Mushroom Classification

109901014 楊芷紜, 109304013張維蓁

一、目的

在世界各地, 誤食蘑菇造成食物中毒的事件層出不窮;輕則導致頭暈昏厥, 重者甚至會有休克的情形發生, 現今仍未有明確辨別毒蘑菇的方法, 因此我們想設計出能夠精準預測蘑菇是否能食用的模型。過程中會將dataset的資料丟入模型訓練後, 再進行是否有毒性的分類, 最後計算出模型的準確率, 並繪製confusion matrix, 讓結果更直觀。
同時也會針對不同模型進行比較, 觀察何者辨別的正確率較高, 並試著找出毒蘑菇的重要特徵。

二、Dataset

從Kaggle上取得"Mushroom Attributes" dataset, 如下圖可見, dataset中共有8124個samples, 每個sample都包含了22種特徵與1種ground truth (class), 而資料的類型為字串。

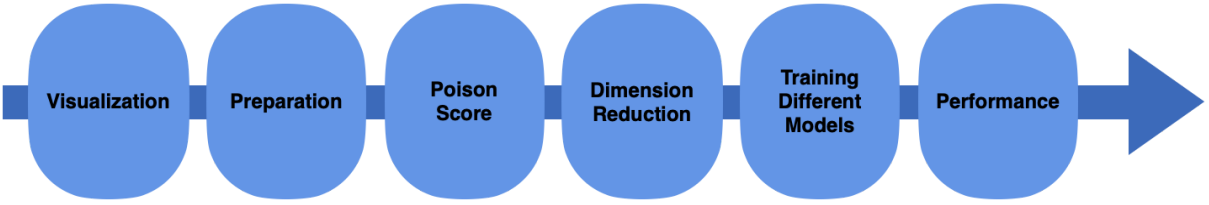
	cap-sh	cap-su	cap-col	bruises	odor	gill-att	gill-spa	gill-size	gill-col	stalk-sl	stalk-r	stalk-st	stalk-st	stalk-c	stalk-c	veil-typ	veil-col	ring-nu	ring-ty	spore-p	popula	habita	class
0	b'x'	b's'	b'n'	b't'	b'p'	b'f'	b'c'	b'n'	b'k'	b'e'	b'e'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'k'	b's'	b'u'	b'p'
1	b'x'	b's'	b'y'	b't'	b'a'	b'f'	b'c'	b'b'	b'k'	b'e'	b'c'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'n'	b'n'	b'g'	b'e'
2	b'b'	b's'	b'w'	b't'	b'l'	b'f'	b'c'	b'b'	b'n'	b'e'	b'c'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'n'	b'n'	b'm'	b'e'
3	b'x'	b'y'	b'w'	b't'	b'p'	b'f'	b'c'	b'n'	b'n'	b'e'	b'e'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'k'	b's'	b'u'	b'p'
4	b'x'	b's'	b'g'	b'f'	b'n'	b'f'	b'w'	b'b'	b'k'	b't'	b'e'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'e'	b'n'	b'a'	b'g'	b'e'
5	b'x'	b'y'	b'y'	b't'	b'a'	b'f'	b'c'	b'b'	b'n'	b'e'	b'c'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'k'	b'n'	b'g'	b'e'
6	b'b'	b's'	b'w'	b't'	b'a'	b'f'	b'c'	b'b'	b'g'	b'e'	b'c'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'k'	b'n'	b'm'	b'e'
7	b'b'	b'y'	b'w'	b't'	b'l'	b'f'	b'c'	b'b'	b'n'	b'e'	b'c'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'n'	b's'	b'm'	b'e'
8	b'x'	b'y'	b'w'	b't'	b'p'	b'f'	b'c'	b'n'	b'p'	b'e'	b'e'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'k'	b'v'	b'g'	b'p'
9	b'b'	b's'	b'y'	b't'	b'a'	b'f'	b'c'	b'b'	b'g'	b'e'	b'c'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'k'	b's'	b'm'	b'e'
10	b'x'	b'y'	b'y'	b't'	b'l'	b'f'	b'c'	b'b'	b'g'	b'e'	b'c'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'n'	b'n'	b'g'	b'e'
11	b'x'	b'y'	b'y'	b't'	b'a'	b'f'	b'c'	b'b'	b'n'	b'e'	b'c'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'k'	b's'	b'm'	b'e'
12	b'b'	b's'	b'y'	b't'	b'a'	b'f'	b'c'	b'b'	b'w'	b'e'	b'c'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'n'	b's'	b'g'	b'e'
13	b'x'	b'y'	b'w'	b't'	b'p'	b'f'	b'c'	b'n'	b'k'	b'e'	b'e'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'n'	b'v'	b'u'	b'p'
14	b'x'	b'f'	b'n'	b'f'	b'n'	b'f'	b'w'	b'b'	b'n'	b't'	b'e'	b's'	b'f'	b'w'	b'w'	b'p'	b'w'	b'o'	b'e'	b'k'	b'a'	b'g'	b'e'
15	b's'	b'f'	b'g'	b'f'	b'n'	b'f'	b'c'	b'n'	b'k'	b'e'	b'e'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'n'	b'y'	b'u'	b'e'
16	b'f'	b'f'	b'w'	b'f'	b'n'	b'f'	b'w'	b'b'	b'k'	b't'	b'e'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'e'	b'n'	b'a'	b'g'	b'e'
17	b'x'	b's'	b'n'	b't'	b'p'	b'f'	b'c'	b'n'	b'n'	b'e'	b'e'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'k'	b's'	b'g'	b'p'
18	b'x'	b'y'	b'w'	b't'	b'p'	b'f'	b'c'	b'n'	b'n'	b'e'	b'e'	b's'	b's'	b'w'	b'w'	b'p'	b'w'	b'o'	b'p'	b'n'	b's'	b'u'	b'p'

Dataset details

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
bruises%3F	bruises=t, no=f
odor	almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
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
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
veil-type	partial=p, universal=u
veil-color	brown=n, orange=o, white=w, yellow=y
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
spore-print-color	black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
population	abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
habitat	grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d
class	edible=e, poisonous=p

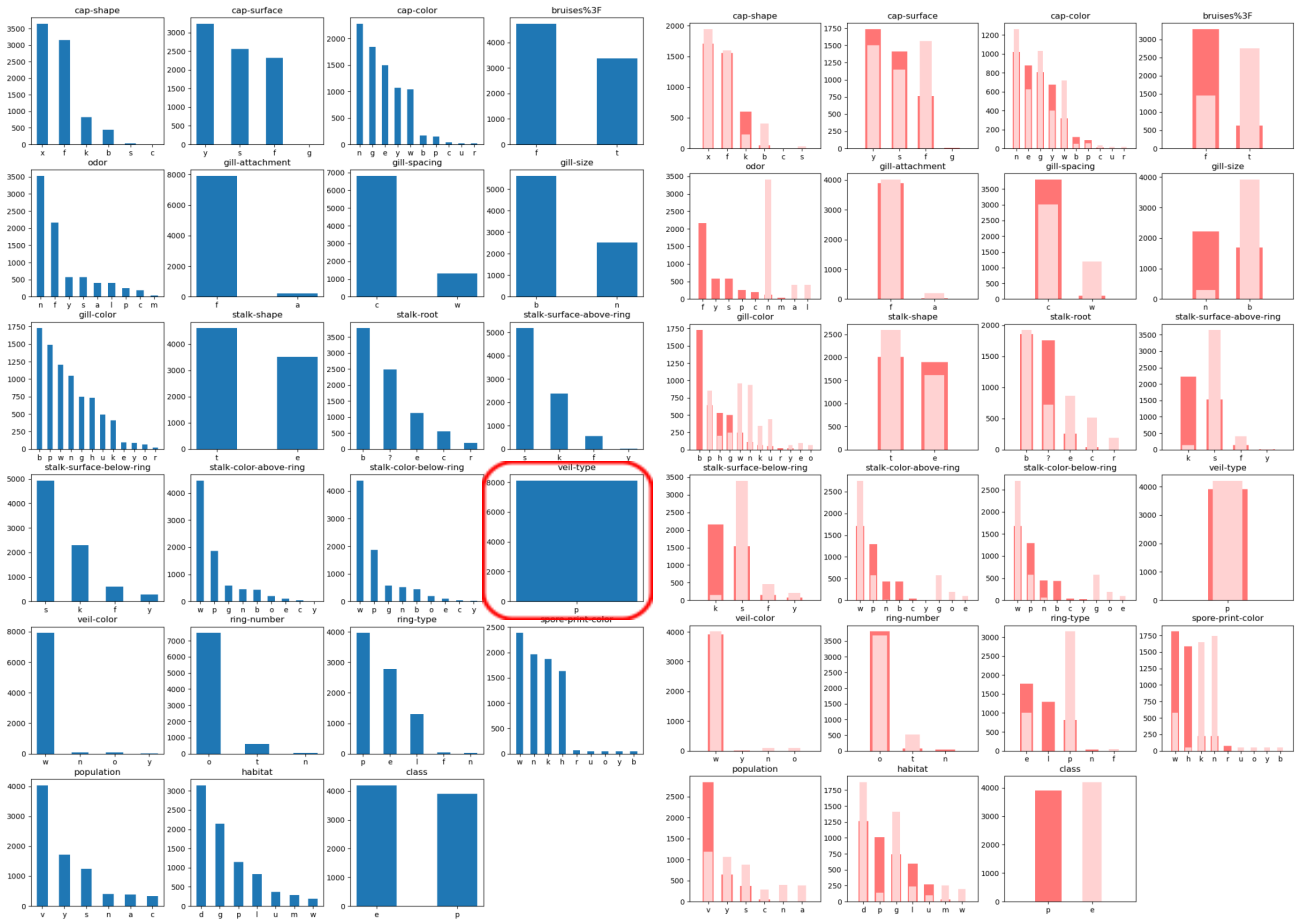
三、流程



四、方法

- Visualization:

為了確認data的分布情形，之後進行了data visualization，左邊為所有特徵的分布圖，右圖的深色柱為毒蘑菇，淺色為可食用蘑菇，進行兩者特徵之間分布的比較，從圖中可以發現到veil-type只有一種類型，因此之後會移除該特徵。



- Preparation:

1. 移除對模型的訓練沒有幫助的unnessasary information, 像是多餘的字串「b' '」, 或者是只有一個種類的特徵(veil-type)。
2. 利用train test split將8124個samples隨機分為training(70%, 共5686個samples)與testing data(30%, 共2438個samples), 並將class當作label。
3. 由於我們的features都是字串, 而非數值, 因此需要用onehot encoder將這些字串轉為數字。每個特徵(feature)都有數個種類(type), 去掉ground truth和veil-type後的所有特徵種類數即是欄數(n_column = 116)。值得注意的是, 這些數值之間沒有連續性, 也沒有相關性。此外, 我們會確定training data與testing data都有116欄, 才會繼續接下來的步驟。

- Poison score:計算各個特徵類別與毒蘑菇的相關程度, 並將分數進行排序, 篩選出重要的特徵種類, 公式如下。

$$\text{poison score} = \text{特徵各類別樣本數} - \frac{\text{poison sample 數}}{\text{特徵類別數量}}$$

- Dimension reduction (Kernel PCA): Onehot encoding完後的data型態是sparse matrix, 將其轉換成dense matrix的形式後, 由於我們的資料是非線性的, 因此我們無法用傳統的PCA降維, 而是需要用kernal PCA。我們分別降到2, 5, 10, 15維, 並使用”param_grid”找出最好的hyperparameter, 想看不同維度對模型表現的影響。

- Training different models (Logistic Regression, KNN, SVC, Naive Bayes, Decision Tree, Random Forest)：用不同模型訓練data，觀察彼此之間的表現差異。

五、結果

- TOP 10 Features

以poison score的分數進行排序，發現veil-color_w是與毒蘑菇最有關聯性的特徵類別。

	n
veil-color_w	2028.250000
ring-number_o	1726.666667
population_v	1524.833333
gill-attachment_f	1339.500000
gill-spacing_c	1279.500000
odor_f	1171.222222
gill-color_b	962.416667
spore-print-color_w	944.222222
stalk-color-above-ring_w	921.222222
stalk-color-below-ring_w	903.222222

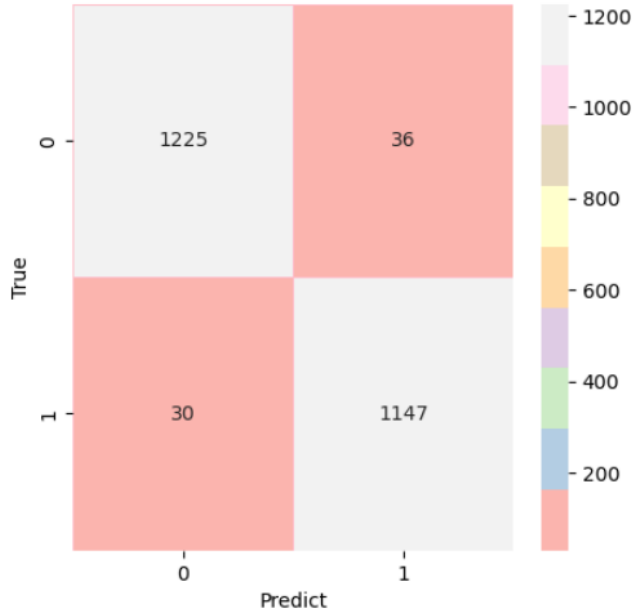
- Classification performance between models

以模型的角度來看，可以清楚看到KNN model的表現最好，即便特徵維度降到2維也還有將近9成的準確率。

降低維度的意義是精簡代入模型的特徵數，希望藉此增加處理效率。從我們的結果可以看到，若擷取少量特徵進行分類，呈現出的結果並不是太好，因此判斷蘑菇是否有毒，不是單一或者少數的特徵類別能夠決定的。綜合來說，我們認為將維度調整至10維是最理想的參數；在n=10時，用最精簡的特徵類別數，幾乎所有的模型都能達到它們最好的準確率，是cp值最高的決定。

模型/維度	n=2	n=5	n=10	n=15
Logistic Regression	87.49%	95.00%	97.29%	97.37%
KNN	89.29%	99.96%	100.00%	100.00%
SVC	87.33%	94.34%	95.37%	95.45%
Naive Bayes	48.28%	89.25%	92.53%	93.07%
Decision Tree	57.88%	89.13%	96.55%	95.45%
Random Forest	49.43%	92.04%	95.69%	94.83%

- Confusion Matrix



Dim-10

Logistic Regression

Accuracy: 97.29%

Specificity: 0.9745

Recall/Sensitivity: 0.9714

六、討論

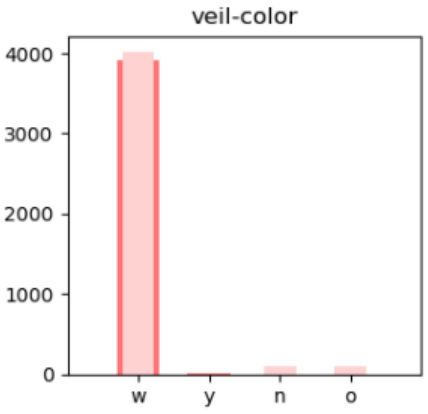
- Is poison score enough to evaluate?

→ 單看poison score會產生偏頗，該類別可能與可食用蘑菇的相關性也很高，不足以篩選出重要特徵類別，因此加入edible score計算平均值，可能會更好的找到分辨毒蘑菇相關的features。

我們使用與計算poison score相同的方式，算出edible score，將poison score與edible score取平均值，公式如下。在進行排序之後發現與毒蘑菇最為相關的特徵類別是odor_f，可以看到原本的veil-color_w已不在前10項重要的特徵中，從先前的data visualization可以推測因edible/posionous的蘑菇都在veil-color_w中占極大的比例，因此調整過後的方法才能較正確的篩選出重要特徵。

$$score = \frac{(poison\ score - edible\ score)}{2}$$

	n
odor_f	1096.222222
stalk-surface-above-ring_k	1078.500000
stalk-surface-below-ring_k	1044.500000
gill-size_n	1041.000000
bruises%3F_f	991.000000
gill-color_b	876.166667
population_v	852.333333
spore-print-color_h	784.222222
ring-type_l	677.200000
spore-print-color_w	634.222222



- 因為**Data**不平均，所以在切分**test, train**時可能會造成兩邊特徵種類數目不同，在訓練模型時會因為欄位數不同而出現錯誤。

→ 套用stratified sampling可以解決這個問題。

- 特徵維度調整到**n=15**時，**Random forest** 和 **Decision tree**的表現反而變差。
→ 我們猜測是因為在15維度時，帶入tree的特徵變多，會導致depth變深，所以結果反而較dim-10更差了。

- 調整參數？
→ 我們在做project時並沒有調整太多模型參數，若能fine tuning 調整某些parameters應該能讓原本表現不是很好的模型更進步，此外我們在訓練模型時也沒有做k-fold validation這個步驟，即使test accuracy成果不錯，但模型訓練效果不得而知。

- 應用：
加入特徵擷取技術再開發app，能夠讓這項辨別技術更易取得、更實用。不需大費周章地開電腦，只要打開手機app、拍照掃描，就能用來即時判斷蘑菇的毒性與否，希望以此減少因為誤食蘑菇造成的死傷人數。

七、資料來源

1. https://www.kaggle.com/datasets/ulrikthygepedersen/mushroom-attributes
2. https://www.kaggle.com/code/turksoyomer/classification-methods-on-mushroom-dataset/notebook#3.-Manipulating-Data
3. https://www.kaggle.com/code/adityapatil673/classification-traits-of-a-poisonous-mushroom#For-more-clarity-on-parts-of-a-mushroom
4. https://github.com/kanchitank/Mushroom-Classification/blob/master/Mushroom-Classification.ipynb