Compte rendu projet Sciences des données

Auteurs :

- QUÉRÉ Anthony
- DEBRAUX Kilian

Introduction

Pour le projet de sciences des données, nous avons réalisé un programme de détection de comestibilité des champignons. Avec Anthony, nous avons choisis de créer ce programme en python. En effet, nous trouvions que c'était le langage le plus approprié pour cette tâche, et python est d'ailleurs le langage le plus utilisé pour du machine learning.

Nous sommes donc fiers de vous présenter notre programme (disponible dans l'archive) ainsi que les résultats que nous avons obtenus présentés sur ce document.

Fonctionnement

Notre programme fonctionne en regardant les différentes propriétés de l'objet dont on veut prédire une classe et regarde avec quels objets du fichier de base il a le plus de propriété en commun (si des propriétés ont une valeur numérique, on peut même être plus précis en calculant aussi la distance au sein même de ces propriétés). Ensuite, on prend les objets du fichier de base qui sont le plus proche de notre objet inconnu (le nombre d'objet pris peut être choisis), et on regarde combien ont quelle valeur pour la classe à prédire, et on prend la valeur la plus présente (si on a un vote pondéré, on prend la valeur qui a obtenu la meilleure note avec la formule de pondération).

Test du programme

Nous avons testé plusieurs options du programme, en modifiant la valeur de K et avec ou sans le vote pondéré.

Pour les tests, nous avons à chaque fois demander à notre programme de prédire la comestibilité de 10 champignons dont on connaît déjà. Ensuite, on regarde si le programme à bien trouvé le bon résultat ou non. Enfin, on calcul la CA, la sensibilité et la confiance de notre programme avec ces résultats.

De plus nous avons réalisé ces tests en double, une fois en se basant sur un jeu de données de 150 champignons, et une fois en se basant sur un jeu de données de 5000 champignons.

Voici les résultats des tests :

En se basant sur le jeu de données de 150 champignons

Paramètres	K=1	K=5, majorité	K=10, majorité	K=5 majorité pondérée	K=10 majorité pondérée
Moyenne Accuracy	1	1	0.8	1	0.7
Moyenne Sensibilité	1	1	0.75	1	0.8571
Moyenne Confiance	1	1	1	1	1

On voit que les paramétrages K=1 et K=5 (pondérés ou non), sont les paramétrages les meilleurs possibles pour notre programme avec le jeu de 150 champignons.

En se basant sur le jeu de données de 5000 champignons

Paramètres	K=1	K=5, majorité	K=10, majorité	K=5 majorité pondérée	K=10 majorité pondérée
Moyenne Accuracy	1	1	1	1	1
Moyenne Sensibilité	1	1	1	1	1
Moyenne Confiance	1	1	1	1	1

Avec les 5000 champignons, on voit que tous les paramétrages sont bons, on peut donc avoir entièrement confiance en ces prédictions.

Fricassée de champignons

Pour connaître la comestibilité des 10 champignons, nous nous sommes basés sur le jeu de données de 5000 champignons car il parait être le plus fiable. On a aussi choisi comme paramétrage K=5 et un vote pondéré, car il fait partie, selon l'étude sur 150 champignons, des plus fiables que l'on ait. Nous précisons que si nous n'avions pas eu de résultats aussi excellents lors des tests, nous aurions choisi le paramétrage avec la meilleure confiance, car la plus important est de surtout ne pas manger de champignons vénéneux, et que l'on peut se permettre de na pas prendre un champignon même s'il est comestible.

Voici donc les résultats :

Champignon	Comestibilité
C1	Non
C2	Non
C3	Oui
C4	Oui
C5	Non
C6	Non
C7	Oui
C8	Non
C9	Non
C10	Oui

Selon les résultats des tests, on peut considérer ces résultats comme fiables, on peut alors utiliser les champignons suivant dans la fricassée :

Résultat de la fricassée

- C3
- C4
- C7
- C10