ỦY BAN NHÂN DÂN TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC SÀI GÒN

-

**TÊN ĐỀ TÀI**

BÁO CÁO MÔN KHAI PHÁ DỮ LIỆU

TÊN SINH VIÊN:**NGUYỄN TRỌNG HIẾU**

MÃ SỐ SINH VIÊN: **3118410123**

TÊN LỚP: **DCT11812**

TP. HỒ CHÍ MINH, THÁNG 5 NĂM 2022

Mục lục

[Chương 1: Giới thiệu 3](#_Toc103111467)

[1.1 Bối cảnh của bài toán 3](#_Toc103111468)

[1.2 Định nghĩa bài toán 3](#_Toc103111469)

[1.3 Các giải pháp hiện tại/thủ công để giải quyết bài toán 3](#_Toc103111470)

[1.4 Vai trò của khai phá dữ liệu trong việc giải quyết bài toán trên 5](#_Toc103111471)

[1.5 Kết quả dự đoán/ứng dụng của khai phá dữ liệu sau khi giải quyết bài toán trên 5](#_Toc103111472)

[Chương 2: Mô tả dữ liệu 6](#_Toc103111473)

[2.1 Giới thiệu về bộ dữ liệu 6](#_Toc103111474)

[2.2 Thống kê mô tả về dữ liệu 7](#_Toc103111475)

[Chương 3: Phương pháp khai phá dữ liệu 12](#_Toc103111476)

[3.1 Quy trình khai phá dữ liệu 12](#_Toc103111477)

[3.2 Nguyên lý hoạt động của mô hình/ thuật toán 13](#_Toc103111478)

[3.3 Cách cài đặt mô hình các tham số (nếu có) của mô hình 13](#_Toc103111479)

[3.4 Tiêu chí đánh giá mô hình 13](#_Toc103111480)

[Chương 4: Thực nghiệm 14](#_Toc103111481)

[4.1 Yêu cầu về chương trình 14](#_Toc103111482)

[4.2 Mô tả chi tiết các bước khai phá dữ liệu 14](#_Toc103111483)

[Chương 5: Kết quả 17](#_Toc103111484)

[5.1 Kết quả khai phá dữ liệu 17](#_Toc103111485)

[5.2 So sánh kết quả thực tế với kết quả dự đoán 18](#_Toc103111486)

[Chương 6: Thảo luận và kết luận 19](#_Toc103111487)

[6.1 Khả năng ứng dụng của giải pháp/mô hình 19](#_Toc103111488)

[6.2 Ưu điểm – nhược điểm của giải pháp/mô hình 19](#_Toc103111489)

[6.3 Đề xuất 19](#_Toc103111490)

[Tài liệu tham khảo 20](#_Toc103111491)

# Chương 1: Giới thiệu

## 1.1 Bối cảnh của bài toán

Nấm là một loại thực phẩm giàu dinh dưỡng, là một nguồn chất xơ, protein và chất chống ôxy hóa dồi dào, ít calo, chúng có đặc tính kháng khuẩn, giảm nồng độ cholesterol trong máu và góp phần tăng cường hiệu quả của hệ thống miễn dịch.Nấm thậm chí còn được sử dụng để phòng ngừa và điều trị một số bệnh như Parkinson,Alzheimer,tăng huyết áp và ung thư..Tuy Nhiên, không phải loại nấm nào cũng ăn được,theo thống kê có tất cả hơn 14.000 loại nấm tuy nhiên chỉ có khoảng một nửa trong số đó có thể được dùng làm thực phẩm phục vụ nhu cầu ăn uống.

Ngộ độc nấm rất nguy hiểm do khó có thể phát hiện do giống với những triệu chứng ngộ độc thường gặp hoặc những triệu chứng lâu dài khó phát hiện. Do đó, nên sớm phát hiện nấm độc để tránh nguy cơ ngộ độc.

Bài toán phân loại nấm nhằm tìm ra loại nấm nào ăn được và nấm nào có độc.

## 1.2 Định nghĩa bài toán

Bài toán phân loại nấm là bài toán xác định một cây nấm có độc hay có thể ăn được dựa vào các đặc điểm của cây nấm như thân, mũ nấm, phiến nấm,…..

input: gồm thông tin về các bộ phận của cây nấm như thân, mũ nấm,........

output: thông tin về loại nấm đã nhập(ăn được hay có độc)

## 1.3 Các giải pháp hiện tại/thủ công để giải quyết bài toán

có thể dựa vào một số đặc điểm cơ bản như sau:

**Nhìn bằng mắt**

Đa phần, nấm độc thường có màu sắc rất sặc sỡ và nổi bật để cuốn hút những côn trùng, động vật khác. Chẳng hạn, phần mũ nấm có nhiều đốm đỏ, trắng hoặc đen, trong khi thân nấm có thể xuất hiện nhiều vằn hay các vết nứt xung quanh. Không những thế, khi hái nấm độc còn xuất hiện nhựa chảy ra.

Trái lại, nấm ăn có màu sắc đơn giản hơn như mũ nấm màu đen hoặc xám, còn thân nấm màu trắng hoặc cùng màu với mũ nấm.

Tuy nhiên một vài loại nấm có màu trắng toàn bộ nhưng vẫn là nấm độc cần lưu ý Nấm có đủ: mũ, phiến nấm, cuống, vòng cuống và bao gốc thường là nấm độc

**Ngửi bằng mũi**

Đối với nấm độc, khi ngắt bạn không chỉ nhìn thấy phần nhựa mủ chảy ra (nếu có) mà còn ngửi được mùi hắc hoặc mùi đắng xộc lên mũi. Tuy nhiên, một số nấm độc vẫn có mùi hương thơm nhẹ.

Trong khi, đối với nấm ăn được thì có hương thơm nhẹ hoặc không mùi.

Cách phân biệt nấm ăn và nấm độc ngử bằng mũi

**Thử nghiệm biến màu**

Bạn có thể sử dụng nguyên liệu hoặc vật dụng khác để kiểm tra nấm có độc hay không? Cụ thể, bạn có thể dùng một trong những cách sau:

+Cách 1: Lấy phần đầu trắng của hành lá chà xát lên phần mũ nấm. Nếu hành lá chuyển sang màu xanh nâu thì chứng tỏ nấm có độc, và ngược lại.

+Cách 2: Dùng muỗng hoặc đũa bằng chất liệu bạc cắm vào mũ hoặc thân nấm. Nếu vật dụng bị đổi màu thì chứng tỏ nấm đó có độc, và ngược lại.

+Cách 3: Nhỏ ít sữa tươi lên phần mũ nấm. Nếu sữa bị vón cục thì chứng tỏ nấm có độc.

**Quan sát cây nấm**

Phần lớn, nấm độc thường có các lá tia (nằm ở phía dưới mũ nấm) màu trắng. Trái lại, nấm ăn thường có phần lá tia này màu da hoặc màu nâu, thậm chí một số loại nấm có tia màu trắng nhưng có thể ăn được chứ không phải là nấm độc.

**màu đỏ trên mũ hoặc thân nấm**

các loại nấm có phần mũ nấm hoặc thân nấm xuất hiện đốm đỏ hoặc có màu đỏ thường không ăn được.

**vẩy trên mũ**

Nấm độc thường có vẩy trên phần mũ nấm, những lớp vẩy đó có thể màu sáng hoặc màu tối, trông như vết đốm. Chẳng hạn, với nấm độc có màu trắng, bạn dễ dàng thấy các mảng vẩy có màu nâu hoặc màu da.

**vòng bao quanh thân nấm**

Nấm độc thường có vòng tròn bao quanh lấy thân nấm, đồng thời bên dưới mũi nấm cũng có thể xuất hiện vòng tròn bao quanh trông như một chiếc mũ nhỏ.

## 1.4 Vai trò của khai phá dữ liệu trong việc giải quyết bài toán trên

Không phải ai cũng có đủ kiến thức để phân biệt các loại nấm hoặc đôi khi thiếu những vật dụng cần thiết để kiểm tra. Khai phá dữ liệu giúp phát hiện nhanh và gần chính xác chỉ cần nhập thông tin về cấu tạo của cây nấm.

## 1.5 Kết quả dự đoán/ứng dụng của khai phá dữ liệu sau khi giải quyết bài toán trên

Về dự đoán, Khai phá dữ liệu sẽ giúp phân loại nhanh chóng số lượng lớn nấm trong thời gian ngắn với độ tin cậy chấp nhận được.

Về ứng dụng trong thực tế: phần lớn người tiêu dùng bình thường không cần lo về vấn đề ngộ độc nấm. Do đó, bài toán ít cần dùng đến khai phá dữ liệu để giải quyết nên tính ứng dụng không cao.

Bài toán mang tính chất nghiên cứu, học thuật, học hỏi thêm về khai phá dữ liệu trong trường học.

# Chương 2: Mô tả dữ liệu

## 2.1 Giới thiệu về bộ dữ liệu

Bộ dữ liệu này bao gồm các mô tả về các mẫu giả định tương ứng với 23 loài nấm mang trong họ Nấm Agaricus và Lepiota được rút ra từ hướng dẫn thực địa của Hiệp hội Audubon về Nấm Bắc Mỹ (1981). Mỗi loài được xác định là chắc chắn ăn được, chắc chắn có độc, hoặc không rõ có thể ăn được và không được khuyến cáo. Sau này gộp thành ăn được và có độc.

Bộ dữ liệu nặng 366kb gồm 23 đặc trưng và 8124 mẫu được đăng ở website kaggle tại địa chỉ: <https://www.kaggle.com/datasets/uciml/mushroom-classification> được bổ sung thêm 1 thuộc tính(veil-type) so với bộ dữ liệu gốc của UCI Machine Learning.

Dữ liệu gốc được đăng tại: https://archive.ics.uci.edu/ml/datasets/Mushroom

Tất cả các thuộc tính đều là kiểu chuỗi dạng category và chỉ có 1 chữ cái

Các đặc trưng của dữ liệu:

**Mũ nấm(Cap)**

+cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s

+cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s

+cap color: brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y

**Phiến nấm(gill)**

+gill-attachment: attached=a,descending=d,free=f,notched=n

+gill-spacing: close=c,crowded=w,distant=d

+gill-size: broad=b,narrow=n

+gill-color: +black=k,brown=n,buff=b,chocolate=h,gray=g,green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y

**Thân nấm(stalk)**

+stalk-shape: enlarging=e,tapering=t

+stalk-root: bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?

+stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s

+stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s

+stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y

+stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y

**Bao gốc nấm(veil)**

veil-type: partial=p,universal=u

veil-color: brown=n,orange=o,white=w,yellow=y

**Vòng cuốn nấm(ring)**

ring-number: none=n,one=o,two=t

ring-type:

cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z

**vết thâm tím(bruises):** bruises=t,no=f

**mùi vị(odor):**

almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s

**Màu bảo tử nấm (Spore Print Color):** black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y

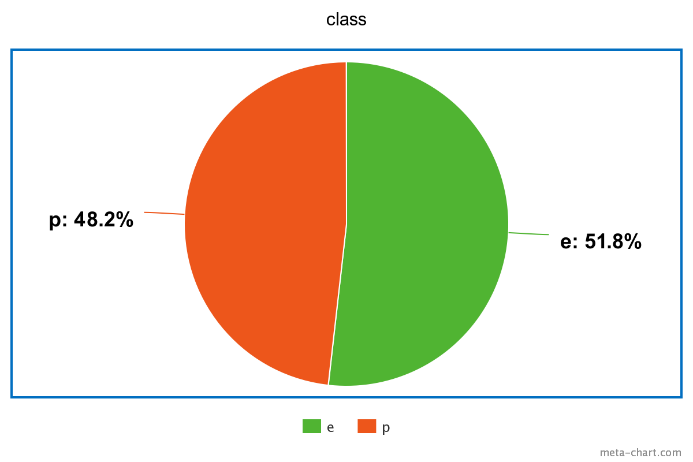
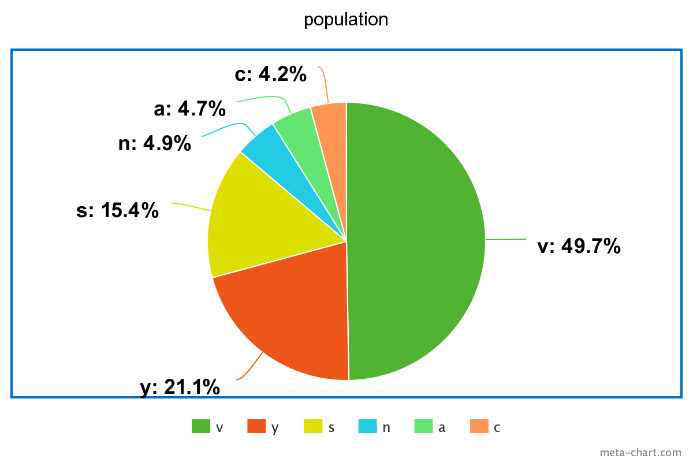
**Quần thể nấm(Population):** abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y

**Môi trường sống(Habitat):**

grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d

## 2.2 Thống kê mô tả về dữ liệu

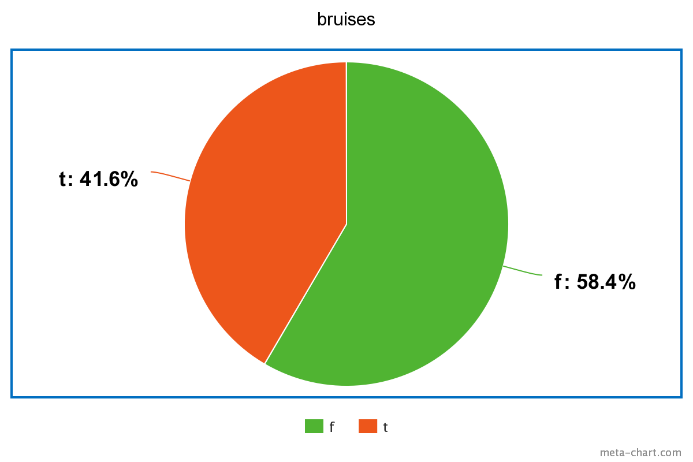
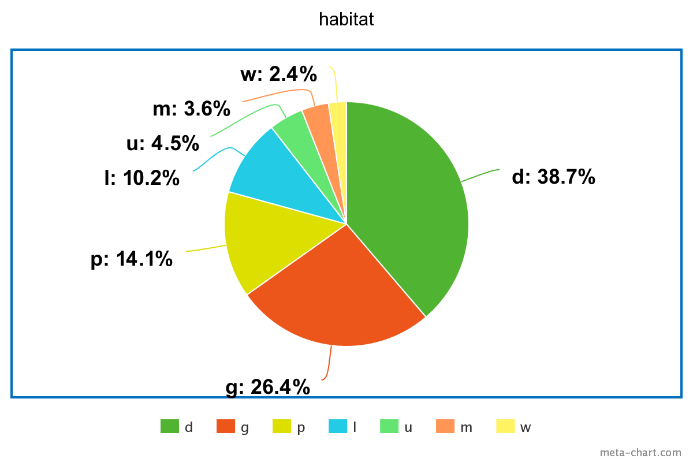
Thống kê phân bố của class Thống kê phân bố của population



Hình 1.1: Thống kê đặc trưng class

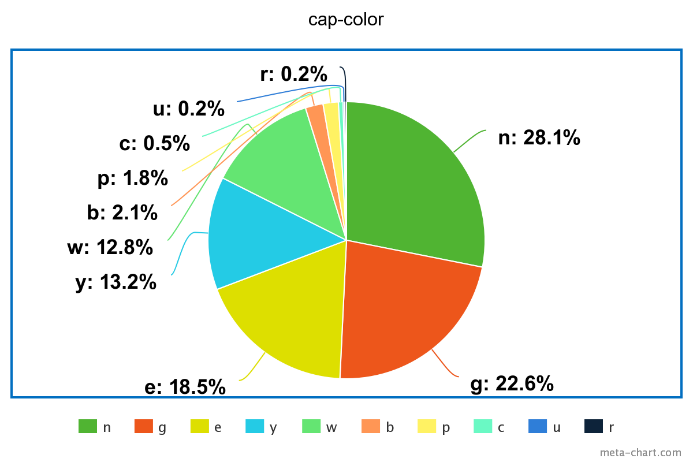
Hình 1.2 Thống kê đặc trưng population

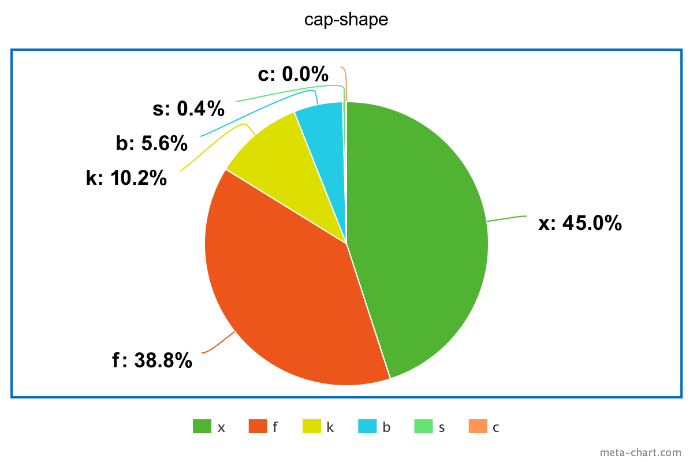
Thống kê phân bố của bruises Thống kê phân bố của habitat



Hình 1.4: Thống kê đặc trưng habitat

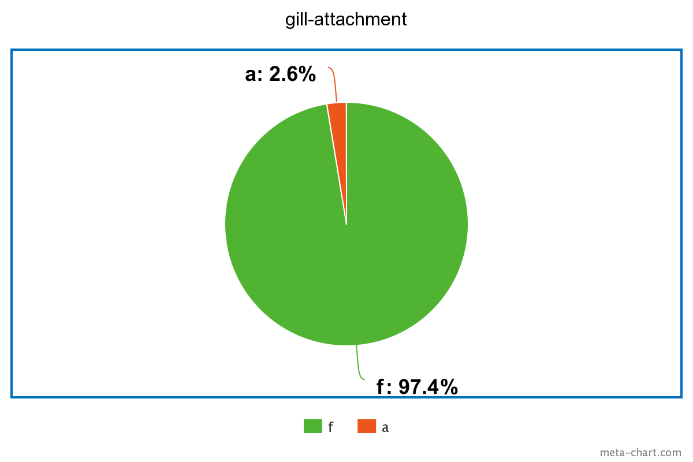
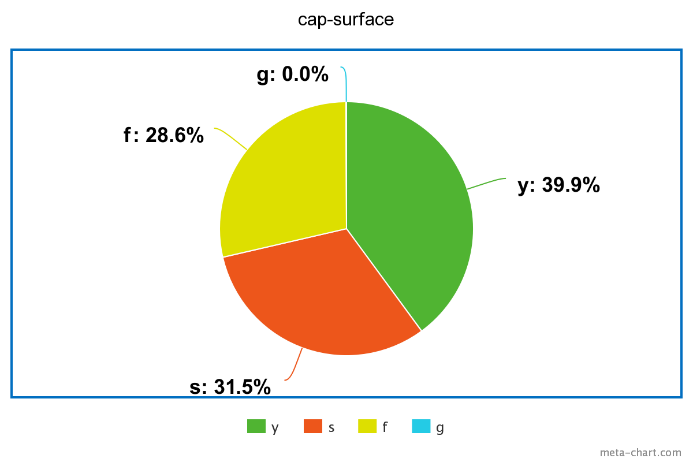
Hình 1.3: Thống kê đặc trưng bruises

Thống kê phân bố của cap-shape Thống kê phân bố của cap-color



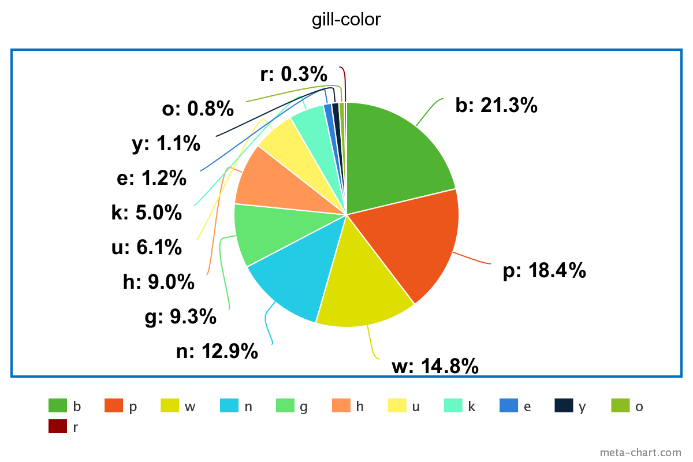
Hình 1.6: Thống kê đặc trưng cap-color

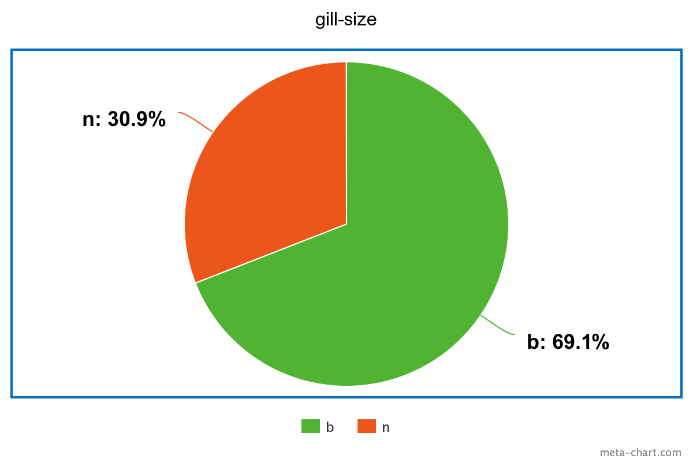
Hình 1.5: Thống kê đặc trưng cap-shape

Thống kê phân bố của cap-shape surface Thống kê phân bố của gill-attachment 

Hình 1.7: Thống kê đặc trưng cap-surface

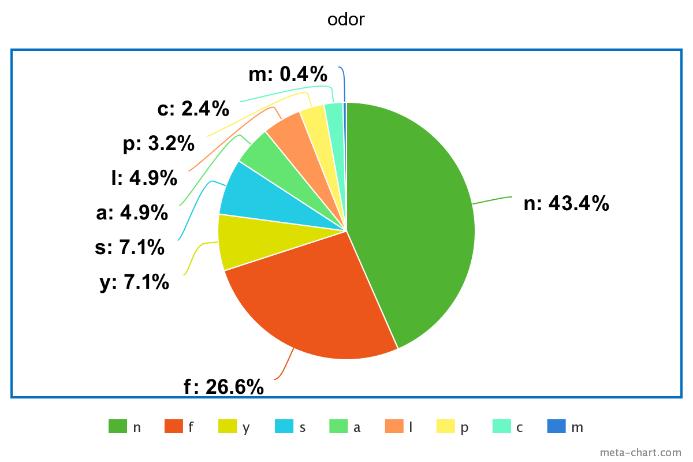
Hình 1.8: Thống kê đặc trưng gill-attachment

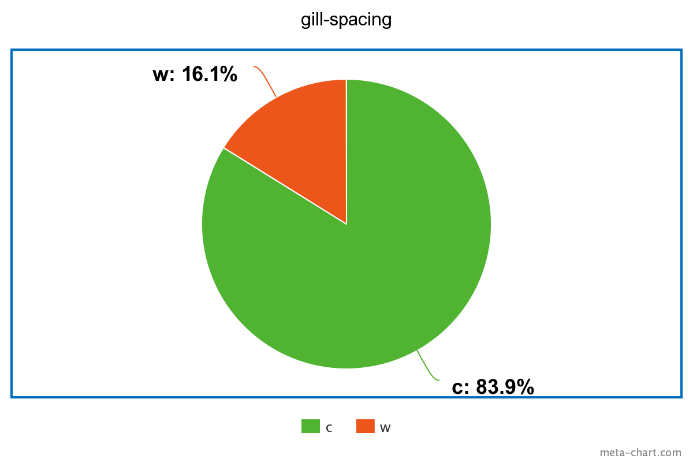
Thống kê phân bố của gill-size Thống kê phân bố của gill-color



Hình 1.9: Thống kê đặc trưng gill-size

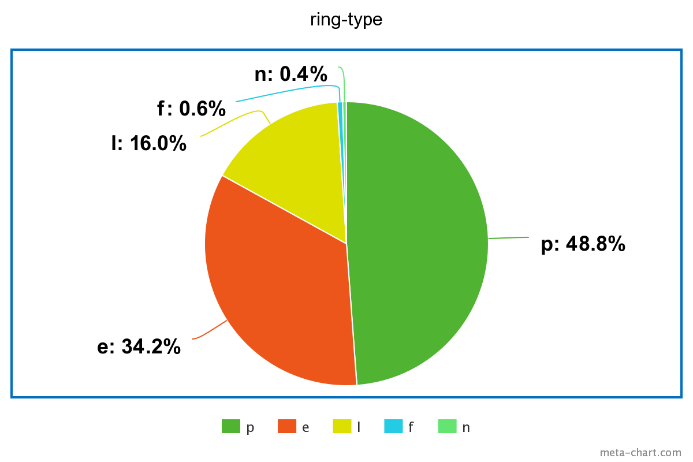
Hình 1.10: Thống kê đặc trưng gill-color

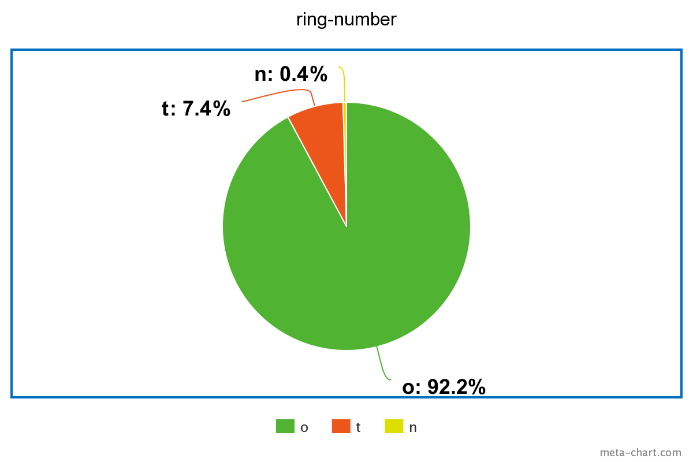
 Thống kê phân bố của gill-spacing Thống kê phân bố của odor



Hình 1.12: Thống kê đặc trưng odor

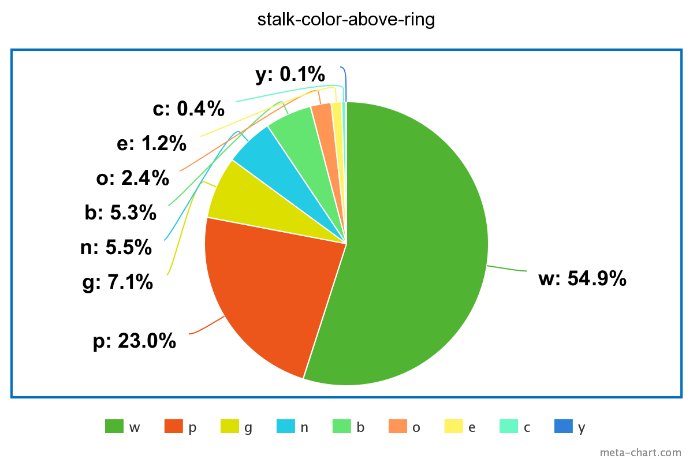
Hình 1.11: Thống kê đặc trưng gill-spacing

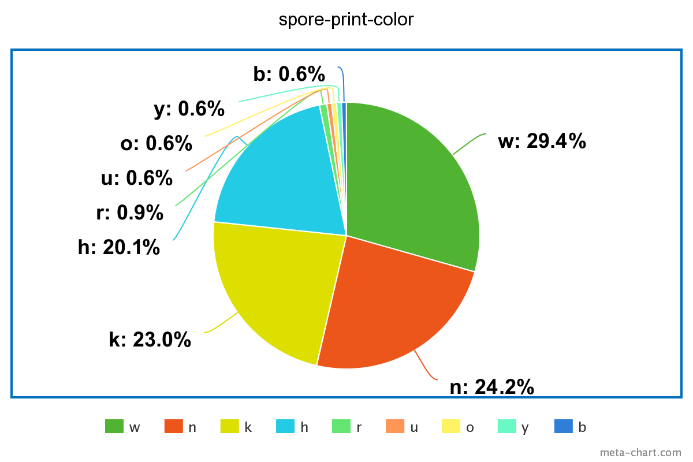
Thống kê phân bố của ring-number Thống kê phân bố của ring-type



Hình 1.14: Thống kê đặc trưng ring-type

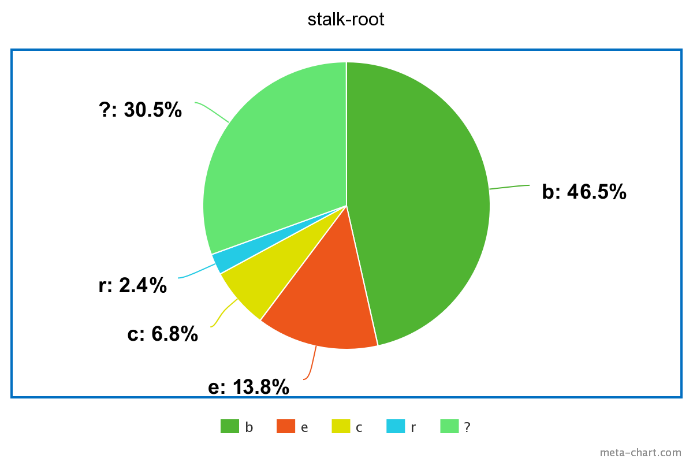
Hình 1.13: Thống kê đặc trưng ring-number

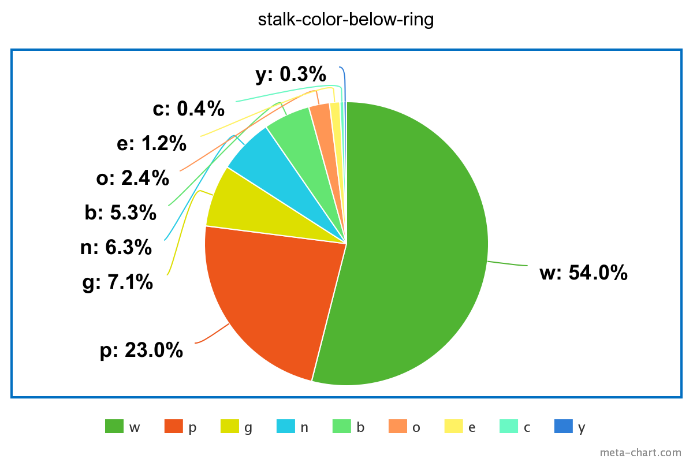
Thống kê phân bố của spore-print-color Thống kê phân bố của stalk-color-above-ring



Hình 1.15: Thống kê đặc trưng spore-print-color

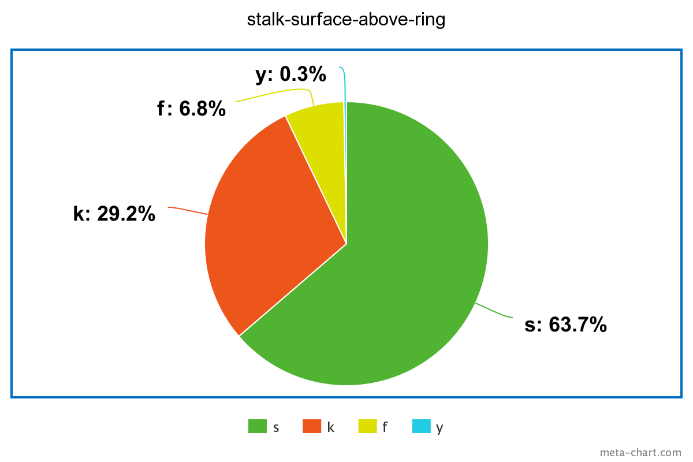
Hình 1.16: Thống kê đặc trưng stalk-color-above-ring

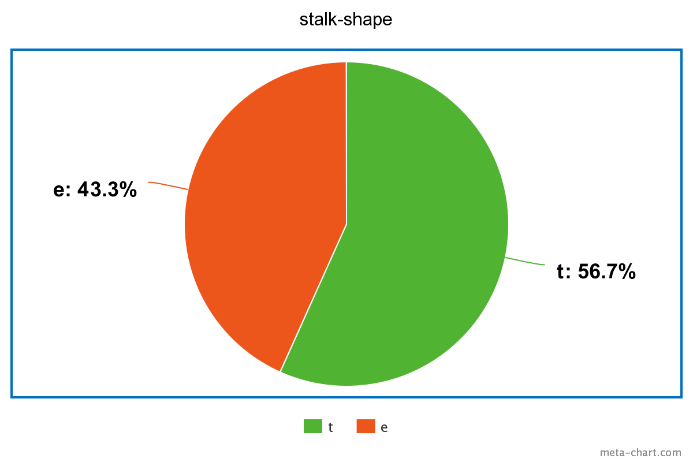
Thống kê phân bố của stalk-color-above-ring Thống kê phân bố của stalk-root



Hình 1.18: Thống kê đặc trưng stalk-root

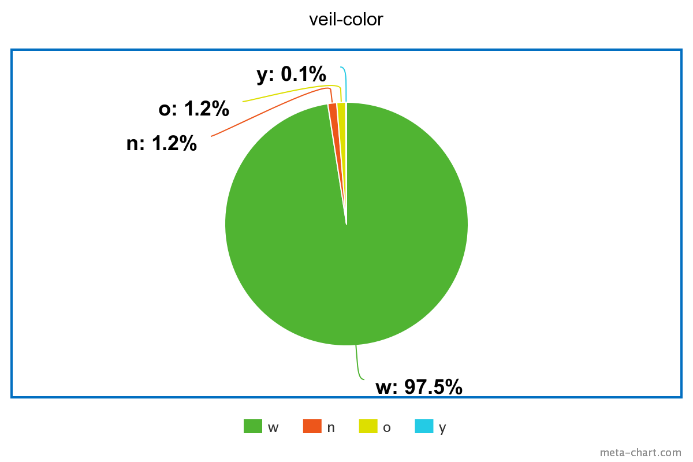
Hình 1.17: Thống kê đặc trưng stalk-color-below-ring

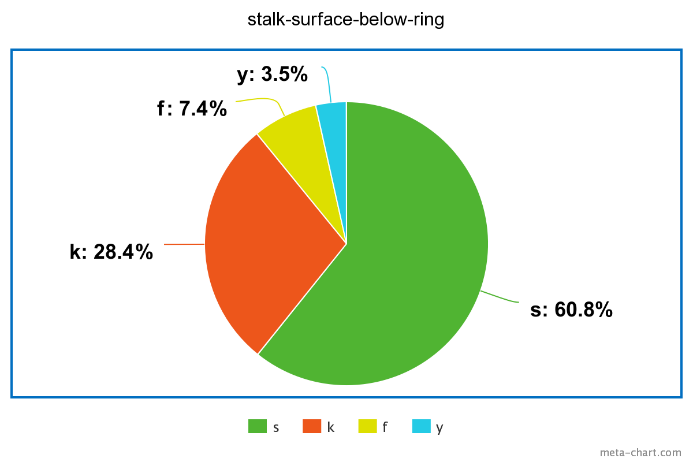
 Thống kê phân bố của stalk-shape Thống kê phân bố của stalk-surface-above-ring



Hình 1.19: Thống kê đặc trưng stalk-shape

Hình 1.20: Thống kê đặc trưng stalk-surface-above-ring

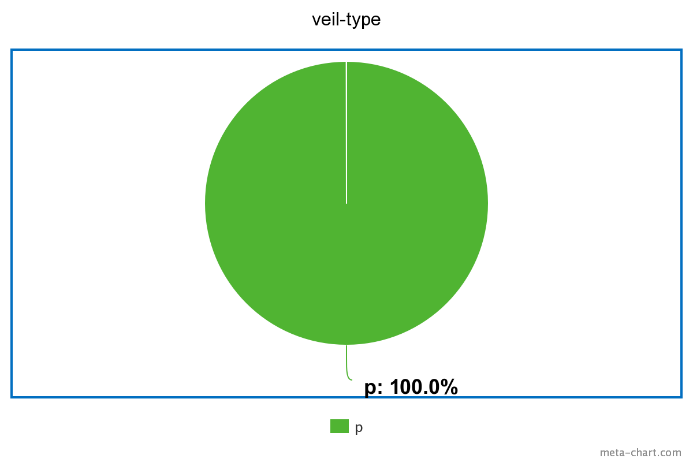
Thống kê phân bố của stalk-surface-above-ring Thống kê phân bố của veil-color



Hình 1.22: Thống kê đặc trưng veil-color

Hình 1.21: Thống kê đặc trưng stalk-surface-below-ring

Thống kê phân bố của veil-type



Hình 1.23: Thống kê đặc trưng veil-type

# Chương 3: Phương pháp khai phá dữ liệu

## 3.1 Quy trình khai phá dữ liệu

**Bước 1:Chuẩn bị dữ liệu**

Ở bước này ta tiến hành tìm hiểu về các tri thức cần thu thập qua đó xác định được các thông tin cần thu thập của bộ dữ liệu. Sau khi đã thu thập đủ dữ liệu, ta tiếp tục tiến hành loại bỏ những dữ liệu trùng lặp, sai định dạng, không chính xác hay đầy đủ về mặt thông tin. Sau khi xử lý ta sẽ được bộ dữ liệu để phục vụ cho bước tiếp theo.

**Bước 2: Tiền xử lý dữ liệu**

Các nhiệm vụ chính trong bước này theo thứ tự gồm:

* Làm sạch dữ liệu(data cleaning) gồm:

+ xử lý dữ liệu bị mất, thiếu

+ xử lý dữ liệu bị nhiễu(dữ liệu vô nghĩa)

* tích hợp dữ liệu(Data integration): kết hợp nhiều bộ dữ liệu
* Giảm dữ liệu(Data reduction): giảm chiều của bộ dữ liệu
* Chuyển đổi dữ liệu(Data Transformation): chuyển dữ liệu về dạng chuẩn.

**Bước 3: Chọn tác vụ, thuật toán khai phá dữ liệu và thực hiện**

Để đạt được mục đích ta cần chọn được tác vụ khai phá dữ liệu. Thông thường có các tác vụ sau:Đặc trưng (feature), Phân biệt (discrimination), Kết hợp (association), Phân lớp (classification),Phân cụm (clusterity), Xu thế (trend analysis), Phân tích độ lệch, Phân tích hiếm.

Sau đó ta sử dụng bộ dữ liệu đã được tiền xử lý và các thuật toán khai phá dữ liệu(ở đây là KNN) để thực hiện tìm kiếm tri thức ẩn dấu.

**Bước 4: Sử dụng các tri thức đã khám phá sau quá trình khai phá**

Sau khi tìm được các tri thức ẩn dấu ta sử dụng chúng vào mục đích thực tế để đạt được mục tiêu ban đầu.

## 3.2 Nguyên lý hoạt động của mô hình/ thuật toán

Nguyên lý hoạt động của K- Nearest Neighbors (kNN): Trong 1 tập dữ liệu D được gắn nhãn muốn xác định một dữ liệu mới A thuộc nhãn nào ta chỉ cần xét K đối tượng gần nhất với A. Khi đó nhãn của A là số lượng nhãn lặp lại nhiều nhất trong K đối tượng.

Các bước:

1.Ta có D là tập các điểm dữ liệu đã được gắn nhãn và A là dữ liệu chưa được phân loại.

2.Đo khoảng cách (Euclidian, Manhattan, Minkowski, Minkowski hoặc Trọng số) từ dữ liệu mới A đến tất cả các dữ liệu khác đã được phân loại trong D.

3.Chọn K (K là tham số mà bạn định nghĩa) khoảng cách nhỏ nhất.

4.Kiểm tra danh sách các lớp có khoảng cách ngắn nhất và đếm số lượng của mỗi lớp xuất hiện.

5.Lấy lớp xuất hiện nhiều lần nhất trong K hàng xóm.

6.Lớp của dữ liệu mới là lớp mà bạn đã nhận được ở bước 5.

## 3.3 Cách cài đặt mô hình các tham số (nếu có) của mô hình

Mô hình chạy trên hệ điều hành windows 10

dùng ngôn ngữ lập trình python phiên bản 3.10.3 có thể tải phiên bản này tại trang:

https://www.python.org/downloads/release/python-3103/.

Sử dụng Ide hỗ trợ lập trình pycharm phiên bản 3/2/2021 Community tải ở trang web: https://www.jetbrains.com/pycharm/download/#section=windows.

Các thư viện hỗ trợ gồm: Numpy, Pandas , Matplotlib , Seaborn , scikit-learn có thể tải ở phần python pakage trong pycharm.

## 3.4 Tiêu chí đánh giá mô hình

* Định nghĩa:

Positive (P-dương tính): là lớp nấm có độc

Negative (N-âm tính): là lớp nấm ăn được

True Positive (TP): là số lượng nấm có độc được dự đoán là có độc

False Positive (FP): là số lượng nấm ăn được dự đoán là có độc

True Negative (TN): là số lượng nấm ăn được được dự đoán là ăn được

False Negative (FN): là số lượng nấm có độc được dự đoán là ăn được

True Positive Rate(TPR) = TP/(TP + FN)

False Positive Rate(FPR) = FP/(FP + TN)

* Các tiểu chí: Accuracy, Precision, Recall, F1-score, AUC

Accuracy = số lượng dự đoán chính xác / tổng số mẫu cần dự đoán

Precision = TP/(TP + FP)

Recall = TP/(TP + FN)

F1-score = 2 \* Precision \* Recall/( Precision + Recall)

AUC được tính bằng diện tích của phần dưới đường cong ROC trong đồ thị có trục tung là TPR và trục hoành là FPR. Giá trị của 2 trục chạy từ 0->1 và TPR+FPR=1

# Chương 4: Thực nghiệm

## 4.1 Yêu cầu về chương trình

Ngôn ngữ lập trình, phiên bản: Ngôn ngữ lập trình python phiên bản 3.10.3

Thư viện, phiên bản: Numpy – 1.22.3, Pandas – 1.4.2, Matplotlib – 3.5.1, Seaborn – 0.11.2, scikit-learn – 1.0.2

Cấu hình máy tính:Intel core i5 8300H, HDD 5400rpm 1T , ram 8gb 2667hz

## 4.2 Mô tả chi tiết các bước khai phá dữ liệu

**Bước 1: Chuẩn bị dữ liệu**

**Giải thích:**Do sử dụng bộ dữ liệu được người khác thu thập và xử lý cho nên ta không cần thực hiện bước này.

**Input**: không có

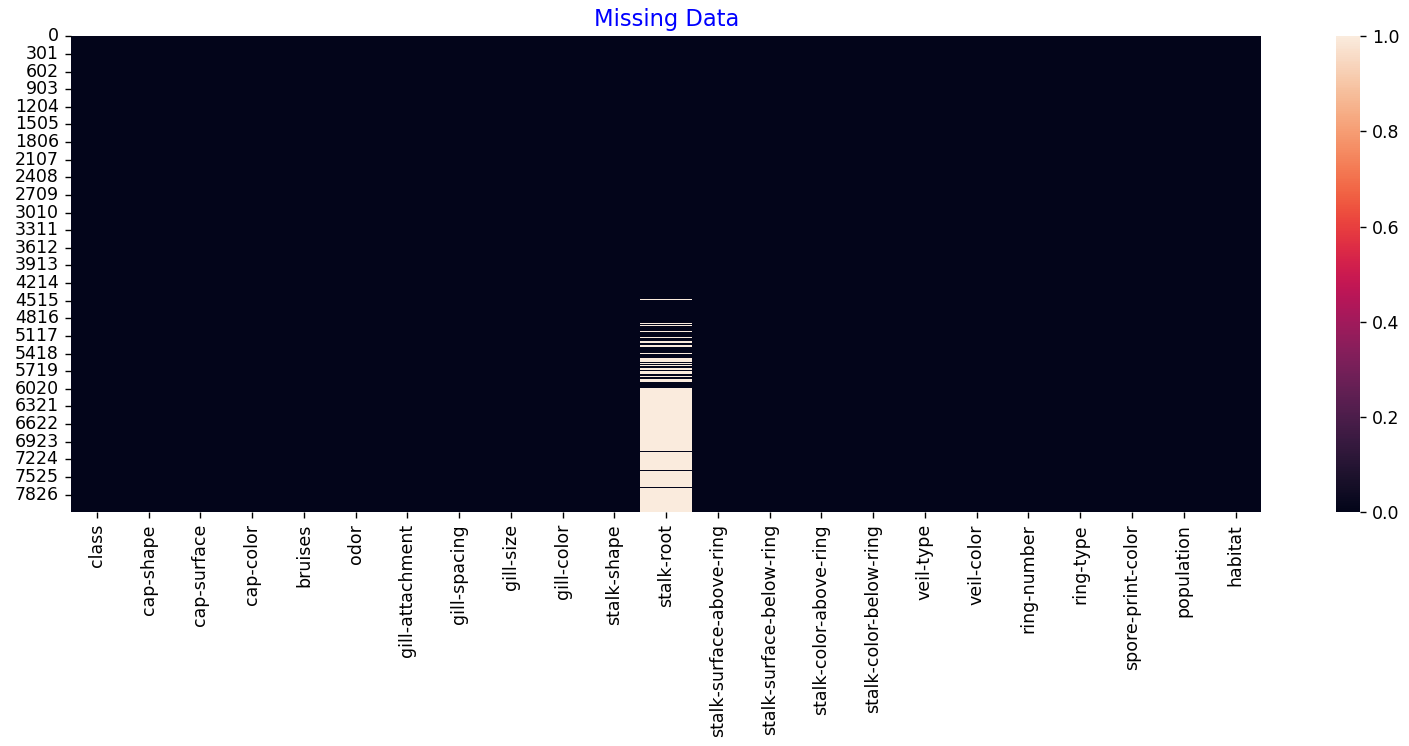
**Output:** Bộ dữ liệu đã thu thập và loại bỏ những dữ liệu trùng lặp, sai định dạng, không chính xác hay đầy đủ về mặt thông tin.

**Bước 2: Tiền xử lý dữ liệu**

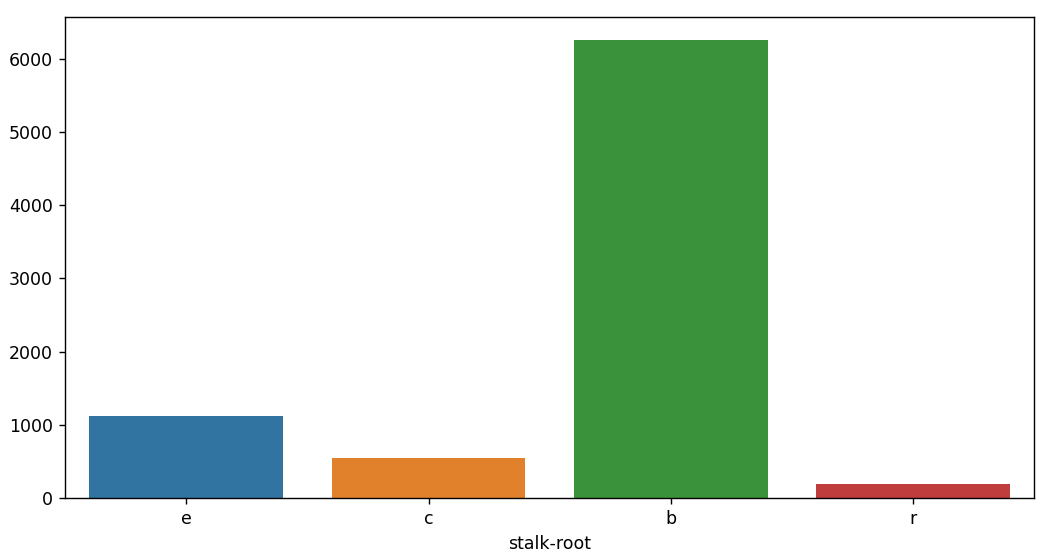
**1. Làm sạch dữ liệu**

**Giải thích:**Bằng cách thống kê ta phát hiện dữ liệu bị thiếu ở thuộc tính thứ 11 stalk-root và giá trị bị thiếu được điền giá trị “?” với tỉ lệ thiếu là 30.5%. Do tỉ lệ thiếu nhiều và chỉ thiếu ở 1 thuộc tính nên ta thực hiện điền giá trị bị thiếu.

**Input**:Bộ dữ liệu gốc có missing data



**Output**:Bộ dữ liệu đã chuyển đổi tất cả giá trị “?” của stalk-root thành b

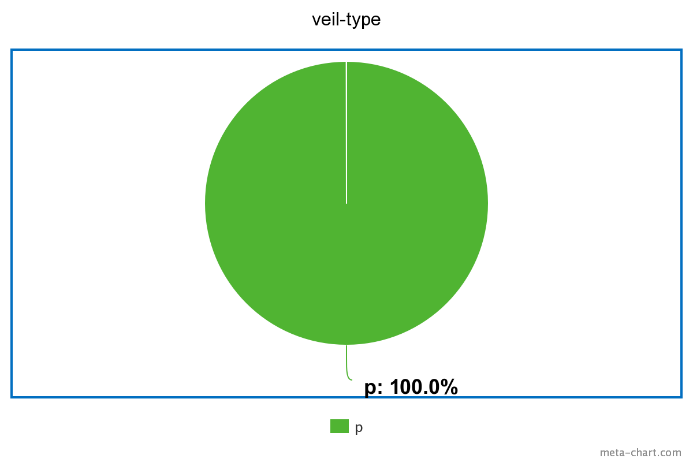


**2. tích hợp dữ liệu**:

**Giải thích:**Do chỉ có 1 bộ dữ liệu nên không thực hiện bước này

**3. Giảm dữ liệu(Data reduction):**

**Giải thích:** Thống kê ta thấy veil-type chỉ có 1 giá trị “p” nên ta có thể bỏ thuộc tính veil-type ra khỏi bộ dữ liệu.



**Input**: Bộ dữ liệu sau khi xử lý missing ở trên

**Output**:Bộ dữ liệu đã bỏ thuộc tính veil-type

**4. Mã hóa và Chuyển đổi dữ liệu(Data Transformation)**:

**Giải thích:**Toàn bộ thuộc tính đều có kiểu dữ liệu là category nên ta sử dụng labelencode chuyển đổi về dạng số tương ứng. ví dụ: có 4 giá trị a,b,c,d thì chuyển về dạng 0,1,2,3

Để tính toán ta cần chuẩn hoá dữ liệu. Do toàn bộ thuộc tính không tuân theo phân phối chuẩn nên ta chuẩn hoá nó về dạng 0->1.

**Input:** Bộ dữ liệu sau khi xử lý ở trên

**Output:** Bộ dữ liệu đã được chuẩn hóa về giá trị 0->1

**Bước 3: Chọn tác vụ, thuật toán khai phá dữ liệu và thực hiện**

**1. Chọn tác vụ, thuật toán:**

Ta đang thực hiện bài toán phân loại nấm nên tác vụ cần thực hiện là phân lớp(classification) để dự đoán nấm có độc hay ăn được.

Ở bài báo cáo này ta chọn thuật toán KNN để thực hiện khai phá dữ liệu

**2. Thực hiện khai phá dữ liệu**

Trước khi khai phá dữ liệu ta chia bộ dữ liệu gốc thành 2 nữa là tập training chiến 1/2 và testing chiếm 1/2 dữ liệu gốc.

Các bước thực hiện khai phá dữ liệu bằng thuật toán KNN:

**1.Tìm k, loại khoảng cách và cách tìm K hàng xóm tốt nhất.**

**Giải thích:**để thực hiện thuật toán KNN ta cần biết k hàng xóm gần nhất, loại khoảng cách muốn dùng và cách tìm k hàng xóm gần nhất. Tuy nhiên ta khó để biết chọn cái nào cho hợp lý, do đó ta thực hiện chạy thử tất cả các bộ lựa chọn và chọn ra lựa chọn có độ chính xác cao nhất dựa trên tập training truyền vào.

Trong python khi sử dụng thư viện scikit-learn thì KNN sẽ có các lựa chọn thường dùng:

Số lượng hàng xóm k: k thuộc khoảng k > 1

Loại khoảng cách: gồm có khoảng cách minkowski, manhattan, euclid, chebyshev

Các thuật toán tính k hàng xóm gần nhất: gồm có ball\_tree, kd\_tree, vét cạn

**Input:** Bộ dữ liệu training.

**Output:**k, loại khoảng cách, cách tính khoảng cách tốt nhất.

**2. thực hiện thuật toán KNN với các thông số trên**

**Input:** bộ dữ liệu dùng để training, bộ dữ liệu dùng để testing,k, loại khoảng cách, cách tính khoảng cách

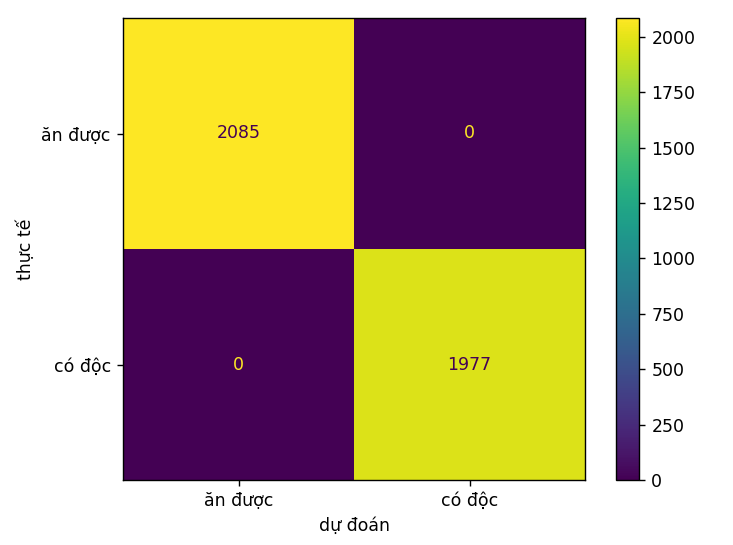
**Output:** Các chỉ số để đánh giá thuật toán: Accuracy, Precision, Recall, F1-score, AUC

# Chương 5: Kết quả

## 5.1 Kết quả khai phá dữ liệu

Tiêu chí đánh giá

tập dữ liệu kiểm thử được tách ra từ tập gốc với kích thước bằng 1/2 ta được:



Ta tính được

Accuracy = (2085+1977) / (2085+1977) \* 100 = 100%

Precision = 1977 / ( 1977 + 0) = 1

Recall = 1977 / (1977 + 0) = 1

F1-score = 2 \* Precision \* Recall/( Precision + Recall) = 2\*1\*1/(1+1) = 1

AUC = 1 do các chỉ số còn lại đều = 1

=> mô hình có độ tinh cậy 100% với tập dữ liệu đang xét.

1. Dựa trên tính giải thích được của mô hình (interpretability). Có thể giải thích tại sao mô hình ra kết quả chính xác được hay không?

## 5.2 So sánh kết quả thực tế với kết quả dự đoán

1. Chất lượng của dữ liệu

Nguyên lý hoạt động của thuật toán có giải quyết được vấn đề hay không?

Nguyên lý của thuật toán giải quyết được vấn đề đặt ra của bài toán.

# Chương 6: Thảo luận và kết luận

## 6.1 Khả năng ứng dụng của giải pháp/mô hình

Về thực tế, mô hình khó áp dụng cho người tiêu dùng bình thường hay các cửa hàng.Về người dùng bình thường, gần như khó xảy ra trường hợp ngộ độc ăn phải nấm nếu mua ở của hàng, siêu thị. Hơn thế nữa, có thể dùng các phương pháp thủ công dể dàng phát hiện nấm độc. Vì thế, mô hình có thể được ứng dụng để phân loại nấm trong môi trường học tập và rèn luyện khai phá dữ liệu.

=> khả năng ứng dụng của mô hình không cao, không có tính ứng dụng.

## 6.2 Ưu điểm – nhược điểm của giải pháp/mô hình

Ưu điểm của mô hình:

- Tỉ lệ chính xác cao

- Tốc độ dự đoán nhanh

Nhược điểm của mô hình:

- Dữ liệu đầu gồm 22 thuộc tính, gây khó khăn cho quá trình nhập thông tin về cây nấm.

- Chưa xử lý được những dữ liệu thực tế, những giá trị mới của thuộc tính nếu có sẽ gây ra sai sót trong khả năng phán đoán.

## 6.3 Đề xuất

Những đề xuất để cải tiến mô hình:

- Chuyển đổi dữ liệu đầu vào thành hình ảnh để tiện nhập liệu và phù hợp với công nghệ hiện nay.

- Thay đổi khởi tạo của bài toán bằng các thuộc tính khác nhau và kiểm tra, so sánh kết quả để chọn giải pháp tối ưu

- Dữ liệu cần lớn hơn nữa để đo đọ chính xác được chuẩn hơn.

-Thường xuyên biểu thị bằng biểu đồ để thể hiện sự tương quan rõ hơn

# Tài liệu tham khảo

|  |  |
| --- | --- |
| [1] | N. T. Hop, “KNN (K-Nearest Neighbors),” 16 7 2019. [Trực tuyến]. Available: https://viblo.asia/p/knn-k-nearest-neighbors-1-djeZ14ejKWz. [Đã truy cập 10 5 2022]. |
| [2] | E. STEVENS, “What Is Nominal Data,” 17 5 2021. [Trực tuyến]. Available: https://careerfoundry.com/en/blog/data-analytics/what-is-nominal-data/. [Đã truy cập 10 5 2022]. |
| [3] | H. Nguyen, “K-Nearest Neighbors (KNN),” 18 4 2021. [Trực tuyến]. Available: https://datasciencebasic.com/?p=134. [Đã truy cập 10 5 2022]. |
| [4] | B. M. Ramageri, “Data mining techniques and applications,” 12 2010. [Trực tuyến]. Available: https://www.researchgate.net/publication/49616224\_Data\_mining\_techniques\_and\_applications. [Đã truy cập 10 5 2022]. |
| [5] | purva91, “Precision vs. Recall – An Intuitive Guide for Every Machine Learning Person,” 4 9 2020. [Trực tuyến]. Available: https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/. [Đã truy cập 10 5 2022]. |