HAX907X TP2 : Arbres

Axel de MONTGOLFIER

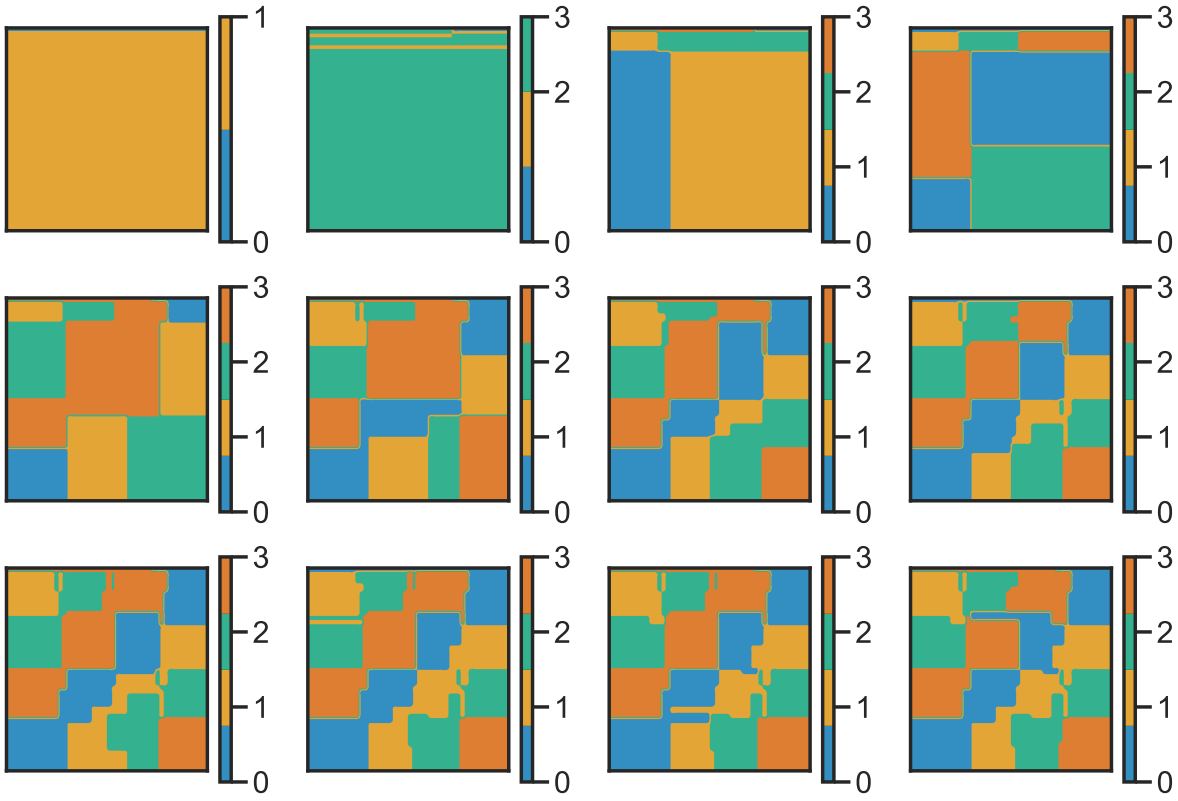
Classification avec les arbres :

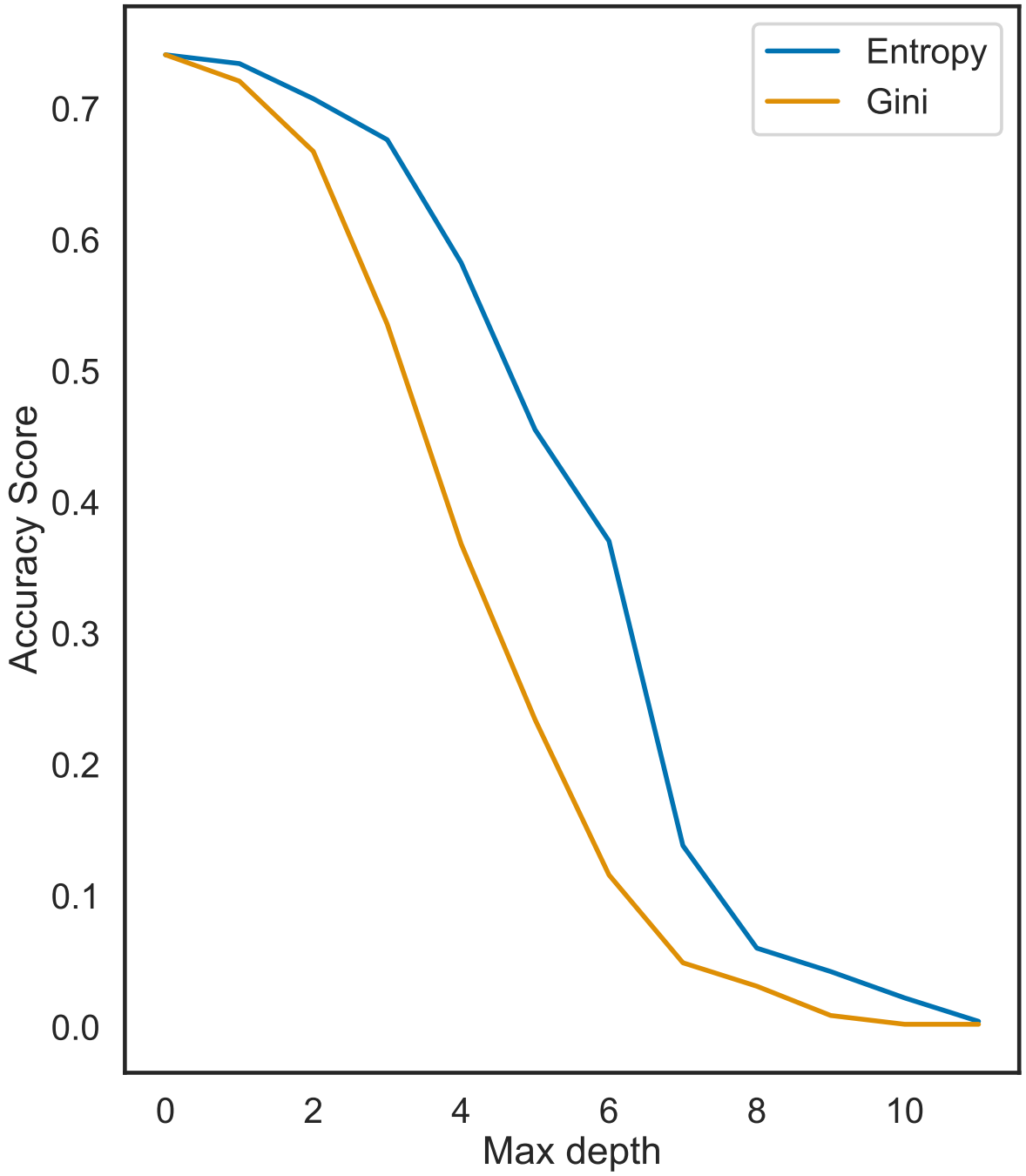
1 - Si on se place dans le cadre d'une régression, un moyen de mesurer l'homogénéité serait de s'intéresser à la variance entre les individus. En effet cette dernière influe sur la proximité ou l'éloignement des individus ce qui nous permet de contrôler ainsi notre répartition. On va ainsi chercher a minimiser la variance entre individus et maximiser la variance entre les groupes d'individus afin de pouvoir réaliser des coupes permettant de rendre le résultat le plus homogène possible.

2 - On va ainsi générer des simulations en utilisant la fonction "rand_checkers" pour créer un échantillon de 456 individus :

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

Maintenant que nous avons effectué cette simulation nous allons pouvoir tester le critère de Gini et le critère d'Entropie en traçant des courbes d'erreurs en fonction de la profondeur maximale de l'arbre :


1
0
3
2
1
0

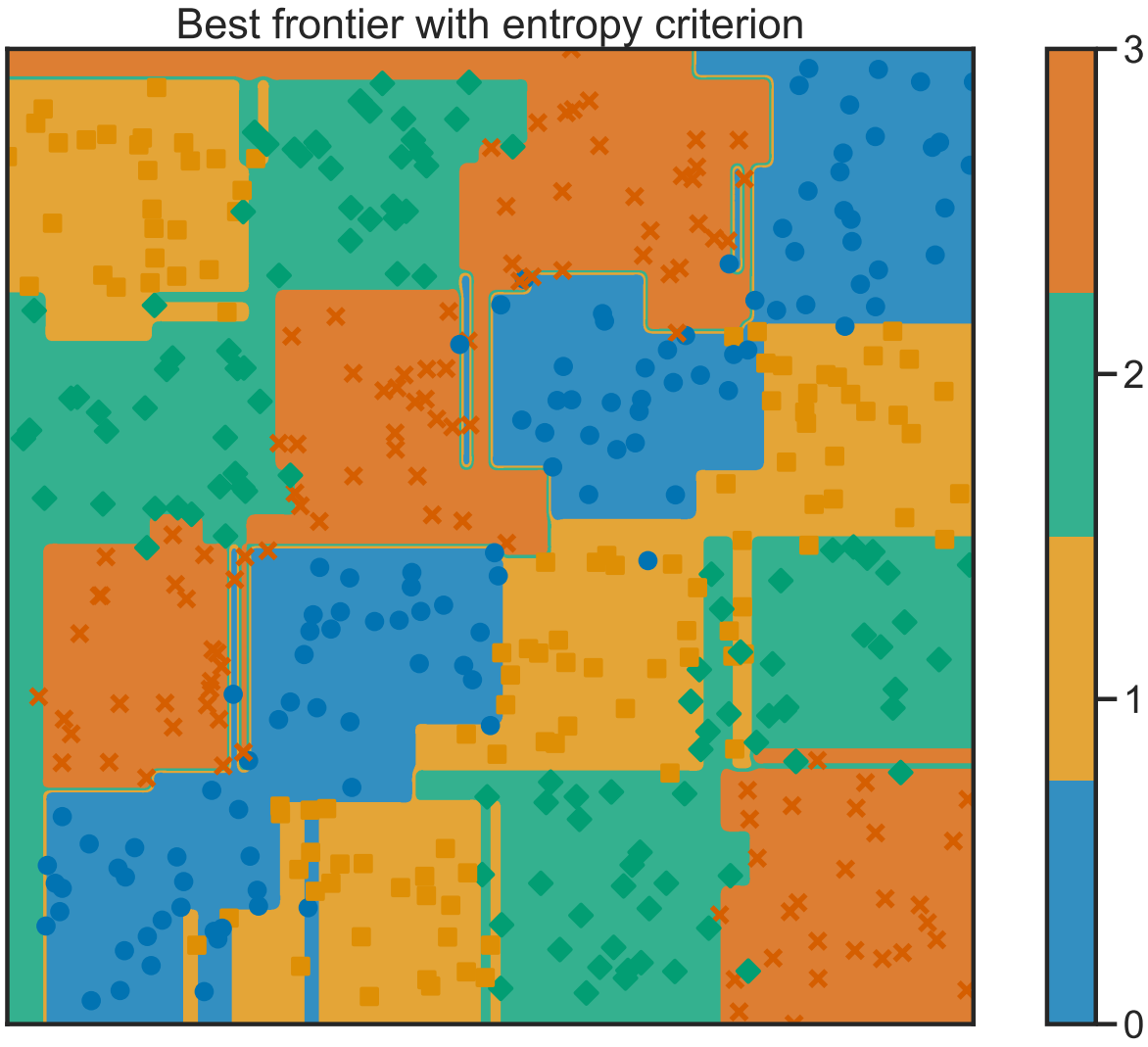


On remarque directement que l'erreur diminue en fonction de la profondeur maximale et tend vers 0, quasiment nul lorsque la profondeur maximale est à 12. On constate également que l'erreur sur l'indice de Gini décroit de la même manière que l'erreur sur l'indice d'Entropie. On en déduit donc que plus notre arbre a une grande profondeur plus on sera précis sur nos

coupes afin de rendre l'erreur de classification minimale. On peut donc choisir arbitrairement une profondeur de 12, la ou les courbes semblent se stabiliser vers 0.

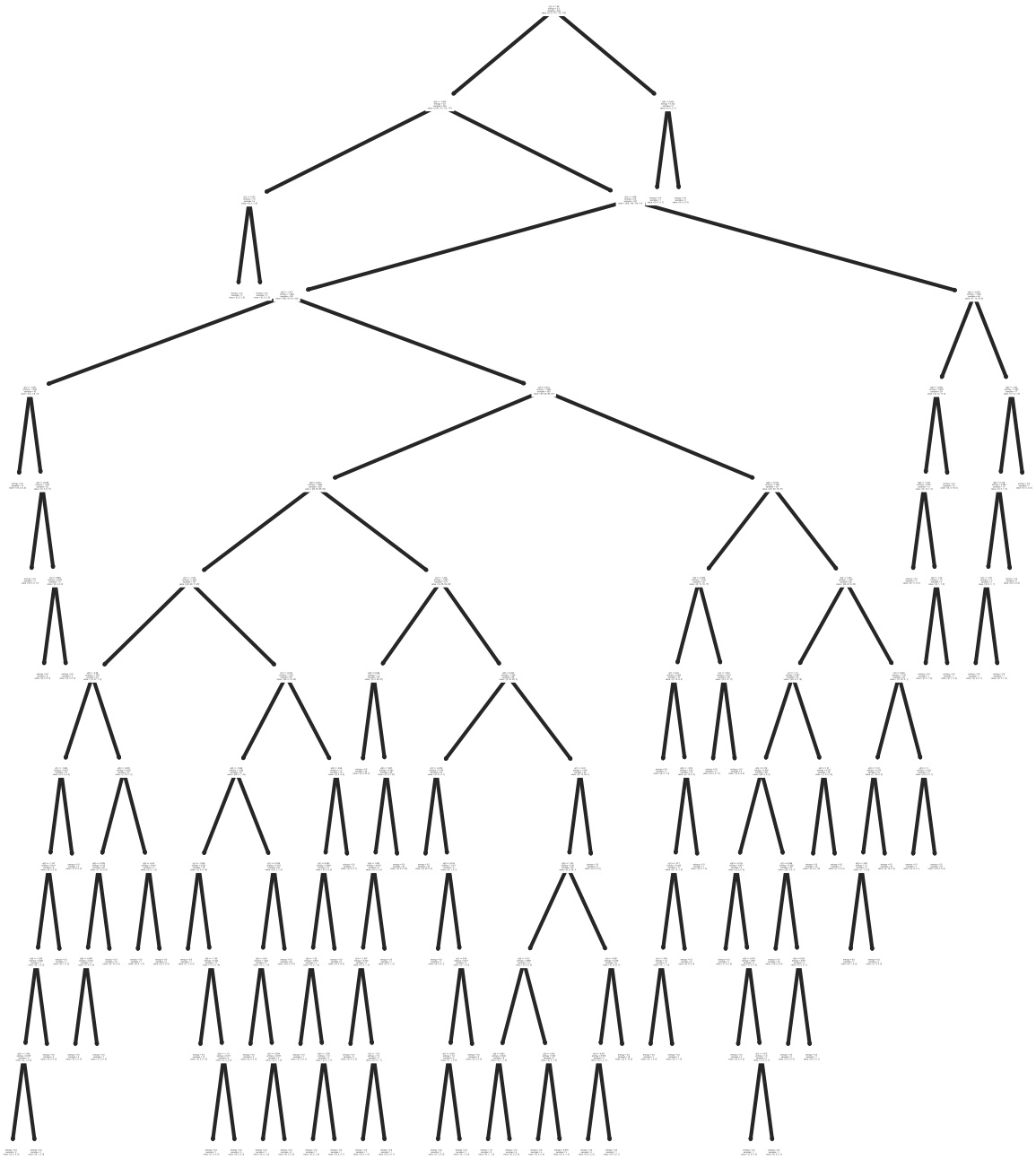
3- On affiche maintenant la classification obtenue pour une profondeur qui minimise l'erreur

```
Best scores with entropy criterion:  0.9955357142857143
```



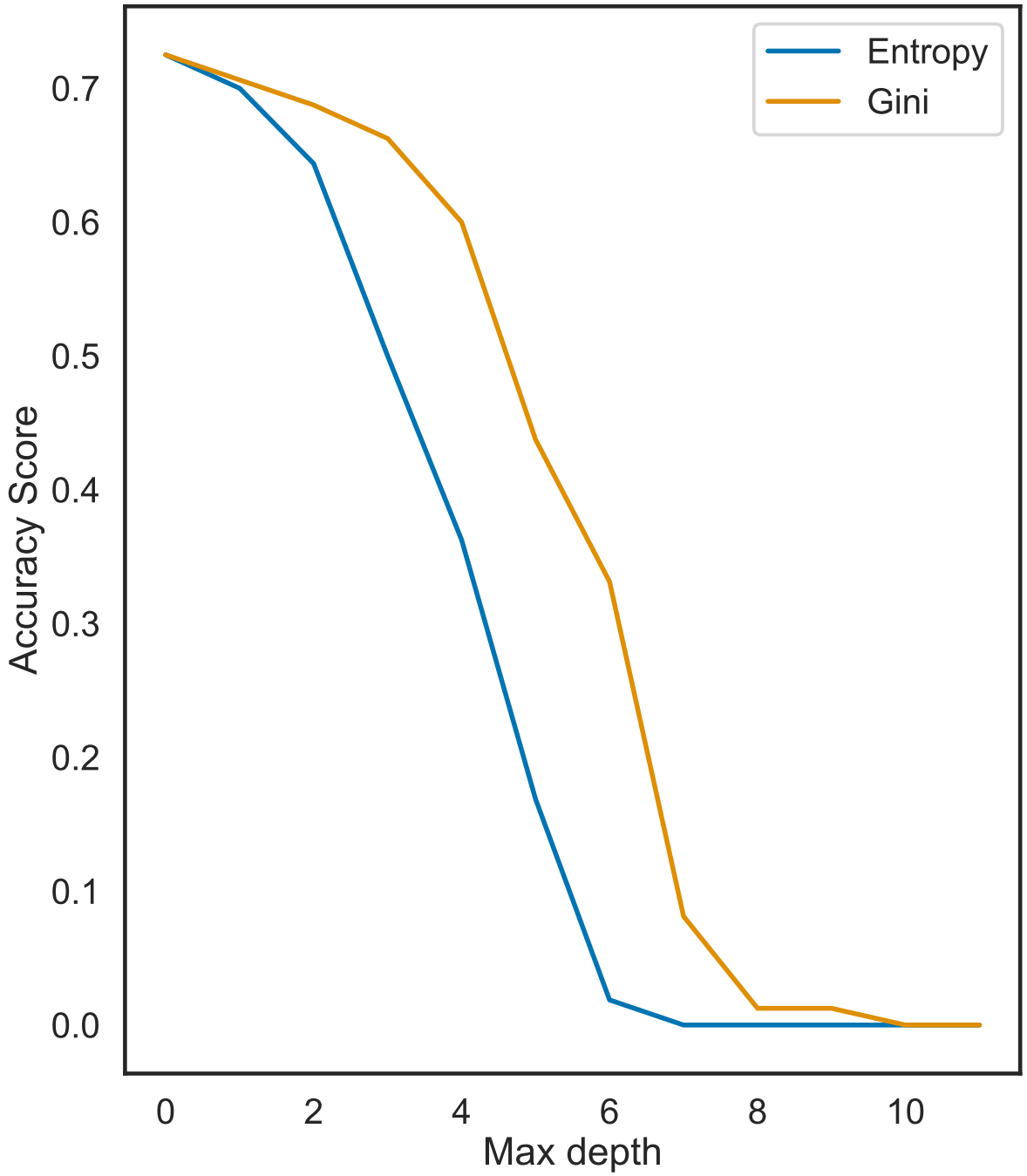
On constate effectivement une répartition plutôt homogène de nos individus.

4- On exporte l'arbre afin de pouvoir l'afficher :



5- On réitère l'expérience avec cette fois ci un jeu de 160 données :

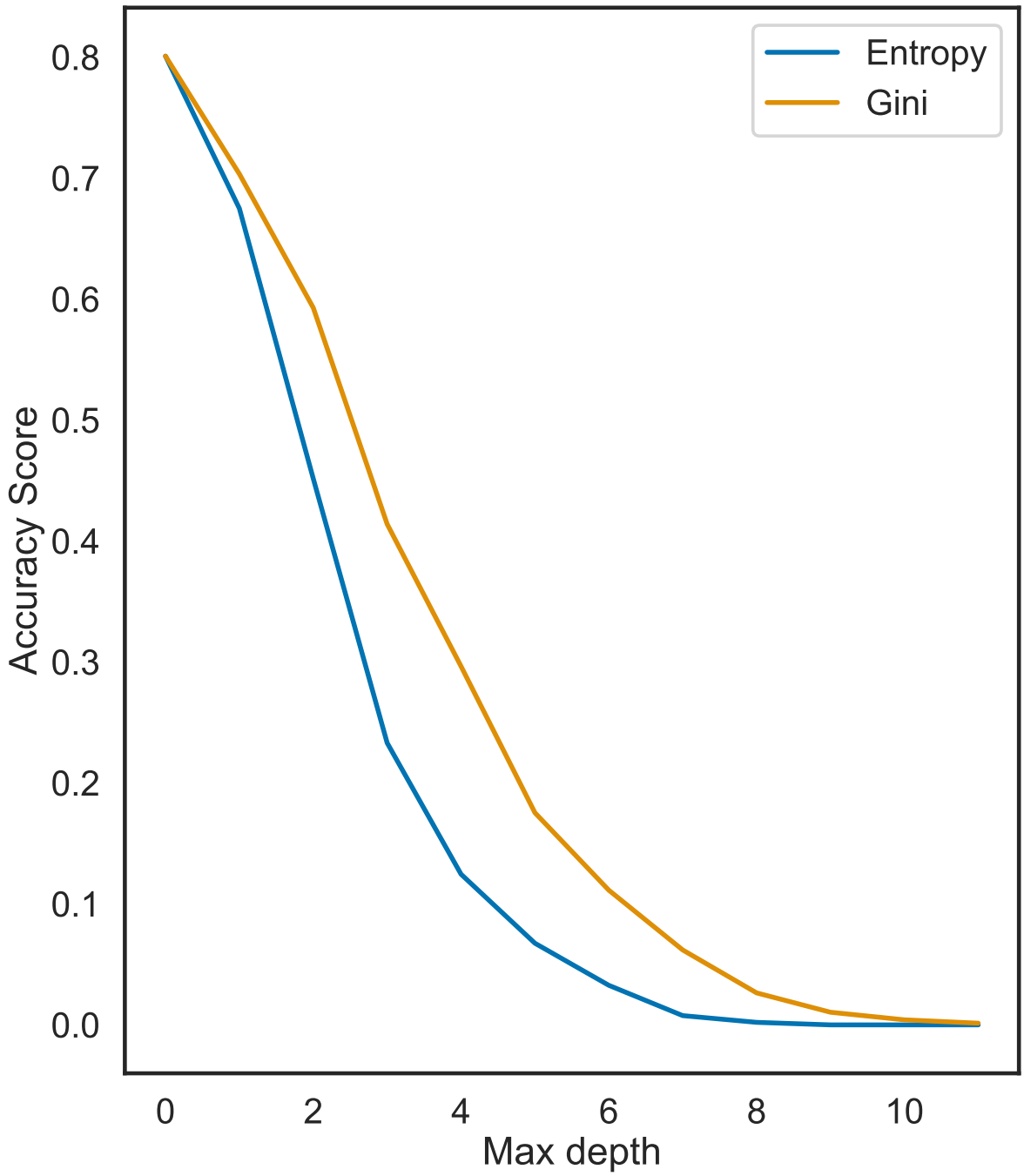
```
<Figure size 4500x3000 with 0 Axes>
```



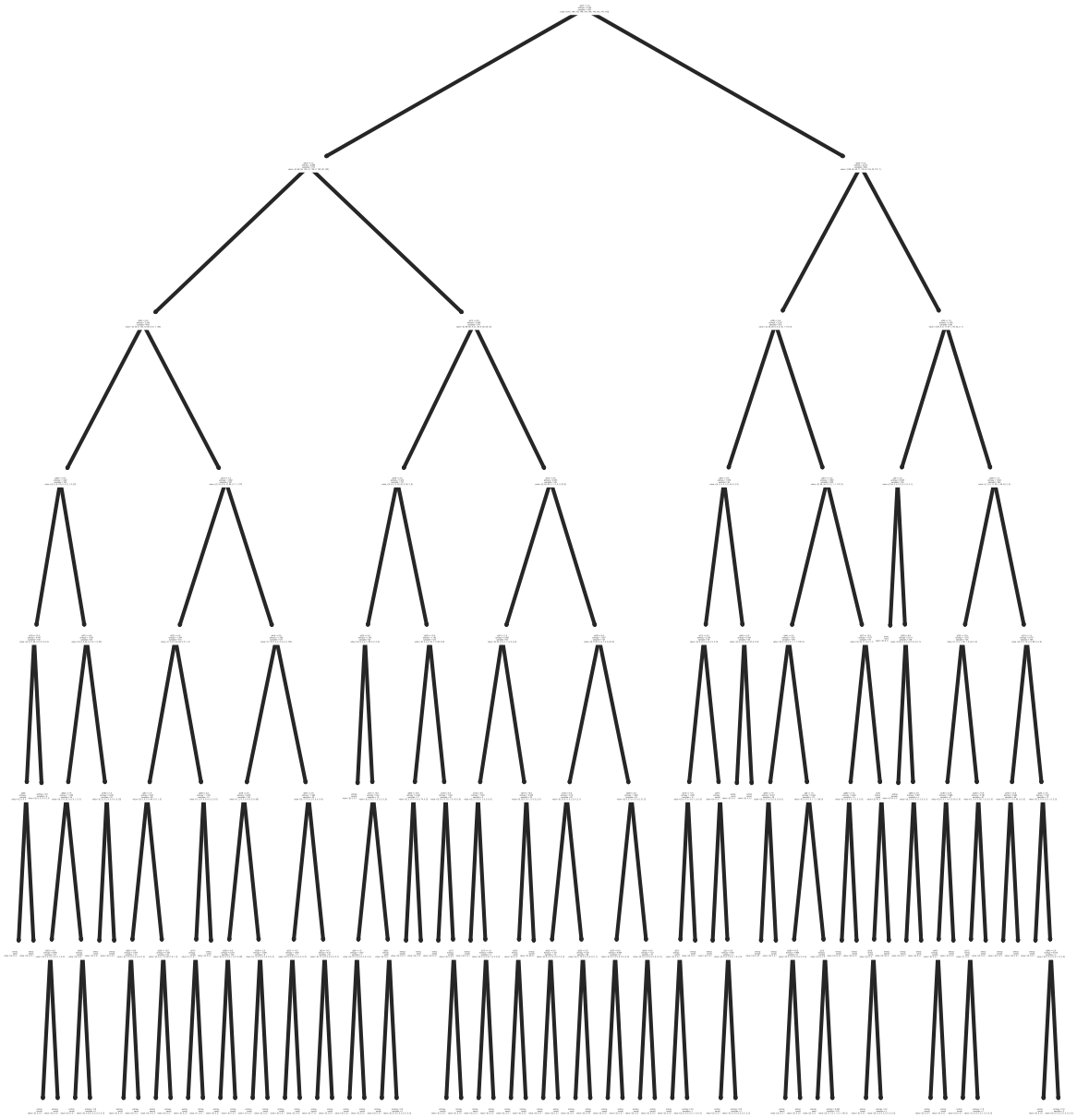
On retrouve une diminution des courbes d'erreur similaire à notre première expérience, cependant les courbes semblent diminuer nettement plus rapidement qu'avant avec cette fois ci une quasi nullité atteinte vers une profondeur de 8. Cependant on constate en faisant plusieurs tentative que les courbes sont nettement plus instable en terme de variance.

6- On réitére notre expérience avec le jeu de données DIGITS :

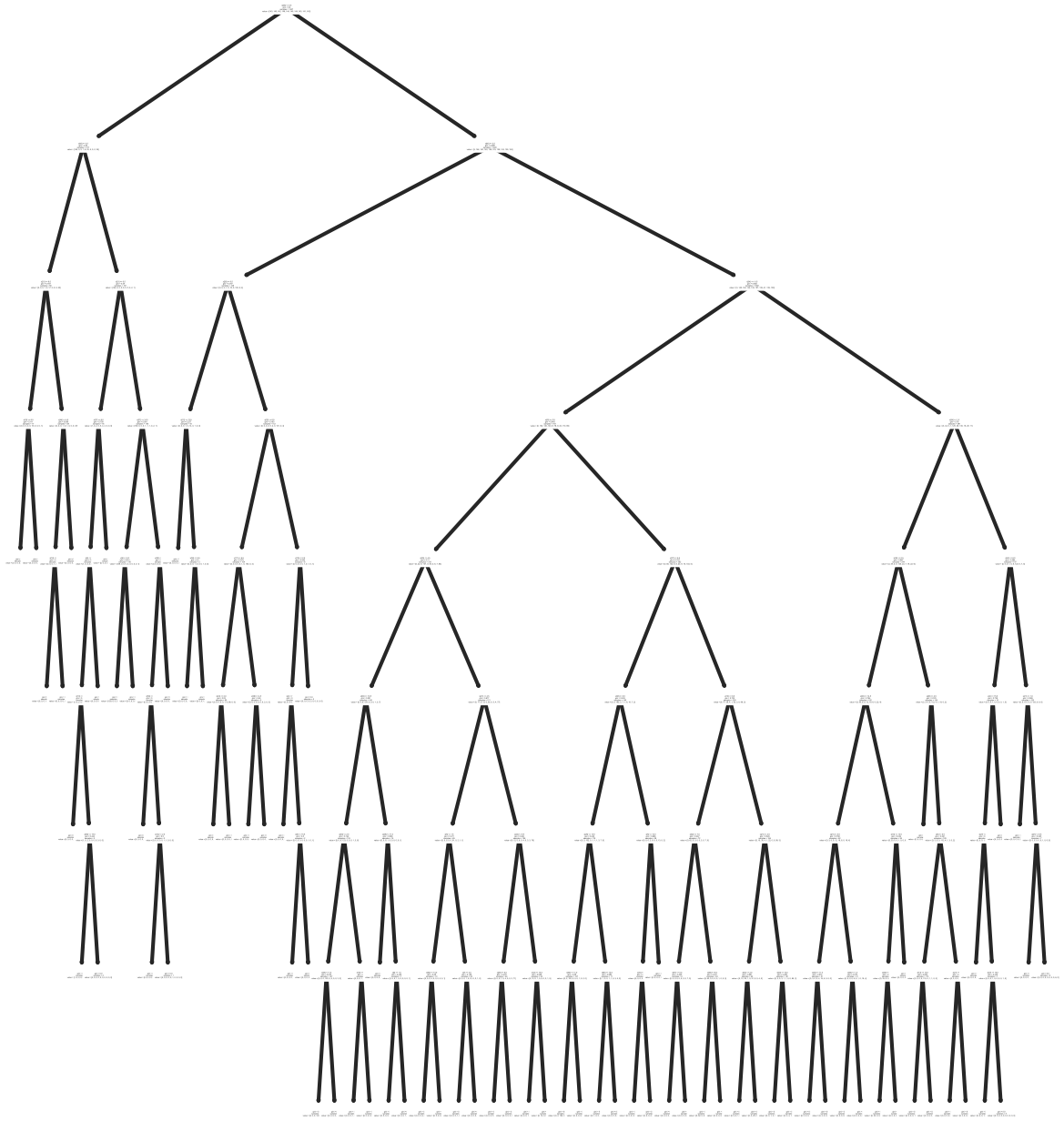
```
<Figure size 4500x3000 with 0 Axes>
```



On constate que la profondeur telle qu'on approche une erreur 0 pour l'entropie est 7 et la profondeur pour l'indice de Gini est 8 on peut donc réaliser des arbres de ces profondeurs pour les deux indices afin d'avoir une bonne étude.



Ici l'arbre de profondeure maximale 7 pour une l'entropie.

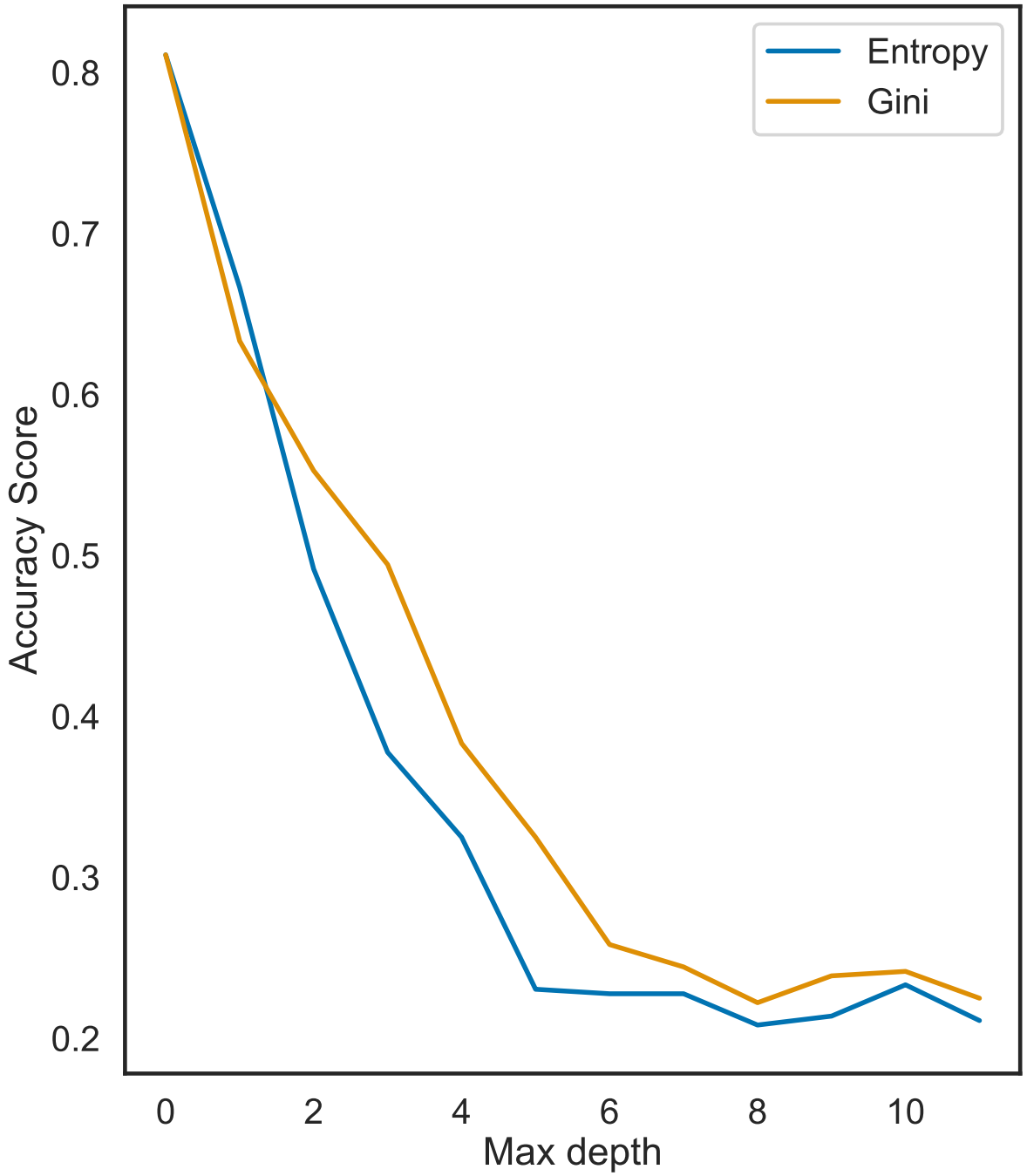


Et ici l'arbre de profondeur maximale 8 pour l'indice de Gini.

Méthode de choix des paramètres - Sèlection de modèle :

7- On applique la validation croisée à l'aide de la fonction "cross_val_score" :

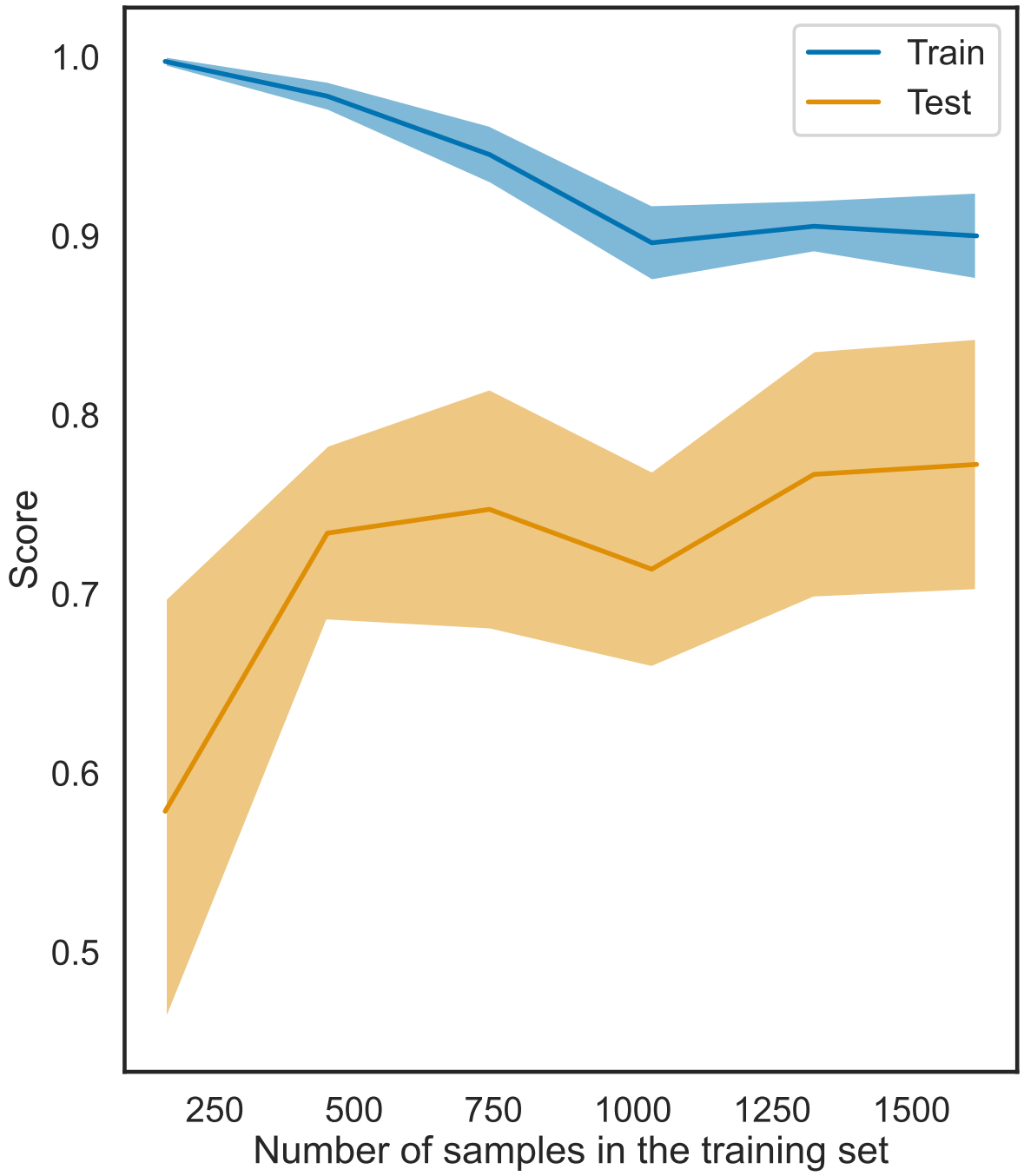
```
<Figure size 4500x3000 with 0 Axes>
```



On constate que L'erreur semble se stabiliser vers 6 pour l'entropie et vers 7 pour l'indice de gini. Cependant cette fois ci nous n'approchons pas 0 comme erreur mais 0,2.

8- On effectue la courbe d'apprentissage de l'arbre ded décision avec une profondeur maximale de 6 pour le critère de l'entropie :

```
<sklearn.model_selection._plot.LearningCurveDisplay at 0x1a27c056460>
```



On constate ainsi que notre courbe des données d'apprentissage "train" est plutôt élevé stagnant a hauteur de 0.9 de plus notre erreur de validation semble croissante et se stabilise au

niveau de 500 données d'entrainement au alentours de 0,7.