|  |
| --- |
| **BỘ LAO ĐỘNG THƯƠNG BINH VÀ XÃ HỘI** |
| **TRƯỜNG ĐẠI HỌC LAO ĐỘNG XÃ HỘI**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**🙞🕮🙜**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |



**BÁO CÁO BÀI TẬP LỚN**

**MÔN KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI : ỨNG DỤNG KỸ THUẬT PHÂN LỚP**

**ĐỂ PHÂN LOẠI CÁC LOÀI HOA .**

|  |  |
| --- | --- |
| **Họ và tên**  **Mã sinh viên** | :Đỗ Thùy Trang  :1116090044 |
| **Lớp** | :D16CN02 |
| **Giảng viên Hướng dẫn** | :Nguyễn Thị Vàn |

Hà nội tháng 05 năm 2023.

**LỜI MỞ ĐẦU**

Trong thời buổi 4.0 hiện đại ngày nay, công nghệ thông tin cũng như những ứng dụng của nó không ngừng phát triển, lượng thông tin và cơ sở dữ liệu được thu thập và lưu trữ cũng tích lũy ngày một nhiều lên. Con người cũng vì thế mà cần có thông tin với tốc độ nhanh nhất để đưa ra quyết định dựa trên lượng dữ liệu khổng lồ đã có. Các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế, vì thế, một khuynh hướng kỹ thuật mới là Kỹ thuật phát hiện tri thức và khai phá dữ liệu nhanh chóng được phát triển.

Khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau ở các nước trên thế giới. Ở Việt Nam, kỹ thuật này đang được nghiên cứu và dần đưa vào ứng dụng. Khai phá dữ liệu là một bước trong quy trình phất hiện tri thức. Hiện nay, mọi người không ngừng tìm tòi các kỹ thuật để thực hiện khai phá dữ liệu một cách nhanh nhất và có được kết quả tốt nhất.

Trong bài tập lớn này, chúng em tìm hiểu và trình bày về một kỹ thuật trong khai phá dữ liệu để phân lớp dữ liệu cũng như tổng quan về khai phá dữ liệu, với đề tài “Ứng dụng kỹ thuật phân lớp để phân loại các loại hoa ”.

Trong quá trình làm bài tập lớn này, em xin gửi lời cảm ơn đến cô Nguyễn Thị Vàn cô đã rất tận tình hướng dẫn chi tiết cho em, những kiến thức cô cung cấp rất hữu ích. Trong quá trình ứng dụng không tránh khỏi những sai sót mong cô góp ý để em có thể hoàn thiện bài báo cáo một cách tốt hơn.

***Em xin chân thành cảm ơn !***

**Sinh viên thực hiện**

Đỗ Thùy Trang

**MỤC LỤC**

**Lời mở đầu 2**

**Chương 1 : Tổng quan Khai phá dữ liệu 5**

1.1 Khai phá dữ liệu là gì? 5

1.2 Khai phá tri thức từ CSDL 5

1.3 Các bước của quá trình KPDL 6

1.4 Ứng dụng của KPDL 7

1.5 Một số kỹ thuật khai phá dữ liệu cơ bản 7

* Kỹ thuật Khai phá luật kết hợp (Associantion Rules) 8
* Kỹ thuật phân lớp 8
* Kỹ thuật phân cụm 8
* Hồi quy 8
* Tổng hợp hóa 8
* Khai thác mẫu tuần tự 9

**Chương 2 : Các kỹ thuật khai phá dữ liệu 10**

2.1 Phân lớp dữ liệu 10

2.2 Các phương pháp phân lớp dữ liệu 12

2.2.1 Kỹ thuật K-NN 12

2.2.2 Kỹ thuật Neural Network 13

2.3 So sánh giữa khai phá dữ liệu và các phương pháp khác. 14

* So sánh với phương pháp học máy 14

**Chương 3 : Demo ứng dụng 16**

3.1 Giới thiệu bài toán 16

* Giới thiệu tập dữ liệu các loài hoa 16
* Sử dụng tập dữ liệu 16
* Định nghĩa bài toán phân loại hoa 16

3.2 Giới thiệu Google Colab 17

* Google Colab là gì ? 17
* Cách sử dụng Colab 17

3.3 Áp dụng thuật toán vào phân loại hoa 18

- Áp dụng thuật toán Neural Network trong phân lớp 18

**KẾT LUẬN 22**

**TÀI LIỆU THAM KHẢO 23**

**CHƯƠNG 1 : TỔNG QUAN KHAI PHÁ DỮ LIỆU**

Trong chương này , trình bày khái niệm của khai phá dữ liệu , các bước của quá trình khai khá và ứng dụng của nó , một số kỹ thuật khai phá dữ liệu cơ bản .

* 1. **Khai phá dữ liệu**

KPDL là một khái niệm ra đời vào những năm cuối của thập kỷ 80. Nó bao hàm một loạt các kỹ thuật nhằm phát hiện ra các thông tin có giá trị tiềm ẩn trong các tập dữ liệu lớn (các kho dữ liệu). Về bản chất, KPDL liên quan đến việc phân tích các dữ liệu và sử dụng các kỹ thuật để tìm ra các mẫu hình có tính chính quy (regularities) trong tập dữ liệu .

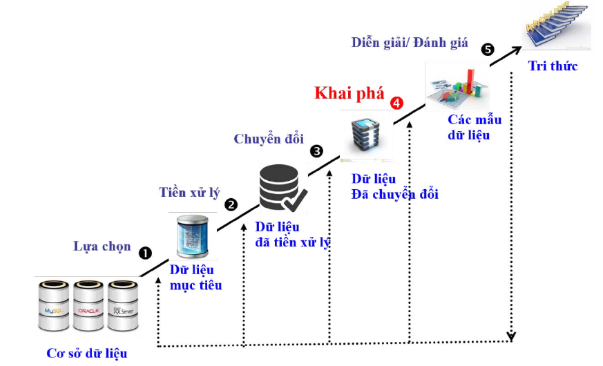
Năm 1989, Fayyad, Piatestsky-Shapiro và Smyth đã dùng khái niệm Phát hiện tri thức trong cơ sở dữ liệu (Knowledge Discovery in Database –KDD) để chỉ toàn bộ quá trình phát hiện các tri thức có ích từ các tập dữ liệu lớn. Trong đó, KPDL là một bước đặc biệt trong toàn bộ tiến trình, sử dụng các giải thuật đặc biệt để chiết xuất ra các mẫu (pattern) (hay các mô hình) từ dữ liệu.

Khai phá dữ liệu là một tiến trình sử dụng các công cụ phân tích dữ liệu khác nhau để khám phá ra các mẫu dưới nhiều góc độ khác nhau nhằm phát hiện ra các mối quan hệ giữa các dữ kiện, đối tượng bên trong CSDL, kết quả của việc khai phá là xác định các mẫu hay các mô hình đang tồn tại bên trong, nhưng chúng nằm ẩn khuất ở các CSDL. Để từ đó rút trích ra được các mẫu, các mô hình hay các thông tin và tri thức từ các CSDL.

* Còn có các tên gọi khác:
* Khám phá tri thức trong các cơ sở dữ liệu ( Knowledge discovery in database KDD)
* Trích rút tri thức ( knowledge extraction)
* Phân tích mẫu / dữ liệu ( data/pattern analysis)
  1. **Khai phá tri thức từ CSDL ( Knowledge Discovery in Database)**

Khai phá tri thức từ CSDL gồm 5 bước :

* B1: Lựa chọn CSDL
* B2: Tiền xử lý
* B3: Chuyển đổi
* B4: Khai phá dữ liệu
* B5: Diễn giải và đánh giá
* Khai phá dữ liệu là 1 bước trong quá trình khai phá tri thức từ CSDL



Hình 1.1 : Quá trình khai phá tri thức

* Chuẩn bị dữ liệu , bao gồm các quá trình làm sạch dữ liệu (data cleaning) , tích hợp dữ liệu (data integration) , chọn dữ liệu ( data selection) , biến đổi dữ liệu ( data transformation).
* Khai thác dữ liệu ( data mining) xác định nhiệm vụ khai thác dữ liệu và lựa chọn kỹ thuật khai thác dữ liẹu . Kết quả cho ta một nguồn tri thức thô.
* Đánh giá dựa trên một số tiêu chí tiến hành kiểm tra và lọc nguồn tri thức thu được

🡺 Quá trình khai thác tri thức không chỉ là một quá trình tuần tự từ bước đầu đến bước cuối mà là một quá trình lặp và có trở lại các bước đã qua .

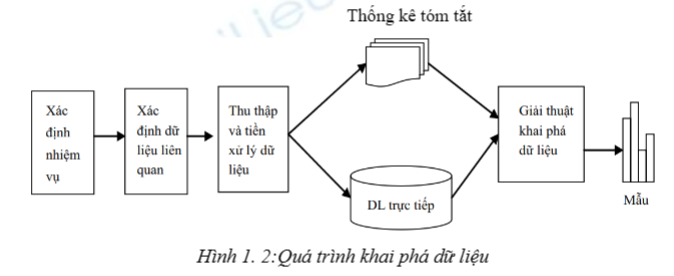
* 1. **Các bước của quá trình Khai phá dữ liệu**

Khai phá dữ liệu là một giai đoạn quan trọng trong quá trình khám phá tri thức.Về bản chất là giai đoạn duy nhất tìm ra được thông tin mới, thông tin tiềm ẩn có trong cơ sở dữ liệu chủ yếu phục vụ cho mô tả và dự đoán.

Mô tả dữ liệu là tổng kết hoặc diễn tả những đặc điểm chung của những thuộc tính dữ liệu trong kho dữl iệu mà con người có thể hiểu được.

Dự đoán là dựa trên những dữ liệu hiện thời để dự đoán những quy luật được phát hiện từ các mối liên hệ giữa các thuộc tính của dữ liệu trên cơ sở đó chiết xuất ra các mẫu, dự đoán được những giá trị chưa biết hoặc những giá trị tương lai của các biến quan tâm.

* Xác định nhiệm vụ: Xác định chính xác các vấn đề cần giải quyết
* Xác định các dữ liệu liên quan: Dùng để xây dựng giải pháp.
* Thu thập và tiền xử lý dữ liệu: Thu thập các dữ liệu liên quan và tiền xử lý chúng thành dạng sao cho thuật toán khai phá dữ liệu có thể hiểu được. Đây là một quá trình rất khó khăn, có thể gặp phải rất nhiều các vướng mắc như: dữ liệu phải được sao ra nhiều bản (nếu được chiết xuất vào các tệp), quản lý tập các dữ liệu, phải lặp đi lặp lại nhiều lần toàn bộ quá trình (nếu mô hình dữ liệu thay đổi)….
* Thuật toán khai phá dữ liệu: Lựa chọn thuật toán khai phá dữ liệu và thực hiện việc khai phá dữ liệu để tìm được các mẫu có ý nghĩa, các mẫu này được biểu diễn dưới dạng luật kết hợp, cây quyết định... tương ứng với ý nghĩa của nó.



* 1. **Ứng dụng của Khai phá dữ liệu .**

Phát hiện tri thức và khai phá dữ liệu liên quan đến nhiều ngành, nhiều lĩnh vực : thống kê, trí tuệ nhân tạo, cơ sở dữ liệu, thuật toán, tính toán song song và tốc độ cao, thu thập tri thức cho các hệ chuyên gia, quan sát dữ liệu.

* Kinh tế - ứng dụng trong kinh doanh , tài chính , tiếp thị bán hàng , bảo hiểm , thương mại , ngân hàng … Đưa ra các bản báo cáo phân tích rủi ro trước khi đưa ra các chiến lược kinh doanh , sản xuất , phân loại khách hàng từ đó phân định thị trường , thị phần…..
* Khoa học : Thiên văn học – dự doán các đường đi các thiên thể , hành tinh…
* Công nghệ sinh học – tìm ra các gen cây mới , cây con giống mới …
* Web : các công cụ tìm kiếm , tóm tắt văn bản …
* Mạng viễn thông – phân tích cuộc gọi và hệ thống giám sát lỗi , sự cố..
  1. **Một số kỹ thuật Khai phá dữ liệu cơ bản**
* Kỹ thuật khai phá luật kết hợp
* Trong khai phá dữ liệu, mục đích của luật kết hợp là tìm ra các mối quan hệ giữa các đối tượng trong khối lượng lớn dữ liệu.
* Để khai phá luật kết hợp có rất nhiều thuật toán, nhưng dùng phổ biến nhất là thuật toán Apriori. Đây là thuật toán khai phá tập phổ biến trong dữ liệu giao dịch để phát hiện các luật kết hợp dạng khẳng định nhị phân và được sử dụng để xác định, tìm ra các luật kết hợp trong dữ liệu giao dịch.

Ngoài ra, còn có các thuật toán FP-growth, thuật toán Partition,…

* Kỹ thuật phân lớp

Trong kỹ thuật phân lớp gồm có các thuật toán:

* Phân lớp bằng cây quyết định (giải thuật ID3, J48): phân lớp dữ liệu dựa trên việc lập nên cây quyết định, nhìn vào cây quyết định có thể ra quyết định dữ liệu thuộc phân lớp nào.
* Phân lớp dựa trên xác suất (Naïve Bayesian): dựa trên việc giả định các thuộc tính độc lập mạnh với nhau qua việc sử dụng định lý Bayes.
* Phân lớp dựa trên khoảng cách (giải thuật K – láng giềng): làm như láng giềng làm, dữ liệu sẽ được phân vào lớp của k đối tượng gần với dữ liệu đó nhất.
* Phân lớp bằng SVM: phân lớp dữ liệu dựa trên việc tìm ra một siêu phẳng “tốt nhất” để tách các lớp dữ liệu trên không gian nhiều chiều hơn.
* Kỹ thuật phân cụm

Phân cụm dữ liệu là cách phân bố các đối tượng dữ liệu vào các nhóm/ cụm sao cho các đối tượng trong một cụm thì giống nhau hơn các phần tử khác cụm, gồm có một số phương pháp phân cụm cơ bản như:

* Phân cụm bằng phương pháp K-mean: tìm ra tâm của các cụm mà khoảng cách của tâm đó đến các đối tượng, dữ liệu khác là ngắn.
* Phân cụm trên đồ thị

Ngoài ra, khai phá dữ liệu có rất nhiều kỹ thuật, nhưng đây là những kỹ thuật cơ bản và đơn giản trong khai phá dữ liệu mà chúng em được tìm hiểu

* Kỹ thuật Hồi quy

Phân tích hồi quy là một kỹ thuật dự báo trong thống kê, được sử dụng để xác định mối quan hệ giữa các biến độc lập và phụ thuộc. Kỹ thuật này được sử dụng để dự báo, mô hình chuỗi thời gian và tìm hiểu mối quan hệ nguyên nhân giữa các biến.

* Kỹ thuật tổng hợp hóa

Là quá trình tạo ra dữ liệu tổng hợp hóa không được tạo ra bởi các sự kiện trong thế giới thực mà được tạo ra bằng các thuật toán hoặc mô phỏng. Nó chủ yếu được sử dụng để xác nhận các mô hình toán học và đào tạo dữ liệu tổng hợp hóa cho các mô hình học sâu. Việc tạo ra dữ liệu tổng hợp hóa thường được thực hiện khi dữ liệu thực không có sẵn hoặc phải được giữ riêng tư vì thông tin nhận dạng cá nhân (PII) hoặc rủi ro tuân thủ. Nó được sử dụng rộng rãi trong các lĩnh vực y tế, sản xuất, nông nghiệp và thương mại điện tử.

* Kỹ thuật khai thác mẫu tuần tự

Là một kỹ thuật khai thác dữ liệu chuyên biệt để phân tích dữ liệu tuần tự, để khám phá các mẫu tuần tự. Cụ thể hơn, nó bao gồm khám phá các chuỗi con thú vị trong một tập hợp các chuỗi, trong đó sự hấp dẫn của một chuỗi con có thể được đo bằng các tiêu chí khác nhau như tần suất xuất hiện, độ dài và hỗ trợ. Khai thác mẫu tuần tự có nhiều ứng dụng đa dạng như phân tích hành vi mua hàng của khách hàng, các mô hình sử dụng web và chẩn đoán y tế.

**CHƯƠNG 2 : CÁC KỸ THUẬT KHAI PHÁ DỮ LIỆU**

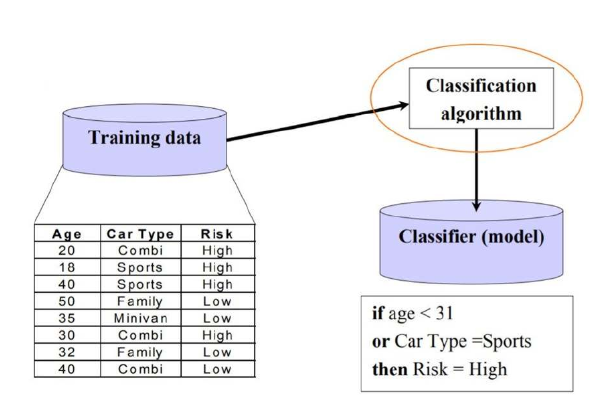
**2.1 Phân lớp dữ liệu**

**a. Khái niệm về phân lớp dữ liệu**

Phân lớp dữ liệu chính là một trong những hướng nghiên cứu chính của khai phá dữ liệu. Phân lớp dữ liệu có thể coi là quá trình học một mô hình mà mô hình này dùng để mô tả các lớp dữ liệu khác nhau. Các lớp dữ liệu đây đã được xác định trước. Khác với phân cụm dữ liệu, phân lớp dữ liệu được xem là một quá trình “học có giám sát”. Sau khi được xây dựng, mô hình phân lớp có thể được sử dụng để phân lớp các dữ liệu mới. Phân lớp và dự đoán là một trong hai dạng của phân tích dữ liệu nhằm rút ra một mô hình mô tả các lớp dữ liệu quan trọng hoặc dự đoán xu hướng của dữ liệu trong tương lai .

**b.Quá trình phân lớp dữ liệu**

* Bước 1 : Xây dựng mô hình (hay giai đoạn “học” hoặc “huấn luyện”) nhằm xây dựng một mô hình mô tả một tập dữ liệu sẵn có. Đầu vào của quá trình này là một tập dữ liệu mẫu đã được gán nhãn và tiền xử lý, mỗi phần tử dữ liệu được giản định thuộc về một lớp phân trước, lớp ở đây là giá trị của một thuộc tính được chọn làm thuộc tính gán nhãn hay thuộc tính phân lớp. Sau đó, sử dụng các quy tắc phân lớp dưới dạng if – then, cây quyết định (Decision tree), hồi quy logistic (Regression Logistic), Mạng lưới Neural (Neural Network).

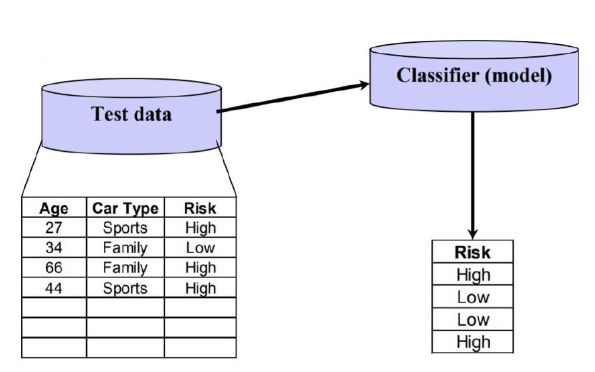


Hình 2.1: Bước 1 của quá trình trình phân lớp: Xây dựng mô hình phân lớp

* Bước 2 : Sử dụng mô hình, được chia thành 2 bước nhỏ

+ Bước 2.1 Đánh giáumô hình (kiểm traytính đúngtđắn củaimô hình)

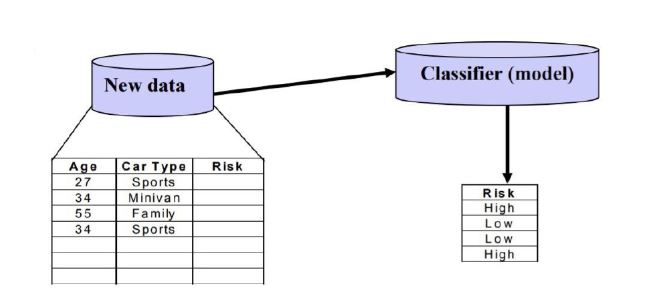
Đầubvào là một tập dữ liệu mẫu đượcbchọn ngẫu nhiêngvà độc lập với cácbmẫu trong tập dữ liệu huấn luyện, cũng đã được gán nhãn và tiền xử lý. Tuy nhiên thuộc tính đãbđược gán nhãndnày bị “lờ” khi đượcbđưa vào mô hình tiếndhànhgphân lớp. Bằng cáchbso sánh thuộcbtính gán nhãn của dữ liệu đầubvào và kết quảhphân lớp từ mô hình, tagdễ dàng xác địnhsđược tính đúngfđắn của mô hình. Holdout làymột kỹ thuật đơn giản để ước lượng tínhtđúng đắn nàyfdựa trên tỷ lệ phầnutrăm các các mẫu trong tập dữ liệu dự báo được mô hình phân lớp đúng (so với thực tế). Kết quả của bướcunày là mô hình sẽ được sử dụng để phânylớp những dữ liệutcần thiếtutrong tương lai, hoặcynhững dữ liệu mà giá trịrcủa thuộc tínhephân lớp là chưatbiết nếu nếu độuchính xác củaimô hình là có thể chấp nhận.



Hình 2.2: Bước 2.1 của quá trình trình phân lớp: Đánh giá mô hình

+ Bước 2.2 : Phân lớp dữ liệu mới

Ở bước này, dữ liệu đầu vào là dữ liệu “khuyết” thuộc tính cần dự đoán lớp (nhãn). Mô hình sẽ tự động phân lớp (gán nhãn) cho các đối tượng dữ liệu này dựa vào những gì được huấn luyện ở bước 1



Hình 2.3: Bước 2.2 của quá trình trình phân lớp: phân lớp dữ liệu mới

**2.2 Các phương pháp phân lớp dữ liệu**

Có rất nhiều phương pháp phân lớp dữ liệu được sử dụng dùng để phân loại các loài hoa có thể kể đến như Neural Network, k-NN(k-Nearest Neighbors)… Các phương pháp sẽ được liệt kê rõ hơn ở bên dưới

**2.2.1 Kỹ thuật K-NN.**

**a. Mô tả thuật toán K-NN**

Các mẫu được mô tả bằng n – chiều thuộc tính số. Mỗi mẫu đại diện cho

một điểm trong một chiều không gian n – chiều. Theo cách này tất cả các mẫu được lưu trữ trong một mô hình không gian n – chiều.

Các bước thực hiện của Thuật toán K-NN được mô tả như sau:

•Xác định giá trị tham số K ( số láng giềng gần nhất).

•Tính khoảng cách giữa đối tượng cần phân lớp (Query Point) với tất

cả các đối tượng trong các mẫu có sẵn (Trainning Data) ( Thường sử dụng khoảng cách Euclidean).

•Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng

gần nhất với Query Point.

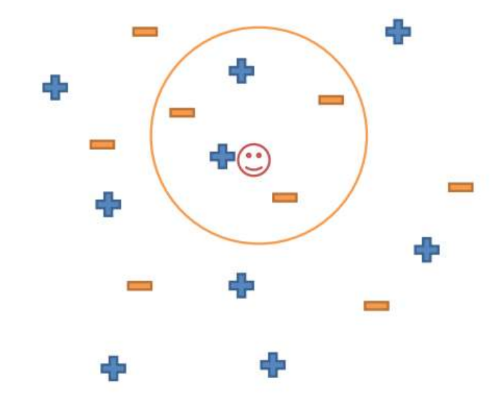
•Lấy tất cả các lớp của K láng giềng gần nhất đã xác định.

•Dựa vào phần lớn lớp của láng giềng gần nhất để xác định lớp cho

Query Ponit.

1. **Minh họa về K -NN**

Trong hình dưới đây, Trainning Data được mô tả bằng dấu (+) và dấu (-), đối tượng cần được xác định lớp cho nó (Query Point) là hình mặt cười đỏ. Nhiệm vụ của ta là ước lượng lớp của Query Point dựa vào việc lựa chọn số láng giềng gần nhất với nó. Nói cách khác ta muốn biết liệu Query Point sẽ được phân vào lớp (+) hay lớp (-)



Hình 2.4 Minh họa về K\_NN

* Ta thấy rằng:

Có 5 Nearest Neightbor: Kết quả là (–) :Query Point được xếp vào lớp dấu (–) vìtrong 5 láng giềng gần nhất với nó thì có 3 đối tượng thuộc lớp (–) nhiều hơn lớp(+) chỉ có 2 đối tượng.)

1. **Đánh giá ưu, nhược điểm của thuật toán**

* Ưu điểm :

+ Tư tưởng đơn giản, thích hợp với hệ thống nhỏ

+ Dễ hiểu, dễ cài đặt

* Nhược điểm :

Giải thuật K-NN thích hợp cho việc phân loại dữ liệu chứ giải thuật này không có khả năng phân tích dữ liệu để tìm ra các thông tin có giá trị. Trong quá trình K-NN hoạt động, nó phải tính toán "khoảng cách" từ dữ liệu cần xác định loại đến tất cảcác dữ liệu trong tập huấn luyện (training set)

=> Nếu tập huấn luyện quá lớn, điềuđó sẽ làm cho thời gian chạy của chương trình sẽ rất lâu

**2.2.2 Kỹ thuật Neural Network**

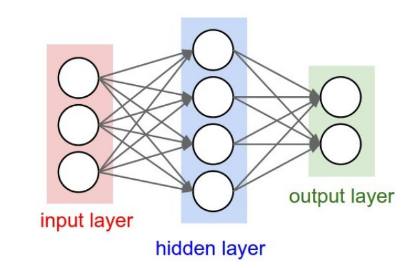
**a. Mô hình khai phá dữ liệu bằng Neural Network**

Neural Network đọc tiếng việt là Mạng nơ-ron nhân tạo, đây là một chuỗi những thuật toán được đưa ra để tìm kiếm các mối quan hệ cơ bản trong tập hợp các dữ liệu. Thông qua việc bắt bước cách thức hoạt động từ não bộ con người. Nói cách khác, mạng nơ ron nhân tạo được xem là hệ thống của các tế bào thần kinh nhân tạo. Đây thường có thể là hữu cơ hoặc nhân tạo về bản chất.

Neural Network chứa các lớp bao hàm các nút được liên kết lại với nhau. Mỗi nút có một cấu trúc tương thích với hàm hồi quy tuyến tính đa biến. Chúng sẽ được sắp xếp với các lớp liên kết với nhau. Lớp đầu vào sẽ thu nhập các dữ liệu đầu vào và các lớp đầu ra sẽ nhận các phân loại hoặc tín hiệu đầu ra mà các mẫu đầu vào có thể phản ánh lại.

Kiến trúc của mạng Neural Network là có sự kết hợp giữa tầng perceptron hay còn gọi là perceptron đa tầng. Mỗi mạng Neural Network sẽ gồm ba kiểu tầng như sau:

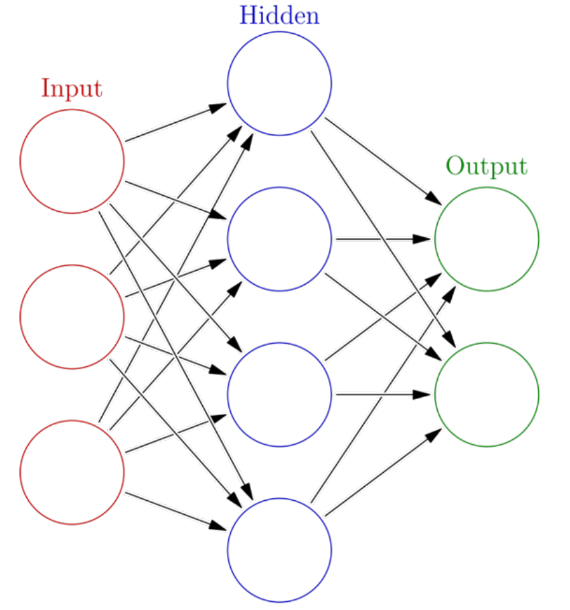
* “Tầng input layer (tầng vào): tầng này nằm phía bên trái của mạng, thể hiệncho các đầu vào của mạng”
* “Tầng output layer (tầng ra): Là tầng bên phải của mạng, nó thể hiện cho cácđầu ra của mạng”
* “Tầng hidden layer (tầng ẩn): Tầng này nằm ở giữa tầng vào và tầng ra thểhiện cho các quá trình suy luận logic của Neural Network”.



Hình 2.5 Mô hình Neural Network

**b. Minh họa về Neural Network**

NeuraloNetwork còn có tên gọi khác là Mạng Neural nhân tạo là một chuỗi thuật toán được sử dụng để đưa ra nhằm xác định chính xác các mối quan hệ trong tập hợp các dự liệu cụ thể, thông qua việc bắt chước bộ não của con người. Một cách dễ hiểu hơn đólà có thể là nhân tạo về bản chất. Neural Network có khả năng tương thích với mọi thứ ngay từ lúc chúng ta thay đổi dữ liệu đầu vào. Thuật toán này có thể đưa ra các kết quả một cách chính xác nhất mà chúng ta không cần phải tiến hành xây dựng các tiêu chí đầu ra.

****

Hình 2.6: Minh họa thuật toán phân lớp Mạng Nơnron nhân tạo (Neural Network )

**2.3 So sánh giữa khai phá dữ liệu và các phương pháp khác.**

* So sánh khai phá dữ liệu với phương pháp học máy .

|  |  |  |
| --- | --- | --- |
|  | Khai phá dữ liệu | Học máy |
| Phương thức hoạt động | Khai phá dữ liệu là một quá trình bắt đầu từ dữ liệu dường như không có cấu trúc để tìm ra các mẫu thú vị. | Sử dụng rất nhiều thuật toán |
| Dữ liệu | Được sử dụng để trích xuất dữ liệu từ bất kỳ kho dữ liệu nào | Máy học là đọc máy liên quan đến phần mềm hệ thống. |
| Ứng dụng | Chủ yếu sử dụng dữ liệu từ một miền cụ thể. | Các kỹ thuật học máy khá chung chung và có thể được áp dụng cho nhiều cài đặt khác nhau |
| Tiêu điểm | Cộng đồng khai thác dữ liệu tập trung chủ yếu vào các thuật toán và ứng dụng. | Các cộng đồng học máy quan tâm nhiều hơn đến lý thuyết |
| Phương pháp luận | Được sử dụng để lấy các quy tắc từ dữ liệu. | Máy học dạy máy tính học và hiểu các quy tắc nhất định. |
| Nghiên cứu | Là một lĩnh vực nghiên cứu sử dụng các phương pháp như học máy. | Học máy là một phương pháp được sử dụng để cho phép máy tính thực hiện các tác vụ thông minh. |

**🡺**Mặc dù học máy hoàn toàn khác với khai thác dữ liệu, nhưng chúng thường tương tự nhau . Cả hai, khai thác dữ liệu và học máy, đều hợp tác bình đẳng để phát triển AI cũng như các lĩnh vực nghiên cứu.

* Ngoài ra còn có một số phương pháp khác như phương pháp thống kê Thống kê, là một lĩnh vực toán học bao gồm tất cả các mô hình toán học, kỹ thuật và định lý đang được sử dụng trong AI.

**CHƯƠNG 3 : DEMO ỨNG DỤNG**

**3.1 Giới thiệu bài toán**

Phân loại là một ứng dụng học máy cổ điển. Phân loại về cơ bản phân loại đầu ra thành hai lớp, tức là đầu ra có thể là một trong hai thứ. Ví dụ, một ngân hàng muốn biết liệu một khách hàng có thể thanh toán các khoản đầu tư hàng tháng của mình hay không? Ta có thể sử dụng các thuật toán học máy để xác định đầu ra của bài toán này, kết quả sẽ là Có hoặc Không (Hai lớp). Nhưng nếu muốn phân loại thứ gì đó có nhiều hơn 2 loại và không đơn giản như vấn đề có / không?

Đây là lúc phân loại nhiều lớp ra đời. Phân loại nhiều lớp có thể được định nghĩa là việc phân loại các cá thể thành một trong ba hoặc nhiều lớp. Trong bài này, chúng ta sẽ phân loại nhiều lớp bằng cách sử dụng  Neural Network.

* **Giới thiệu tập dữ liệu**

Tập dữ liệu đang sử dụng ở đây cho tác vụ nhận dạng hoa chứa 250 hình ảnh hoa. Thu thập dữ liệu dựa trên dữ liệu Flickr, hình ảnh google, hình ảnh Yandex. Sử dụng bộ dữ liệu này để nhận dạng những bông hoa trong ảnh.

* **Sử dụng tập dữ liệu**

Các hình ảnh được chia thành năm lớp: hoa cúc, hoa tulip, hoa hồng, hướng dương, bồ công anh. Đối với mỗi lớp, có khoảng 50 bức ảnh. Các bức ảnh không có độ phân giải cao, khoảng 320×240 pixel. Hình ảnh không được giảm xuống một kích thước, chúng có tỷ lệ khác nhau.

* **Định nghĩa bài toán**

Khi thông qua tất cả những điều cơ bản bắt đầu một số triển khai. Sử dụng nhiều thư viện python (Để đọc tập dữ liệu), Sklearn (Để đào tạo tập dữ liệu và triển khai mô hình) và các thư viện như Seaborn và Matplotlib (Để trực quan hóa dữ liệu).

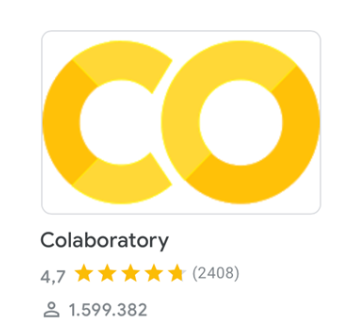
Một cách khác là sử dụng google colab để thực hiện thử nghiệm trực tuyến với tất cả các thư viện được cài đặt sẵn. Tập dữ liệu sắp sử dụng được gọi là tập dữ liệu hoa .

Sau khi đã có dữ liệu rồi, hoàn toàn có thể chạy thuật toán Neural Network ngay để tiến hành phân loại.

**3.2 Giới thiệu Google Colab**

* **Google Colab là gì?**

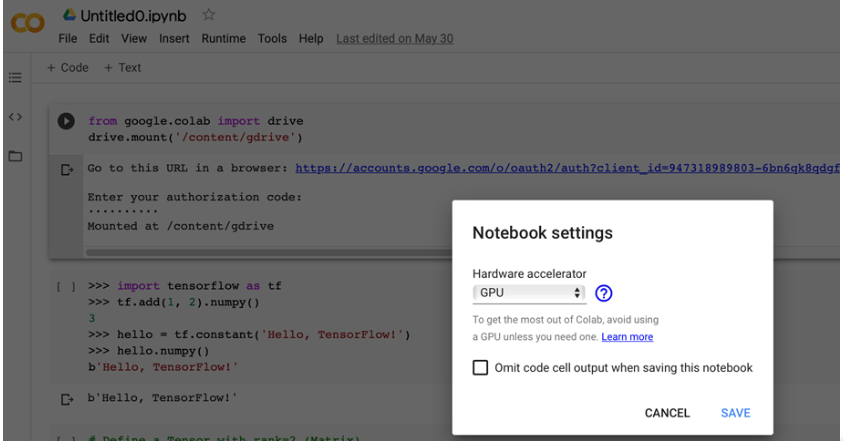
Colaboratory hay còn gọi là Google Colab, là một sản phẩm từ Google Research, nó cho phép chạy các dòng code python thông qua trình duyệt, đặc biệt phù hợp với Data analysis, machine learning và giáo dục. Colab không cần yêu cầu cài đặt hay cấu hình máy tính, mọi thứ có thể chạy thông qua trình duyệt, bạn có thể sử dụng tài nguyên máy tính từ CPU tốc độ cao và cả GPUs và cả TPUs đều được cung cấp.



Hình 3.1 Ảnh Google Colab

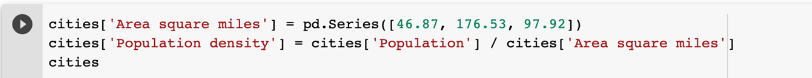
* **Cách sử dụng Colab**

Việc sử dụng cũng đơn giản như ứng dụng Google Docs, hay Google Sheet, chỉ cần truy cập vào Google Drive, tạo thư mục vào chọn tạo mới tạo tin Google Colab là có thể sử dụng.



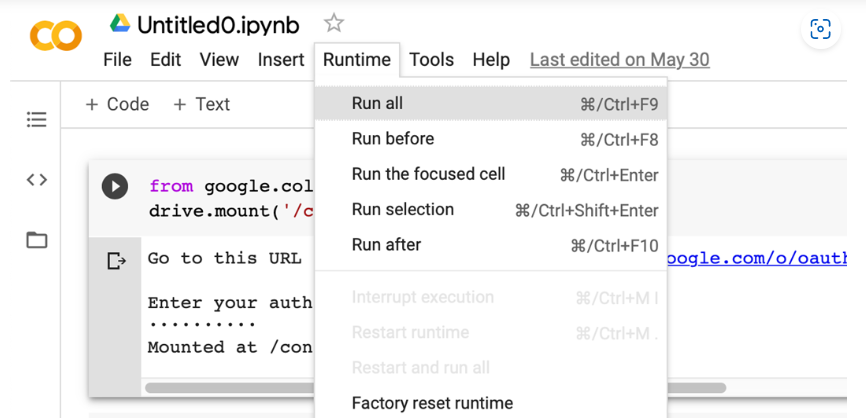
Hình 3.2 Thiết lập tùy chọn GPU hay TPU để chạy đoạn code

Vào Edit\Notebook Setting để có thể tùy chọn phần cứng để chạy đoạn code của mình, có thể chọn GPU, TPU.



Hình 3.3 Chạy code từng đoạn

Ngoài cách click chuột vào biểu tượng mũi tên tam giác ở phía trên, chúng ta có cách khác để chương trình chạy tất cả, chạy từng cái hay chỉ chạy những đoạn code mình chọn, thì chọn Runtime trên Toolbar như hình dưới đây.

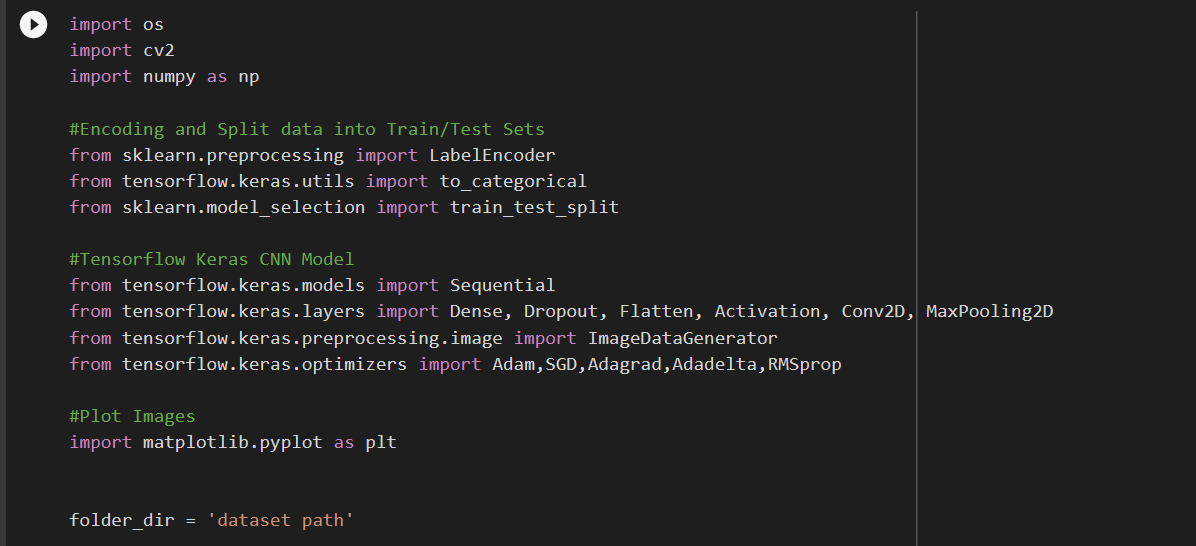


Hình 3.4 Chạy tất cả chương trình

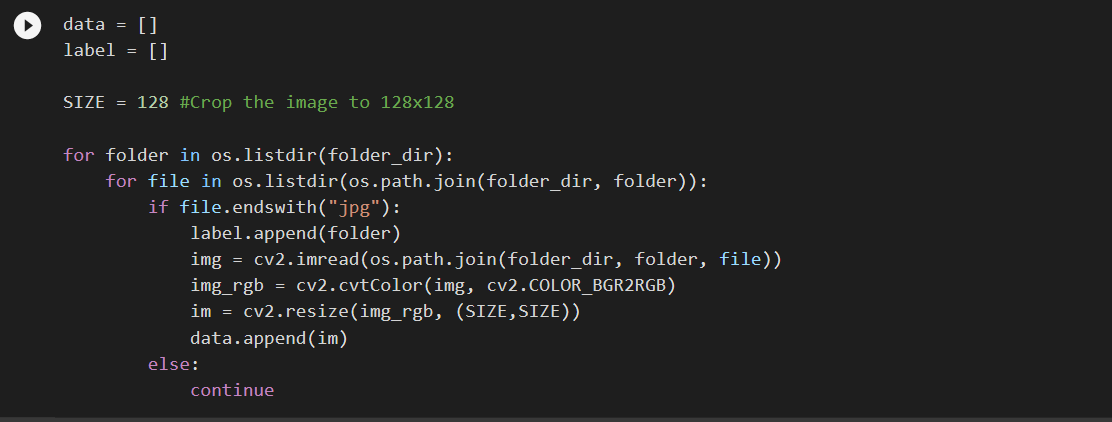
**3.3 Áp dụng thuật toán Neural Network vào phân loại hoa.**

* Thu thập dữ liệu: Thu thập dữ liệu về các loài hoa cần phân loại.
* Xử lý dữ liệu: Xử lý dữ liệu bằng cách loại bỏ các giá trị không hợp lệ và điền các giá trị bị thiếu.
* Phân chia tập dữ liệu: Phân chia tập dữ liệu thành 2 phần: tập huấn luyện và tập kiểm tra. Tập huấn luyện được sử dụng để huấn luyện mô hình và tập kiểm tra được sử dụng để đánh giá mô hình .
* Trích xuất đặc trưng: Trích xuất đặc trưng của các loài hoa từ tập huấn luyện.
* Huấn luyện mô hình: Sử dụng các thuật toán học máy để huấn luyện mô hình phân loại.
* Đánh giá mô hình: Đánh giá mô hình phân loại bằng cách sử dụng tập kiểm tra.

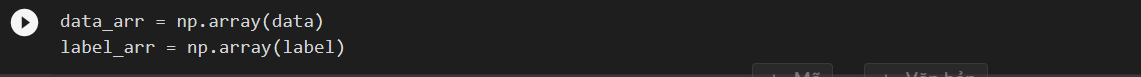
Nhập các thư viện Python cần thiết để bắt đầu với nhiệm vụ Nhận dạng phân loại hoa với Python:



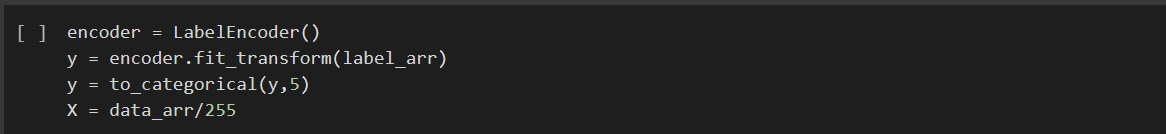
Đọc từng hình ảnh trong dữ liệu và tạo nhãn cho mỗi hình ảnh có tên của thư mục:



Chuyển đổi dữ liệu thành các giá trị số:



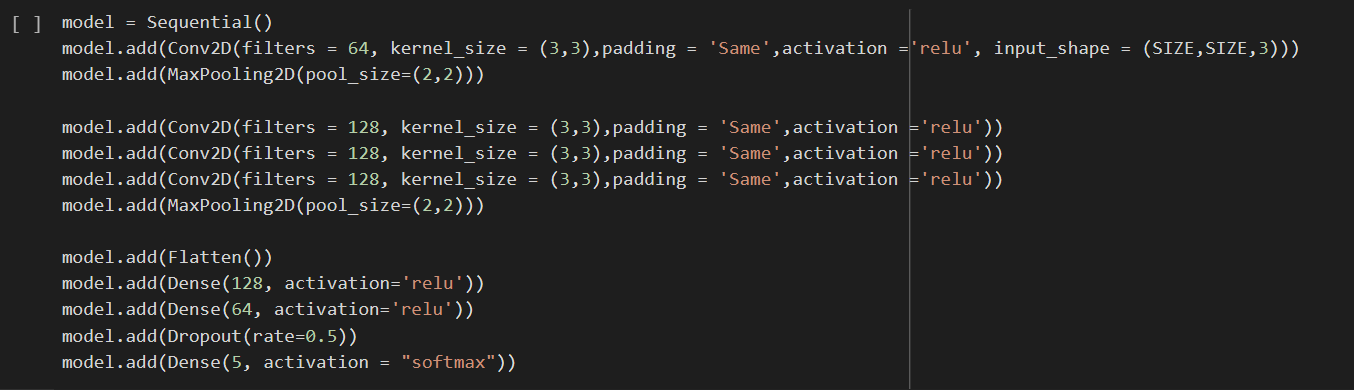
Sử dụng bộ mã hóa Label và chuẩn hóa dữ liệu:



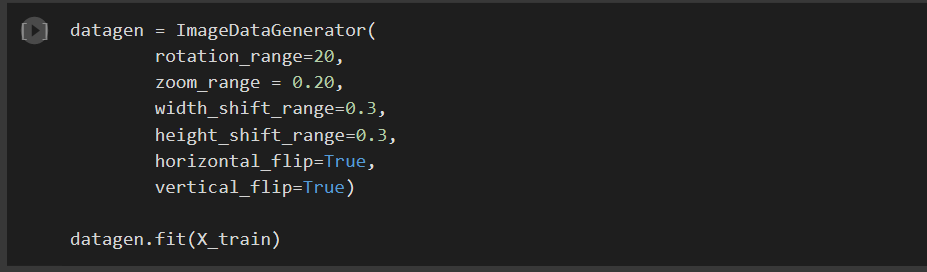
Chia tập dữ liệu thành 80% đào tạo và 20% bộ thử nghiệm:



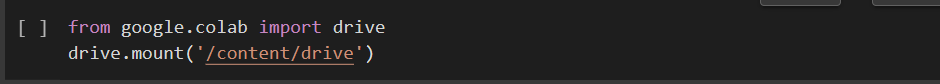
Xây dựng một mô hình mạng nơ-ron cho nhiệm vụ Nhận dạng hoa:



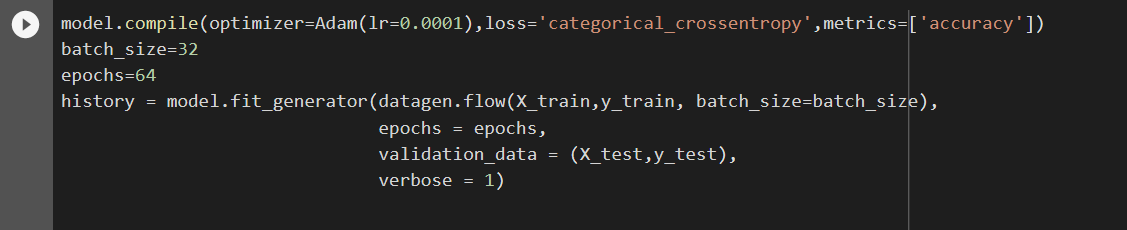
Trước khi biên dịch mô hình, cần tạo thêm hình ảnh đào tạo để tránh quá tải:



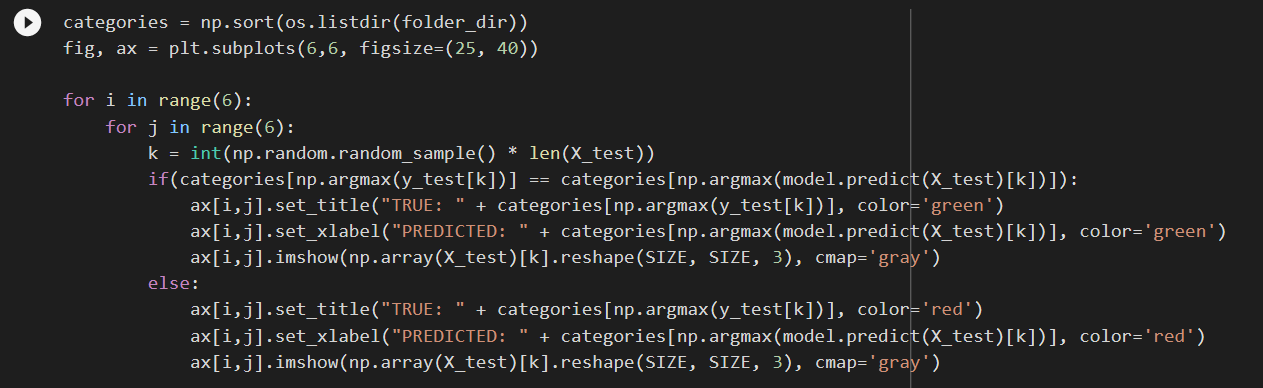
Liên kết với google drive



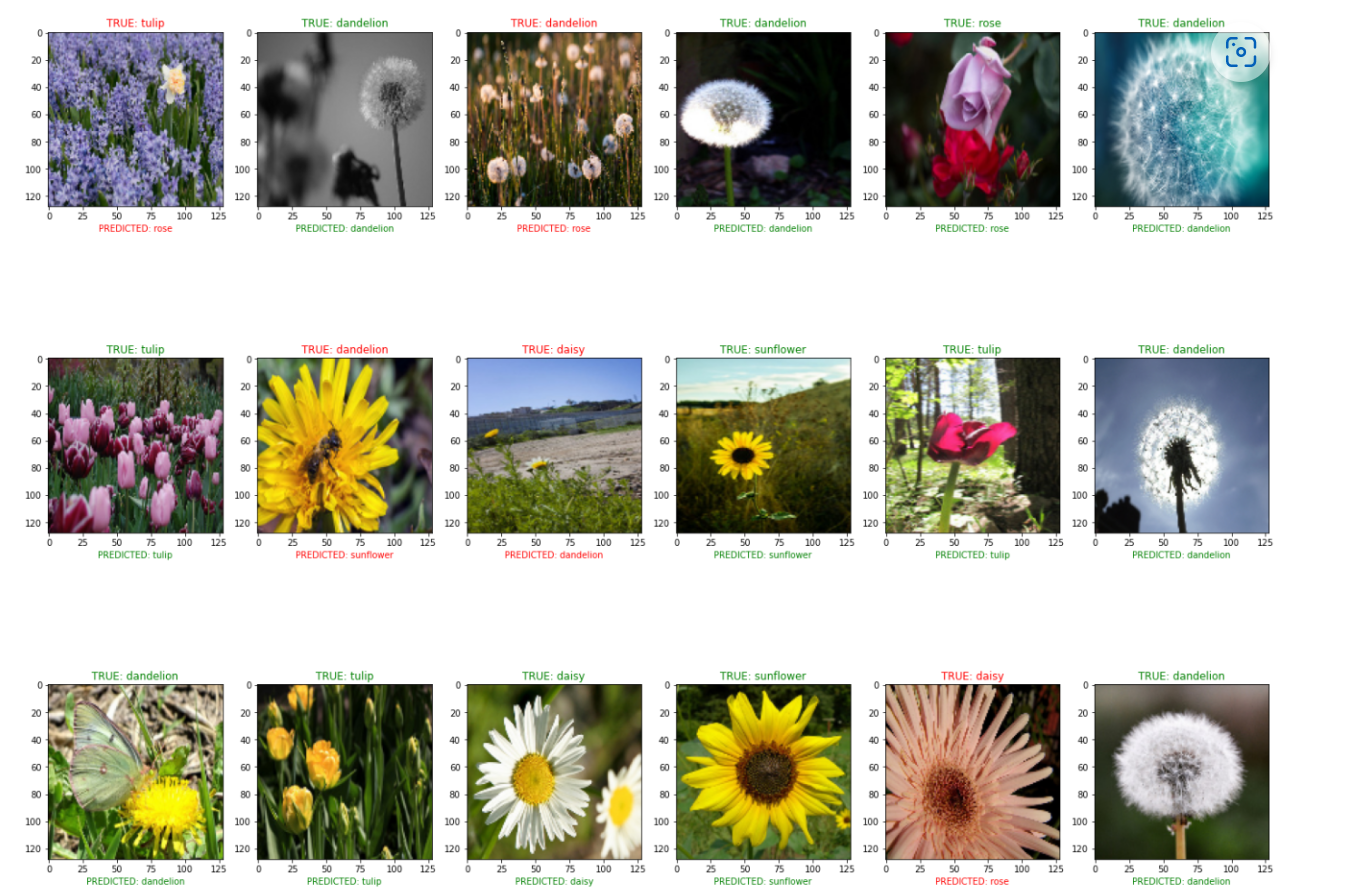
Biên dịch mô hình mạng nơ-ron:

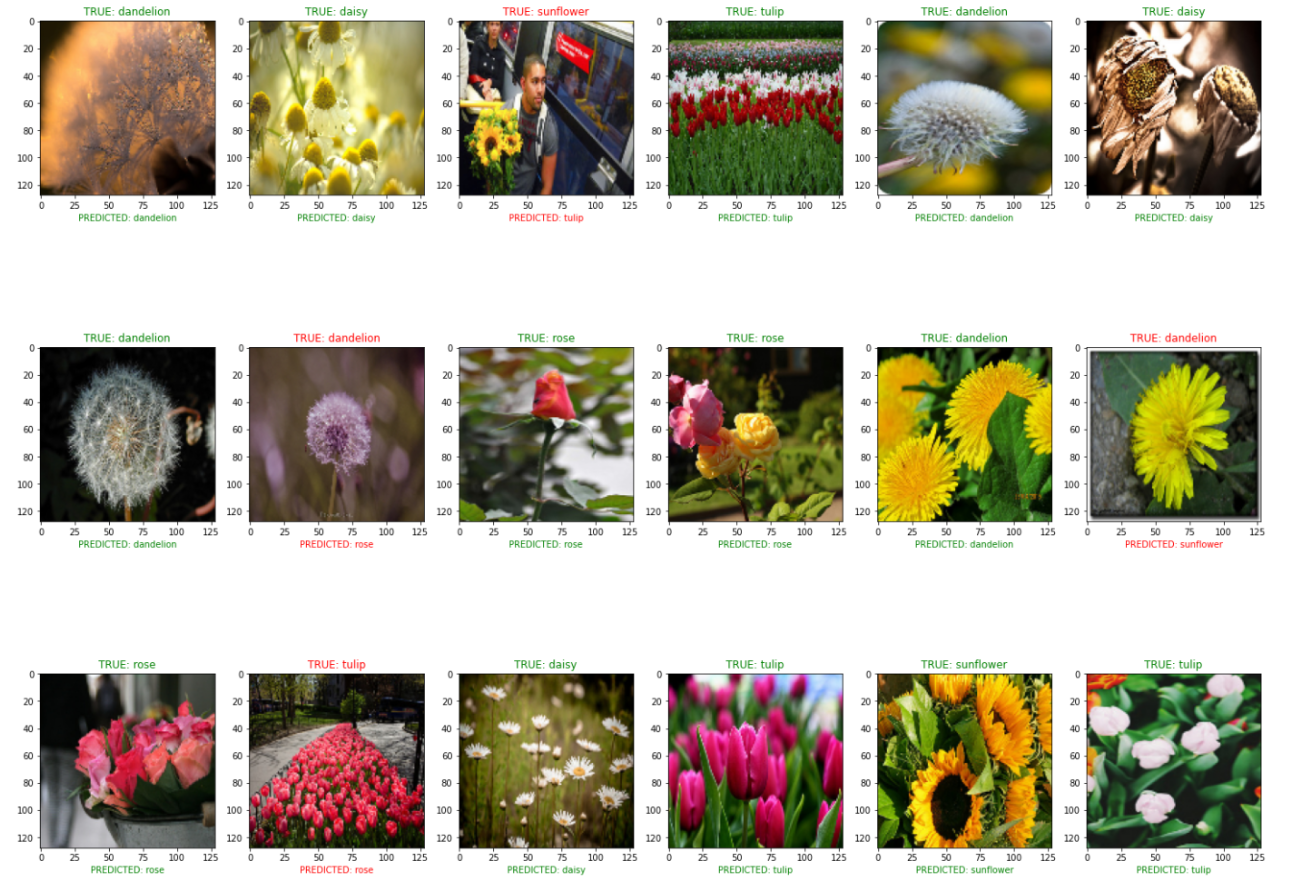


Mô hình kiểm tra xem mô hình dự đoán của chương trình có khớp với nhãn thực tế của hình ảnh hay không. Nếu đúng, nó sẽ hiển thị tiêu đề và nhãn dự đoán với màu xanh lá cây và nếu sai, nó sẽ hiển thị tiêu đề và nhãn dự đoán với màu đỏ.



🡺Kết quả thu được Hình ảnh được hiển thị bên dưới tiêu đề và nhãn.:





**Kết luận**

Bài báo cáo nêu lên những nét cơ bản về khai phá dữ liệu, khám phá tri thức và những vấn đề liên quan, kỹ thuật phân lớp dữ liệu và đi sâu vào một số phương pháp phân lớp như phân lớp bằng kỹ thuật KNN và kỹ thuật Neural Network ngoài ra còn một số kỹ thuật phân lớp khác... Các quá trình để phân lớp các dữ liệu hoa . Tìm hiểu bài toán tự động nhận dạng phân loại hoa và thực hiện phát triển . Cài đặt phương án giải quyết cho bài toán . Hoàn thiện xây dựng bộ cơ sở dữ liệu ảnh phụ vụ nhận dạng hoa với số lượng ảnh có trong tập dữ liệu. Thực nghiệm với bộ dữ liệu test và cho ra kết quả. Hệ thống nhận dạng hoa vẫn còn rất nhiều điều cần cải thiện đặc biệt là màu sắc và kích thước cũng như chất lượng của bộ cơ sở dữ liệu ảnh .

Em xin bày lòng biết ơn sau sắc đến cô Nguyễn Thị Vàn vì đã truyền đạt vốn kiến thức quý báu cho em trong suốt thời gian học tập . Nhờ có những lời hướng dẫn của cô nên e mới có thể hoàn thành đề tài này . Mặc dù đã cố gắng, nỗ lực hết mình song do thời gian nghiên cứu, trình độ của bản thân có hạn và điều kiện nghiên cứu còn nhiều hạn chế nên không thể tránh khỏi những khuyết thiếu và hạn chế, rất mong nhận được những góp ý, nhận xét quý báu của quý thầy cô và bạn bè để kết quả của đề tài hoàn thiện hơn. Em xin chân thành cảm ơn .

**Tài liệu tham khảo**

[Thuật Toán K-Nearest Neighbors (KNN) Siêu Cơ Bản (codelearn.io)](https://codelearn.io/sharing/thuat-toan-k-nearest-neighbors-knn)

[Ứng dụng Support Vector Machine trong bài toán phân loại hoa (viblo.asia)](https://viblo.asia/p/ung-dung-support-vector-machine-trong-bai-toan-phan-loai-hoa-PdbGnLXBkyA)

[Google Colab Là Gì Và Dùng Để Làm Gì? (codelearn.io)](https://codelearn.io/sharing/google-colab-la-gi)

[Iris datasets | Kaggle](https://www.kaggle.com/datasets/chuckyin/iris-datasets?resource=download)

**--**-Khai phá tri thức từ CSDL ( Knowledge Discovery in Database)

-Khai phá tri thức từ CSDL ( Knowledge Discovery in DatabKhai phá tri thức từ CSDL ( Knowledge Discovery in Database)