cột có giá trị bất thường

Với mỗi cột có kiểu dữ liệu dạng phân loại, ta sẽ tính:

- Tỷ lệ thiếu giá trị (từ 0 đến 100).
- Số lượng các giá trị khác nhau.
- Tỷ lệ xuất hiện (từ 0 đến 100) của mỗi giá trị.

Tên cột	Tỷ lệ thiếu giá trị (%)	Số lượng giá trị khác nhau	Tỷ lệ xuất hiện (%) của mỗi giá trị
Species	0.0	3	{'Iris-setosa': 33.3, 'Iris-versicolor': 33.3,...

Bảng 4.6: Phân bố của các giá trị trong mỗi cột có kiểu dữ liệu dạng phân loại.

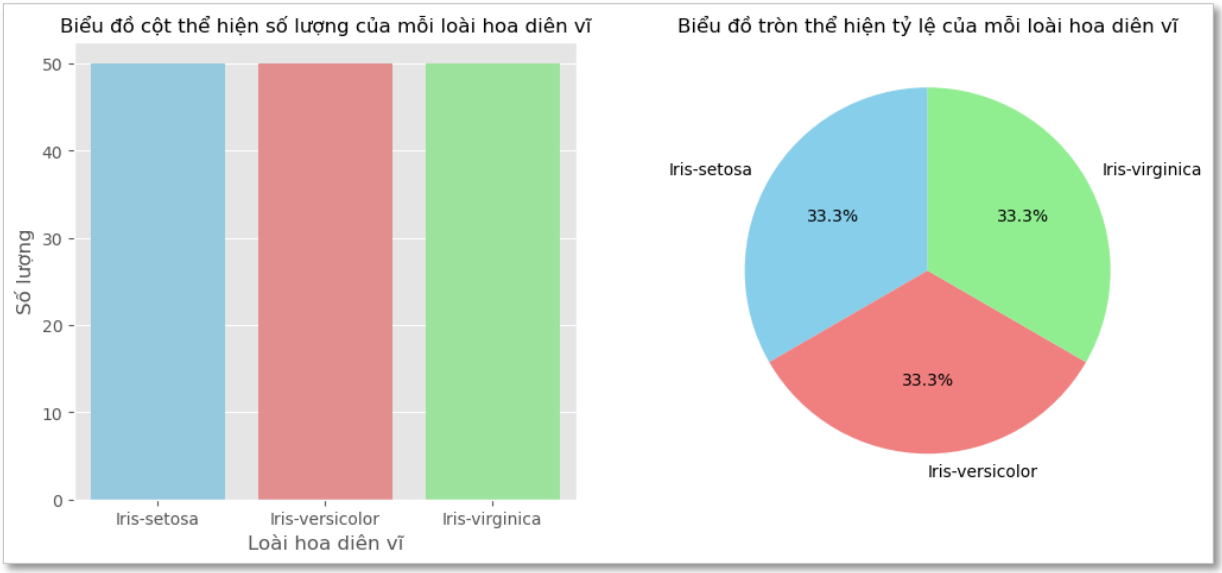
Nhận xét:

- Trong bộ dữ liệu mà ta đang xem xét, nhóm thuộc tính phân loại chỉ bao gồm một cột "Species". Sau khi quan sát bảng mô tả cho các cột có kiểu dữ liệu dạng phân loại, ta thấy không có cột nào bị thiếu giá trị và có vẻ như dữ liệu cũng không có gì bất thường.

5. PHÂN TÍCH THỐNG KÊ CƠ BẢN CHO MỖI BIẾN

5.1. Phân tích tỷ lệ đối với các biến định tính

Trong phần này, ta sẽ phân tích tỷ lệ của mỗi loài hoa có trong cột "Species".



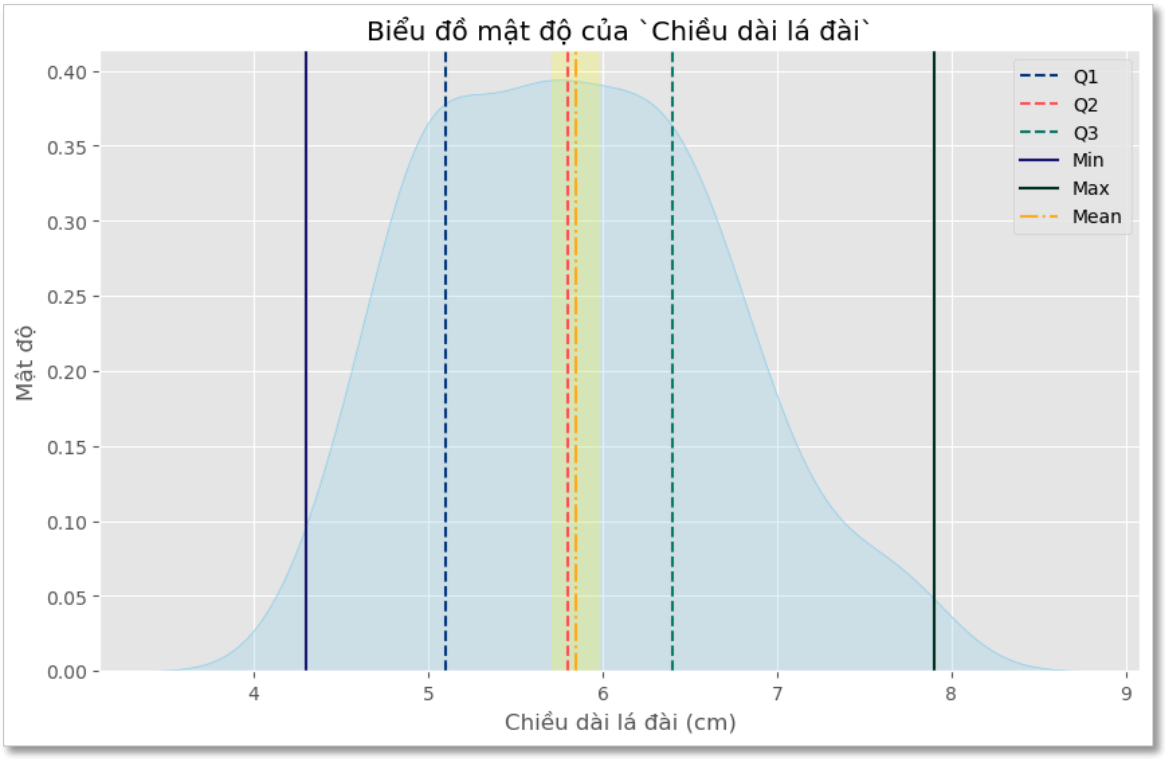
Hình 5.1: Biểu đồ thể hiện tỷ lệ của các loài hoa Iris (diên vĩ).

Nhận xét:

- Trong bộ dữ liệu mà ta đang làm việc, ta thấy có sự xuất hiện của ba loài hoa khác nhau là: "Iris-setosa", "Iris-versicolor" và "Iris-virginica". Đặc biệt, số lượng mẫu dữ liệu thuộc về mỗi loài hoa là bằng nhau, mỗi loài có 50 mẫu dữ liệu.
- Việc các loài hoa có số lượng mẫu dữ liệu tương đương nhau là một điều rất tốt, giúp hạn chế tình trạng kết quả phân tích bị lệch hẳn về một loài hoa có nhiều mẫu dữ liệu.

5.2. Phân tích phân phối đối với các biến định lượng

5.2.1. Phân tích phân phối của chiều dài lá đài (SepalLengthCm)



Hình 5.2.1: Biểu đồ thể hiện mật độ phân bố của các giá trị trong cột "SepalLengthCm".

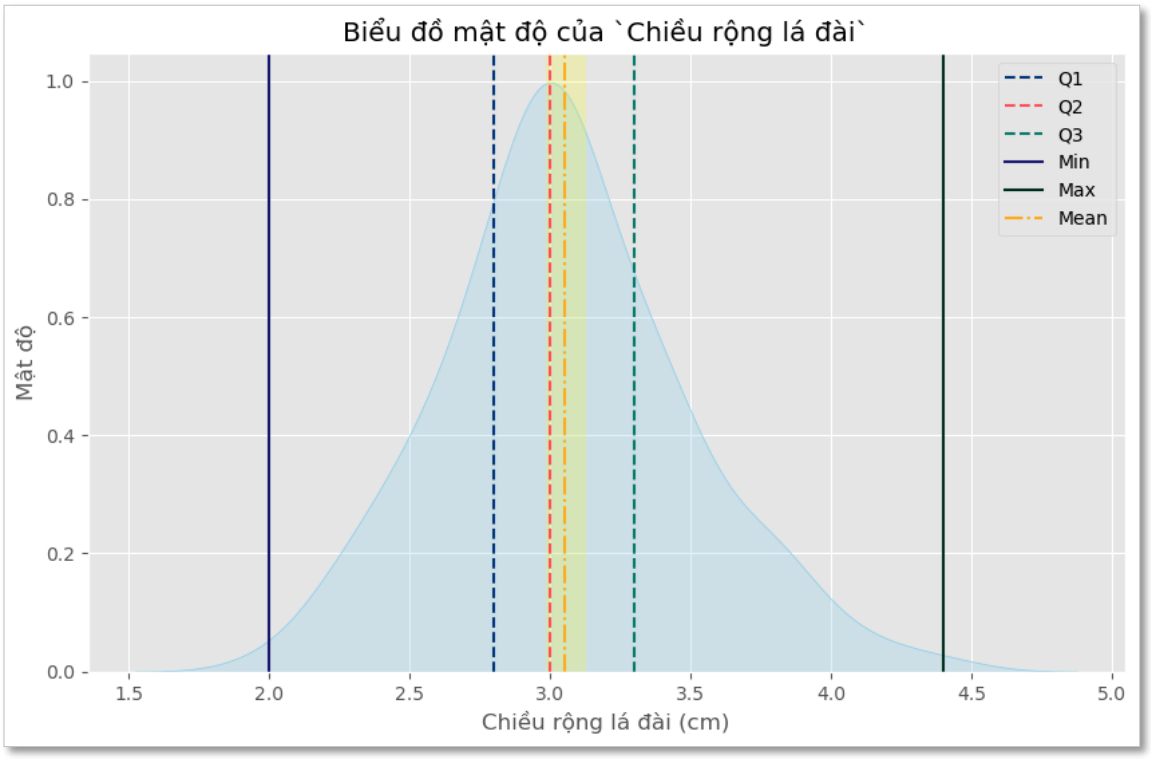
Đại lượng thống kê cho `Chiều dài lá đài (cm)`	Giá trị
min	4.300000
lower_quartile	5.100000
median	5.800000
upper_quartile	6.400000
max	7.900000
mean	5.843333
std	0.828066
skew	0.314911
kurt	-0.552064

Bảng 5.2.1: Bảng mô tả các đại lượng thống kê cho cột "SepalLengthCm".

Nhận xét:

- Ta thấy chiều dài lá đài (SepalLengthCm) có phân bố hình chuông, khá đối xứng (|skew| = 0.31 < 0.5) và có phần đuôi khá mỏng (kurt = -0.55 < 0):
  - Tuy giá trị trung bình có lớn hơn giá trị trung vị một chút nhưng không đáng kể.
  - Trong 150 mẫu dữ liệu đang được phân tích, thuộc tính chiều dài lá đài có giá trị nhỏ nhất là 4.3 (cm) và có giá trị lớn nhất là 7.9 (cm).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [5.1; 6.4] (đơn vị: cm).
  - Với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều dài trung bình của lá đài trong bộ dữ liệu này là: (5.71; 5.98) (đơn vị: cm).
- Như vậy, ta thấy rằng các điểm dữ liệu của thuộc tính "SepalLengthCm" tập trung khá nhiều ở khoảng trung tâm. Mật độ dữ liệu giảm mạnh sau khi rời khỏi khoảng trung tâm được giới hạn bởi giá trị tứ phân vị thứ nhất và thứ ba. Điều này có thể xuất phát từ phân bố chiều dài lá đài khác nhau của từng loài hoa diên vĩ khác nhau và ta cần phân tích sâu hơn để làm rõ đặc điểm này.

5.2.2. Phân tích phân phối của chiều rộng lá đài (SepalWidthCm)



Hình 5.2.2: Biểu đồ thể hiện mật độ phân bố của các giá trị trong cột "SepalWidthCm".

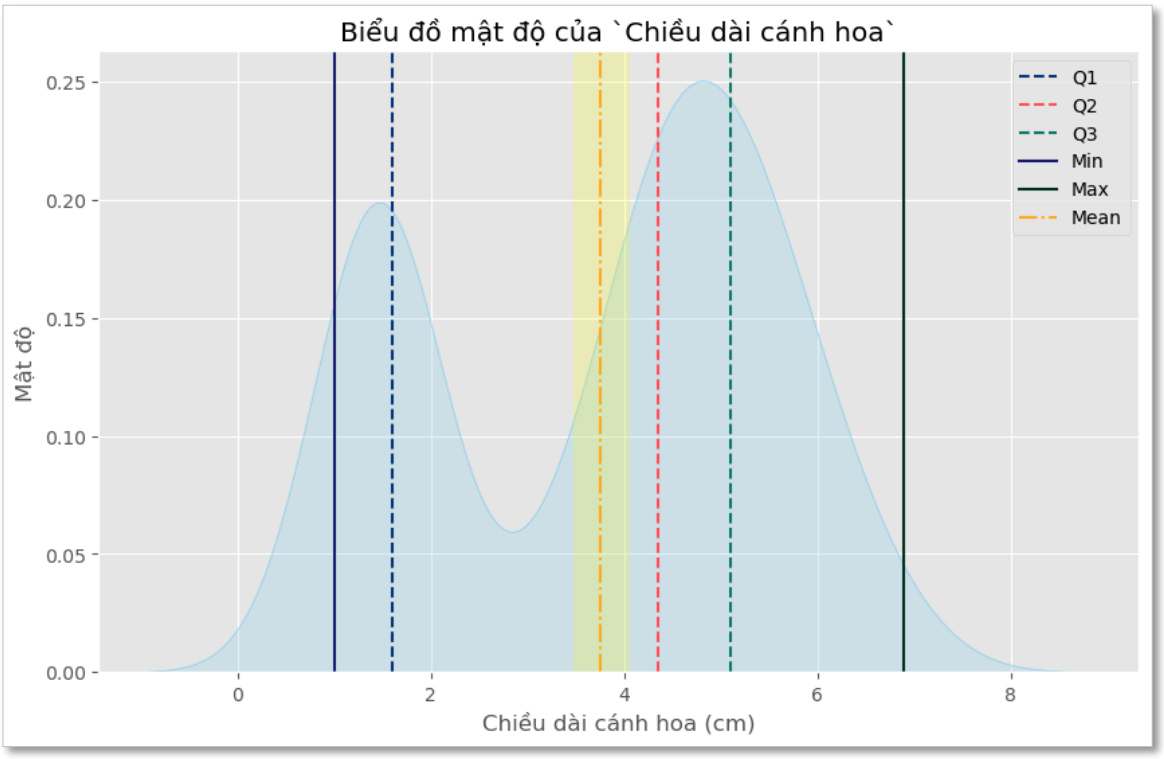
Đại lượng thống kê cho `Chiều rộng lá đài (cm)`	Giá trị
min	2.000000
lower_quartile	2.800000
median	3.000000
upper_quartile	3.300000
max	4.400000
mean	3.054000
std	0.433594
skew	0.334053
kurt	0.290781

Bảng 5.2.2: Bảng mô tả các đại lượng thống kê cho cột "SepalWidthCm".

Nhận xét:

- Ta thấy phân bố của chiều rộng lá đài (SepalWidthCm) khá đối xứng (|skew| = 0.33 < 0.5) và có phần đuôi hơi đậm (kurt = 0.29 > 0):
  - Các giá trị chiều rộng lá đài có phạm vi phân bố nằm trong đoạn [2; 4.4] (đơn vị: cm).
  - 25% chiều rộng lá đài có giá trị nhỏ hơn 2.8 (cm).
  - 50% chiều rộng lá đài có giá trị nhỏ hơn 3 (cm).
  - 75% chiều rộng lá đài có giá trị nhỏ hơn 3.3 (cm).
  - Với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều rộng trung bình của lá đài trong bộ dữ liệu này là: (2.98; 3.12) (đơn vị: cm).
- Như vậy, phân bố của chiều rộng lá đài khá thưa thớt ở khoảng giá trị trung tâm và tập trung nhiều ở phần đuôi. Điều này có thể xuất phát từ những đặc điểm phân bố giống nhau của thuộc tính chiều rộng lá đài ở các loài hoa khác nhau. Nếu giả thuyết này là đúng thì đây là một dấu hiệu cho thấy "SepalWidthCm" có thể không phải là một công cụ độc lập đủ mạnh để giúp ta dễ dàng phân biệt giữa các loài hoa diên vĩ khác nhau. Tuy nhiên, ta có thể tập trung phân tích sâu hơn vào thuộc tính "SepalWidthCm" ở các bước tiếp theo để làm rõ giả thuyết này.

5.2.3. Phân tích phân phối của chiều dài cánh hoa (PetalLengthCm)



Hình 5.2.3: Biểu đồ thể hiện mật độ phân bố của các giá trị trong cột "PetalLengthCm".

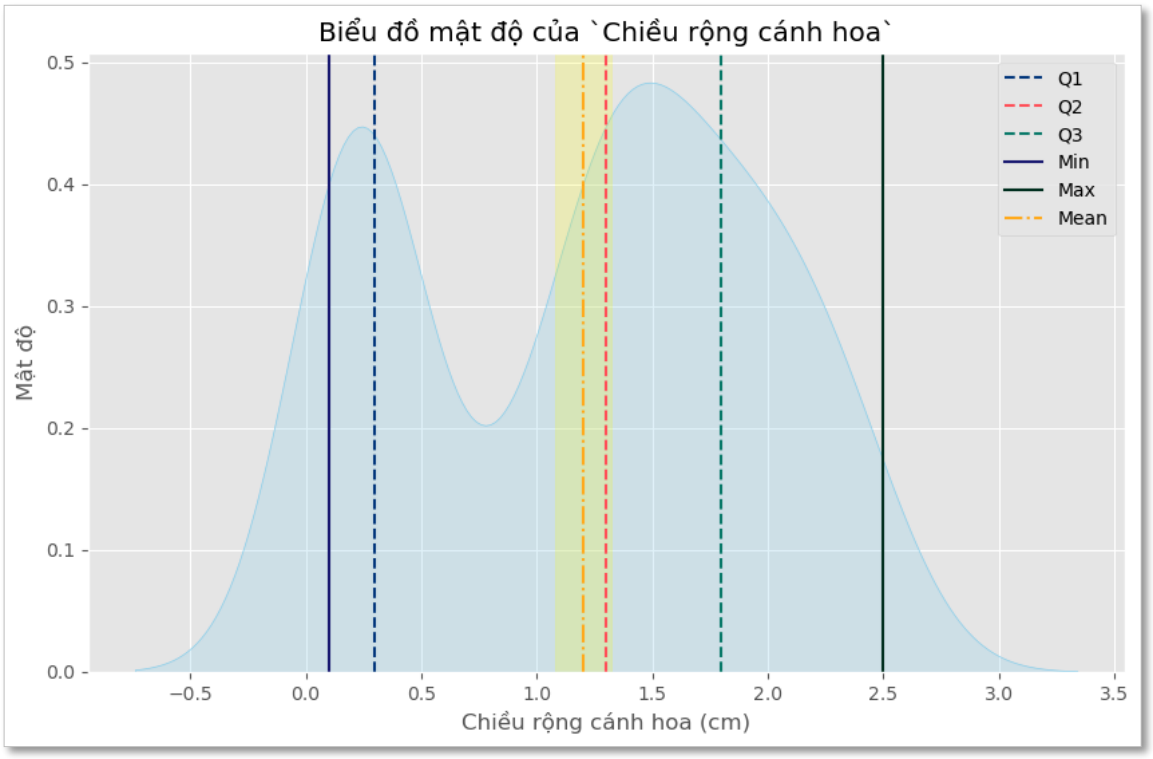
Đại lượng thống kê cho `Chiều dài cánh hoa (cm)`	Giá trị
min	1.000000
lower_quartile	1.600000
median	4.400000
upper_quartile	5.100000
max	6.900000
mean	3.758667
std	1.764420
skew	-0.274464
kurt	-1.401921

Bảng 5.2.3: Bảng mô tả các đại lượng thống kê cho cột "PetalLengthCm".

Nhận xét:

- Phân bố của chiều dài cánh hoa (PetalLengthCm) nổi bật với việc xuất hiện hai đỉnh giá trị khác nhau, đây có thể là một manh mối "quý giá" giúp ta hoàn thành mục tiêu đã đề ra. Từ bảng thống kê, ta thấy rằng phân bố của cột "PetalLengthCm" khá đối xứng (|skew| = 0.27 < 0.5) và có phần đuôi khá mỏng (kurt = -1.4 < 0):
  - Các giá trị trong cột "PetalLengthCm" có phạm vi phân bố khá rộng, từ 1.0 đến 6.9 (cm).
  - Khoảng 50% giá trị sẽ tập trung trong đoạn [1.6; 5.1] (đơn vị: cm).
  - 25% giá trị nhỏ nhất có phạm vi phân bố hẹp hơn đáng kể so với phạm vi phân bố của 25% giá trị lớn nhất.
  - Giá trị trung vị của chiều dài cánh hoa (4.4) lớn hơn giá trị trung bình (3.8) và lệch về phía đỉnh cao hơn là một điểm thú vị mà ta cần quan tâm.
  - Với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều dài trung bình của cánh hoa trong bộ dữ liệu này là: (3.47; 4.04) (đơn vị: cm).
- Như vậy, ta thấy dữ liệu chiều dài cánh hoa xuất hiện hai đỉnh khác nhau và giá trị trung vị lệch về phía đỉnh cao hơn. Đây có thể là dấu hiệu cho thấy tồn tại một loài hoa diên vĩ có chiều dài cánh hoa thường ngắn hơn đáng kể so với hai loài hoa còn lại. Do đó, ta nên tập trung phân tích sâu hơn về sự khác biệt trong chiều dài cánh hoa ở các loài hoa diên vĩ khác nhau để làm rõ xu hướng này.

5.2.4. Phân tích phân phối của chiều rộng cánh hoa (PetalWidthCm)



Hình 5.2.4: Biểu đồ thể hiện mật độ phân bố của các giá trị trong cột "PetalWidthCm".

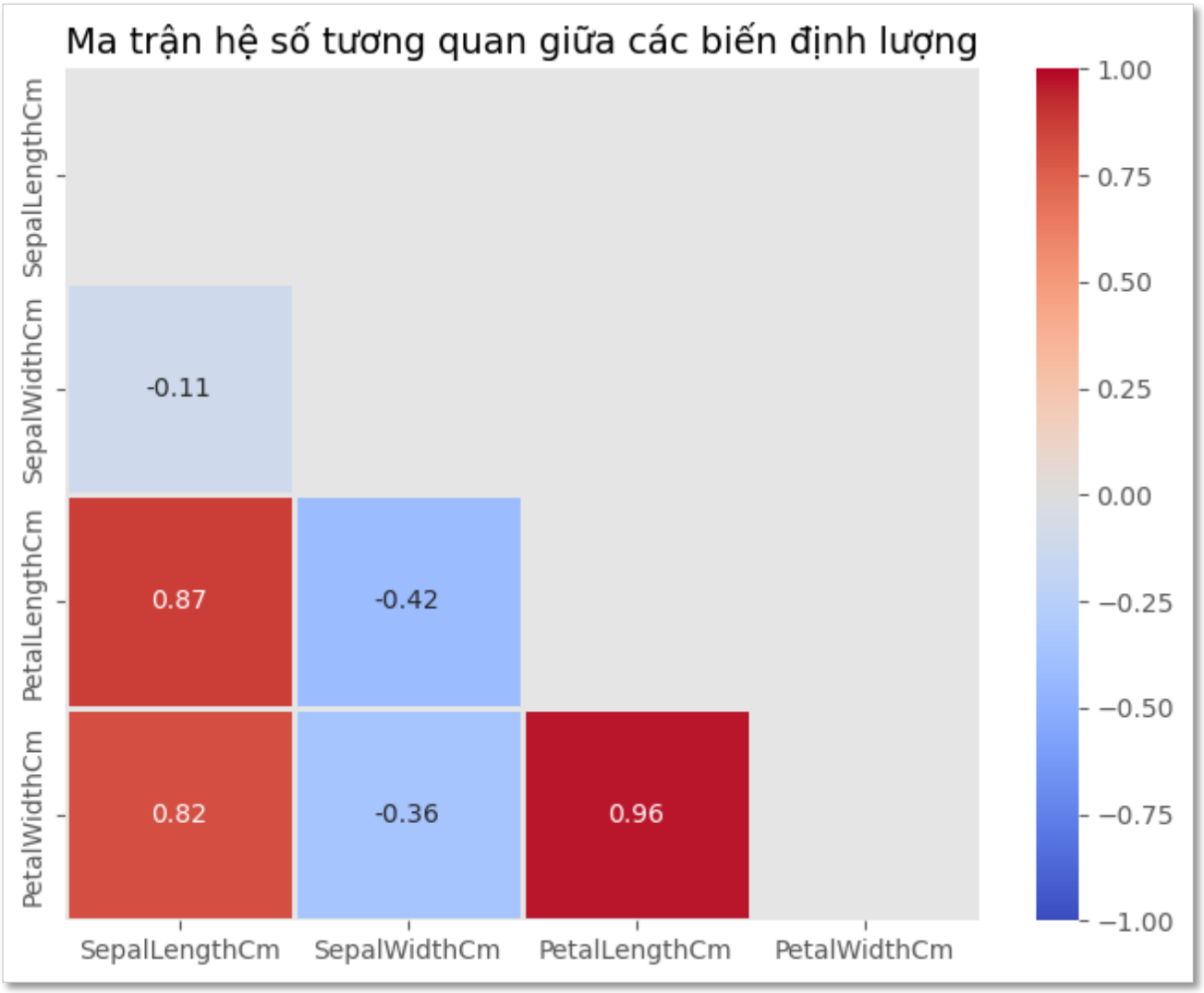
Đại lượng thống kê cho `Chiều rộng cánh hoa (cm)`	Giá trị
min	0.100000
lower_quartile	0.300000
median	1.300000
upper_quartile	1.800000
max	2.500000
mean	1.198667
std	0.763161
skew	-0.104997
kurt	-1.339754

Bảng 5.2.4: Bảng mô tả các đại lượng thống kê cho cột "PetalWidthCm".

Nhận xét:

- Ta thấy biểu đồ mật độ của cột "PetalWidthCm" có khá nhiều nét tương đồng với biểu đồ mật độ của cột "PetalLengthCm".
- Phân bố của chiều rộng cánh hoa (PetalWidthCm) xuất hiện hai đỉnh khác nhau. Phân bố này khá đối xứng (|skew| = 0.1 < 0.5) và có phần đuôi khá mỏng (kurt = -1.34 < 0):
  - Thuộc tính chiều rộng cánh hoa có phạm vi phân bố trong đoạn [0.1; 2.5] (đơn vị: cm).
  - Có khoảng 50% giá trị chiều rộng cánh hoa tập trung trong đoạn [0.3; 1.8] (đơn vị: cm).
  - 25% giá trị nhỏ nhất có phạm vi phân bố hẹp hơn đáng kể so với phạm vi phân bố của 25% giá trị lớn nhất.
  - Giá trị trung vị của chiều rộng cánh hoa lệch về phía đỉnh cao hơn.
  - Với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều rộng trung bình của cánh hoa trong bộ dữ liệu này là: (1.08; 1.32) (đơn vị: cm).
- Như vậy, ta thấy dữ liệu chiều rộng cánh hoa xuất hiện hai đỉnh khác nhau và giá trị trung vị lệch về phía đỉnh cao hơn. Đây có thể là dấu hiệu cho thấy tồn tại một loài hoa diên vĩ có chiều rộng cánh hoa thường ngắn hơn đáng kể so với hai loài hoa còn lại. Do đó, ta có thể tiếp tục phân tích sâu hơn về sự khác biệt trong chiều rộng cánh hoa giữa các loài hoa diên vĩ khác nhau để làm rõ xu hướng này.
- Trong trường hợp đỉnh bên trái ở hai biểu đồ mật độ của "PetalLengthCm" và "PetalWidthCm" là xuất phát từ cùng một loài hoa, thì ta chỉ cần sử dụng hai thông tin về kích thước của cánh hoa là đã có thể nhận biết chính xác phần lớn mẫu dữ liệu thuộc về loài hoa này. Nếu giả thuyết trên là đúng thì đây sẽ là một bước tiến lớn giúp ta đạt được mục tiêu đã đề ra.

6. PHÂN TÍCH HỆ SỐ TƯƠNG QUAN GIỮA CÁC BIẾN ĐỊNH LƯỢNG BẰNG BẢN ĐỒ NHIỆT



Hình 6: Bản đồ nhiệt thể hiện hệ số tương quan Pearson giữa các cặp biến định lượng.

Nhận xét:

- Quan sát biểu đồ, ta thấy ba thuộc tính "SepalLengthCm", "PetalLengthCm" và "PetalWidthCm" có mức độ tương quan thuận rất mạnh với nhau:
  - Chiều dài cánh hoa (PetalLengthCm) và chiều rộng cánh hoa (PetalWidthCm) có mức độ tương quan thuận gần như hoàn hảo (0.96).
  - Chiều dài lá đài (SepalLengthCm) và chiều dài cánh hoa (PetalLengthCm) có mức độ tương quan thuận rất mạnh (0.87).
  - Hơn thế nữa, chiều dài lá đài (SepalLengthCm) và chiều rộng cánh hoa (PetalWidthCm) cũng có mức độ tương quan thuận rất mạnh (0.82).
- Tuy nhiên, chiều rộng lá đài có mức độ tương quan nghịch với cả ba thuộc tính còn lại:
  - Chiều rộng lá đài (SepalWidthCm) có mức độ tương quan nghịch khá yếu với chiều dài cánh hoa (PetalLengthCm) và chiều rộng cánh hoa (PetalWidthCm), hệ số tương quan có giá trị lần lượt là -0.42 và -0.36.
  - Đặc biệt, ta nhận thấy giữa chiều rộng và chiều dài của lá đài (giữa "SepalWidthCm" và "SepalLengthCm") có mức độ tương quan nghịch rất yếu (-0.11).
- Sau khi phân tích hệ số tương quan giữa các cặp biến định lượng, ta thấy rằng nếu một thuộc tính bất kỳ trong nhóm "SepalLengthCm", "PetalLengthCm", "PetalWidthCm" tăng lên thì hai thuộc tính còn lại cũng có xu hướng tăng lên và mức độ quan hệ tuyến tính giữa ba thuộc tính này là rất mạnh. Ngược lại, khi chiều rộng lá đài (SepalWidthCm) tăng lên thì ba thuộc tính còn lại thường có xu hướng giảm xuống và mức độ quan hệ giữa "SepalWidthCm" với các thuộc tính khác là khá yếu. Tuy nhiên, ta cần thực hiện nhiều phân tích chi tiết hơn nữa để hiểu rõ mối quan hệ giữa các biến định lượng. Đồng thời, việc tập trung phân tích vào các biến có mức độ tương quan cao có thể giúp ta khám phá ra mối quan hệ thú vị ẩn sau chúng.

7. ĐƯA RA CÁC CÂU HỎI MÀ TA CÓ THỂ TRẢ LỜI BẰNG BỘ DỮ LIỆU

Sau khi đã hiểu hơn về dữ liệu, ta có thể đưa ra các câu hỏi cần trả lời để giúp ta đạt được mục tiêu đã đề ra.

Mục tiêu "khám phá ra các thuộc tính hữu ích giúp ta giải quyết bài toán phân lớp các loài hoa Iris khác nhau" có thể được cụ thể hóa ra thành các câu hỏi sau:

7.1. Câu hỏi 1: Liệu chiều dài lá đài (SepalLengthCm) có phải là yếu tố đủ để phân loại chính xác các loài hoa Iris (diên vĩ) khác nhau hay không?

Việc trả lời câu hỏi này có ý nghĩa:

- Nếu phân bố chiều dài lá đài giữa các loài hoa Iris có sự khác biệt đáng kể, thì việc sử dụng thông tin này có thể giúp ta phân loại chính xác các loài hoa khác nhau. Điều này sẽ hỗ trợ rất nhiều trong việc xây dựng quy trình nhận diện và phân loại tự động các loài hoa Iris.
- Nếu chiều dài lá đài là yếu tố có vai trò quyết định trong việc phân biệt các loài hoa Iris, thì thông tin này có thể được áp dụng vào các ứng dụng thực tiễn như trong nông nghiệp, sinh học, hoặc các lĩnh vực khác đòi hỏi phải phân loại hoa Iris.
- Sử dụng kết quả phân tích dữ liệu, kết hợp với các kiến thức về sinh thái học, ta có thể khám phá ra sự biến đổi và thích nghi của các loài hoa diên vĩ trong môi trường tự nhiên. Điều này có thể hỗ trợ trong việc nghiên cứu về tiến hóa và đưa ra chính sách bảo tồn các loài hoa diên vĩ trong tương lai.

7.2. Câu hỏi 2: Liệu chiều rộng lá đài (SepalWidthCm) có phải là yếu tố đủ để phân loại chính xác các loài hoa Iris (diên vĩ) khác nhau hay không?

Việc trả lời câu hỏi này có ý nghĩa:

- Nếu chiều rộng lá đài là yếu tố đủ để phân loại chính xác các loài hoa Iris khác nhau, thì ta có thể tối ưu hóa quá trình phân loại bằng cách chỉ tập trung vào đặc điểm này. Đồng thời, việc tập trung phân tích, xử lý trên một đặc điểm quan trọng nhất sẽ giúp ta tiết kiệm "chi phí" (như: thời gian, tiền bạc, công sức, v.v.) trong việc thu thập và xử lý dữ liệu, cũng như trong quá trình nghiên cứu và phát triển mô hình phân loại hoa Iris.
- Nếu chiều rộng lá đài là yếu tố mang tính chất quyết định trong việc phân loại các loài hoa Iris khác nhau, thì thông tin này có thể được áp dụng vào thực tiễn như trong nông nghiệp, công nghiệp, sinh học, hoặc các lĩnh vực khác đòi hỏi phải phân loại hoa Iris.
- Kết quả từ quá trình phân tích dữ liệu chiều rộng lá đài có thể được sử dụng làm nền tảng cho các nghiên cứu tiếp theo về hoa Iris. Trong trường hợp ta khám phá ra những đặc điểm riêng biệt chỉ có ở loài hoa Iris, thì kết quả này có thể cung cấp nhiều thông tin quan trọng cho các nghiên cứu liên quan đến việc phân loại cây cối dựa trên đặc điểm hình thái.

7.3. Câu hỏi 3: Liệu chiều dài cánh hoa (PetalLengthCm) có phải là yếu tố đủ để phân loại chính xác các loài hoa Iris (diên vĩ) khác nhau hay không?

Việc trả lời câu hỏi này có ý nghĩa:

- Thông qua quy trình phân tích phân bố chiều dài cánh hoa ở các loài hoa Iris khác nhau, ta sẽ có cái nhìn tổng quát và sâu sắc hơn về sự đa dạng, những đặc trưng riêng biệt của mỗi loài hoa. Từ đó, ta có thể khám khá ra cách phân loại các loài hoa Iris khác nhau thông qua đặc điểm chiều dài cánh hoa của chúng.
- Nếu chiều dài cánh hoa là yếu tố đủ để phân loại chính xác các loài hoa Iris, thì ta có thể tối ưu hóa quá trình phân loại bằng cách chỉ tập trung vào đặc điểm này. Việc tập trung phân tích, xử lý trên một đặc điểm duy nhất sẽ giúp giảm bớt sự phức tạp trong việc thu thập và tiền xử lý dữ liệu.
- Trả lời được câu hỏi này có thể giúp ta tiết kiệm thời gian và công sức trong quá trình nghiên cứu và phát triển mô hình phân loại hoa Iris, bằng cách tập trung vào một đặc điểm quan trọng nhất.
- Kết quả từ quá trình phân tích dữ liệu chiều dài cánh hoa có thể được dùng làm nền tảng cho các nghiên cứu tiếp theo về hoa Iris. Trong một số trường hợp, các đặc trưng riêng biệt về hoa Iris có thể cung cấp những thông tin quan trọng cho các nghiên cứu liên quan đến phân loại thực vật dựa trên đặc điểm hình thái của chúng.

7.4. Câu hỏi 4: Liệu chiều rộng cánh hoa (PetalWidthCm) có phải là yếu tố đủ để phân loại chính xác các loài hoa Iris (diên vĩ) khác nhau hay không?

Việc trả lời câu hỏi này có ý nghĩa:

- Tương tự như với chiều dài cánh hoa, quá trình phân tích phân bố chiều rộng cánh hoa ở các loài hoa Iris khác nhau sẽ giúp ta có cái nhìn tổng quát và sâu sắc hơn về sự đa dạng, những đặc trưng riêng biệt của mỗi loài hoa. Từ đó, ta có thể khám khá ra cách phân loại các loài hoa Iris khác nhau thông qua đặc trưng chiều rộng cánh hoa của chúng.
- Nếu chiều rộng cánh hoa là yếu tố đủ để phân loại chính xác các loài hoa Iris khác nhau, thì ta có thể tối ưu hóa quá trình phân loại bằng cách chỉ tập trung vào đặc điểm này, từ đó giảm bớt sự phức tạp trong việc thu thập và xử lý dữ liệu.
- Trả lời được câu hỏi này có thể giúp ta tiết kiệm thời gian và công sức trong quá trình nghiên cứu và phát triển mô hình phân loại hoa Iris, bằng cách tập trung vào một đặc điểm quan trọng nhất.
- Kết quả từ quá trình phân tích chiều rộng cánh hoa ở ba loài hoa diên vĩ khác nhau có trong bộ dữ liệu sẽ cung cấp một nền tảng vững chắc để ta tiếp tục các nghiên cứu, phân tích chuyên sâu trong tương lai. Từ những nền tảng này, ta có thể mở rộng nghiên cứu để giải quyết bài toán phân loại thêm nhiều loài hoa diên vĩ hơn nữa.

7.5. Câu hỏi 5: Liệu kích thước (chiều dài và chiều rộng) lá đài có phải là yếu tố đủ để phân loại chính xác các loài hoa Iris (diên vĩ) khác nhau hay không?

Việc trả lời câu hỏi này có ý nghĩa:

- Bằng cách xem xét cả chiều dài và chiều rộng của lá đài, ta sẽ có cái nhìn toàn diện hơn về đặc điểm hình thái của các loài hoa Iris khác nhau. Thông qua đó, ta sẽ có thêm cơ sở, công cụ để giải quyết bài toán phân loại các loài hoa Iris.
- Nếu kích thước lá đài là yếu tố đủ để phân loại chính xác các loài hoa Iris khác nhau, thì ta có thể tối ưu hóa quá trình phân loại bằng cách chỉ tập trung vào các đặc điểm này. Từ đó, ta có thể hạn chế "chi phí" trong việc thu thập và xử lý dữ liệu so với khi phải xem xét nhiều đặc điểm khác nhau.
- Trả lời được câu hỏi này có thể giúp ta tiết kiệm thời gian, công sức trong quá trình nghiên cứu và phát triển mô hình phân loại hoa Iris, bằng cách tập trung vào một số đặc điểm quan trọng nhất và bỏ qua các đặc điểm không có nhiều ý nghĩa trong quá trình phân loại.
- Kết quả phân tích dữ liệu kích thước lá đài có thể giúp ta khám phá sự biến đổi và thích nghi của các loài hoa diên vĩ trong môi trường tự nhiên. Điều này sẽ hỗ trợ phần nào trong việc nghiên cứu về tiến hóa và sinh thái học.

7.6. Câu hỏi 6: Liệu kích thước (chiều dài và chiều rộng) cánh hoa có phải là yếu tố đủ để phân loại chính xác các loài hoa Iris (diên vĩ) khác nhau hay không?

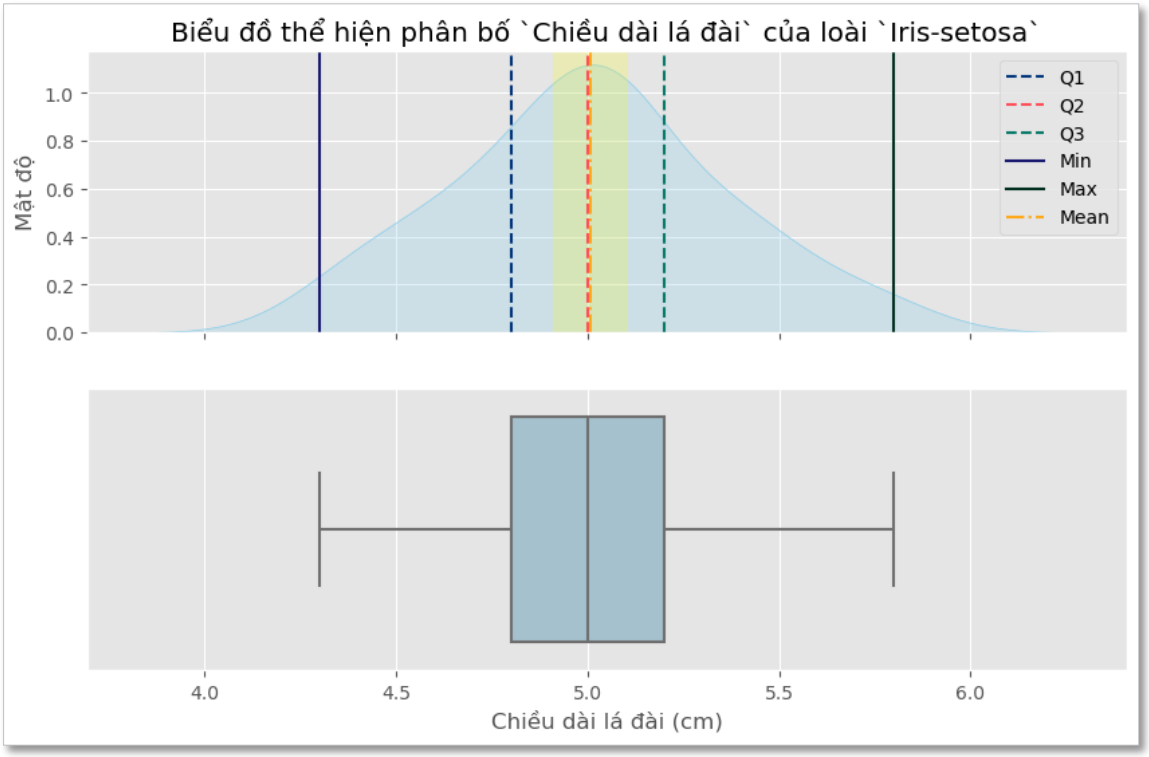
Việc trả lời câu hỏi này có ý nghĩa:

- Bằng cách xem xét cả chiều dài và chiều rộng của cánh hoa, ta sẽ có cái nhìn toàn diện hơn về đặc điểm hình thái của các loài hoa Iris và cách chúng có thể được phân loại dựa trên kích thước cánh hoa.
- Nếu kích thước cánh hoa là yếu tố đủ để phân loại chính xác các loài hoa Iris, thì ta có thể tối ưu hóa quá trình phân loại bằng cách chỉ tập trung vào các đặc điểm này. Từ đó, ta có thể hạn chế "chi phí" trong việc thu thập và xử lý dữ liệu so với khi phải xem xét nhiều đặc điểm khác nhau.
- Trả lời được câu hỏi này có thể giúp ta tiết kiệm thời gian, công sức trong quá trình nghiên cứu và phát triển mô hình phân loại hoa Iris, bằng cách tập trung vào một số đặc điểm quan trọng nhất và bỏ qua các đặc điểm không có nhiều ý nghĩa trong quá trình phân loại.
- Nếu kích thước cánh hoa là yếu tố có vai trò quyết định trong việc phân loại các loài hoa Iris khác nhau, thì thông tin này có thể được áp dụng vào thực tiễn như trong nông nghiệp, sinh học, hoặc các lĩnh vực mà đòi hỏi phải phân loại hoa Iris.
- Kết quả từ việc trả lời câu hỏi này có thể làm nền tảng cho các nghiên cứu tiếp theo về hoa Iris. Đồng thời, kết quả này có thể cung cấp một số thông tin hữu ích cho các nghiên cứu liên quan đến phân loại thực vật dựa trên đặc điểm hình thái của chúng.

8. TIỀN XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU ĐỂ TRẢ LỜI CHO CÂU HỎI 1

Câu hỏi 1: Liệu chiều dài lá đài (SepalLengthCm) có phải là yếu tố đủ để phân loại chính xác các loài hoa Iris (diên vĩ) khác nhau hay không?

8.1. Phân tích phân bố chiều dài lá đài của loài "Iris-setosa"



Hình 8.1: Biểu đồ thể hiện phân bố chiều dài lá đài của loài "Iris-setosa".

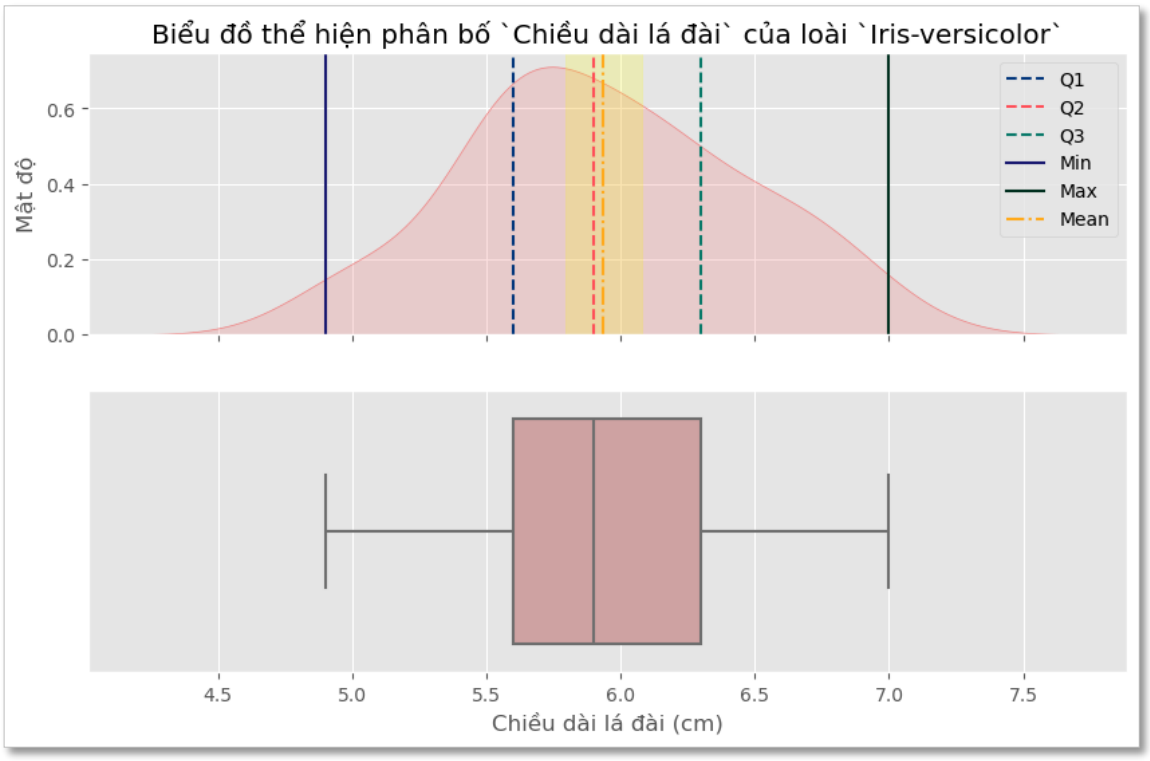
Đại lượng thống kê cho `Chiều dài lá đài (cm)` của loài `Iris-setosa`	Giá trị
min	4.300000
lower_quartile	4.800000
median	5.000000
upper_quartile	5.200000
max	5.800000
mean	5.006000
std	0.352490
skew	0.120087
kurt	-0.252689

Bảng 8.1: Bảng mô tả các đại lượng thống kê cho chiều dài lá đài của loài "Iris-setosa".

Nhận xét:

- Ta thấy chiều dài lá đài (SepalLengthCm) của loài "Iris-setosa" có phân bố hình chuông, khá đối xứng (|skew| = 0.12 < 0.5) và có phần đuôi hơi mỏng (kurt = -0.25 < 0):
  - Chiều dài lá đài của loài "Iris-setosa" có giá trị nhỏ nhất là 4.3 (cm) và giá trị lớn nhất là 5.8 (cm).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [4.8; 5.2] (đơn vị: cm).
  - Nhìn vào biểu đồ mật độ, ta không cảm nhận được có sự khác biệt đáng kể nào giữa giá trị trung bình và giá trị trung vị của chiều dài lá đài.
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều dài trung bình của lá đài ở loài hoa "Iris-setosa" là: (4.91; 5.10) (đơn vị: cm).
- Như vậy, chiều dài lá đài của loài "Iris-setosa" tập trung đông đúc ở các khoảng giá trị trung tâm của phân bố, càng tiến về hai đuôi, mật độ dữ liệu giảm dần. Việc tập trung phân tích các mẫu dữ liệu có chiều dài lá đài nằm trong khoảng trung tâm có thể là một yếu tố giúp ta phân biệt loài "Iris-setosa" với các loài hoa diên vĩ khác. Tuy nhiên, ta vẫn cần nhiều phân tích chi tiết hơn để làm rõ giả thuyết này.

8.2. Phân tích phân bố chiều dài lá đài của loài "Iris-versicolor"



Hình 8.2: Biểu đồ thể hiện phân bố chiều dài lá đài của loài "Iris-versicolor".

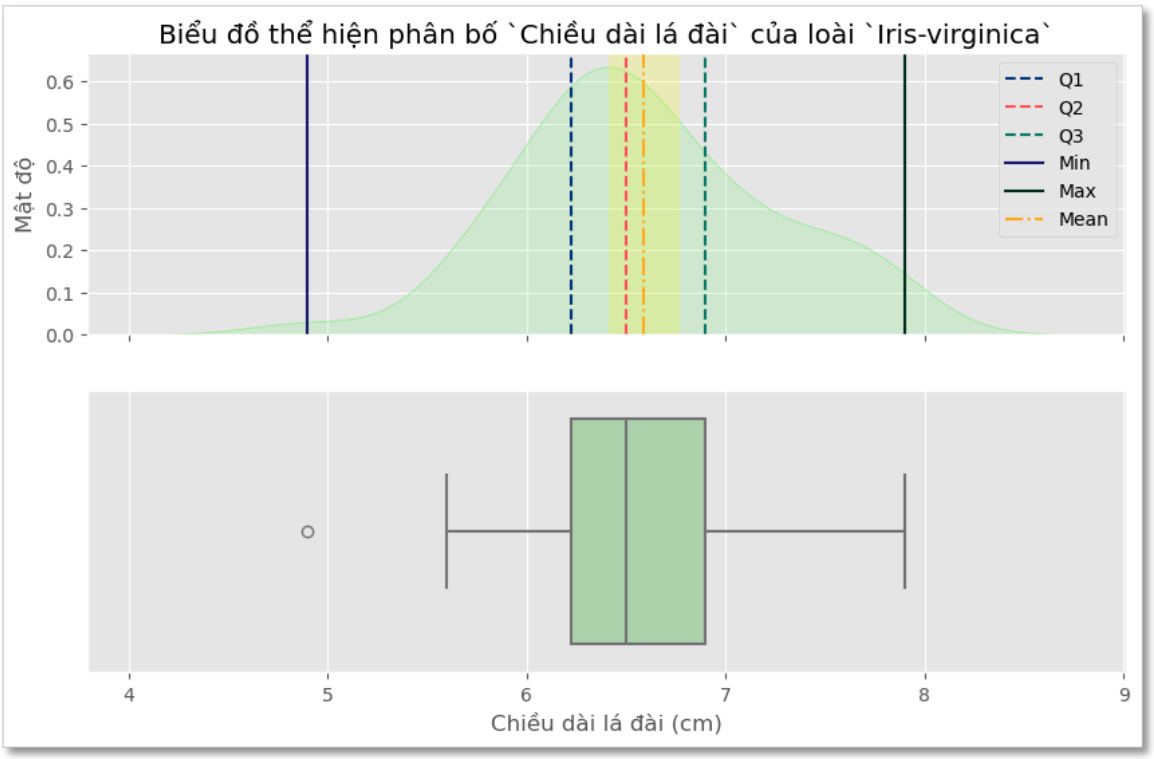
Đại lượng thống kê cho `Chiều dài lá đài (cm)` của loài `Iris-versicolor`	Giá trị
min	4.900000
lower_quartile	5.600000
median	5.900000
upper_quartile	6.300000
max	7.000000
mean	5.936000
std	0.516171
skew	0.105378
kurt	-0.533010

Bảng 8.2: Bảng mô tả các đại lượng thống kê cho chiều dài lá đài của loài "Iris-versicolor".

Nhận xét:

- Ta thấy phân bố chiều dài lá đài của loài "Iris-versicolor" khá đối xứng (|skew| = 0.11 < 0.5) và có phần đuôi khá mỏng (kurt = -0.53 < 0):
  - Các giá trị mà ta quan sát được sẽ có phạm vi phân bố nằm trong đoạn [4.9; 7.0] (đơn vị: cm).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [5.6; 6.3] (đơn vị: cm).
  - 25% giá trị nhỏ nhất có phạm vi phân bố tương đương với phạm vi phân bố của 25% giá trị lớn nhất. Và phạm vi này cũng không chênh lệch quá nhiều so với phạm vi phân bố của 50% điểm dữ liệu ở khu vực trung tâm.
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều dài trung bình của lá đài ở loài hoa "Iris-versicolor" là: (5.79; 6.08) (đơn vị: cm).
- Như vậy, chiều dài lá đài của loài "Iris-versicolor" tập trung khá nhiều ở các khoảng trung tâm của phân bố. Nhìn sơ qua biểu đồ mật độ, ta có cảm giác các điểm dữ liệu có xu hướng lệch về phía bên phải nhưng hiện tại ta chưa có đủ cơ sở để khẳng định điều này. Do đó, ta có thể phân tích trên nhiều mẫu dữ liệu hơn để có thể đưa ra các kết luận chính xác hơn về phân bố chiều dài lá đài của loài "Iris-versicolor".

8.3. Phân tích phân bố chiều dài lá đài của loài "Iris-virginica"



Hình 8.3: Biểu đồ thể hiện phân bố chiều dài lá đài của loài "Iris-virginica".

Đại lượng thống kê cho `Chiều dài lá đài (cm)` của loài `Iris-virginica`	Giá trị
min	4.900000
lower_quartile	6.200000
median	6.500000
upper_quartile	6.900000
max	7.900000
mean	6.588000
std	0.635880
skew	0.118015
kurt	0.032904

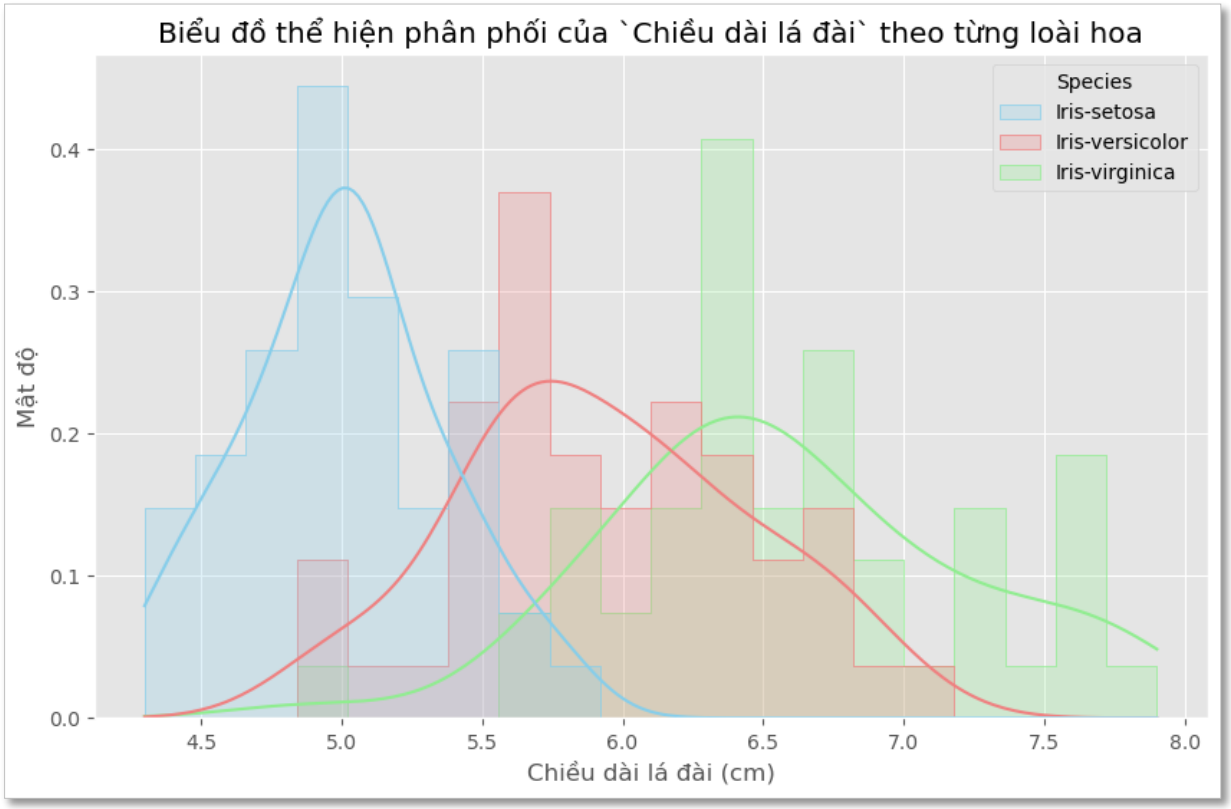
Bảng 8.3: Bảng mô tả các đại lượng thống kê cho chiều dài lá đài của loài "Iris-virginica".

Nhận xét:

- Ta thấy phân bố chiều dài lá đài của loài "Iris-virginica" khá đối xứng (|skew| = 0.12 < 0.5) và có phần đuôi không quá đậm (kurt = 0.03 > 0):
  - Các giá trị mà ta quan sát được sẽ có phạm vi phân bố nằm trong đoạn [4.9; 7.9] (đơn vị: cm).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [6.2; 6.9] (đơn vị: cm).
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều dài trung bình của lá đài ở loài hoa "Iris-virginica" là: (6.41; 6.76) (đơn vị: cm).
- Quan sát biểu đồ hộp, ta thấy giá trị tối thiểu (4.9) đang được đánh dấu là giá trị ngoại lai theo phương pháp IQR. Vì số lượng giá trị ngoại lai là không quá lớn nên ta có thể loại bỏ điểm dữ liệu này trong quá trình phân tích để thu được kết quả có độ chính xác cao hơn.
- Mặt khác, hình dạng phân bố của chiều dài lá đài ở hai loài hoa "Iris-virginica" và "Iris-versicolor" có nhiều nét tương đồng. Đây có thể là một đặc điểm đáng chú ý trong việc phân biệt giữa hai loài hoa này thông qua thuộc tính chiều dài lá đài.

8.4. So sánh phân bố chiều dài lá đài ở ba loài hoa và đưa ra kết luận

Để dễ dàng so sánh phân bố chiều dài lá đài của ba loài hoa, ta sẽ trực quan hóa dữ liệu trên cùng một biểu đồ mật độ và dùng màu sắc để chú thích cho phân bố của mỗi loài hoa.



Hình 8.4: Biểu đồ thể hiện phân bố chiều dài lá đài ở ba loài hoa Iris.

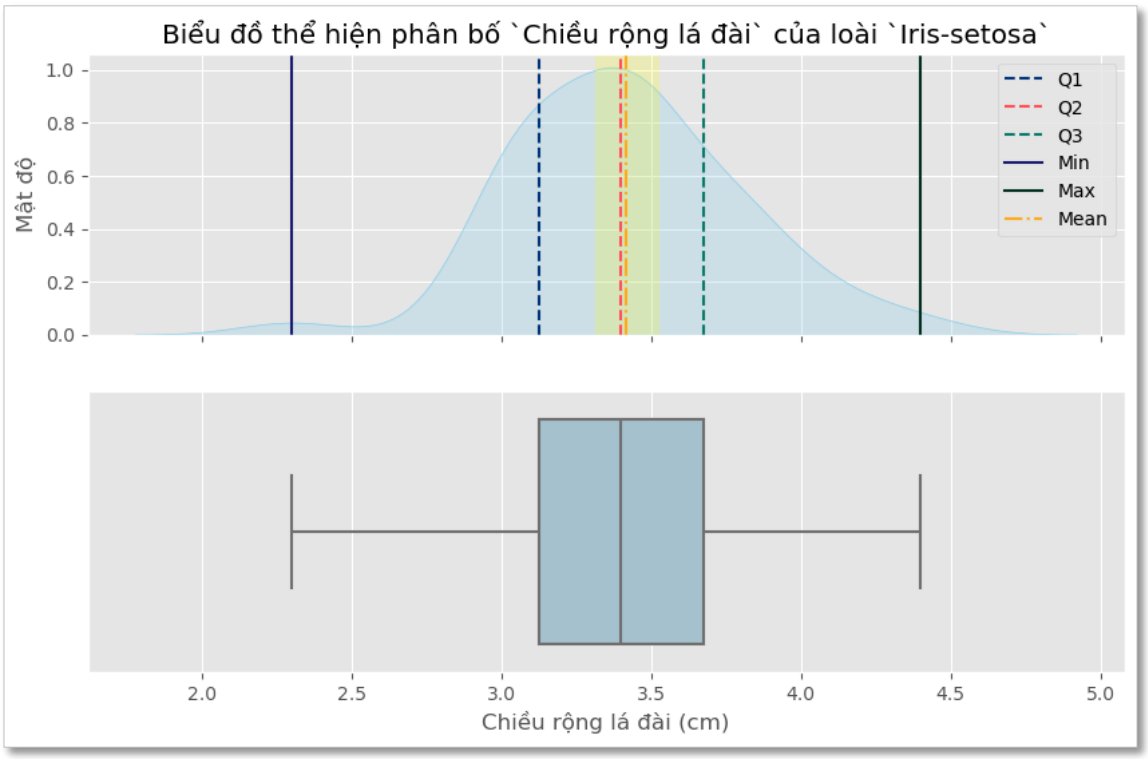
Nhận xét:

- Trong bộ dữ liệu mà ta đang phân tích, "Iris-setosa" là loài hoa diên vĩ có chiều dài lá đài ngắn nhất, sau đó là đến loài "Iris-versicolor", và "Iris-virginica" là loài hoa có chiều dài lá đài dài nhất. Tuy nhiên, đây chỉ là nhận định khách quan dựa trên quan sát từ biểu đồ, ta sẽ cần phân tích thêm nhiều mẫu dữ liệu hơn để có thể đưa cái nhìn tổng quát hơn.
- Điểm nổi bật nhất mà ta thấy được chính là tình trạng chồng chéo của các điểm dữ liệu ở các loài hoa khác nhau. Chính điều này làm cho phần giao trong tập giá trị chiều dài lá đài ở các loài hoa trở nên sát lại gần nhau và không thể tách biệt chúng bằng một ranh giới rõ ràng. Do đó, ta có thể khẳng định rằng: ta không thể sử dụng đặc điểm chiều dài lá đài ở mỗi loài hoa một cách độc lập để có thể phân loại chính xác các loài hoa diên vĩ khác nhau.
- Như vậy, ta sẽ cần phân tích thêm các đặc trưng khác ở loài hoa diên vĩ để có thể tìm ra cách phân loại các loài hoa với độ chính xác thỏa một ngưỡng được xác định trước.

9. TIỀN XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU ĐỂ TRẢ LỜI CHO CÂU HỎI 2

Câu hỏi 2: Liệu chiều rộng lá đài (SepalWidthCm) có phải là yếu tố đủ để phân loại chính xác các loài hoa Iris (diên vĩ) khác nhau hay không?

9.1. Phân tích phân bố chiều rộng lá đài của loài "Iris-setosa"



Hình 9.1: Biểu đồ thể hiện phân bố chiều rộng lá đài của loài "Iris-setosa".

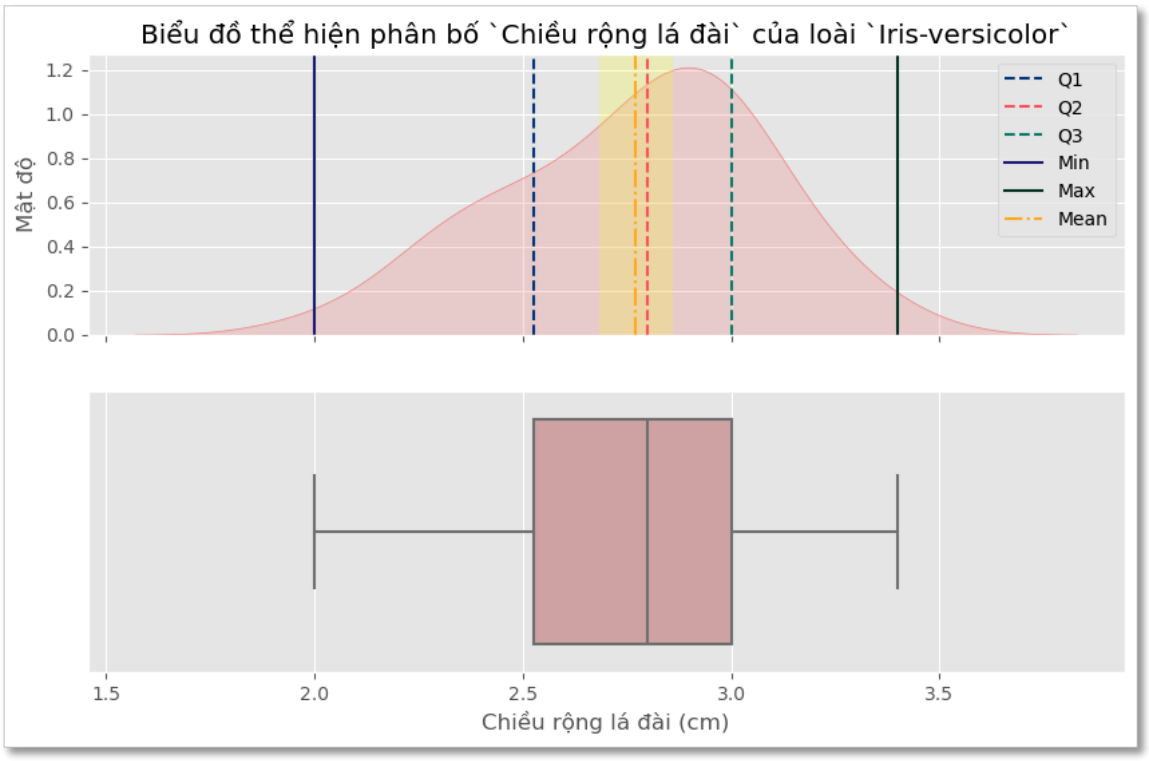
Đại lượng thống kê cho `Chiều rộng lá đài (cm)` của loài `Iris-setosa`	Giá trị
min	2.300000
lower_quartile	3.100000
median	3.400000
upper_quartile	3.700000
max	4.400000
mean	3.418000
std	0.381024
skew	0.107053
kurt	0.889251

Bảng 9.1: Bảng mô tả các đại lượng thống kê cho chiều rộng lá đài của loài "Iris-setosa".

Nhận xét:

- Ta thấy chiều rộng lá đài (SepalWidthCm) của loài "Iris-setosa" có phân bố được xem là đối xứng (|skew| = 0.11 < 0.5) và có phần đuôi khá đậm (kurt = 0.89 > 0):
  - Các giá trị ta quan sát được sẽ phân bố trong đoạn [2.3; 4.4] (đơn vị: cm).
  - 50% điểm dữ liệu nằm chính giữa khoảng phân bố sẽ có giá trị nằm trong đoạn [3.1; 3.7] (đơn vị: cm).
  - 25% giá trị nhỏ nhất và 25% giá trị lớn nhất có phạm vi phân bố rộng hơn một chút so với phạm vi phân bố của 50% giá trị ở khu vực trung tâm. Điều này cho thấy các điểm dữ liệu có xu hướng tập trung ở phần đuôi của phân phối.
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều rộng trung bình của lá đài ở loài hoa "Iris-setosa" là: (3.31; 3.52) (đơn vị: cm).
- Như vậy, ta thấy chiều rộng lá đài của loài "Iris-setosa" không tập trung quá nhiều ở phần trung tâm của phân phối mà có xu hướng phân tán rộng sang hai bên. Đây có thể là một bất lợi trong việc sử dụng chiều rộng lá đài để phân biệt "Iris-setosa" với các loài hoa diên vĩ khác vì mức độ phân tán càng lớn thì tình trạng "chồng chéo" sẽ có khả năng xảy ra cao hơn. Nhưng trước hết, ta cần phân tích thêm phân bố chiều rộng lá đài ở các loài hoa khác để có cái nhìn chính xác hơn.

9.2. Phân tích phân bố chiều rộng lá đài của loài "Iris-versicolor"



Hình 9.2: Biểu đồ thể hiện phân bố chiều rộng lá đài của loài "Iris-versicolor".

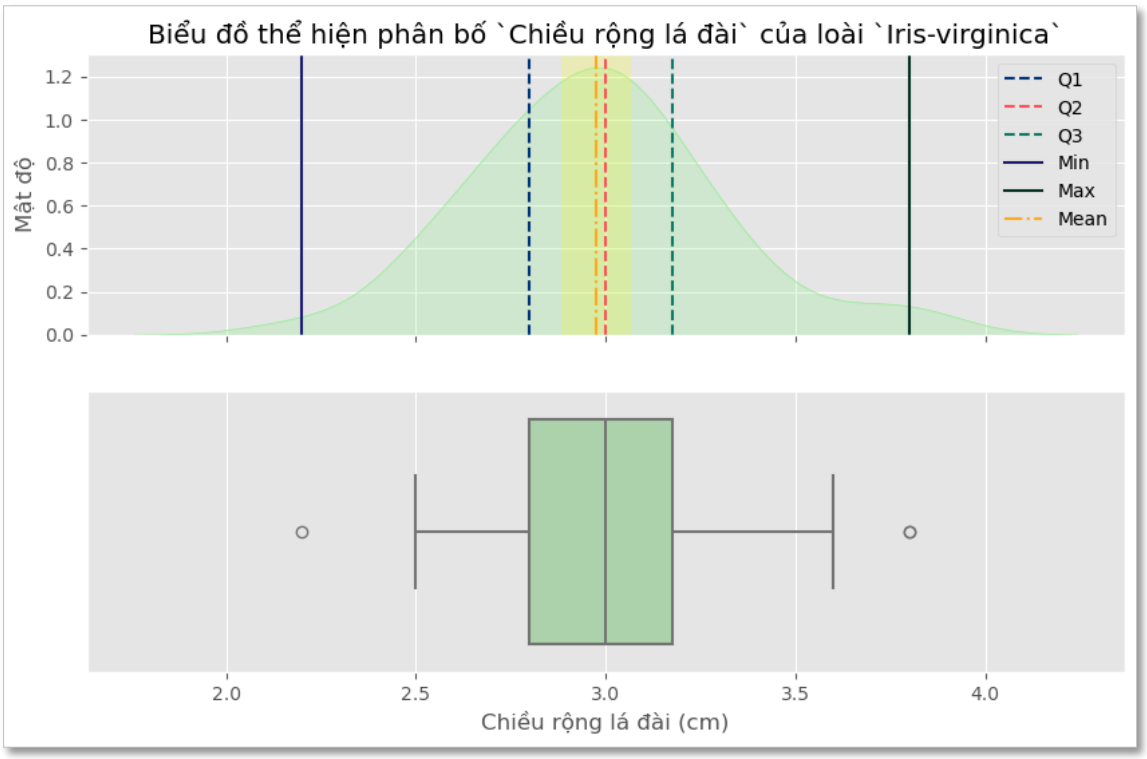
Đại lượng thống kê cho `Chiều rộng lá đài (cm)` của loài `Iris-versicolor`	Giá trị
min	2.000000
lower_quartile	2.500000
median	2.800000
upper_quartile	3.000000
max	3.400000
mean	2.770000
std	0.313798
skew	-0.362845
kurt	-0.366237

Bảng 9.2: Bảng mô tả các đại lượng thống kê cho chiều rộng lá đài của loài "Iris-versicolor".

Nhận xét:

- Ta thấy phân bố chiều rộng lá đài của loài "Iris-versicolor" có mức độ đối xứng không quá cao (|skew| = 0.36 < 0.5) và có phần đuôi khá mỏng (kurt = -0.37 < 0):
  - Các giá trị mà ta quan sát được sẽ có phạm vi phân bố nằm trong đoạn [2.0; 3.4] (đơn vị: cm).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [2.5; 3.0] (đơn vị: cm).
  - 25% giá trị nhỏ nhất, 50% giá trị ở trung tâm và 25% giá trị lớn nhất đều có phạm vi phân bố tương đương nhau, không chênh lệch quá nhiều.
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều rộng trung bình của lá đài ở loài hoa "Iris-versicolor" là: (2.68; 2.86) (đơn vị: cm).
- Như vậy, các giá trị chiều rộng lá đài ở loài "Iris-versicolor" thường có xu hướng tập trung ở khoảng giữa của phân bố. Đây có thể là một yếu tố then chốt giúp tạo ra một vùng "biên giới" để có thể phân biệt "Iris-versicolor" với các loài hoa khác. Tuy nhiên, ta cần so sánh phân bố chiều rộng lá đài ở các loài hoa khác nhau để có thể đưa ra kết luận cuối cùng.

9.3. Phân tích phân bố chiều rộng lá đài của loài "Iris-virginica"



Hình 9.3: Biểu đồ thể hiện phân bố chiều rộng lá đài của loài "Iris-virginica".

Đại lượng thống kê cho `Chiều rộng lá đài (cm)` của loài `Iris-virginica`	Giá trị
min	2.200000
lower_quartile	2.800000
median	3.000000
upper_quartile	3.200000
max	3.800000
mean	2.974000
std	0.322497
skew	0.365949
kurt	0.706071

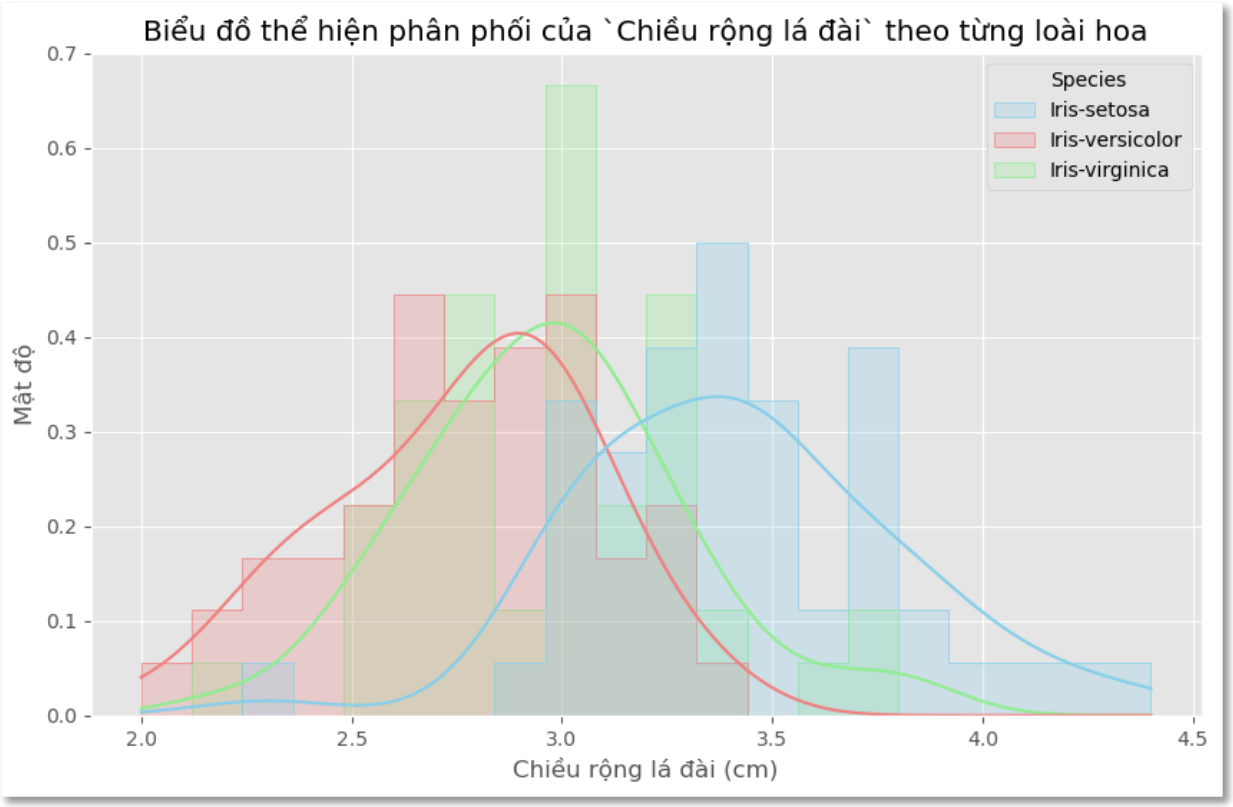
Bảng 9.3: Bảng mô tả các đại lượng thống kê cho chiều rộng lá đài của loài "Iris-virginica".

Nhận xét:

- Ta thấy phân bố chiều rộng lá đài của loài "Iris-virginica" khá đối xứng (|skew| = 0.37 < 0.5) và có phần đuôi khá đậm (kurt = 0.71 > 0):
  - Các giá trị mà ta quan sát được sẽ có phạm vi phân bố nằm trong đoạn [2.2; 3.8] (đơn vị: cm).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [2.8; 3.2] (đơn vị: cm).
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều rộng trung bình của lá đài ở loài hoa "Iris-virginica" là: (2.88; 3.06) (đơn vị: cm).
- Quan sát biểu đồ hộp, theo phương pháp IQR, ta phát hiện có hai điểm dữ liệu ngoại lai nằm ở hai đầu của phân phối. Đó đồng thời là giá trị tối thiểu và giá trị tối đa của thuộc tính mà ta đang quan sát. Vì số lượng giá trị ngoại lai là không quá lớn nên ta có thể loại bỏ các điểm dữ liệu này trong quá trình phân tích để thu được kết quả có độ chính xác cao hơn.

9.4. So sánh phân bố chiều rộng lá đài ở ba loài hoa và đưa ra kết luận

Để dễ dàng so sánh phân bố chiều rộng lá đài của ba loài hoa, ta sẽ trực quan hóa dữ liệu trên cùng một biểu đồ mật độ và dùng màu sắc để chú thích cho phân bố của mỗi loài hoa.



Hình 9.4: Biểu đồ thể hiện phân bố chiều rộng lá đài ở ba loài hoa Iris.

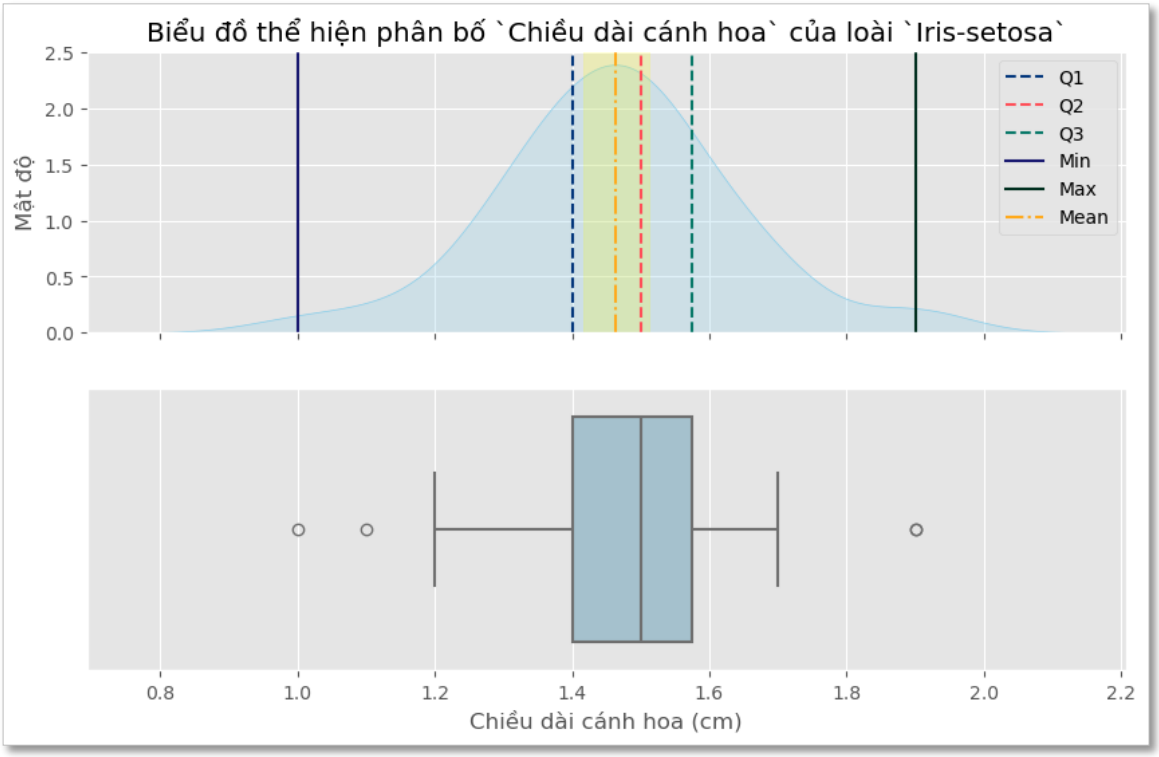
Nhận xét:

- Trong bộ dữ liệu mà ta đang phân tích, "Iris-setosa" là loài hoa diên vĩ có chiều rộng lá đài lớn nhất, kế đến là loài "Iris-virginica", và "Iris-versicolor" là lòai hoa có chiều rộng lá đài nhỏ nhất. Tuy nhiên, do số lượng mẫu dữ liệu mà ta đang phân tích là không quá lớn nên các quan sát ở trên chỉ mang tính tham khảo, ta cần tiến hành phân tích trên bộ dữ liệu lớn và đa dạng hơn để có thể đưa ra kết luận khách quan hơn.
- Tương tự như khi so sáng chiều dài lá đài ở các loài hoa khác nhau, ta thấy phân bố chiều rộng lá đài của các loài hoa diên vĩ khác nhau không có tách biệt hoàn toàn mà có xu hướng chồng chéo, đan xen lẫn nhau rất phức tạp. Sự chồng chéo này không chỉ xuất hiện ở phần đuôi của các phân bố mà chúng có sự giao thoa rất lớn. Đặc biệt là ở hai loài "Iris-versicolor" và "Iris-virginica", ta thấy phân bố của chúng gần như nằm chồng lên nhau. Như vậy, dù có loại bỏ các giá trị ngoại lai trong bộ dữ liệu của loài "Iris-virginica" thì tình trạng này cũng không có sự cải thiện đáng kể. Do đó, ta có thể khẳng định rằng: ta không thể sử dụng đặc điểm chiều rộng lá đài ở mỗi loài hoa diên vĩ như một phép thử độc lập để có thể phân loại chính xác các loài hoa khác nhau.
- Như vậy, ta sẽ cần phân tích thêm các đặc trưng khác ở loài hoa diên vĩ để có thể tìm ra cách phân loại các loài hoa với độ chính xác đáp ứng một ngưỡng cho trước.

10. TIỀN XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU ĐỂ TRẢ LỜI CHO CÂU HỎI 3

Câu hỏi 3: Liệu chiều dài cánh hoa (PetalLengthCm) có phải là yếu tố đủ để phân loại chính xác các loài hoa Iris (diên vĩ) khác nhau hay không?

10.1. Phân tích phân bố chiều dài cánh hoa của loài "Iris-setosa"



Hình 10.1: Biểu đồ thể hiện phân bố chiều dài cánh hoa của loài "Iris-setosa".

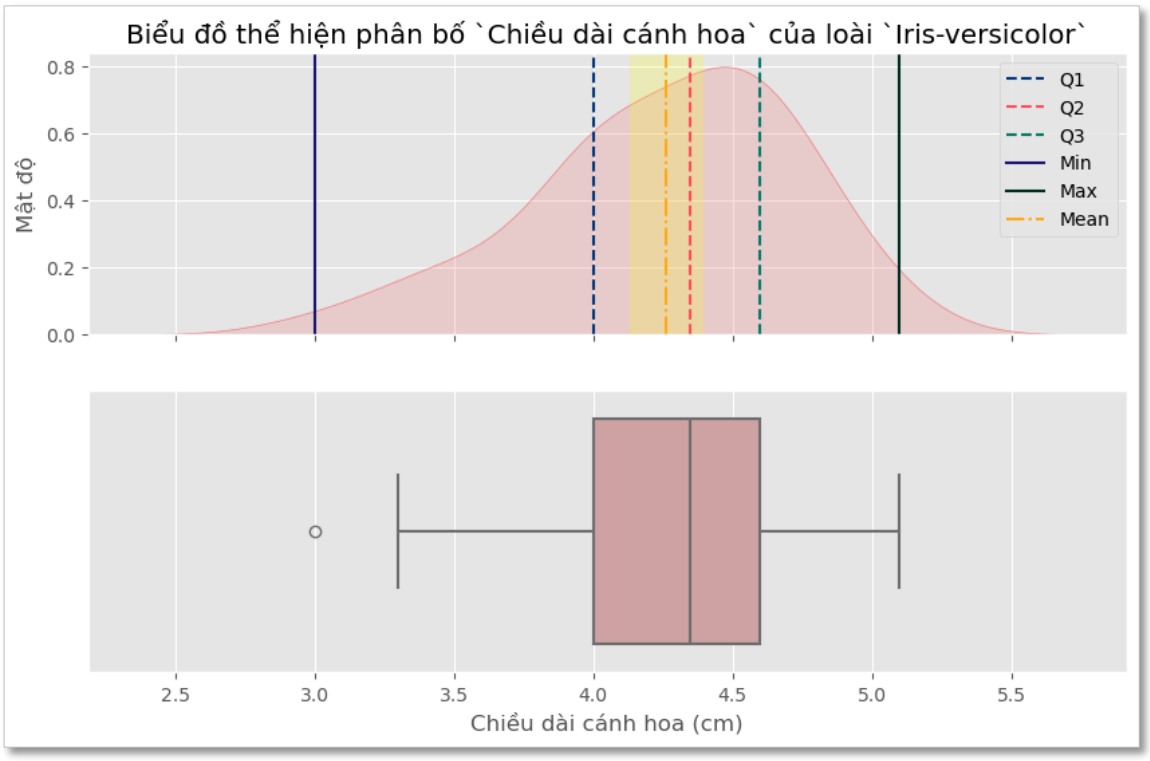
Đại lượng thống kê cho `Chiều dài cánh hoa (cm)` của loài `Iris-setosa`	Giá trị
min	1.000000
lower_quartile	1.400000
median	1.500000
upper_quartile	1.600000
max	1.900000
mean	1.464000
std	0.173511
skew	0.071846
kurt	1.031626

Bảng 10.1: Bảng mô tả các đại lượng thống kê cho chiều dài cánh hoa của loài "Iris-setosa".

Nhận xét:

- Ta thấy chiều dài cánh hoa (PetalLengthCm) của loài "Iris-setosa" có phân bố hình chuông, khá đối xứng (|skew| = 0.07 < 0.5) và có phần đuôi khá đậm (kurt = 1.03 > 0):
  - Chiều dài cánh hoa của loài "Iris-setosa" có giá trị nhỏ nhất là 1.0 (cm) và giá trị lớn nhất là 1.9 (cm).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [1.4; 1.6] (đơn vị: cm).
  - Nhìn vào biểu đồ mật độ, ta thấy có sự khác biệt đáng kể trong phạm vi phân bố của 25% giá trị nhỏ nhất và 25% giá trị lớn nhất so với phạm vi phân bố của 50% giá trị nằm ở trung tâm.
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều dài trung bình của cánh hoa ở loài "Iris-setosa" là: (1.42; 1.51) (đơn vị: cm).
- Quan sát biểu đồ hộp, ta thấy phương pháp IQR giúp ta chỉ ra khoảng ba điểm dữ liệu ngoại lai, đó là các điểm có khoảng cách quá xa so với phạm vi phân bố của phần lớn giá trị còn lại. Điều này có thể lý giải phần nào cho việc biểu đồ mật độ có phần đuôi trải ra khá xa so với phần trung tâm được giới hạn của các giá trị phân vị. Như vậy, việc tìm cách xử lý các giá trị ngoại lai này (có thể là loại bỏ hoàn toàn hoặc thay thế chúng bằng các giá trị biên, v.v.) sẽ giúp các phân tích của ta có độ tin cậy cao hơn.

10.2. Phân tích phân bố chiều dài cánh hoa của loài "Iris-versicolor"



Hình 10.2: Biểu đồ thể hiện phân bố chiều dài cánh hoa của loài "Iris-versicolor".

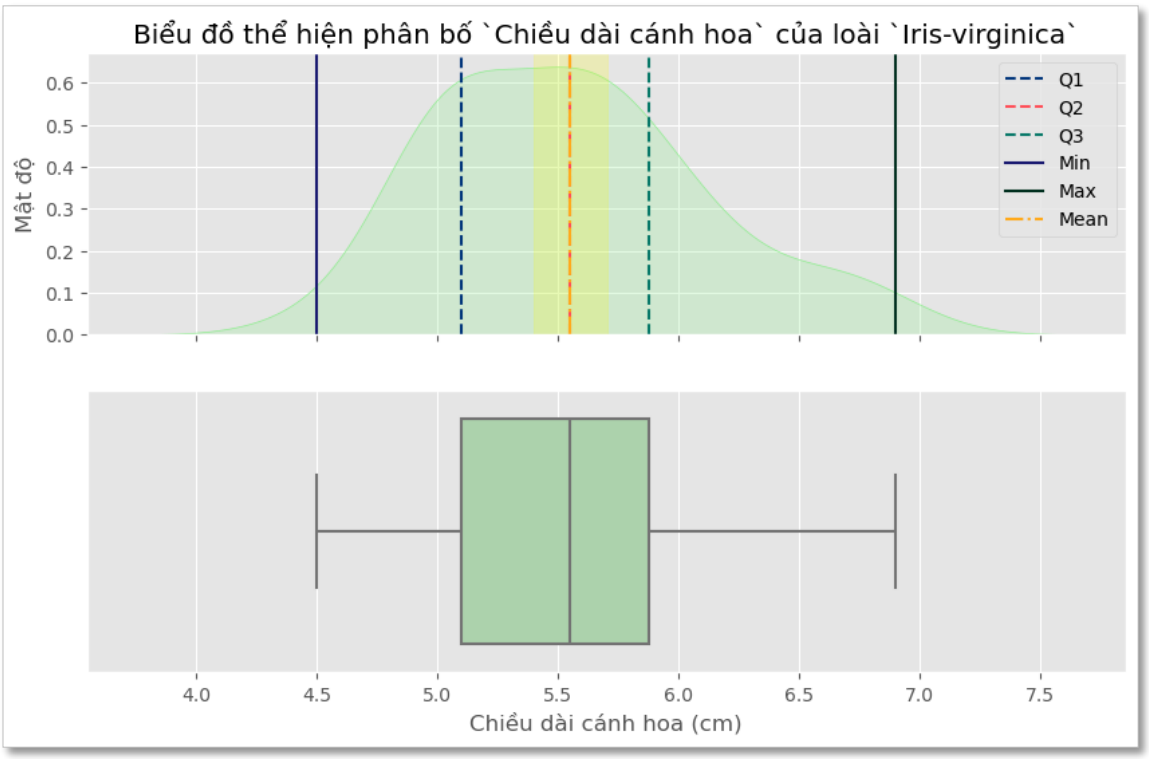
Đại lượng thống kê cho `Chiều dài cánh hoa (cm)` của loài `Iris-versicolor`	Giá trị
min	3.000000
lower_quartile	4.000000
median	4.400000
upper_quartile	4.600000
max	5.100000
mean	4.260000
std	0.469911
skew	-0.606508
kurt	0.047903

Bảng 10.2: Bảng mô tả các đại lượng thống kê cho chiều dài cánh hoa của loài "Iris-versicolor".

Nhận xét:

- Ta thấy phân bố chiều dài cánh hoa của loài "Iris-versicolor" bị lệch trái (skew = -0.61 < -0.5) và có phần đuôi hơi đậm (kurt = 0.05 > 0):
  - Các giá trị mà ta quan sát được sẽ có phạm vi phân bố nằm trong đoạn [3.0; 5.1] (đơn vị: cm).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [4.0; 4.6] (đơn vị: cm).
  - Việc 25% giá trị nhỏ nhất có phạm vi phân bố rộng hơn khá nhiều so với phạm vi phân bố của 50% giá trị ở khu vực trung tâm và 25% giá trị lớn nhất là một điều cần được phân tích chi tiết.
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều dài trung bình của cánh hoa ở loài "Iris-versicolor" là: (4.13; 4.39) (đơn vị: cm).
- Quan sát biểu đồ hộp, ta thấy xuất hiện một giá trị ngoại lai nằm lệch về phía bên trái của phân phối, đây đồng thời là giá trị nhỏ nhất mà ta ghi nhận được. Điều này có thể lý giải phần nào cho phạm vi phân bố lớn bất thường của 25% giá trị nhỏ nhất như đã phân tích bên trên. Như vậy, việc tiền xử lý giá trị ngoại lai này sẽ đóng vai trò rất lớn trong việc tăng cường tính đối xứng của phân phối và giúp các phân tích có độ tin cậy cao hơn.

10.3. Phân tích phân bố chiều dài cánh hoa của loài "Iris-virginica"



Hình 10.3: Biểu đồ thể hiện phân bố chiều dài cánh hoa của loài "Iris-virginica".

Đại lượng thống kê cho `Chiều dài cánh hoa (cm)` của loài `Iris-virginica`	Giá trị
min	4.500000
lower_quartile	5.100000
median	5.600000
upper_quartile	5.900000
max	6.900000
mean	5.552000
std	0.551895
skew	0.549445
kurt	-0.153779

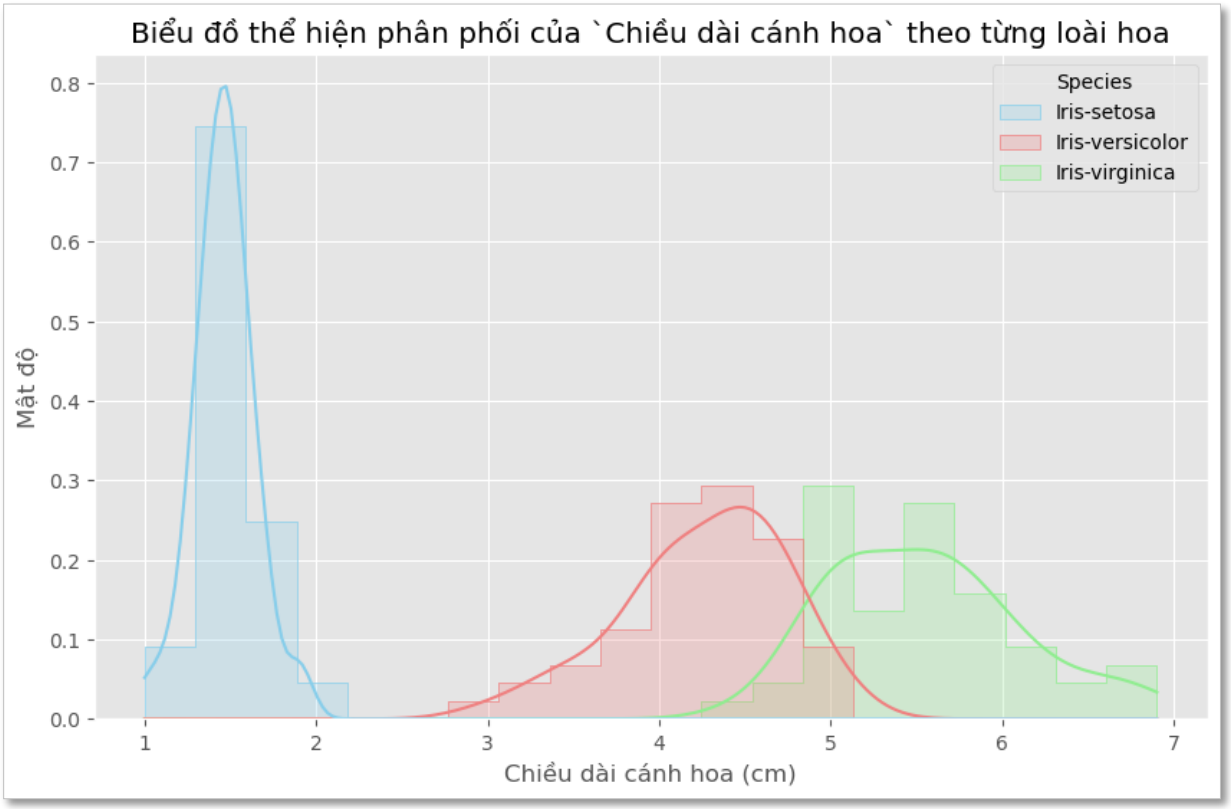
Bảng 10.3: Bảng mô tả các đại lượng thống kê cho chiều dài cánh hoa của loài "Iris-virginica".

Nhận xét:

- Ta thấy phân bố chiều dài cánh hoa của loài "Iris-virginica" hơi lệch về phía bên phải (skew = 0.55 > 0.5) và có phần đuôi hơi mỏng (kurt = -0.15 < 0):
  - Các giá trị mà ta quan sát được sẽ có phạm vi phân bố nằm trong đoạn [4.5; 6.9] (đơn vị: cm).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [5.1; 5.9] (đơn vị: cm).
  - 25% giá trị lớn nhất có phạm vi phân bố lớn hơn đáng kể so với phạm vi phân bố của 25% giá trị nhỏ nhất và 50% giá trị ở khu vực trung tâm.
  - Các điểm dữ liệu có xu hướng tập trung đông đúc hơn ở khu vực trung tâm, càng tiến về hai biên, mật độ dữ liệu có xu hướng giảm mạnh.
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều dài trung bình của cánh hoa ở loài "Iris-virginica" là: (5.40; 5.70) (đơn vị: cm).
- Quan sát biểu đồ hộp, theo phương pháp IQR, ta không thấy có sự xuất hiện của bất kỳ giá trị ngoại lai nào. Điều này cho thấy tình trạng phân phối của dữ liệu bị lệch có thể xuất phát từ chính bản chất của dữ liệu, hoặc có thể là do số lượng mẫu dữ liệu của ta chưa đủ lớn, v.v.. Đây là một đặc điểm thú vị mà ta có thể nghiên cứu, phân tích sâu hơn để phát hiện ra các nguyên nhân tiềm ẩn phía sau hiện tượng này. Đó có thể là một manh mối quý giá giúp ta dễ dàng phân biệt "Iris-virginica" với các loài hoa diên vĩ khác.
- Mặt khác, tình trạng phân phối chiều dài cánh hoa của loài "Iris-virginica" bị lệch cũng mang lại một vài thử thách trong quá trình xây dựng mô hình học máy giúp phân lớp các loài hoa. Do đó, ta có thể phải sử dụng một số kỹ thuật giúp tăng cường tính đối xứng của dữ liệu này trước khi tiến hành huấn luyện mô hình để có được kết quả thực sự tốt và khách quan.

10.4. So sánh phân bố chiều dài cánh hoa ở ba loài hoa và đưa ra kết luận

Để dễ dàng so sánh phân bố chiều dài cánh hoa của ba loài hoa, ta sẽ trực quan hóa dữ liệu trên cùng một biểu đồ mật độ và dùng màu sắc để chú thích cho phân bố của mỗi loài hoa.



Hình 10.4: Biểu đồ thể hiện phân bố chiều dài cánh hoa ở ba loài hoa Iris.

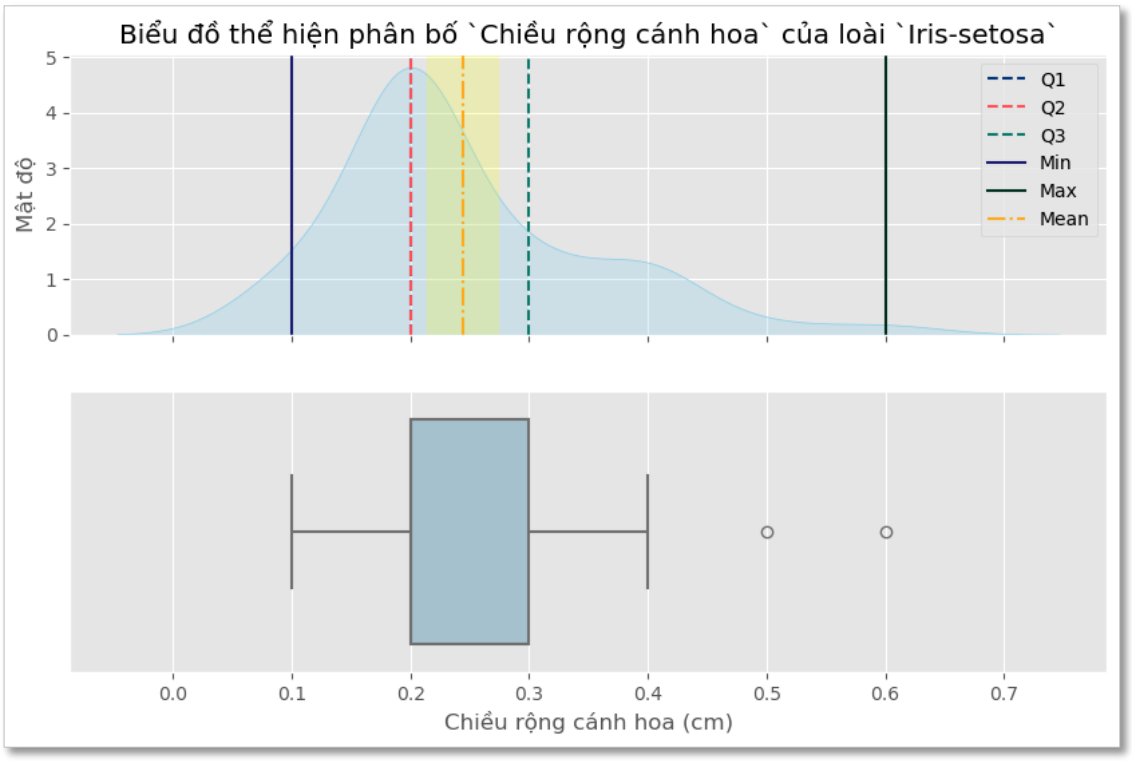
Nhận xét:

- Trong bộ dữ liệu mà ta đang phân tích, cánh hoa của loài "Iris-setosa" thường có chiều dài ngắn hơn đáng kể so với các loài khác. Loài "Iris-versicolor" có chiều dài cánh hoa ở mức trung bình và "Iris-virginica" là loài có cánh hoa dài nhất. Tuy nhiên, đây chỉ là nhận định khách quan dựa trên việc phân tích dữ liệu và quan sát biểu đồ, ta sẽ cần phân tích thêm nhiều mẫu dữ liệu hơn để có thể đưa cái nhìn tổng quát và chính xác hơn.
- Khác với khi phân tích trên kích thước (chiều dài và chiều rộng) của lá đài, ở đặc trưng chiều dài cánh hoa, ta thấy có sự tách biệt khá rõ rệt giữa các mẫu dữ liệu của loài "Iris-setosa" với hai loài còn lại. Dường như chỉ với một đường thẳng là ta đã có thể xác định miền ranh giới giữa các bông hoa thuộc lớp "Iris-setosa" với các lớp khác.
- Tuy nhiên, mọi thứ sẽ phức tạp hơn trong trường hợp ta muốn phân biệt rạch ròi giữa các mẫu dữ liệu thuộc hai loài "Iris-versicolor" và "Iris-virginica". Mặc dù tình trạng các phân bố chồng chéo lên nhau đã có phần cải thiện nhưng chúng vẫn không tách biệt hoàn toàn. Do đó, ta chưa có một cơ sở đủ mạnh mẽ để có thể giải quyết bài toán phân lớp giữa hai loài "Iris-versicolor" và "Iris-virginica".
- Như vậy, ta đã tìm ra manh mối đầu tiên trên con đường hoàn thành mục tiêu đã đề ra. Chỉ với một đặc trưng chiều dài cánh hoa, ta đã có thể nhận biết chính xác phần lớn các mẫu dữ liệu thuộc loài "Iris-setosa". Tuy nhiên, để có thể phân biệt rạch ròi giữa hai loài "Iris-versicolor" và "Iris-virginica" vẫn là một bài toán nan giải mà ta cần tiếp tục giải quyết.
- Khi này, ta sẽ cần phân tích thêm đặc trưng cuối cùng trong bộ dữ liệu (chiều rộng cánh hoa) với hy vọng tìm ra cách để phân biệt giữa hai loài "Iris-versicolor" và "Iris-virginica". Đồng thời, ta cũng hy vọng rằng đặc trưng này có thể là một "công cụ đắc lực" giúp tăng cường khả năng phân lớp chính xác các mẫu dữ liệu thuộc loài "Iris-setosa".

11. TIỀN XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU ĐỂ TRẢ LỜI CHO CÂU HỎI 4

Câu hỏi 4: Liệu chiều rộng cánh hoa (PetalWidthCm) có phải là yếu tố đủ để phân loại chính xác các loài hoa Iris (diên vĩ) khác nhau hay không?

11.1. Phân tích phân bố chiều rộng cánh hoa của loài "Iris-setosa"



Hình 11.1: Biểu đồ thể hiện phân bố chiều rộng cánh hoa của loài "Iris-setosa".

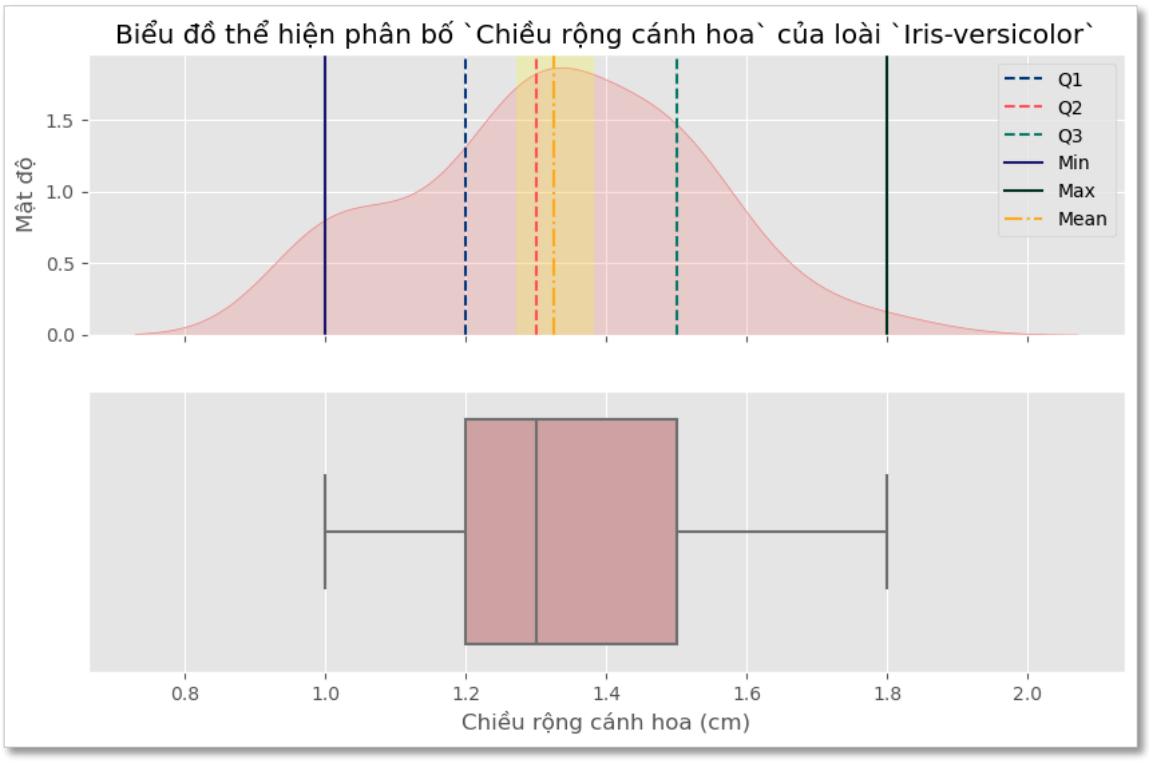
Đại lượng thống kê cho `Chiều rộng cánh hoa (cm)` của loài `Iris-setosa`	Giá trị
min	0.100000
lower_quartile	0.200000
median	0.200000
upper_quartile	0.300000
max	0.600000
mean	0.244000
std	0.107210
skew	1.197243
kurt	1.566442

Bảng 11.1: Bảng mô tả các đại lượng thống kê cho chiều rộng cánh hoa của loài "Iris-setosa".

Nhận xét:

- Ta thấy chiều rộng cánh hoa của loài "Iris-setosa" có phân bố bị lệch sang bên phải (skew = 1.20 > 0.5) và có phần đuôi đậm (kurt = 1.57 > 0):
  - Các giá trị ta quan sát được sẽ phân bố trong đoạn [0.1; 0.6] (đơn vị: cm).
  - 50% điểm dữ liệu nằm chính giữa khoảng phân bố sẽ có giá trị nằm trong đoạn [0.2; 0.3] (đơn vị: cm).
  - Việc 25% giá trị lớn nhất có phạm vi phân bố rộng một cách bất thường so với phạm vi phân bố của 25% giá trị nhỏ nhất và 50% giá trị ở khu vực trung tâm là một điểm đáng chú ý.
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều rộng trung bình của cánh hoa ở loài "Iris-setosa" là: (0.21; 0.27) (đơn vị: cm).
- Quan sát biểu đồ hộp, sử dụng phương pháp IQR, ta thấy xuất hiện các giá trị ngoại lai nằm lệch về bên phải của phân phối. Điều này có thể phần nào giải thích cho việc 25% giá trị lớn nhất có phạm vi phân bố rộng bất thường như đã phân tích ở trên. Đồng thời, ta còn thấy giá trị trung vị hoàn toàn trùng khớp với giá trị tứ phân vị thứ nhất. Điều này càng chứng tỏ rằng phần lớn mẫu dữ liệu thuộc loài "Iris-setosa" thường có cánh hoa rất hẹp. Do đó, ta có thể áp dụng các phương pháp xử lý giá trị ngoại lại để tăng cường tính đối xứng và chất lượng cho tập dữ liệu này.
- Như vậy, ta thấy chiều rộng cánh hoa của loài "Iris-setosa" có xu hướng tập trung rất nhiều ở phần trung tâm của phân phối. Việc tập trung phân tích vào các mẫu dữ liệu này có thể giúp ta phát hiện ra nguyên nhân ẩn sau hiện tượng này.

11.2. Phân tích phân bố chiều rộng cánh hoa của loài "Iris-versicolor"



Hình 11.2: Biểu đồ thể hiện phân bố chiều rộng cánh hoa của loài "Iris-versicolor".

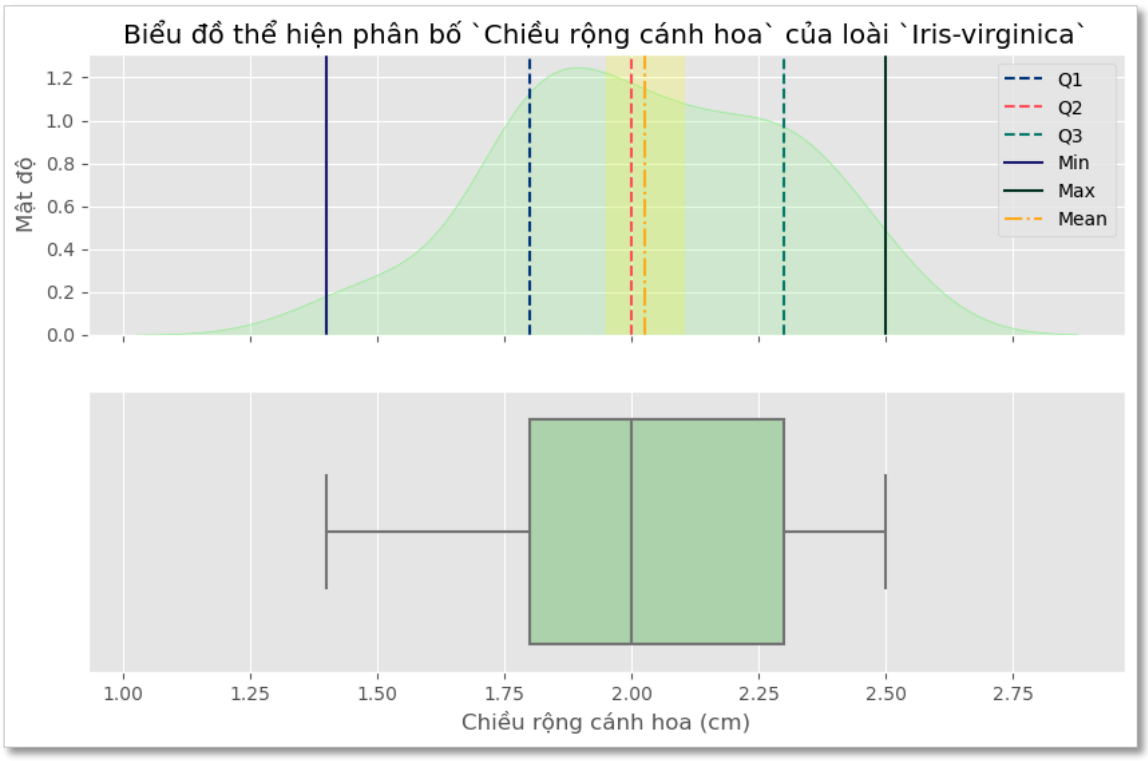
Đại lượng thống kê cho `Chiều rộng cánh hoa (cm)` của loài `Iris-versicolor`	Giá trị
min	1.000000
lower_quartile	1.200000
median	1.300000
upper_quartile	1.500000
max	1.800000
mean	1.326000
std	0.197753
skew	-0.031180
kurt	-0.410059

Bảng 11.2: Bảng mô tả các đại lượng thống kê cho chiều rộng cánh hoa của loài "Iris-versicolor".

Nhận xét:

- Ta thấy phân bố chiều rộng cánh hoa của loài "Iris-versicolor" khá đối xứng (|skew| = 0.03 < 0.5) và có phần đuôi hơi mỏng (kurt = -0.41 < 0):
  - Các giá trị mà ta quan sát được sẽ có phạm vi phân bố nằm trong đoạn [1.0; 1.8] (đơn vị: cm).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [1.2; 1.5] (đơn vị: cm).
  - 25% giá trị nhỏ nhất, 50% giá trị ở trung tâm và 25% giá trị lớn nhất đều có phạm vi phân bố tương đương nhau, không chênh lệch quá nhiều.
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều rộng trung bình của cánh hoa ở loài "Iris-versicolor" là: (1.27; 1.38) (đơn vị: cm).
- Như vậy, các giá trị chiều rộng cánh hoa ở loài "Iris-versicolor" thường có xu hướng tập trung ở khoảng giữa của phân bố, càng tiến về hai đuôi, mật độ dữ liệu càng giảm. Đây có thể là một yếu tố then chốt giúp tạo ra một vùng "biên giới" để có thể phân biệt "Iris-versicolor" với các loài hoa khác. Tuy nhiên, ta cần có sự so sánh phân bố chiều rộng cánh hoa ở các loài khác nhau để có thể đưa ra kết luận cuối cùng.

11.3. Phân tích phân bố chiều rộng cánh hoa của loài "Iris-virginica"



Hình 11.3: Biểu đồ thể hiện phân bố chiều rộng cánh hoa của loài "Iris-virginica".

Đại lượng thống kê cho `Chiều rộng cánh hoa (cm)` của loài `Iris-virginica`	Giá trị
min	1.400000
lower_quartile	1.800000
median	2.000000
upper_quartile	2.300000
max	2.500000
mean	2.026000
std	0.274650
skew	-0.129477
kurt	-0.602264

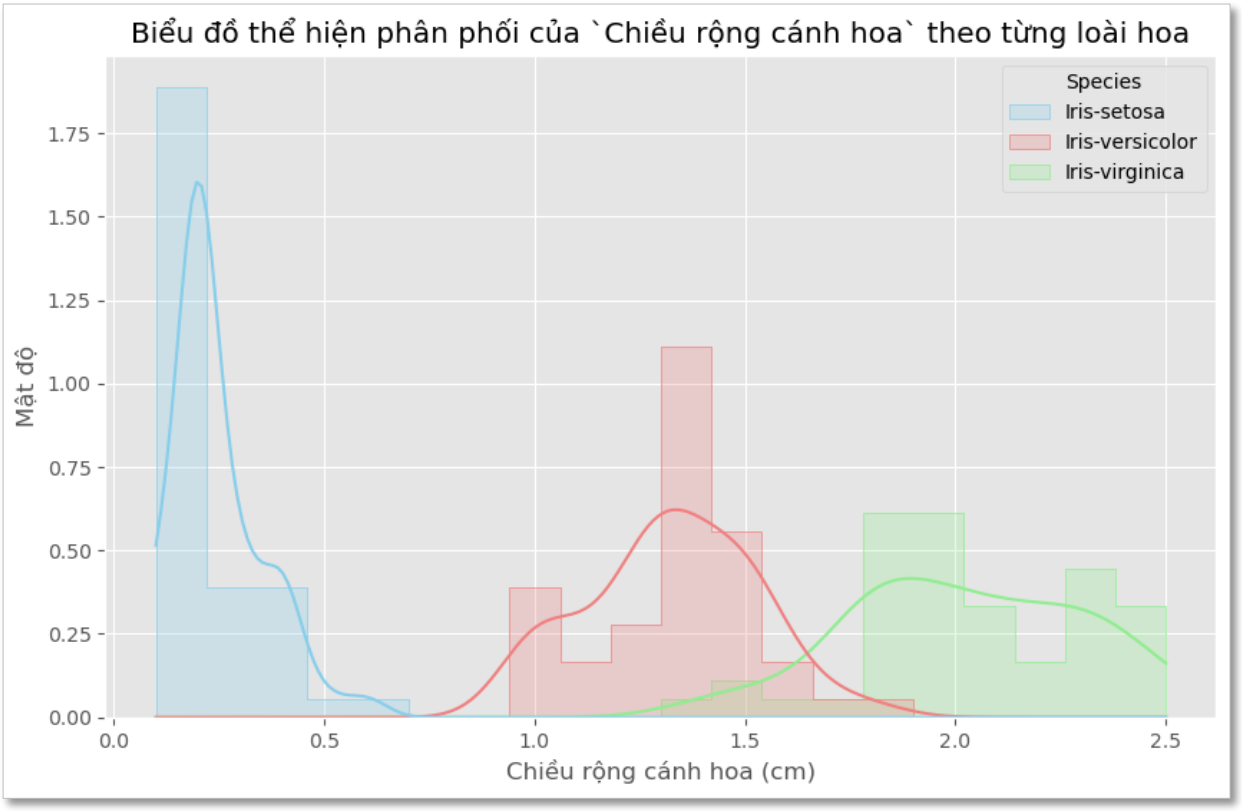
Bảng 11.3: Bảng mô tả các đại lượng thống kê cho chiều rộng cánh hoa của loài "Iris-virginica".

Nhận xét:

- Ta thấy phân bố chiều rộng cánh hoa của loài "Iris-virginica" khá đối xứng (|skew| = 0.13 < 0.5) và có phần đuôi khá mỏng (kurt = -0.60 < 0):
  - Các giá trị mà ta quan sát được sẽ có phạm vi phân bố nằm trong đoạn [1.4; 2.5] (đơn vị: cm).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [1.8; 2.3] (đơn vị: cm).
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị chiều rộng trung bình của cánh hoa ở loài "Iris-virginica" là: (1.95; 2.10) (đơn vị: cm).
- Quan sát biểu đồ hộp, sử dụng phương pháp IQR, ta không phát hiện có giá trị ngoại lai nào. Tuy nhiên, việc 25% giá trị nhỏ nhất có phạm vi phân bố rộng hơn đáng kể so với phạm vi phân bố của 25% giá trị lớn nhất cũng phản ánh một xu hướng thú vị. Phần lớn mẫu dữ liệu thuộc loài "Iris-virginica" sẽ có độ rộng cánh hoa từ mức trung bình đến lớn, chỉ có một ít mẫu dữ liệu có chiều rộng cánh hoa ở mức trung bình - nhỏ. Tuy nhiên, ta cần phân tích trên các bộ dữ liệu lớn và phong phú hơn để có thể làm sáng tỏ xu hướng này.

11.4. So sánh phân bố chiều rộng cánh hoa ở ba loài hoa và đưa ra kết luận

Để dễ dàng so sánh phân bố chiều rộng cánh hoa của ba loài hoa, ta sẽ trực quan hóa dữ liệu trên cùng một biểu đồ mật độ và dùng màu sắc để chú thích cho phân bố của mỗi loài hoa.



Hình 11.4: Biểu đồ thể hiện phân bố chiều rộng cánh hoa ở ba loài hoa Iris.

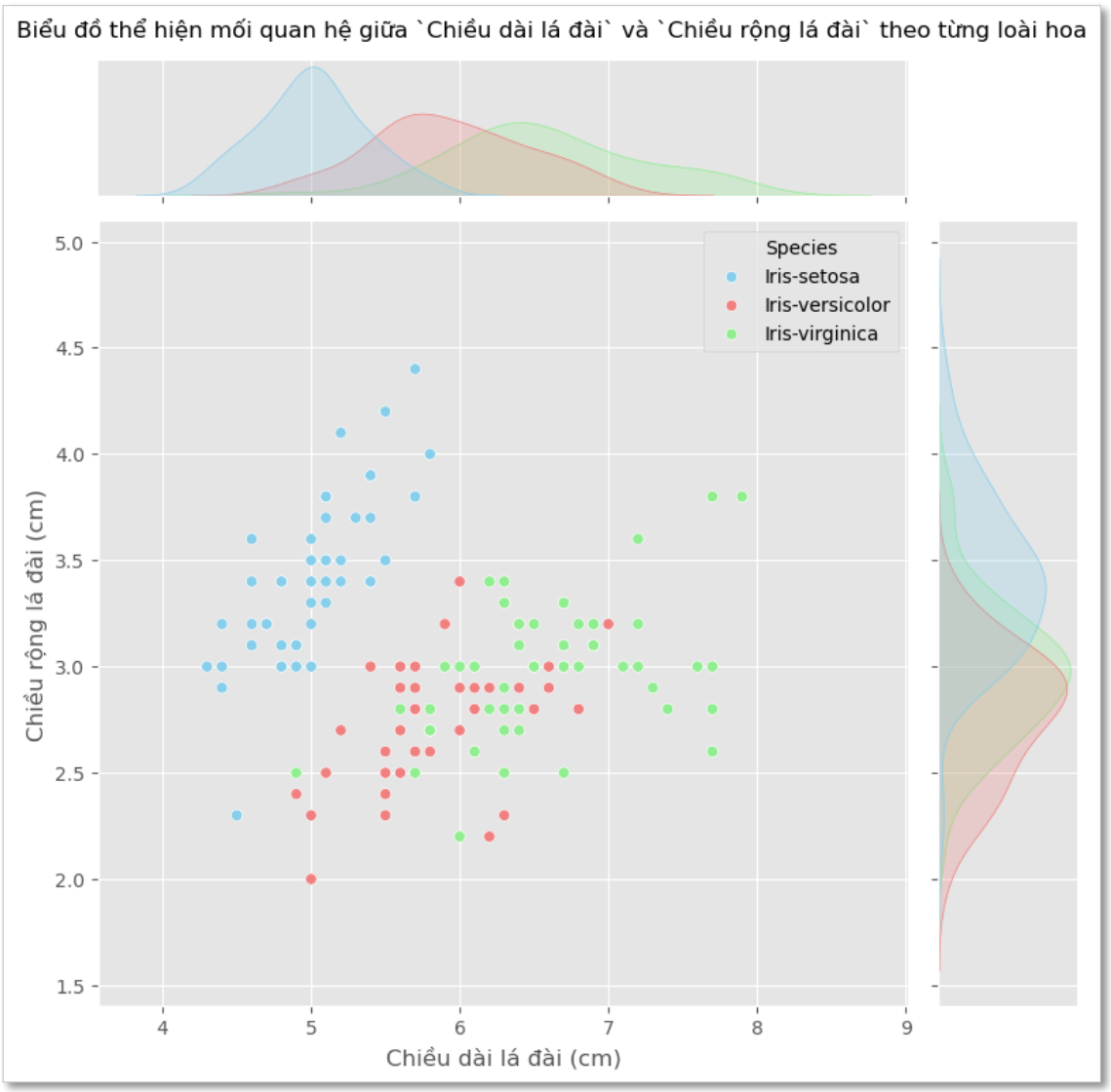
Nhận xét:

- Trong bộ dữ liệu mà ta đang phân tích, "Iris-setosa" là loài hoa diên vĩ có chiều rộng cánh hoa hẹp hơn đáng kể so với mức trung bình của các loài hoa khác. Sau đó là đến loài "Iris-versicolor", và "Iris-virginica" là loài có cánh hoa rộng nhất trong số ba loài hoa mà ta đang phân tích. Tuy nhiên, do số lượng mẫu dữ liệu mà ta đang phân tích là không quá lớn nên các quan sát ở trên có thể không phản ánh được thực tế khách quan. Do đó, ta có thể tiến hành phân tích trên bộ dữ liệu lớn và đa dạng hơn để có thể đưa ra kết luận có độ chính xác và độ tin cậy hơn.
- Tương tự như khi phân tích trên đặc trưng chiều dài cánh hoa, ở đặc trưng chiều rộng cánh hoa, ta thấy có sự tách biệt khá rõ rệt giữa các mẫu dữ liệu thuộc loài "Iris-setosa" với hai loài còn lại. Dường như chỉ với một đường thẳng là ta đã có thể xác định miền ranh giới giữa các bông hoa thuộc lớp "Iris-setosa" với các bông hoa thuộc lớp khác.
- Cũng tương tự như các phân tích trước đó, mọi thứ sẽ khó khăn hơn trong trường hợp ta muốn phân biệt rạch ròi giữa các mẫu dữ liệu thuộc hai loài "Iris-versicolor" và "Iris-virginica". Mặc dù tình trạng các phân bố chồng chéo lên nhau đã được cải thiện rất nhiều nhưng chúng vẫn không thể tách biệt một cách hoàn toàn. Do đó, với bộ dữ liệu hiện có, ta gần như phải chấp nhận việc không thể nào phân biệt rạch ròi giữa các mẫu dữ liệu thuộc hai lớp "Iris-versicolor" và "Iris-virginica".
- Như vậy, đặc trưng chiều rộng cánh hoa là manh mối tiếp theo giúp ta hoàn thành mục tiêu đã đề ra. Bằng cách kết hợp các thông số kích thước (chiều dài và chiều rộng) của cánh hoa, ta đã có trong tay một công cụ đủ mạnh mẽ để phân biệt chính xác hầu hết các bông hoa thuộc loài "Iris-setosa". Còn đối với hai lớp còn lại, ta sẽ cần thu thập thêm nhiều mẫu dữ liệu cũng như các đặc trưng nổi bật khác để có thể cải thiện hiệu suất hoạt động của mô hình phân lớp. Đây sẽ là một bài toán thú vị mà ta có thể tiếp tục giải quyết trong tương lai.

12. TIỀN XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU ĐỂ TRẢ LỜI CHO CÂU HỎI 5

Câu hỏi 5: Liệu kích thước (chiều dài và chiều rộng) lá đài có phải là yếu tố đủ để phân loại chính xác các loài hoa Iris (diên vĩ) khác nhau hay không?

Ta sẽ dùng biểu đồ phân tán để thể hiện mối quan hệ giữa các thông số kích thước của lá đài theo từng loài hoa diên vĩ khác nhau. Thông qua đó, ta sẽ có thể xác định xem liệu các loài hoa có phân tách tuyến tính với nhau hay không.



Hình 12: Biểu đồ thể hiện mối quan hệ giữa chiều dài và chiều rộng lá đài theo các loài hoa Iris khác nhau.

Sau đó, ta tạo bảng mô tả hệ số tương quan Pearson giữa chiều dài và chiều rộng lá đài ở các loài hoa diên vĩ khác nhau.

Loài hoa	Hệ số tương quan Pearson giữa chiều dài và chiều rộng lá đài
Iris-setosa	0.750000
Iris-versicolor	0.530000
Iris-virginica	0.460000

Bảng 12: Bảng mô tả hệ số tương quan Pearson giữa chiều dài và chiều rộng lá đài ở các loài hoa diên vĩ khác nhau.

Nhận xét:

- Nhìn chung, ở cả ba loài hoa, ta thấy có mối tương quan thuận khá mạnh giữa chiều dài và chiều rộng của lá đài. Khi chiều dài lá đài ở một bông hoa tăng lên thì chiều rộng lá đài của bông hoa đó cũng có xu hướng tăng lên (và ngược lại). Điều này không có gì bất thường. Trong đó, quan hệ tuyến tính giữa chiều dài và chiều rộng lá đài ở loài "Iris-setosa" có mức độ tương quan cao nhất (0.75), sau đó đến loài "Iris-versicolor" (0.53) và cuối cùng là loài "Iris-virginica" (0.46).
- Mặc dù giá trị hệ số tương quan ở loài "Iris-setosa" có lớn hơn một chút so với giá trị hệ số tương quan ở hai loài còn lại nhưng kết quả này có thể bị ảnh hưởng phần nào bởi kích thước mẫu khá nhỏ cũng như các giá trị ngoại lai mà ta đã phân tích bên trên. Do đó, ta cần thực hiện thêm phân tích trên các bộ dữ liệu lớn và đa dạng hơn để có thể đưa ra kết luận khách quan nhất.
- Quan sát kích thước lá đài của các loài hoa diên vĩ, ta thấy rằng:
  - Lá đài ở loài "Iris-setosa" thường không quá dài nhưng lại rộng hơn khá nhiều so với hai loài khác.

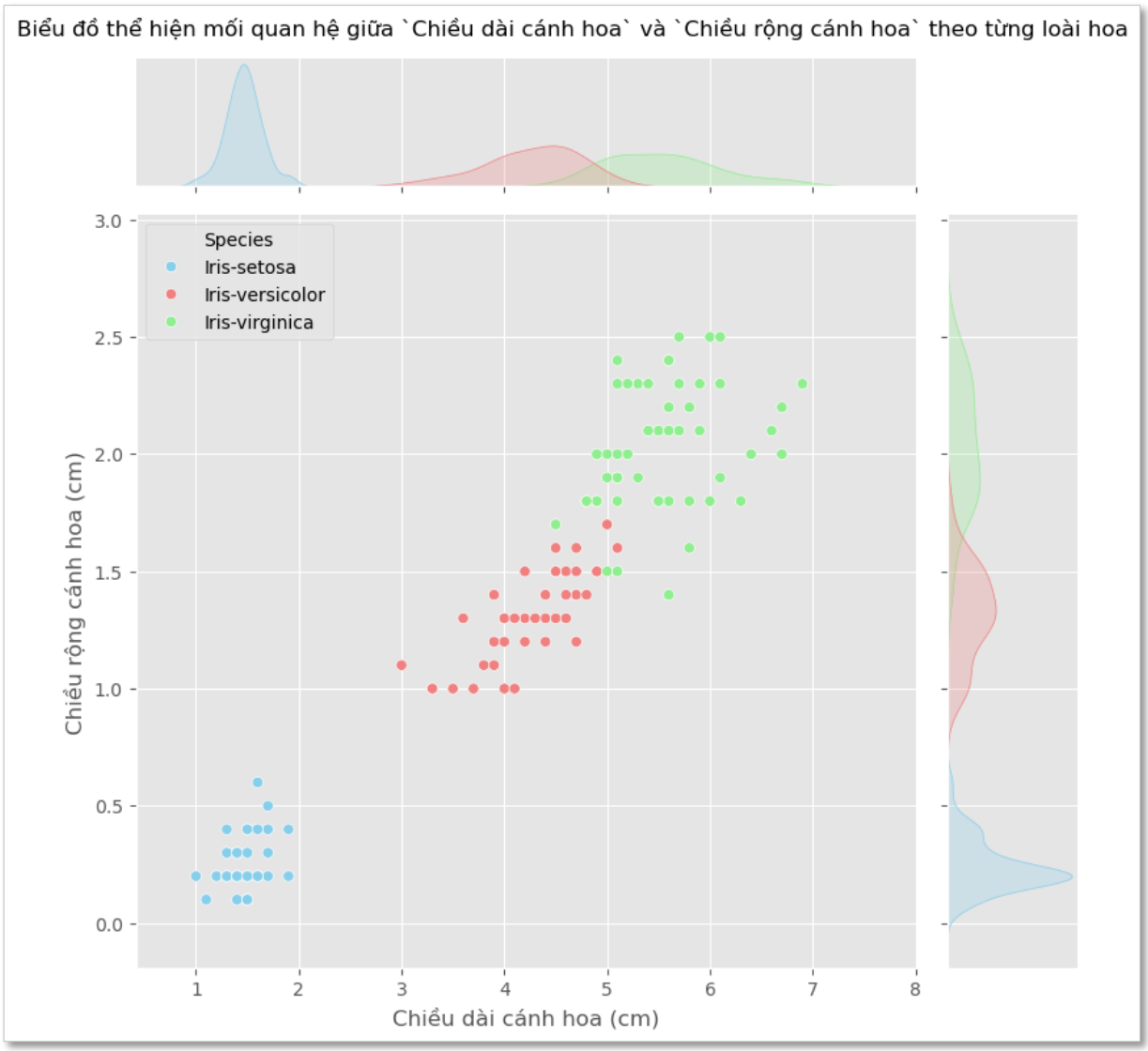
- Ngược lại, lá đài ở loài "Iris-virginica" tuy không quá rộng nhưng lại dài hơn tương đối so với hai loài khác.
- Trong khi đó, loài "Iris-versicolor" thường có kích thước lá đài ở mức trung bình so với hai loài khác.

- Thông qua biểu đồ phân tán, chỉ với các thông số liên quan đến kích thước lá đài, ta có thể dễ dàng phân biệt giữa "Iris-setosa" với hai loài hoa diên vĩ khác.
- Nhưng việc phân biệt giữa hai loài "Iris-versicolor" và "Iris-virginica" sẽ khó khăn hơn rất nhiều vì các điểm dữ liệu của hai loài này có mức độ chồng chéo đáng kể. Ta không thể tạo ra các đường phân tách tuyến tính hoặc các bộ luật "if-else" để xây dựng một mô hình đơn giản giúp phân biệt giữa hai loài hoa này.
- Như vậy, kích thước (chiều dài và chiều rộng) lá đài không phải là các đặc trưng đủ mạnh mẽ giúp ta phân loại các loài hoa một cách chính xác nhất có thể. Do đó, ta cần phân tích các thuộc tính khác của các loài hoa diên vĩ để tìm ra các đặc trưng nổi bật cho mỗi loài hoa khác nhau. Đó sẽ là chìa khóa giúp ta đạt được mục tiêu đã đề ra.

13. TIỀN XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU ĐỂ TRẢ LỜI CHO CÂU HỎI 6

Câu hỏi 6: Liệu kích thước (chiều dài và chiều rộng) cánh hoa có phải là yếu tố đủ để phân loại chính xác các loài hoa Iris (diên vĩ) khác nhau hay không?

Ta sẽ dùng biểu đồ phân tán để thể hiện mối quan hệ giữa các thông số kích thước của cánh hoa theo từng loài hoa diên vĩ khác nhau. Thông qua đó, ta sẽ có thể xác định xem liệu các loài hoa có thể được phân tách tuyến tính với nhau hay không.



Hình 13: Biểu đồ thể hiện mối quan hệ giữa chiều dài và chiều rộng cánh hoa theo các loài hoa Iris khác nhau.

Sau đó, ta tạo bảng mô tả hệ số tương quan Pearson giữa chiều dài và chiều rộng cánh hoa ở các loài hoa diên vĩ khác nhau.

Loài hoa	Hệ số tương quan Pearson giữa chiều dài và chiều rộng cánh hoa
Iris-setosa	0.310000
Iris-versicolor	0.790000
Iris-virginica	0.320000

Bảng 13: Bảng mô tả hệ số tương quan Pearson giữa chiều dài và chiều rộng cánh hoa ở các loài hoa diên vĩ khác nhau.

Nhận xét:

- Nhìn chung, ở cả ba loài hoa, ta thấy chiều dài và chiều rộng cánh hoa có mối quan hệ tuyến tính theo chiều dương (tương quan thuận). Khi chiều dài cánh hoa của một bông hoa tăng lên thì chiều rộng cánh hoa của bông hoa đó cũng có xu hướng tăng lên (và ngược lại). Điều này không có gì bất thường.
- Trong đó, quan hệ tuyến tính giữa chiều dài và chiều rộng cánh hoa ở loài "Iris-versicolor" có mức độ tương quan rất mạnh (0.79). Tuy nhiên, giá trị hệ số tương quan ở hai loài hoa còn lại chỉ đạt mức trung bình (khoảng hơn 0.30). Kết quả này có thể xuất phát từ việc bộ dữ liệu của ta có kích thước khá nhỏ, không phản ánh tốt dữ liệu trong thực tế. Do đó, ta cần thực hiện thêm phân tích trên các bộ dữ liệu lớn và đa dạng hơn để có thể đưa ra kết luận khách quan nhất.
- Quan sát kích thước cánh hoa của các loài hoa diên vĩ, ta thấy rằng:
  - Các cánh hoa của "Iris-setosa" thường có kích thước bé nhất so với hai loài còn lại (cả về chiều dài lẫn chiều rộng).
  - Ngược lại, các cánh hoa ở loài "Iris-virginica" thường dài và rộng hơn khá nhiều so với cánh hoa ở hai loài còn lại.

- Trong khi đó, loài "Iris-versicolor" thường có kích thước cánh hoa ở mức trung bình. Cánh hoa của loài "Iris-versicolor" thường sẽ dài và rộng hơn cánh hoa của loài "Iris-setosa", nhưng sẽ ngắn và hẹp hơn một chút so với cánh hoa của loài "Iris-virginica".

- Quan sát các cụm có màu sắc khác nhau trên biểu đồ phân tán, ta có thể đưa ra kết luận rằng: chỉ cần sử dụng chiều dài và chiều rộng của cánh hoa là ta đã có thể dễ dàng nhận diện hầu hết bông hoa thuộc loài "Iris-setosa" trong bộ dữ liệu.
- Tuy các mẫu dữ liệu ở hai loài "Iris-versicolor" và "Iris-virginica" không tách biệt hoàn toàn với nhau, nhưng tình trạng chồng chéo đã giảm đi rất nhiều so với khi phân tích trên kích thước của lá đài. Đây là một dấu hiệu tích cực cho thấy ta có thể sử dụng kích thước cánh hoa như là các đặc trưng để xây dựng mô hình phân lớp. Ngay cả khi không đạt độ chính xác 100%, nhưng mô hình vẫn có thể tạo ra các đường phân tách tuyến tính hoặc các bộ luật "if-else" giúp phân loại các loài hoa.
- Từ kết quả phân tích dữ liệu, ta có thể tạo ra một thuật toán đơn giản thông qua các câu lệnh "if-else" để xây dựng một bộ phân lớp từ bộ dữ liệu đang phân tích:
  - Nếu (0 <= [Chiều dài cánh hoa] < 2) và (0 <= [Chiều rộng cánh hoa] < 0.8) thì [Loài hoa] = "Iris-setosa".
  - Nếu (2.5 <= [Chiều dài cánh hoa] < 5) và (1 <= [Chiều rộng cánh hoa] < 1.7) thì [Loài hoa] = "Iris-versicolor".
  - Nếu (5 <= [Chiều dài cánh hoa] < 7) và (1.7 <= [Chiều rộng cánh hoa] < 2.6) thì [Loài hoa] = "Iris-virginica".
  - Nếu không thỏa trường hợp nào thì đưa ra thông báo: "Không nhận diện được mẫu dữ liệu thuộc về loài hoa diên vĩ nào".
- Như vậy, kích thước (chiều dài và chiều rộng) cánh hoa là một trong các đặc trưng nổi bật nhất mà ta có thể dùng để phân biệt các loài hoa diên vĩ khác nhau. Tuy vẫn chưa thể phân loại chính xác 100% các bông hoa trong tập dữ liệu, nhưng bộ phân lớp mà ta xây dựng vẫn có hiệu suất rất tốt và có thể tiếp tục chỉnh sửa, phát triển khi kích

thước bộ dữ liệu tăng lên. Trong tương lai, ta có thể thu thập thêm nhiều mẫu dữ liệu cũng như các đặc trưng khác về loài hoa diên vĩ để có thể tiếp tục phân tích, chọn ra các đặc trưng nổi bật và nâng cao hiệu suất tổng thể của mô hình.

14. KẾT LUẬN

- Sau khi phân tích các đặc trưng về kích thước của cánh hoa và lá đài trong bộ dữ liệu hoa Iris, ta thấy rằng chiều dài và chiều rộng cánh hoa là các đặc trưng hữu ích giúp phân loại các loài hoa Iris một cách khá chính xác. Trong khi đó, kích thước lá đài không thực sự là một đặc trưng đủ tốt để giải quyết bài toán phân lớp các loài hoa Iris do mức độ chồng chéo dữ liệu ở các thuộc tính này diễn ra khá phức tạp.
- Bằng cách kết hợp chiều dài và chiều rộng cánh hoa, ta có thể phân loại chính xác hầu hết mẫu dữ liệu thuộc loài "Iris-setosa". Riêng với hai loài "Iris-versicolor" và "Iris-virginica", do phân bố kích thước cánh hoa ở hai loài này không tách biệt hoàn toàn với nhau nên ta không thể phân biệt chính xác hai loài này bằng một mô hình đơn giản. Mặc dù ta có thể xây dựng một mô hình phức tạp để đạt độ chính xác tuyệt đối trên bộ dữ liệu đang phân tích nhưng mô hình đó sẽ có không có tính tổng quát hóa và không phải là mục tiêu mà ta muốn đạt được.
- Như vậy, để tiếp tục cải thiện hiệu suất của mô hình, ta có thể thu thập thêm dữ liệu mới cũng như các đặc trưng khác ở hoa Iris. Sau đó, ta thực hiện quy trình phân tích dữ liệu để chọn ra các đặc trưng nổi bật giúp phân biệt các loài hoa khác nhau. Khi này, thay vì xây dựng một mô hình phức tạp chỉ đơn thuần "học thuộc lòng" trên dữ liệu huấn luyện, ta có thể sử dụng một mô hình đơn giản hơn với khả năng giải thích và tổng quát hóa cao, đảm bảo hiệu suất hoạt động tốt trên cả dữ liệu huấn luyện và dữ liệu mới trong tương lai.

15. TÀI LIỆU THAM KHẢO

[1]: Basic visualization techniques - kaggle.com/upadorprofzs.

[2]: Exploratory Data Analysis on Iris Dataset - geeksforgeeks.org.

[3]: Exploratory Data Analysis on IRIS Dataset - github.com/abhikumar22.

[4]: Compute a confidence interval from sample data - stackoverflow.com.