TRỰC QUAN HOÁ TẬP DỮ LIỆU LOÀI HOA IRIS

Nguyễn Hoài Nam

Ngày 2 tháng 11 năm 2023

Tóm tắt nội dung

Dự án này tập trung vào việc trực quan hoá dữ liệu liên quan đến loài hoa Iris bằng cách sử dụng các biểu đồ và công cụ trực quan hóa dữ liệu. Loài hoa Iris là một loài thực vật phổ biến trong nghiên cứu học máy và phân loại dữ liệu. Dự án này sẽ tạo ra một loạt biểu đồ như biểu đồ đường, biểu đồ cột và biểu đồ điểm để hiểu rõ hơn về đặc điểm và sự đa dạng của các loài hoa Iris.

Dự án sẽ sử dụng các thư viện trực quan hóa dữ liệu như Matplotlib, Seaborn để tạo ra các biểu đồ đẹp mắt và dễ đọc. Kết quả của dự án này sẽ giúp người dùng hiểu rõ hơn về các đặc điểm của các loài hoa Iris và cách chúng có thể được phân loại dựa trên các đặc điểm này.

1 Giới thiệu

Tập dữ liệu loài hoa Iris là một trong những tập dữ liệu phổ biến và kinh điển trong lĩnh vực học máy và thống kê. Nó đã được giới thiệu bởi nhà thống kê và nhà sinh học người Anh Ronald A. Fisher vào năm 1936, và từ đó đã trở thành một phần quan trọng của nghiên cứu và giảng dạy về phân loại dữ liệu.

Tập dữ liệu Iris chứa thông tin về ba loài hoa Iris khác nhau: Iris setosa, Iris versicolor và Iris virginica. Mỗi loài có 50 mẫu, tổng cộng tập dữ liệu chứa 150 mẫu. Đối với mỗi mẫu, có bốn đặc điểm đo lường: chiều dài và chiều rộng của lá đài và cánh hoa (sepal length, sepal width, petal length, petal width).

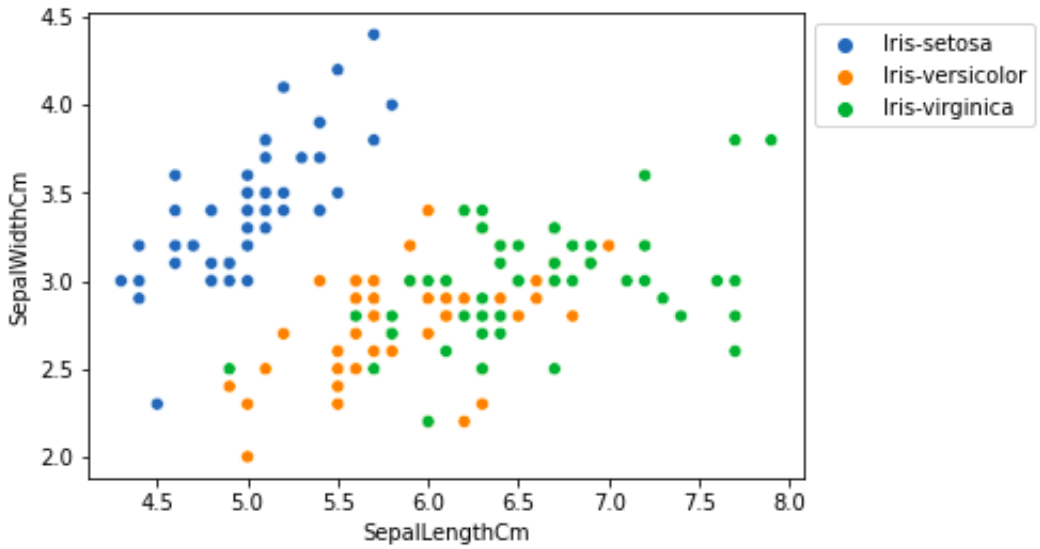
Tập dữ liệu Iris thường được sử dụng để minh họa các khái niệm cơ bản trong học máy và thống kê, chẳng hạn như phân loại dữ liệu, phân tích phân phối, phân tích tương quan và trực quan hóa dữ liệu. Nó cũng là một tập dữ liệu thử nghiệm phổ biến để kiểm tra hiệu suất của các mô hình học máy và thuật toán phân loại.

2 Phương pháp

Để thu thập được tập dữ liệu loài hoa Iris, dự án sử dụng thư viện Scikit-learn.

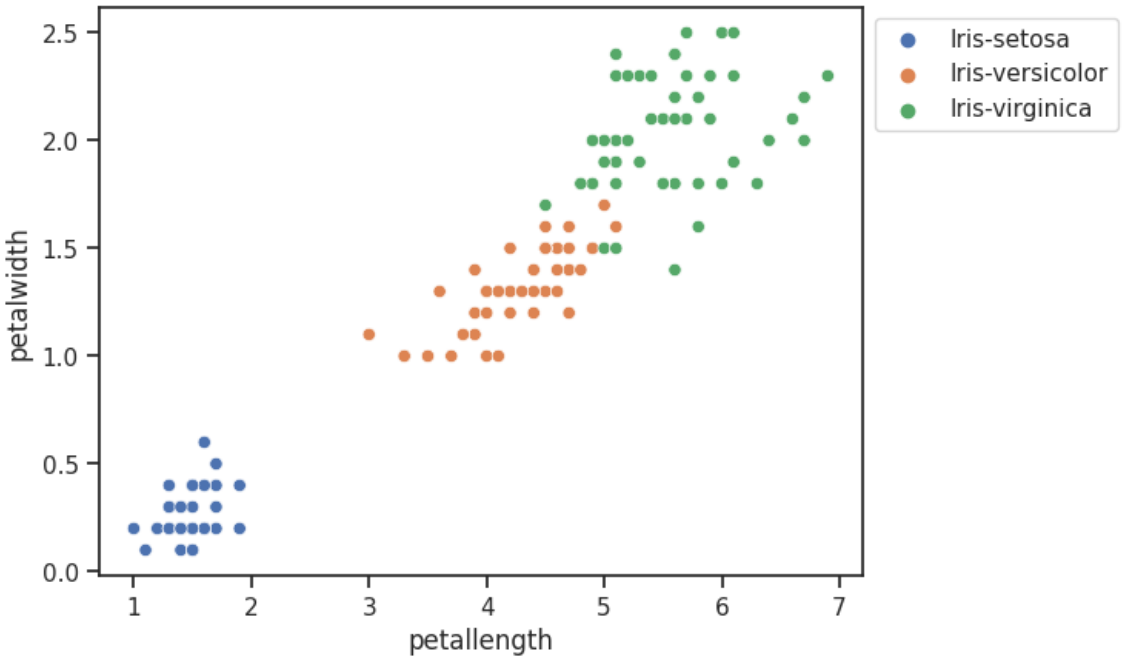
Để trực quan hoá tập dữ liệu loài hoa Iris, dự án sử dụng các thư viện trực quan hóa dữ liệu như Matplotlib, Seaborn.

3 Kết quả

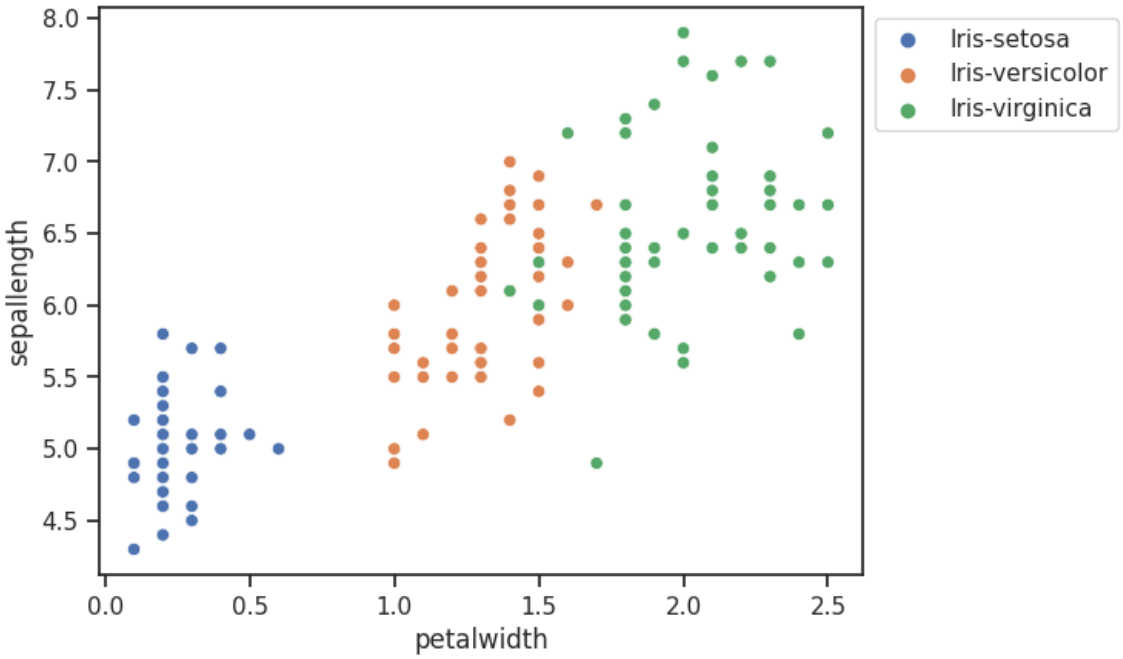


Hình 1: So sánh giữa 2 đặc trưng Sepal Length and Sepal Width

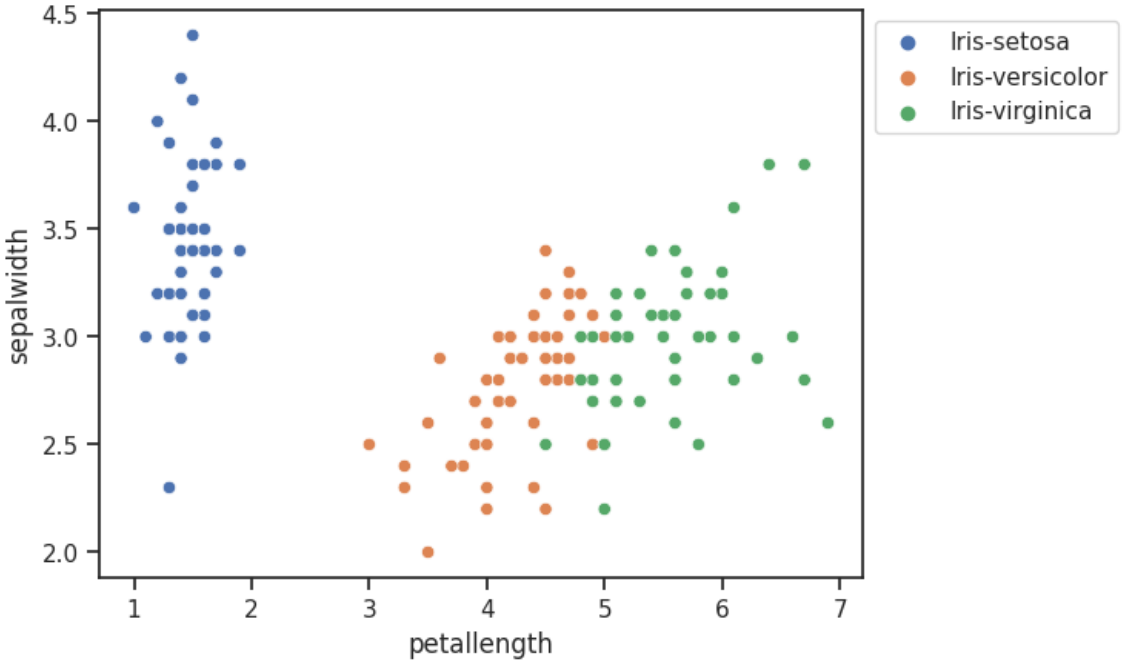
Từ đồ thị trên, ta có thể suy ra rằng: Loài Setosa có chiều dài đài hoa nhỏ hơn nhưng chiều rộng đài hoa lớn hơn; Loài Versicolor nằm ở giữa hai loài còn lại về chiều dài và chiều rộng của đài hoa; Loài Virginica có lá đài dài hơn nhưng chiều rộng lá đài nhỏ hơn.



Hình 2: So sánh giữa 2 đặc trưng Pepal length and Pepal Width



Hình 3: So sánh giữa 2 đặc trưng Sepal Length and Petal Width

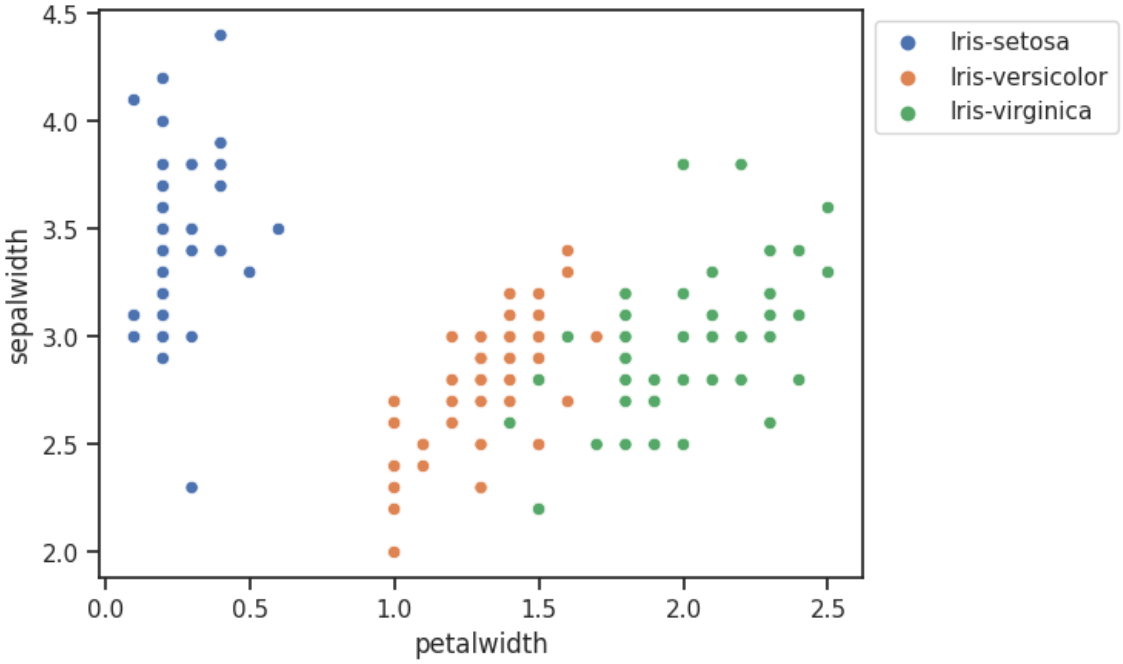


Hình 4: So sánh giữa 2 đặc trưng Sepal Width and Petal Length

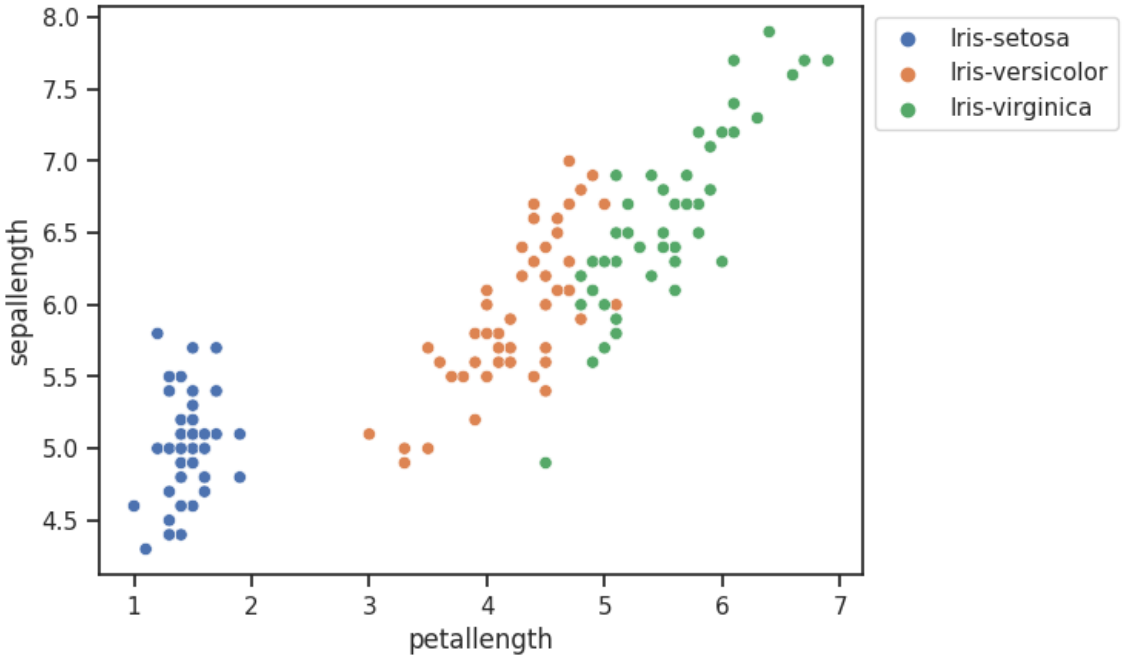
4 Thảo luận

4.1 Hạn chế

Mặc dù tập dữ liệu cung cấp sự phân biệt trực quan rõ ràng, cần lưu ý rằng nó đại diện cho một tình huống đơn giản hóa. Các tập dữ liệu thực tế có thể không cho thấy những mẫu rõ ràng như vậy và có thể cần các kỹ thuật phân tích tiên tiến hơn.



Hình 5: So sánh giữa 2 đặc trưng Sepal Width and Petal Width



Hình 6: So sánh giữa 2 đặc trưng Sepal Length and Pepal Length

4.2 Công việc trong tương lai

Dựa trên những phát hiện từ việc trực quan hoá tập dữ liệu, trong tương lai dự án có thể tập trung vào việc phát triển các mô hình phân loại dựa trên các đặc trưng đã nêu. Hơn nữa, việc tìm hiểu sâu hơn về mối quan hệ giữa các đặc trưng và việc áp dụng các kỹ thuật xử lý dữ liệu tiên tiến hơn có thể giúp cải thiện hiệu suất phân loại.

Bạn có muốn xem lại nội dung ở phần Sec. 1?