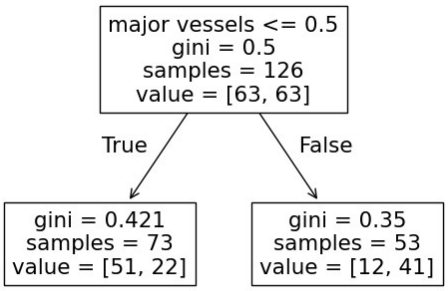
# Machine Learning Project 1
# Predicting Heart Disease

## Tache 1 : Entrainer le premier Arbre de décision.

A la vue du résultat de l'arbre de décision, il semble que le modèle soit en « underfitting ».
En effet, le résultat final montre que les échantillons sont encore fortement mélangés entre les personnes souffrant d'une maladie cardiaque et les autres.
Il est déconseillé d'utiliser ce modèle. Le paramètre « major vessels » ne permet à lui seul une séparation assez nette pour séparer les possibilités de l'une ou l'autre catégorie.

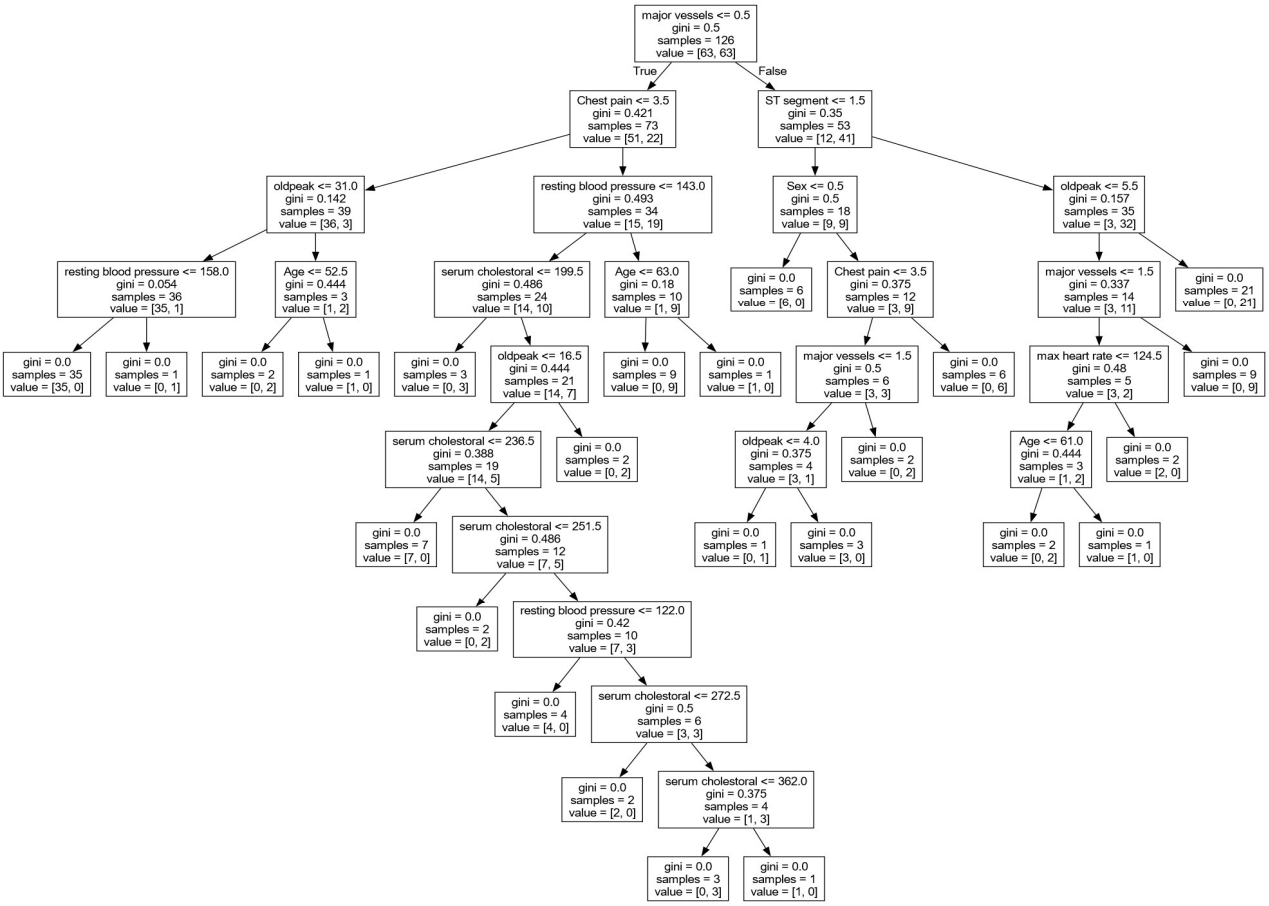



```
Accuracy for the training tree is 0.73
Accuracy for the testing tree is 0.78
```

## Tache 2 : Entrainer un autre Arbre de décision.

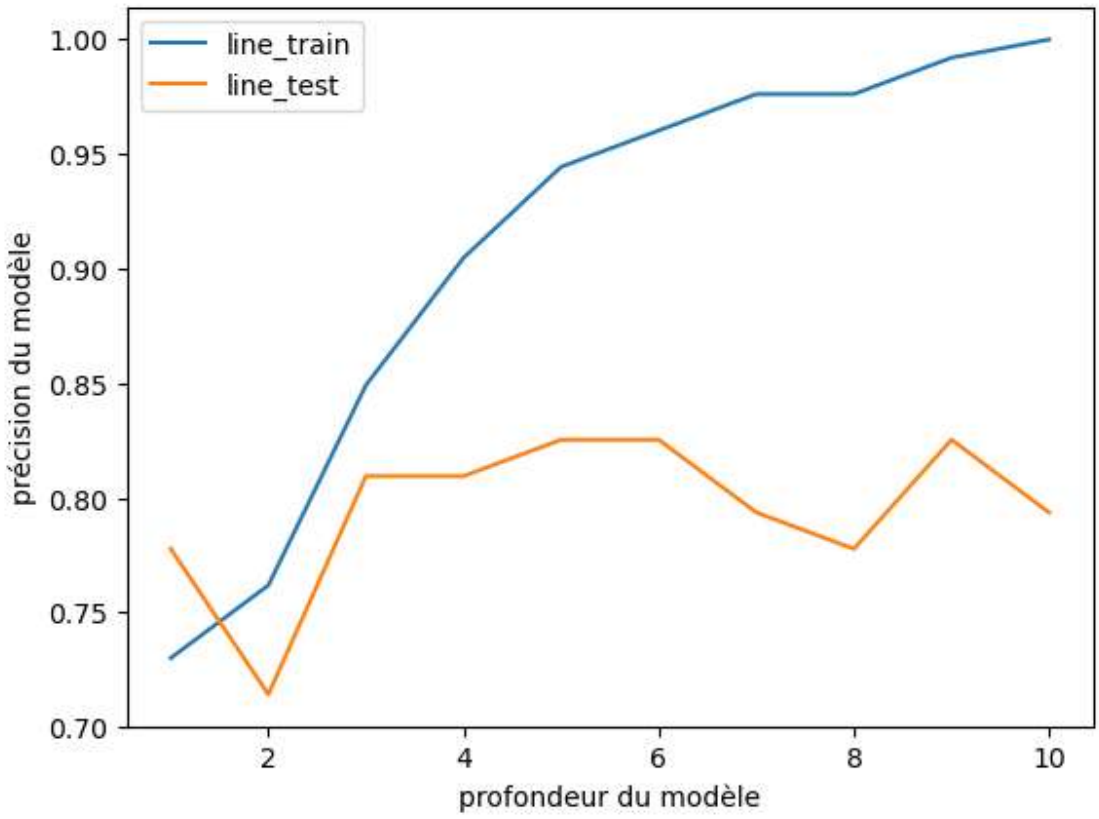
Nous sommes ici confrontés au problème inverse, l'overfitting. Le modèle a appris par cœur pour atteindre un niveau de précision de 100% sur le modèle d'entrainement mais bien inférieur dans le cas du dataset de test. On arrive à avoir des feuilles qui sont unitaires et totalement pures (GINI=0.00).
Il est déconseillé d'utiliser ce modèle car du fait de son apprentissage « par cœur », il n'est plus assez apte pour généraliser la prédiction sur d'autres données inconnues.

```
Accuracy for the training tree is 1.00
Accuracy for the testing tree is 0.79
```




## Tache 3 : Trouver le meilleur Arbre de décision :

Si l'on se base sur le graphique qui compare les résultats du test aux résultats d'entrainement, nous pouvons remarqué que le modèle commence à atteindre son niveau optimum à partir de la profondeur 3.
En deçà, le modèle n'est pas assez complexe que pour aider à la généralisation.
Au-delà, on remarque que le niveau de prédiction reste plus ou moins stable malgré un légère augmentation à la profondeur 5 mais un chute à partir de la profondeur 7.
De plus, l'écart entre le modèle d'entrainement et le modèle de test ne fait que croître à partir de la profondeur 3 ce qui pourrait s'expliquer par le fait que le modèle d'entrainement commence par apprendre par cœur.
Pour rejoindre la théorie du cours, nous apprenons que seules les prédictions du modèle sur un set de test est important. De plus, à niveau de prédiction équivalent, il est toujours préférable de prendre le modèle le plus simple. Ce qui dans le cas présent est le modèle 3.




## Tache 5 : Réflexion sur le travail

Le fait de garder un fichier de test séparer, permet de s'assurer que le modèle n'a pas pu s'entrainer sur les données présentes. De ce fait, il ne peut pas y avoir de biais qui permettrait au modèle de donner la réponse qu'il aurait apprise par cœur.
Le score de Kaggle représente la précision de mon modèle sur le fichier test de Kaggle.
Le score entre Kaggle et le score que j'ai obtenu, peut-être dû aux « bruits » des données qui peuvent influencer dans une certaines mesures les résultats de prédictions. Il se peut aussi que les formules pour calculer la précision du modèle diffèrent entre celui que j'ai utilisé et celui utilisé sur le site.
Quoiqu'il en soit, le modèle actuellement présenté pourrait être largement améliorer si une recherche sur les meilleurs méta-paramètre avait été faites préalablement notamment avec Grid Search