Mushroom Database

Created by UCI

Victoire Linder & Charles Benizri

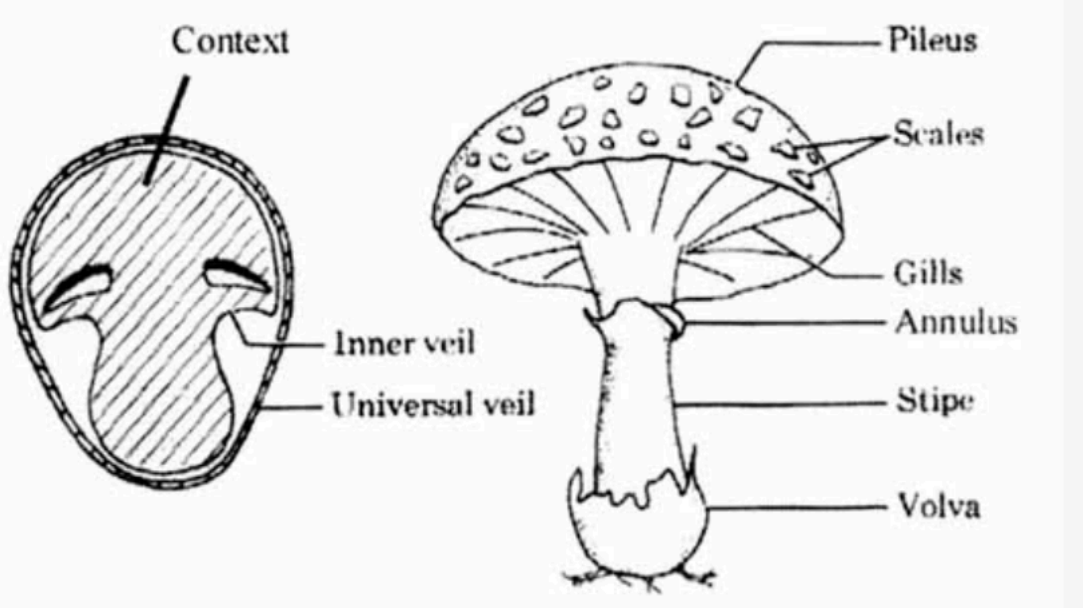
Introduction

- Nous allons explorer et analyser la base de donnée de « UCI Machine Learning » qui permet de classifier des champignons.

- Le but est de classifier ces champignons pour savoir s'ils sont comestibles (e) ou empoisonnés (p)

- Commençons par analyser cette data, puis nous allons la classifier en essayant d'avoir la meilleure précision possible.

Data

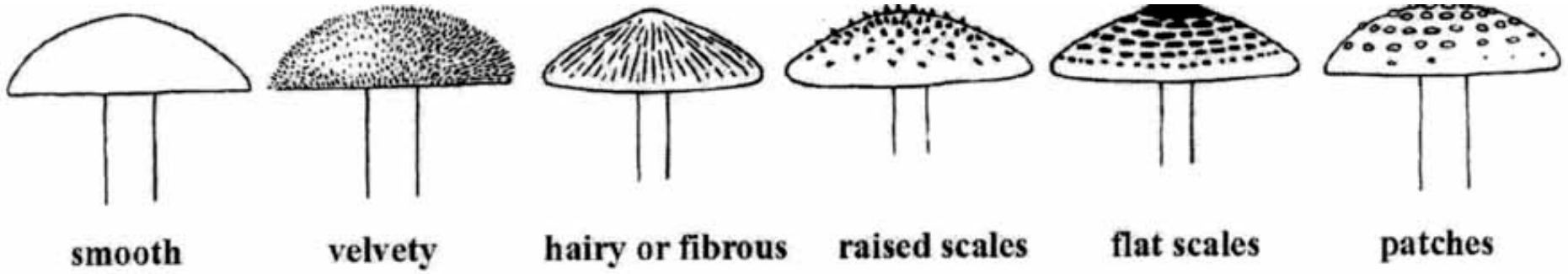


Les différentes colonnes représentent les informations des champignons.

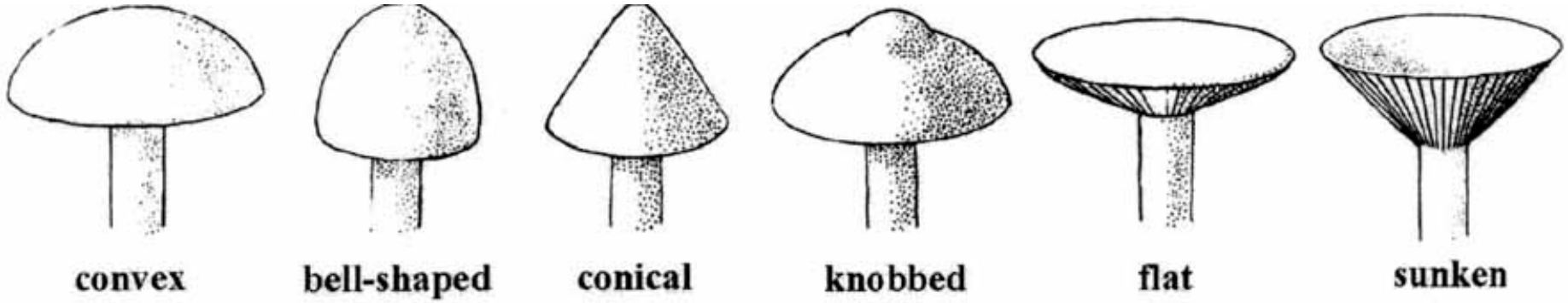
Sa taille, sa forme, son odeur, sa texture…

Nous avons 23 colonnes et 8124 lignes.

Mushroom cap surface:



Mushroom cap shape:



Notre Dataset est faite de caractères

	target	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-above-ring	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-color	ring-number	ring-type	spore-print-color	po
0	p	x	s	n	t	p	f	c	n	k	...	s	s	w	w	w	o	p	k	
1	e	x	s	y	t	a	f	c	b	k	...	s	s	w	w	w	o	p	n	
2	e	b	s	w	t	l	f	c	b	n	...	s	s	w	w	w	o	p	n	
3	p	x	y	w	t	p	f	c	n	n	...	s	s	w	w	w	o	p	k	
4	e	x	s	g	f	n	f	w	b	k	...	s	s	w	w	w	o	e	n	

5 rows × 22 columns

On utilise un label encoder pour transformer ces lettres en chiffres. On associe un entier à une valeur.

	target	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type	spore-print-color	population
0	1	5	2	4	1	6	1	0	1	4	...	2	7	7	0	2	1	4	2	3
1	0	5	2	9	1	0	1	0	0	4	...	2	7	7	0	2	1	4	3	2
2	0	0	2	8	1	3	1	0	0	5	...	2	7	7	0	2	1	4	3	2
3	1	5	3	8	1	6	1	0	1	5	...	2	7	7	0	2	1	4	2	3
4	0	5	2	3	0	5	1	1	0	4	...	2	7	7	0	2	1	0	3	0

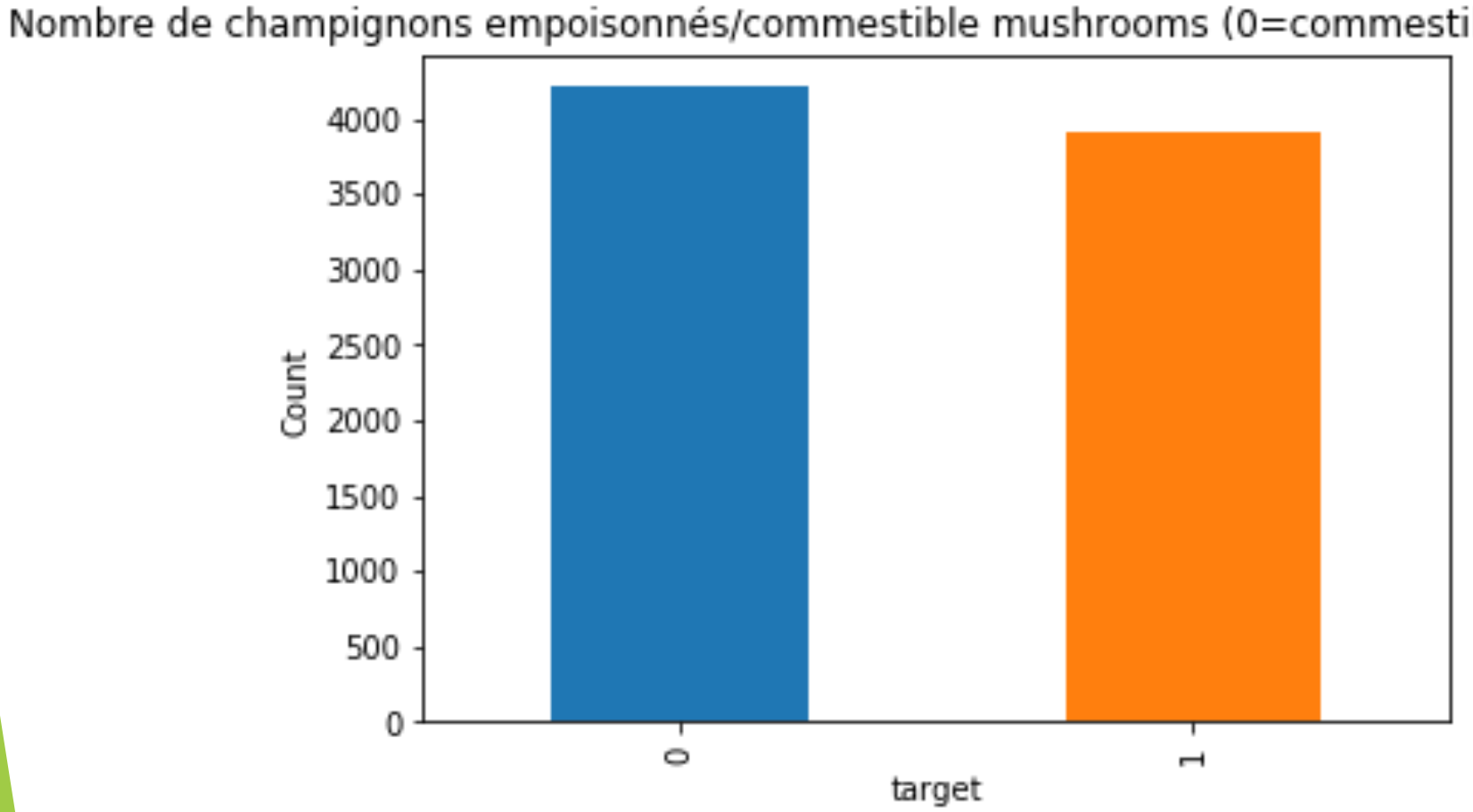
Data Cleaning

	veil-type	
0	8124.0	812
2	0.0	
1	0.0	
0	0.0	
0	0.0	
0	0.0	
0	0.0	
0	0.0	

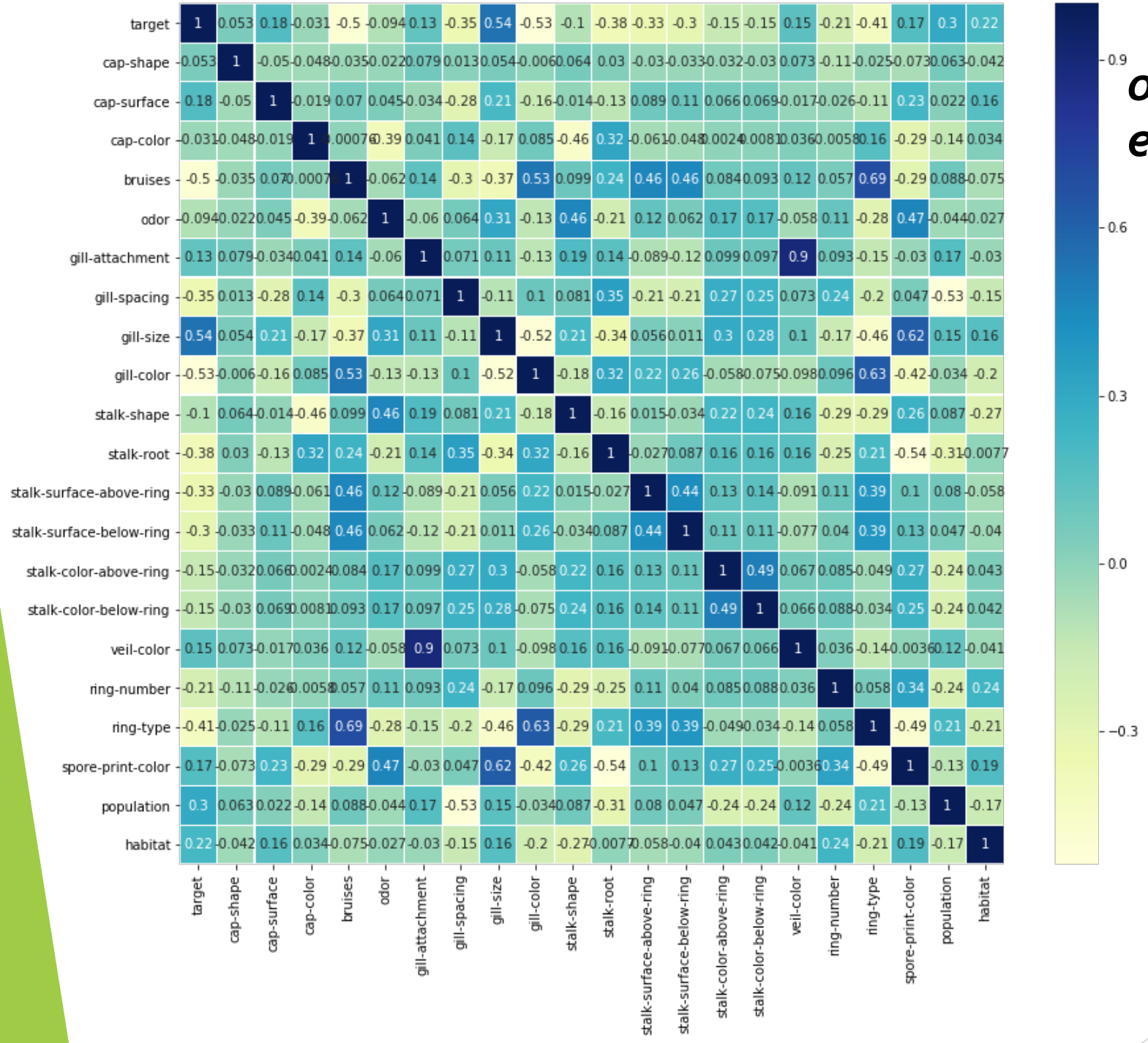
On va supprimer la colonne Veil-type car elle n'a aucun intérêt pour nos prédictions

```
b     3776
?     2480
e     1120
c      556
r      192
Name: stalk-root
```

Stalk Root contient 2480 « ? »
On va donc se débarrasser de ces valeurs

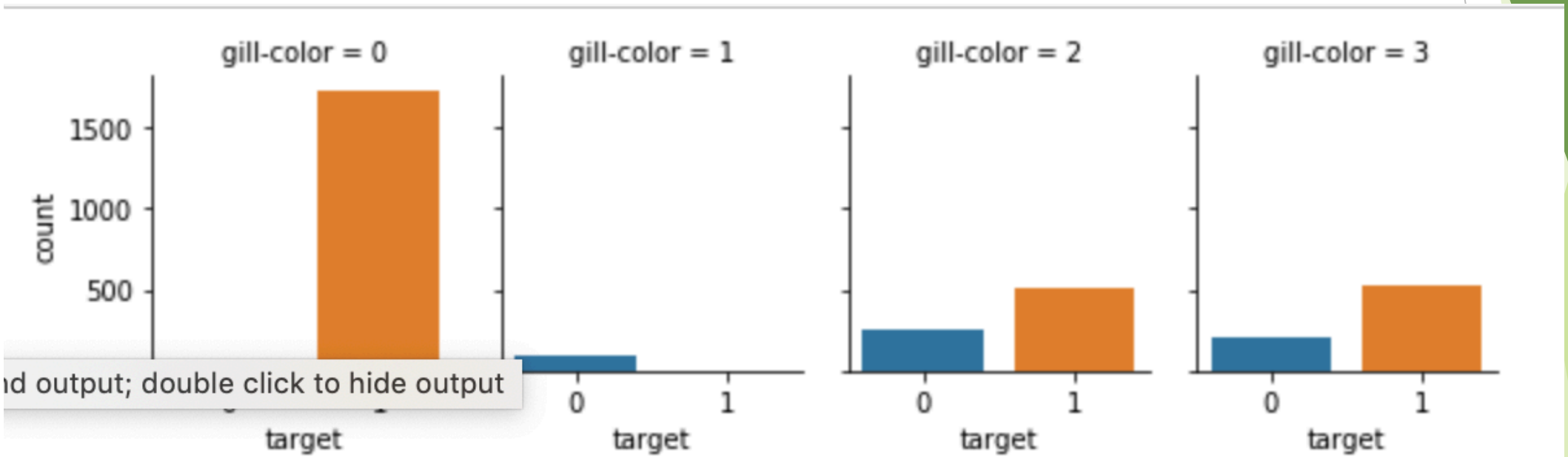


On remarque que la répartition de champignon empoisonnés /comestibles est homogène.
C'est important pour avoir un modèle d'entrainement non biaisé.



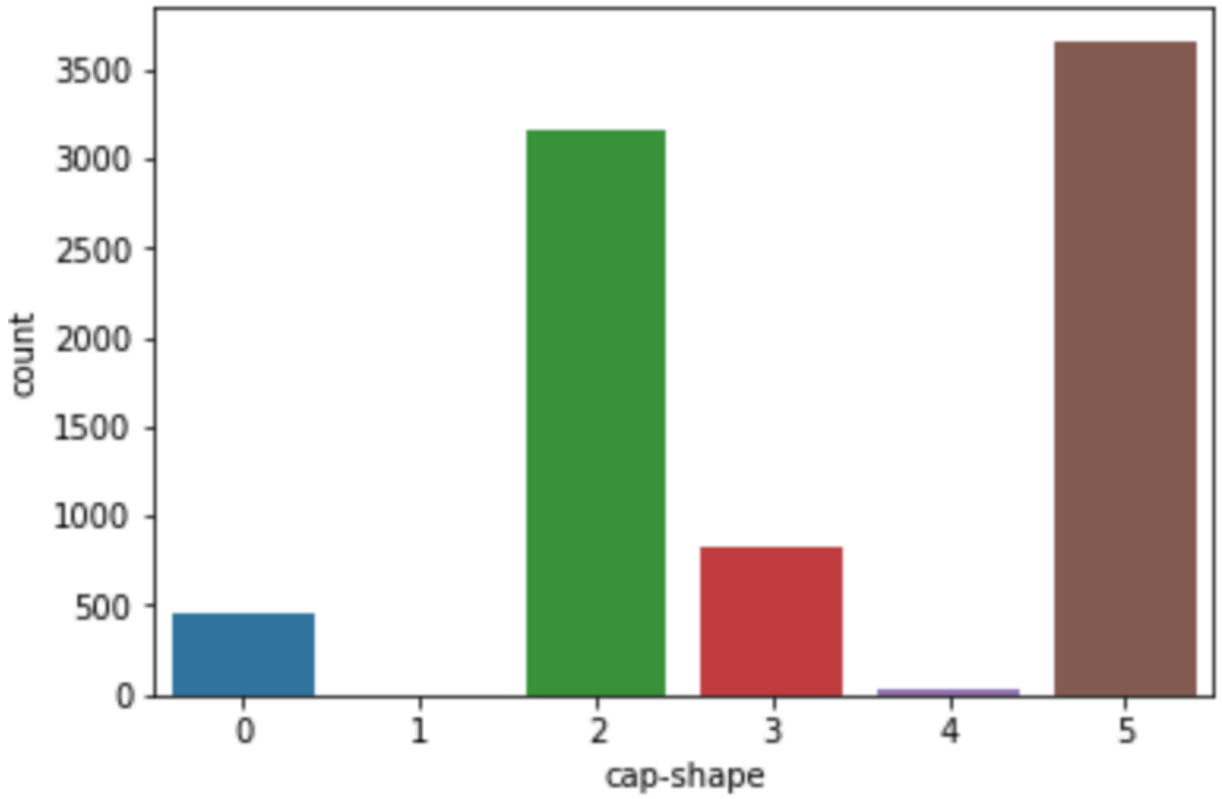
on remarque que gill color est la moins corrélé avec "-0.53"

Observons donc Gill-color



Cette variable n’aura pas grand intérêt dans notre algorithme

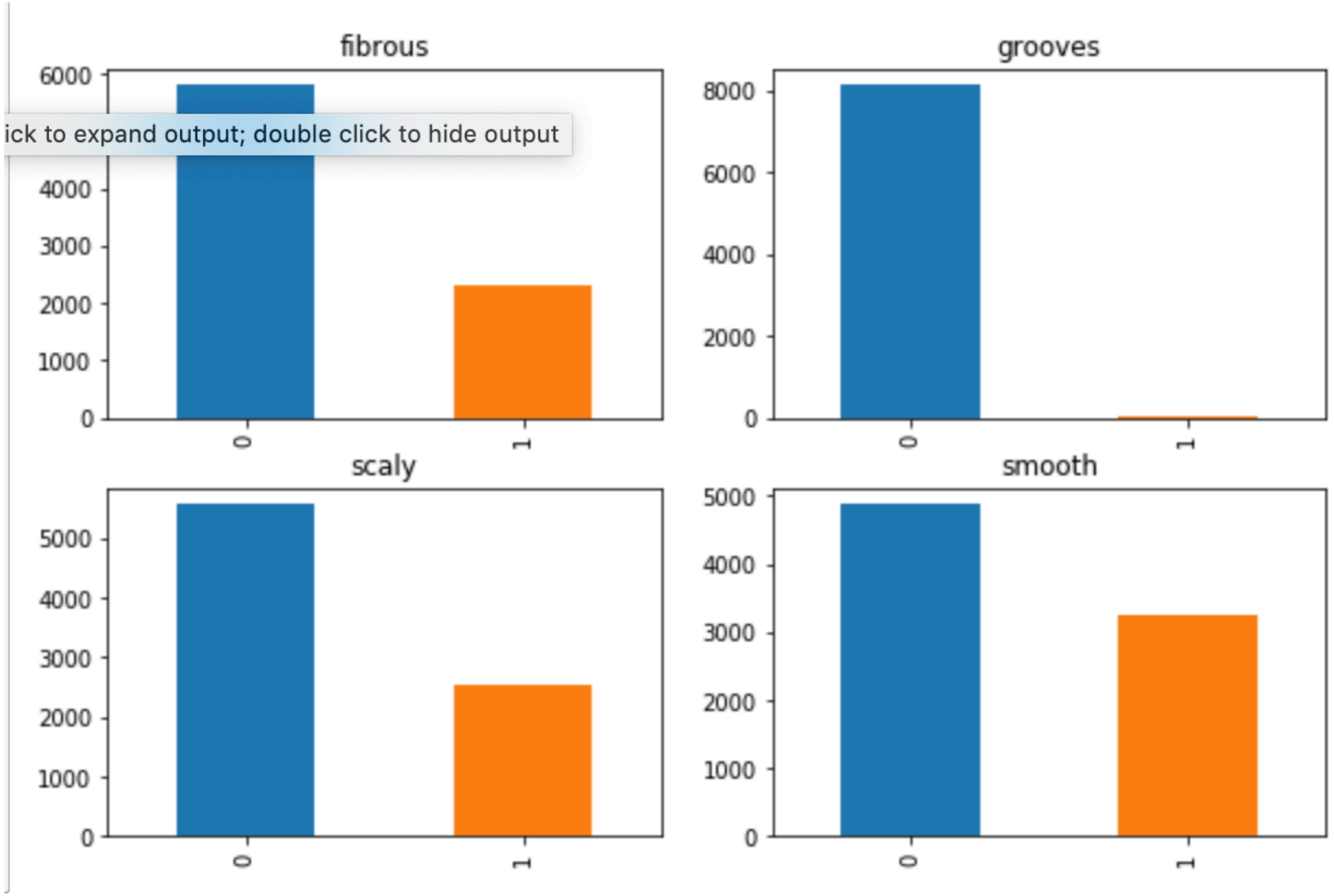
Répartition des formes de champignons



cap-shape	
0	452
1	4
2	3152
3	828
4	32
5	3656

3656 champignons sont de forme convex, suivit de près par les champignons plats.
Ici 5 correspond à convex et 2 correspond à plat

Surface des champignons



Prédictions

On commence par effectuer un Dummies et normaliser nos variables via un scaler.

	cap-shape_b	cap-shape_c	cap-shape_f	cap-shape_k	cap-shape_s	cap-shape_x	cap-surface_f	cap-surface_g	cap-surface_s	cap-surface_y	...	population_n	population_s	population_v	p
0	0	0	0	0	0	1	0	0	1	0	...	0	1	0	
1	0	0	0	0	0	1	0	0	1	0	...	1	0	0	
2	1	0	0	0	0	0	0	0	1	0	...	1	0	0	
3	0	0	0	0	0	1	0	0	0	1	...	0	1	0	
4	0	0	0	0	0	1	0	0	1	0	...	0	0	0	

5 rows × 97 columns

Logistic regression

- On effectue une regression Logistic sur notre training set en utilisant d'abbord les hyperparametes par défaut

```
metrics.f1_score(y_test, y_pred)
```

```
1.0
```

- On obtient alors un F1-score parfait

Nested cross-validation

- Cela permer de séparer notre jeu de donnée en training/ testing set pour ne pas biaiser notre model.
-
- Un grid search est également effectué sur chacun des 10 splits pour trouver les meilleurs hyper paramètres (10 test par split)

- Nos performances:

```
scores
```

```
array([1.        , 0.99767442, 1.        , 1.        , 0.99767442,
       1.        , 1.        , 1.        , 1.        , 1.        ])
```

- Hyper paramètres:

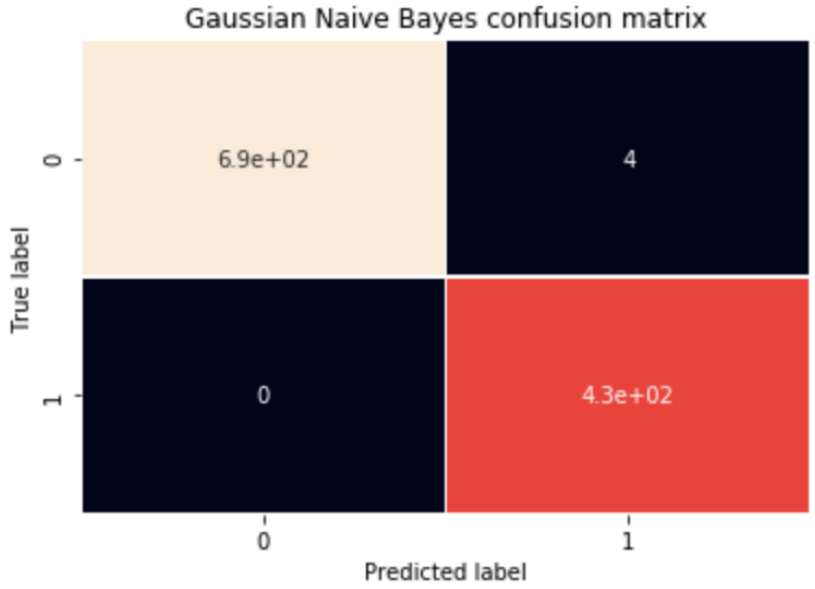
```
{'C': 0.1}
```

Corrélation des variables

	parameter value
odor_n	-1.043962
odor_l	-0.553868
odor_a	-0.553868
stalk-root_c	-0.481090
spore-print-color_n	-0.448844

On remarque que l'odeur du champignon est un facteur prédominant pour déterminer si un champignon est comestible.

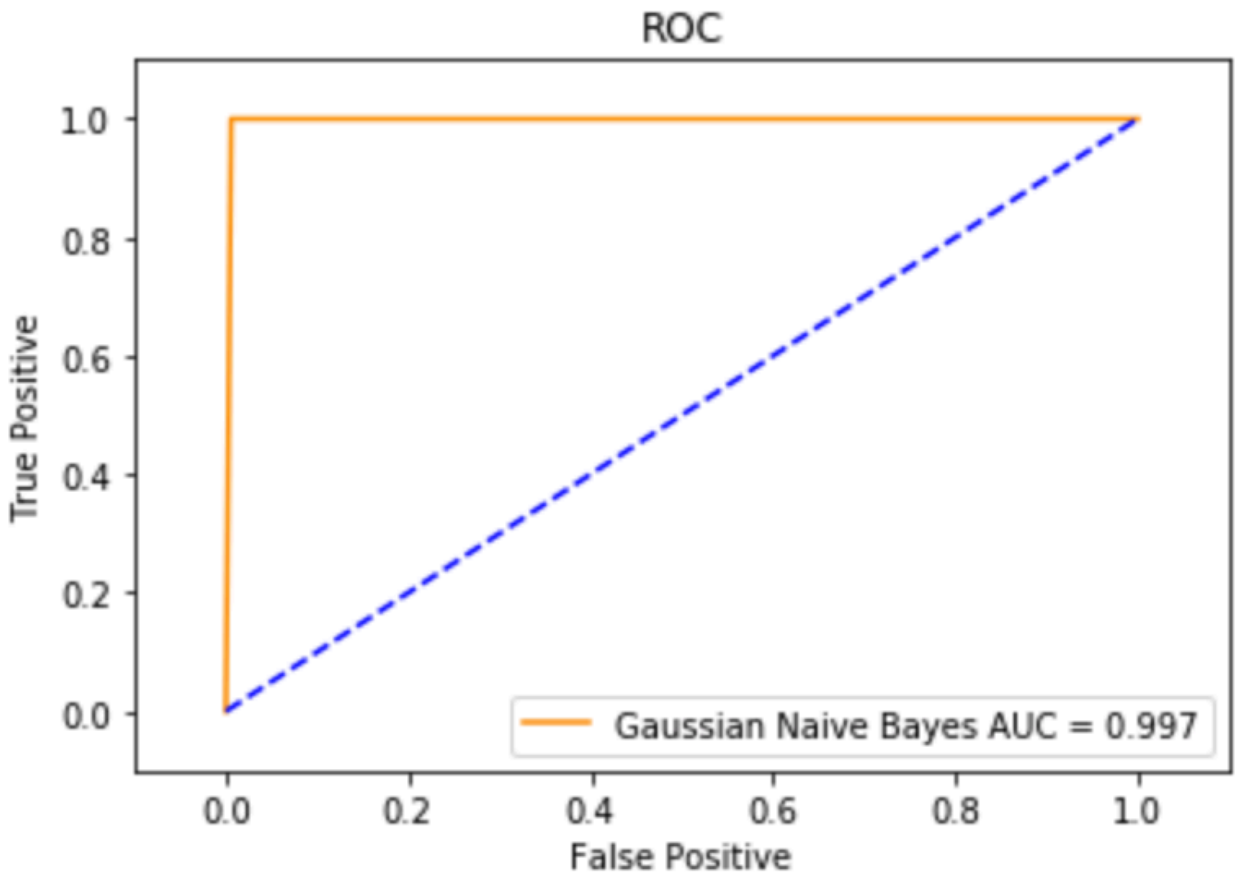
Naive bayes



Gaussian Naive Bayes report

	precision	recall	f1-score	support
0	1.00	0.99	1.00	698
1	0.99	1.00	1.00	431
avg / total	1.00	1.00	1.00	1129

Evaluation


ROC
True Positive
False Positive
0.0
0.2
0.4
0.6
0.8
1.0
Gaussian Naive Bayes AUC = 0.997

Conclusion

Nos deux models sont excellent.

On aurait pu s’arrêter à la régression logistique mais il est toujours bon de tester plusieurs model.