# **1.1.** **Mô tả tập dữ liệu**

Tập dữ liệu gồm 8416 bản ghi với 23 cột dữ liệu

Bộ dữ liệu này bao gồm các mô tả về các mẫu giả định tương ứng với 23 loài nấm trong họ Agaricus và Lepiota. Mỗi loài được xác định là chắc chắn ăn được, chắc chắn có độc. Hướng dẫn nêu rõ rằng không có quy tắc đơn giản nào để xác định khả năng ăn được của nấm.

Các thuộc tính được mô tả như dưới đây

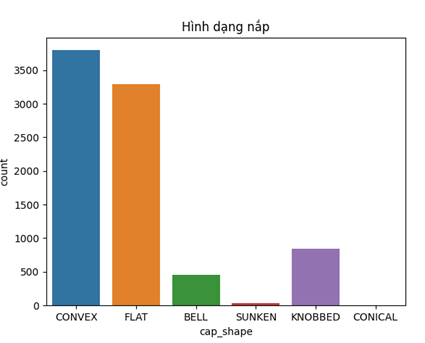
|  |  |  |
| --- | --- | --- |
| **STT** | **Tên** | **Mô tả** |
| 1 | cap\_shape | Hình dạng nắp: (convex:lồi lõm ,flat: phẳng , knobbed: có núm ,bell: chuông , sunken: sâu thỏm , conical: hình nón ) |
| 2 | cap\_surface | Bề mặt nắp (scaly: có vảy , smooth: trơn tru , fibrous: dạng sợi , grooves: rãnh) |
| 3 | cap\_color | Màu nắp (brown: nâu, gray: xám, red:đỏ, yellow: vàng, white: trắng, buff: màu da, pink: hồng, cinnamon: quế, purple: tím, green:xanh lá) |
| 4 | bruises | Vết đốm (no: không có, bruises: có vết đốm) |
| 5 | odor | Mùi (none: không mùi, foul: hôi, spicy: cay, fishy: tanh, almond: mùi hạnh nhân, anise: mùi hồi, pungent: mùi hăng, creosote: mùi dầu creosote, musty: mùi mốc) |
| 6 | gill\_attachment | Chẻ nấm đính kèm (free: không, attached: có đính kèm) |
| 7 | gill\_spacing | Khoảng cách chẻ nấm (close: gần, crowded: chen chúc) |
| 8 | gill\_size | Kích thước chẻ nấm (broad: rộng lớn, narrow: hẹp) |
| 9 | gill\_color | Màu sắc chẻ nấm (buff: màu da, pink: hồng, white: trắng, brown: nâu, chocolate: màu socola, gray: xám tro, purple: tím, black: đen, red: đỏ, yellow: vàng, orange: cam, green: xanh lá) |
| 10 | stalk\_shape | Hình dạng cuống (tapering: hình nón, enlargin: mở rộng) |
| 11 | stalk\_root | Gốc thân cây (bulbous: có hình củ hành, club: chùm, rooted: vững chắc, equal: cân đều nhau) |
| 12 | stalk\_surface\_above\_ring | Bề mặt cuống trên vòng (smooth: trơn tru, silky: mướt, fibrous: có sợi, thớ; scaly: có vảy) |
| 13 | stalk\_surface\_below\_ring | Bề mặt cuống bên dưới vòng (smooth: trơn tru, silky: mướt, fibrous: có sợi, thớ; scaly: có vảy) |
| 14 | stalk\_color\_below\_ring | Màu cuống bên dưới vòng (buff: màu da, pink: hồng, white: trắng, brown: nâu, black: đen, red: đỏ, yellow: vàng, orange: cam, green: xanh lá) |
| 15 | veil\_type | Loại mạng che mặt: từng phần |
| 16 | veil\_color | Màu màn che (white: trắng, brown: nâu, orange: cam, yellow: vàng) |
| 17 | ring\_number | Số lượng vòng (one: một, two: hai, none: không có) |
| 18 | ring\_type | Kiểu vòng (pendant: bền bỉ, evanescent: chóng phai, large: lớn, flaring: hoa hồng, none: không) |
| 19 | spore\_print\_color | Màu bào tử (buff: màu da, pink: hồng, white: trắng, brown: nâu, chocolate: màu socola, gray: xám tro, purple: tím, black: đen, red: đỏ, yellow: vàng, orange: cam, green: xanh lá) |
| 20 | population | Mật độ (several: riêng biệt, cá nhân; solitary: cô độc;  Scattered: rải rác, abundant: nhiều, numerous: rất nhiều, clustered: nhóm thành cụm) |
| 21 | habitat | Môi trường sống (woods: thân gỗ, grasses: bãi cỏ, paths: lối đi, leaves: lá, urban:đô thị, meadows: đồng cỏ, waste: đất bỏ hoang) |
| 22 | mushroom | Phân loại nấm (edible: ăn được, poisonous: độc) |

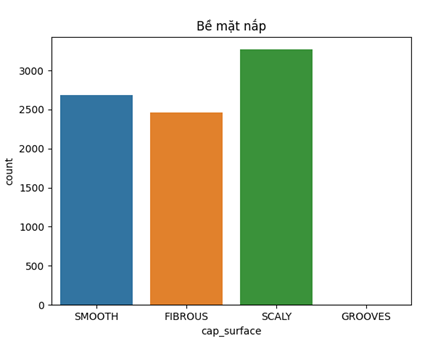
# **1.2. Mô tả bài toán:**

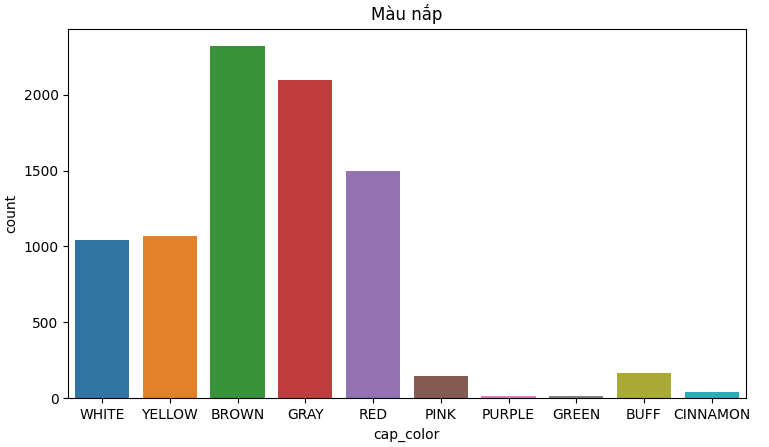
- Bài toán phân lớp nhị phân dựa vào đặc tính của loài nấm để phân lớp nấm ăn được hay có độc.

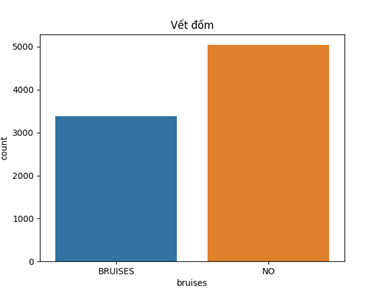
- Kỹ thuật sử dụng: Logistic Regression

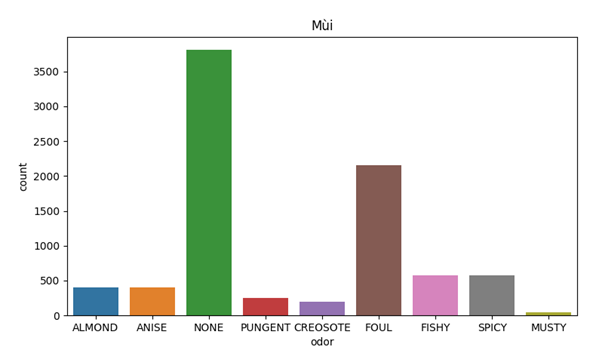
## - Đầu vào bài toán: 21 thuộc tính đầu tiên mô tả về đặc tính của loài nấm như ở bảng trên(cap\_shape, cap\_surface, cap\_color, bruises, odor, gill\_attachment, gill\_spacing, gill\_size, gill\_color, stalk\_shape, stalk\_root, stalk\_surface\_above\_ring, stalk\_surface\_below\_ring, stalk\_color\_below\_ring, veil\_type, veil\_type, veil\_color, ring\_number, ring\_type, spore\_print\_color, population, habitat). Dưới đây là hình biểu diễn giá trị của một vài thuộc tính tiêu biểu.



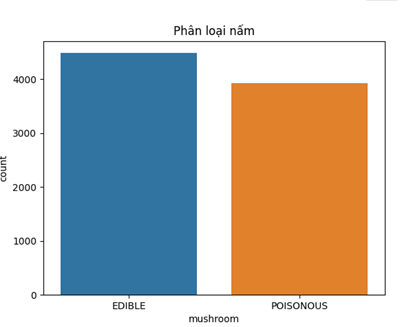








## - Đầu ra: nhãn ‘mushroom’ phân loại ăn được và độc.



# **1.3.** **Tiền xử lý dữ liệu**

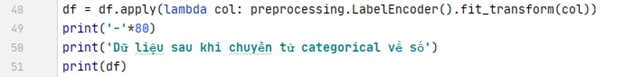
- Xóa dữ liệu NAN



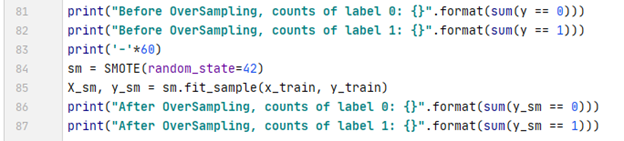
- Xóa cột stalk\_root do có nhiều giá trị bị thiếu.



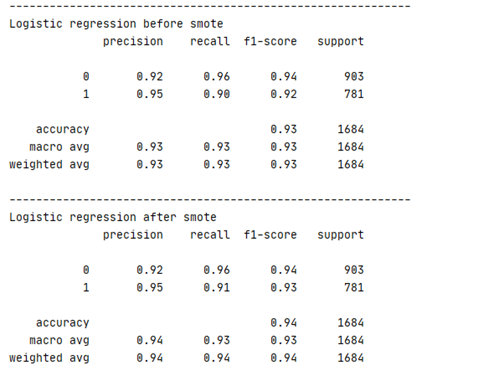
- Dùng LabelEncoder để mã hóa nhãn dữ liệu từ categorical về dạng số.



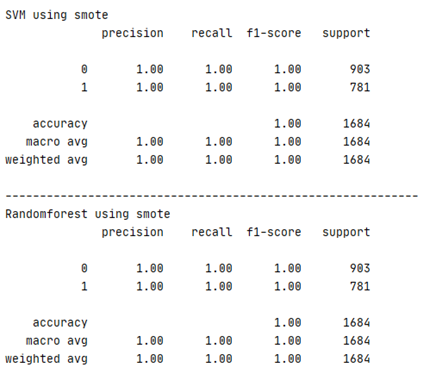
- Dùng smote để xử lý mất cân bằng dữ liệu (do dữ liệu khá đẹp, gần như cân bằng, nên việc xử lý mất cân bằng với bộ dữ liệu này không có ảnh hưởng nhiều đến kết quả đầu ra)

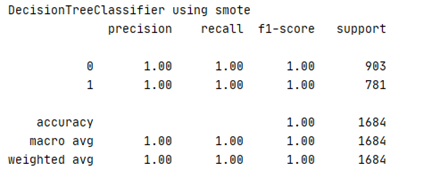


- Kết quả chương trình sau khi sử dụng kỹ thuật logistic regression



- So sánh kết quả khi chạy với một số kỹ thuật phân lớp khác





# **1.4.** **Kết luận**

Logistic Regression chạy khá nhanh và chuẩn xác, tuy nhiên nhược điểm của nó cũng như các thuật toán phân lớp khác như SVM, DecisionTreeClassifier, Random Forest là rất nhạy cảm với dữ liệu mất cân bằng. Kết quả khi chạy có thể rất đẹp nhưng mô hình dự đoán thường không đáng tin cậy. Chính vì thế, để đảm bảo dữ liệu được huấn luyện tốt và cho ra dự đoán chính xác, cần tiến hành xử lý mất cân bằng ngay từ khâu ban đầu trước khi chạy mô hình dự đoán kết quả.