Semaine 7B - Notes du cours

# 

[**Ce que vous apprendrez pendant ce cours**](#_2et92p0) **2**

[**Arbre de décision**](#_1t3h5sf) **2**

[**Explication intuitive**](#_4d34og8) **2**

[**Noeux (ou Nodes)**](#_ps42zzkwclby) **4**

[**Branches**](#_94riqdhz9h3o) **4**

[**Feuilles (ou Leaf Node)**](#_9frw63yqqdya) **4**

[**Avantages et inconvénients des arbres**](#_r89w8z3yr5yd) **5**

[**Avantages**](#_8wxsoahcq1aa) **5**

[**Inconvénients**](#_piheyhm1ievj) **5**

[**Random Forest**](#_c00drejuyr2j) **6**

[**Explication intuitive**](#_d4ye9am6dfu0) **6**

[**Avantages**](#_cmxggmfgs0r9) **6**

[**Inconvénients**](#_tevnj3asi1b4) **6**

[**Ressources**](#_ywvx2tjsvnfb) **7**

# 

Modèles de ML : Decision Trees

Il existe un type d’algorithme très puissants en Machine Learning et qu’il est bon de connaître : Les arbres de décisions (ou *Decision Trees)*. Ceux-ci ont été popularisés depuis l’apparition de la Forêt d’arbres aléatoires (*Random Forest*) qui s’avèrent être extrêmement performante. Dans ce cours, nous verrons comment aborder cet algorithme de manière intuitive.

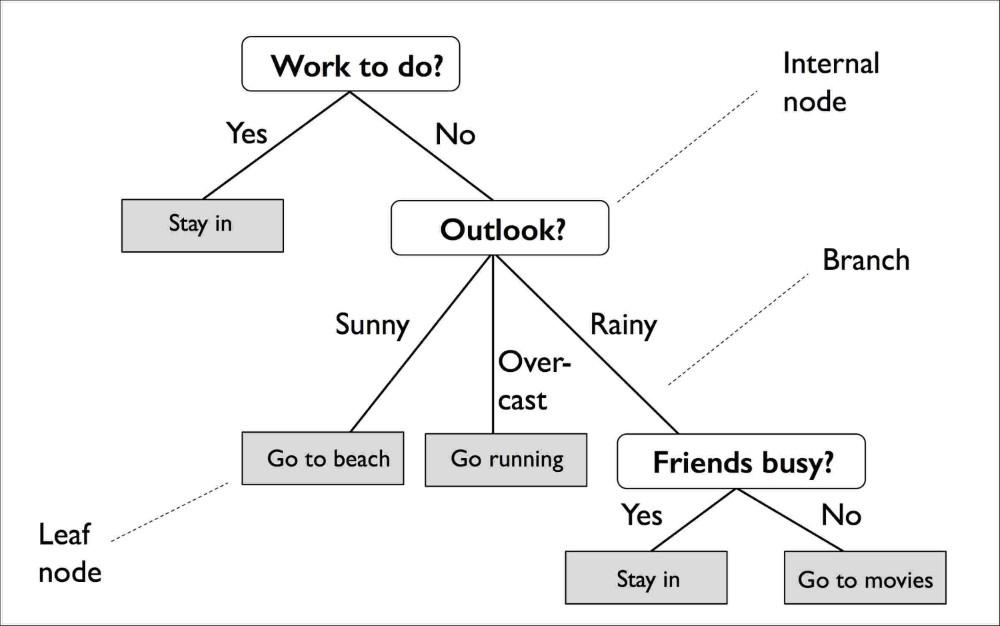
# Ce que vous apprendrez pendant ce cours

* Qu’est ce qu’un arbre de décision
* Qu’est ce qu’une forêt d’arbre aléatoire
* Définir la profondeur d’un arbre aléatoire
* Définir une nombre d’arbres aléatoires dans une forêt

# Arbre de décision

## Explication intuitive

Un arbre de décision n’est ni plus ni moins qu’une suite de conditions qui nous permettent, une fois vérifiées, d’arriver à une décision. Pour notre algorithme, cette décision sera la prédiction. On parle d’arbre puisque la représentation graphique y ressemble grandement. Voyons un exemple :



Credit : <https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054>

Comme nous pouvons le voir dans cette image, nous essayons de prédire ce que nous allons faire de notre journée. Parmi les possibilités nous avons :

* *Stay In* (Rester chez soi)
* *Go to the beach* (Aller à la plage)
* *Go Running* (Aller courir)
* *Go to the movies* (Aller au cinéma)

Pour prédire nos différentes possibilités, nous avons plusieurs variables :

* *Work to do (*est ce que j’ai du travail à faire ?)
* *Outlook (*quel temps fait-il ?)
* *Friends busy* (Est ce que mes amis sont occupés)

En fonction des différentes réponses à ces variables, nous allons pouvoir émettre une décision.

Maintenant que l’on a compris le fonctionnement intuitif de l’arbre, nous allons pouvoir définir les termes présents dans le graphique.

## Noeux (ou *Nodes)*

Les noeux dans un arbre de décision correspondent aux différentes variables explicatives. Par exemple, *Work to Do* est un noeud de la même manière que *Outlook ?*

## Branches

Dans chacune des variables explicatives, il y a différentes catégories possibles. Par exemple dans *Outlook ?* nous avons les catégories :

* Sunny
* Over-cast
* Rainy

Ces différentes catégories sont ce qu’on appelle des branches.

## Feuilles (ou Leaf Node)

Enfin les feuilles sont les différents résultats de votre arbre de décision. Dès qu’on arrive au bout d’une chaîne de conditions, nous tombons sur une feuille. Dans le cas du graphique au dessus, nous avons 5 feuilles.

## Avantages et inconvénients des arbres

### Avantages

* Les arbres de décision facilement explicable et compréhensibles, ce qui en fait un algorithme de Machine Learning très populaire.
* Vous n’avez pas besoin d’énormément de preprocessing pour faire fonctionner un arbre de décision
* L’arbre de décision est capable de gérer des variables numériques mais aussi catégoriques. A l’inverse, la régression logistique gère moins bien les variables catégoriques

### Inconvénients

* L’inconvénient principal est que les arbres de décisions overfit souvent.
* Si certaines classes dans votre variable cible est plus présente que les autres, vos arbres seront souvent biaisés

# 

# Random Forest

## Explication intuitive

Pour contrecarrer le problème d’overfitting, les Data Scientists utilisent le concept de forêt aléatoires. Au lieu d’utiliser un seul arbre de décision, nous allons créer aléatoirement un nombre X d’arbres aléatoires et regarder les prédictions de chacun d’eux.

Nous allons ensuite faire un vote à majorité des décisions de chacun des arbres pour faire notre prédiction.

## Avantages

* La forêt aléatoire s’avère être l’un des algorithme de Machine Learning les plus performant en termes de prédiction. Plus le nombre d’arbres augmentent meilleurs seront les résultats
* Malgré le fait qu’utiliser une nombre aléatoires d’arbres rendent le modèle plus complexe à expliquer, celui-ci reste très *white box* et compréhensible pour un néofite

## Inconvénients

* L’inconvénient principal des forêts aléatoires est le fait qu’ils peuvent être assez lourd en puissance calculatoire si vous demandez d’utiliser un très grand nombre d’arbres et que vous souhaitez tester un grand nombre de paramètres.

# 

# Ressources

Decision Trees - [https://bit.ly/2FFUjAn](https://scikit-learn.org/stable/modules/tree.html)

Classification & Regression Analysis with Decision Trees - [http://bit.ly/2bdDELb](https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054)