Rapport TP2 Arbres

Thibault FERRETTI

2023-09-28

Question 1

Dans le cadre de la régression, on peux choisir comme fonction de perte l'erreur quadratique moyenne qui équivaut à une réduction de la variance et qui minimise la perte L2 en effectuant la moyenne des noeux terminaux.

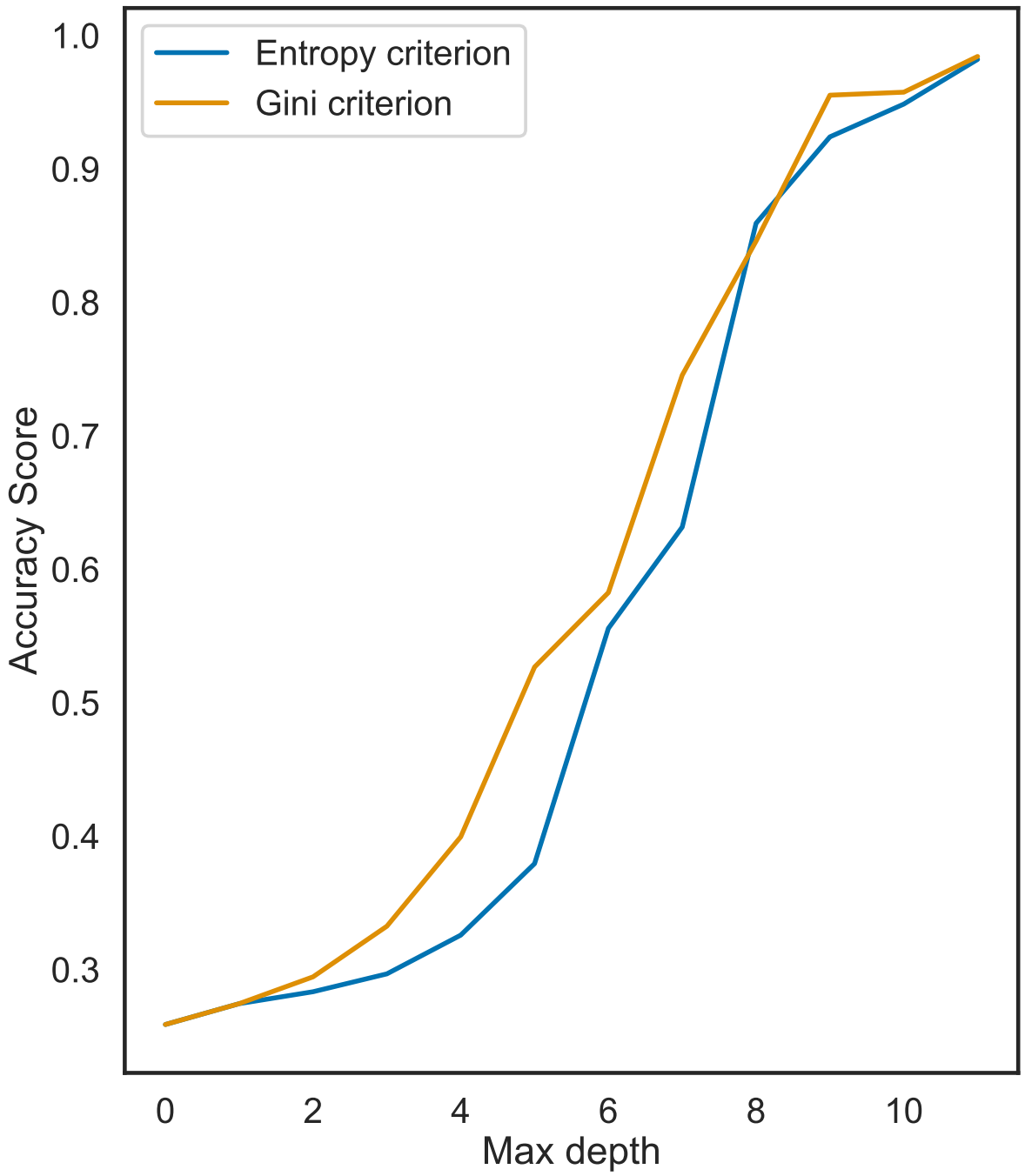
On peut aussi choisir de minimiser la perte L1 en utilisant cette fois ci la médiane des noeux terminaux.

Il existe aussi d'autres critères comme l'erreur quadratique moyenne de Friedman, ou le critère de Poisson.

Question 2

On simule à l'aide de rand_checkers des échantillons de taille $n = 456$. Ces données serviront à construire deux arbres, l'un utilisant le critère de Gini et l'autre utilisant l'entropie. On affiche ensuite les courbes donnant le pourcentage d'erreur en fonction de la profondeur de l'arbre.

```
Gini criterion
1.0
Entropy criterion
1.0
```



Sans grande surprise on trouve que la profondeur qui minimise le pourcentage d'erreur (ou maximise le score) est tout simplement la profondeur maximale de notre boucle, ici sa valeur est de 12.

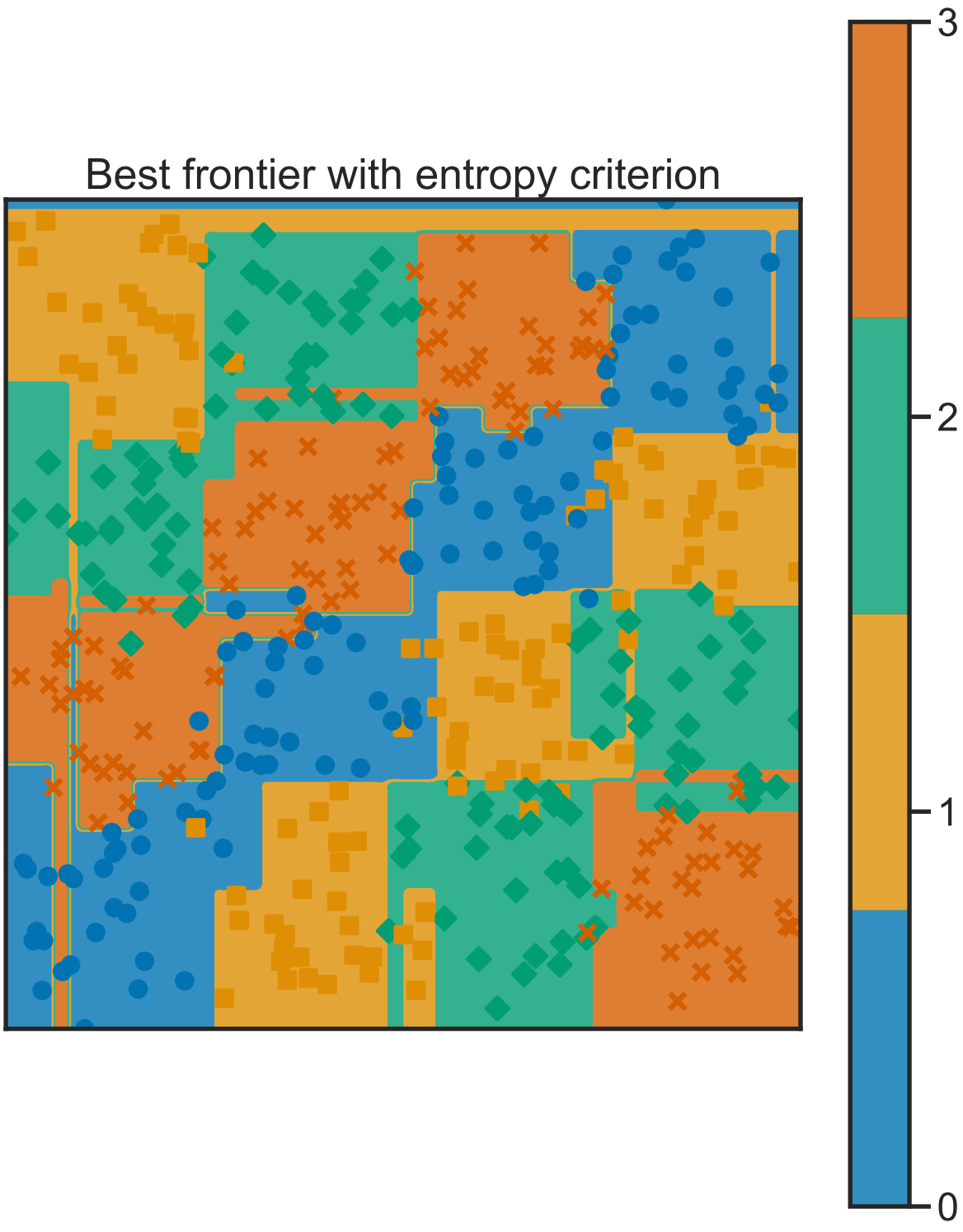
Question 3

En utilisant la profondeur qui minimise le pourecntage d'erreur (ou maximise le score) on trouve la classification suivante:

```
dt_entropy.max_depth = 12

plt.figure()
frontiere(lambda x: dt_entropy.predict(x.reshape((1, -1))), X, Y, step=100)
plt.title("Best frontier with entropy criterion")
plt.draw()
print("Best scores with entropy criterion: ", dt_entropy.score(X, Y))
```

```
Best scores with entropy criterion:  0.9821428571428571
```


Best frontier with entropy criterion
3
2
1
0

Le score étant très proche de 1, le modèle est 'parfait', on est dans le cas de l'overfitting, ce modèle n'est donc pas adapté. Comme on peut le voir dans la figure ci-dessus, certaines partitions sont de trop et se généraliseront très mal sur de nouvelles données.

Question 4

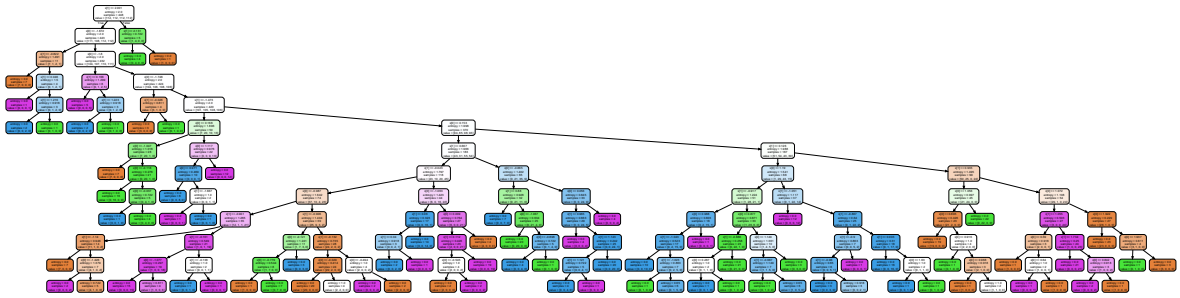


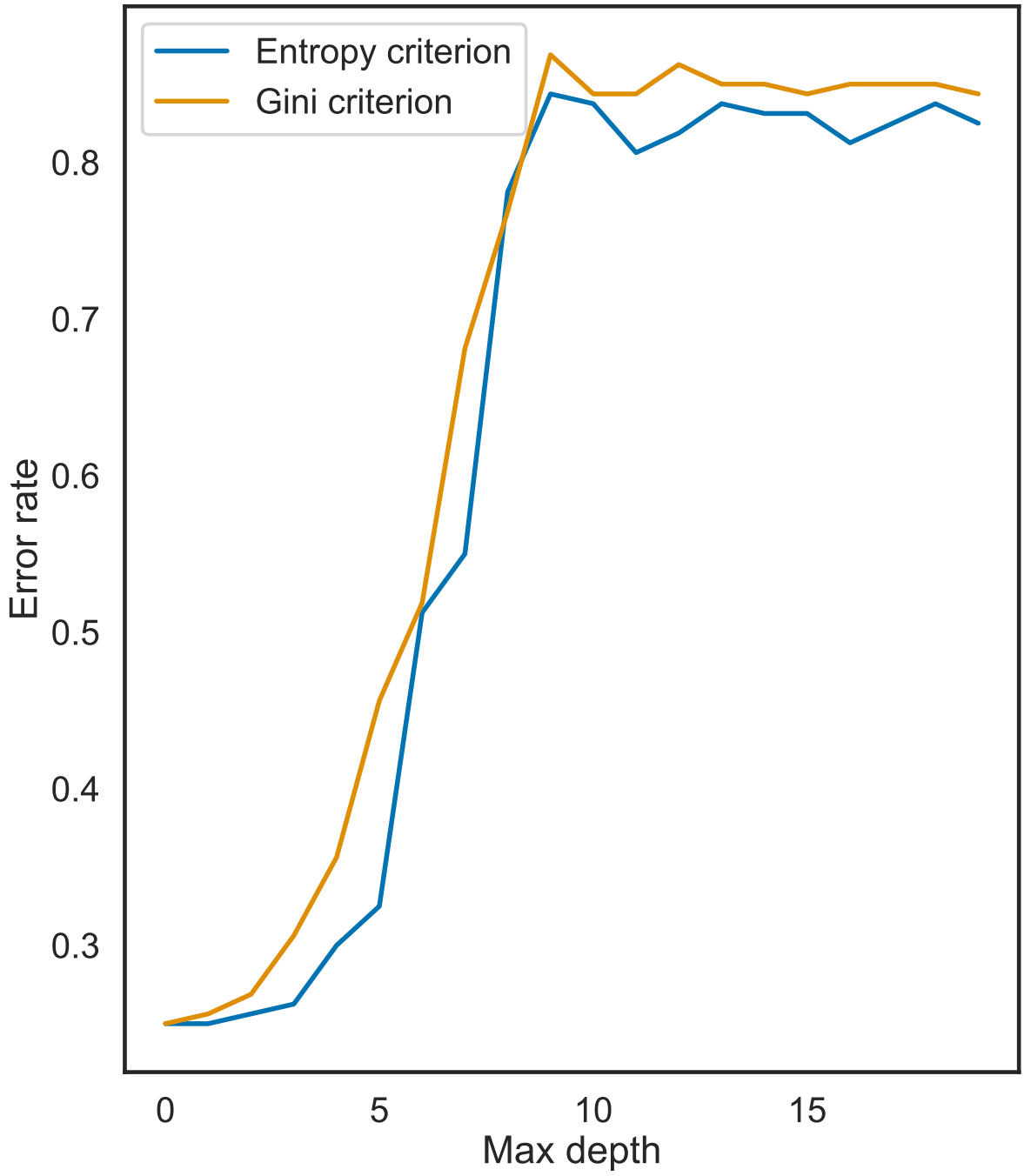
Figure 1: Arbre de décision

Voici l'arbre de décision que l'on obtient, chaque noeud correspond à une condition Vrai/Faux qui nous donnera un partionnement de l'ensemble des données. Si la valeur du booléen est vrai on passera au noeud suivant correspondant, sinon on passera au noeud correspond à Faux. On continue ce processus jusqu'à atteindre la profondeur de l'arbre.

Question 5

On va maintenant utiliser les arbres précedemment trouvés pour calculer le score sur un nouvel échantillon test avec $n = 160$

```
data_test = rand_checkers(40, 40, 40, 40, sigma)
n_test_samples = len(data_test)
X_test = data_test[:, : -1]
Y_test = data_test[:, -1].astype(int)
```

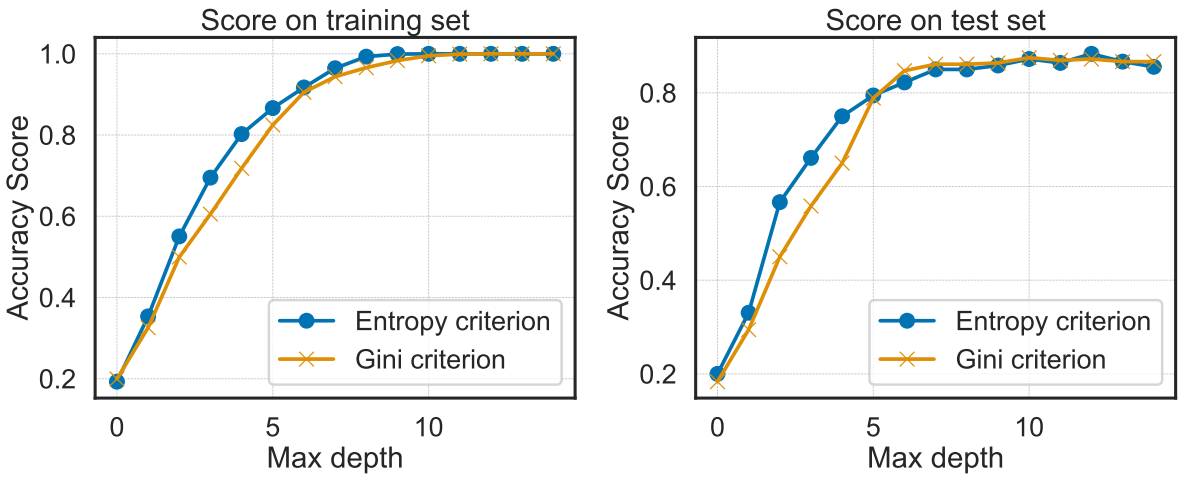


On trouve ici que la profondeur qui maximise le score est de 10, pour le critère de Gini ou l'entropie. On a donc trouvé un arbre de profondeur plus faible que précédemment, réduisant ainsi le risque d'overfitting.

Dataset DIGITS

Dans la suite du TP on s'intéresse au jeu de données DIGITS qui contient des images 8x8 de chiffres. On sépare les données en train/test split de 0.8/0.2.

Question 6



Comme dans le cas précédent on obtient les courbes de score en fonction de la profondeur de l'arbre. Le score sur l'échantillon test semble atteindre un plateau lorsque la profondeur dépasse 7, on préfèrera alors le modèle le plus parsimonieux.

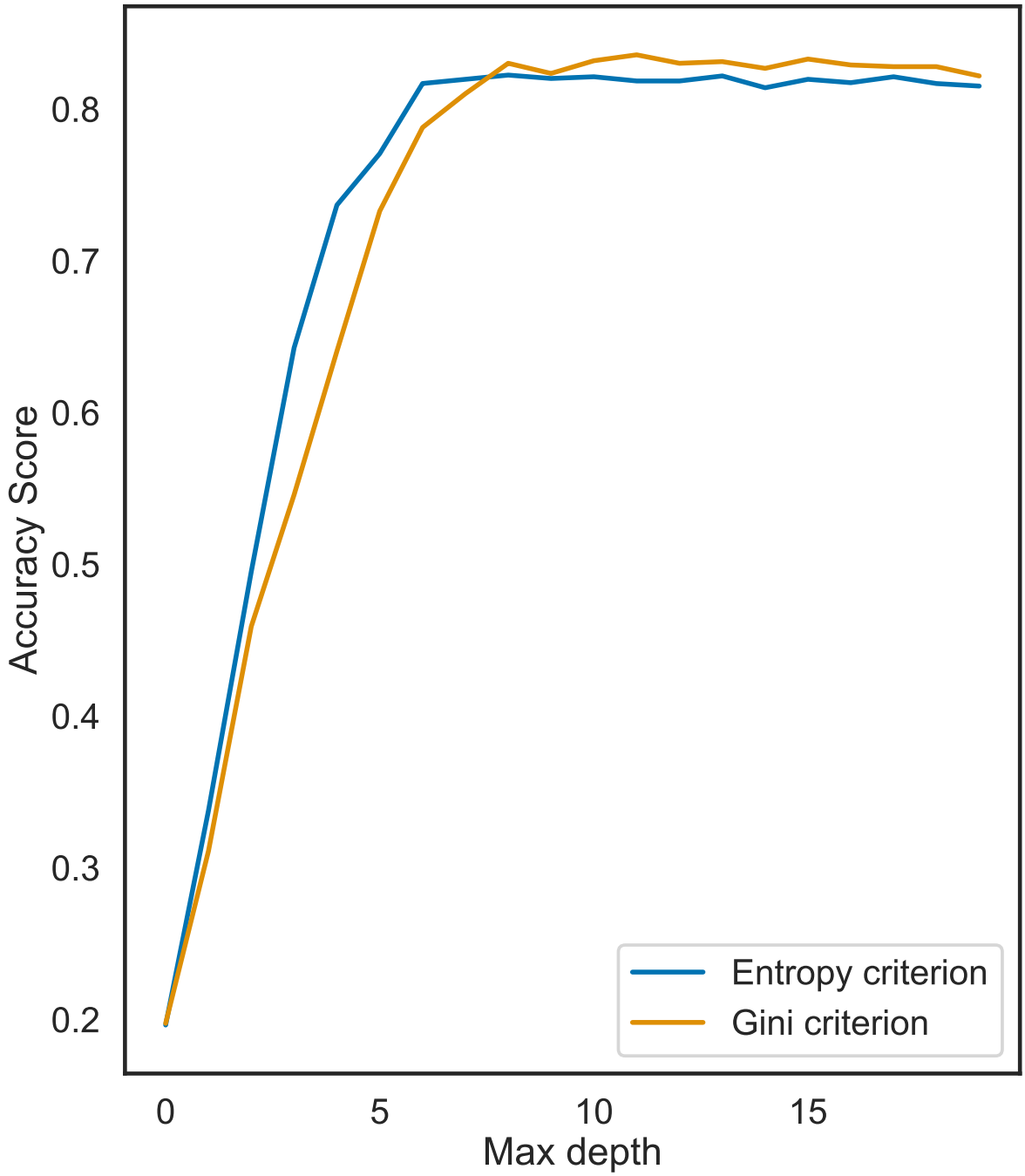
Sélection de modèle

Question 7

On utilise maintenant la validation croisée pour effectuer le choix du paramètre de profondeur.

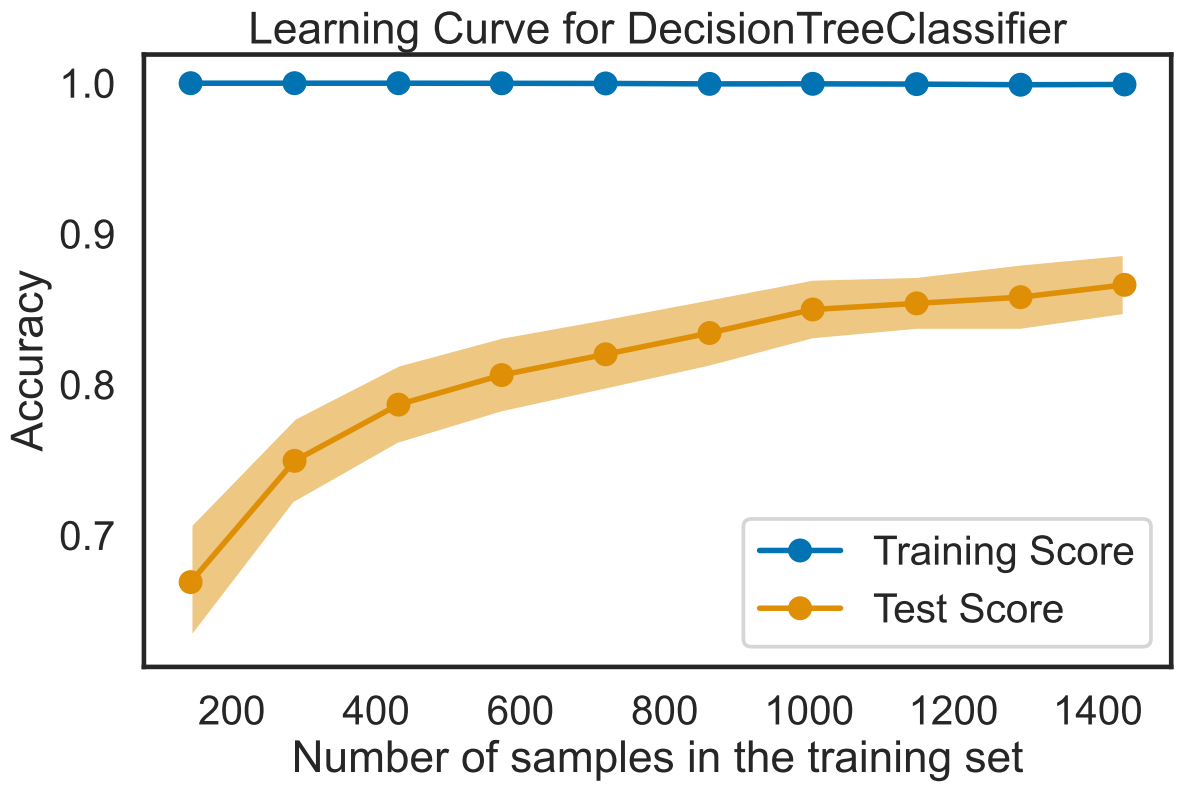
```
0.819 accuracy with a standard deviation of 0.055
```

```
The maximum cross-validation score for entropy is 0.8225 at depth 9.
The maximum cross-validation score for gini is 0.8358 at depth 12.
```


0.8
0.7
0.6
0.5
0.4
0.3
0.2
Accuracy Score
0
5
10
15
Max depth
Entropy criterion
Gini criterion

Question 8

Dans cette question on affiche la courbe d'apprentissage qui mesure l'effet du score en fonction du nombre de données durant la période d'apprentissage du modèle.



Dans notre cas, le score d'apprentissage reste relativement élevé peu importe la taille de l'échantillon d'apprentissage. Cependant, le score sur l'échantillon de test augmente en fonction de la taille de l'échantillon d'apprentissage jusqu'à un certain plateau. L'ajout de nouvelles données aura un effet de moins en moins prononcé.