2021

Machine Learning

VEILLE TECHNOLOGIQUE

KHEIRA GUEDDOURA

I. Intelligence Artificiel :

Définition de l'intelligence artificielle **:** L'intelligence artificielle (IA, ou AI en anglais pour *Artificial Intelligence*) consiste à mettre en œuvre un certain nombre de techniques visant à permettre aux machines d'imiter une forme d'intelligence réelle. L'IA est un vaste domaine touchant autant à l'informatique qu'aux mathématiques, mais également à la neuroscience et même à la philosophie. L'IA s'appuie sur des algorithmes en mesure d'ajuster leurs calculs en fonction des traitements qu'ils ont à effectuer. Ces réseaux de neurones artificiels, constitués de serveurs puissants, permettent de traiter de nombreuses sources d'informations issues de gigantesques bases de données en effectuant de lourds calculs.
L'intelligence artificielle présente deux machines : Learning et Deep.
Dans le cas présent nous allons nous intéresser à la machine Learning .

II. Machine Learning :

Définition :

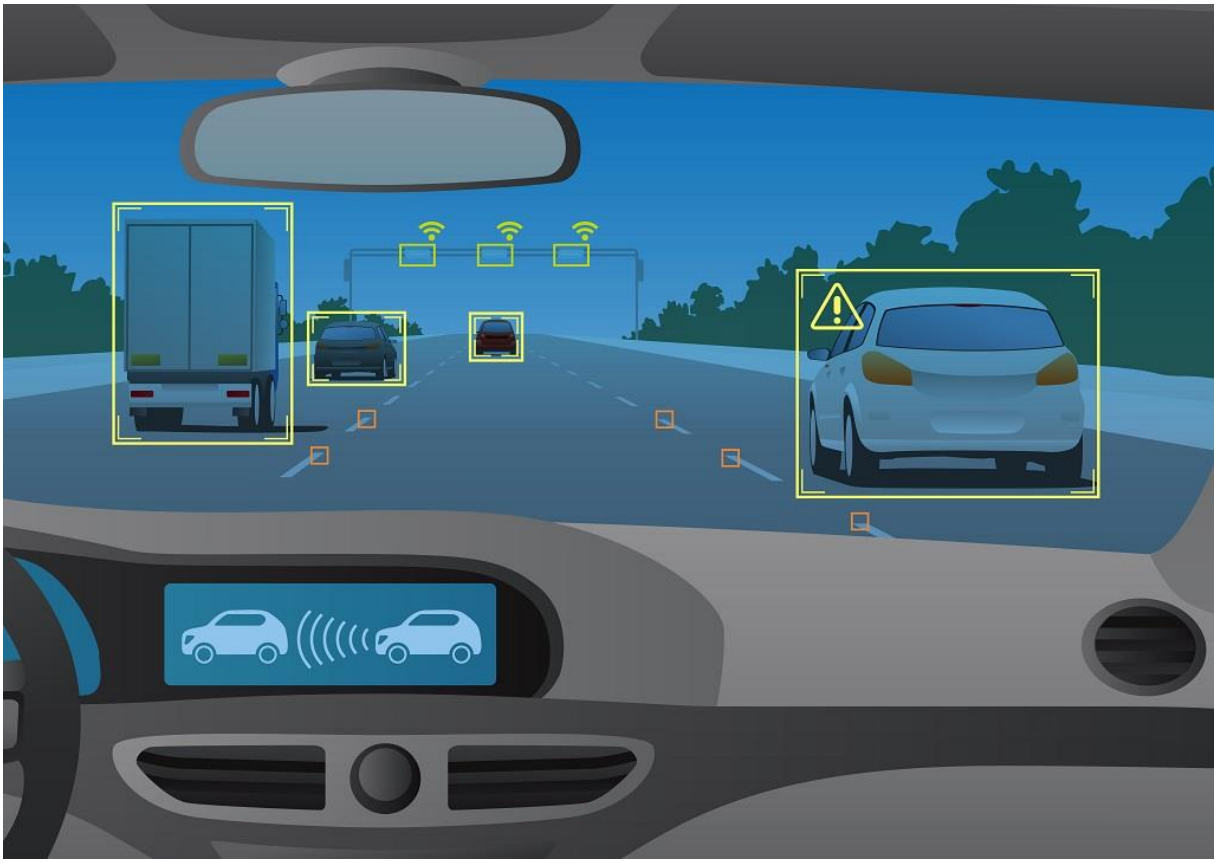
L'apprentissage automatique est une technique de programmation informatique qui utilise la probabilité statistique pour permettre aux ordinateurs d'apprendre par eux-mêmes sans programmation explicite. L'objectif fondamental de l'apprentissage automatique est d ' « apprend à apprendre » les ordinateurs, puis d'agir et de réagir comme des humains, améliorant automatiquement leur style d'apprentissage et leurs connaissances au fil du temps.
Leur objectif ultime serait que les ordinateurs agissent et réagissent sans être explicitement programmés pour ces actions et réaction. Le machine Learning utilise des programmes de développement qui s'ajustent chaque fois qu'ils sont exposés à différents types de données en entrée.

Exemple de machine Learning :

Exemple de machine Learning : la voiture autonome est un très bon exemple comme machine Learning. Une voiture autonome est équipée de plusieurs caméras, plusieurs radars et d'un capteur lidar. Ces différents équipements assurent les fonctions suivantes :

- Utiliser le GPS pour déterminer l'emplacement de la voiture en permanence et avec précision.
- Analyse la section de route située sur l'arrière en avant de la voiture.
- Détecter les objets mobiles ou fixes situés sur l'arrière ou les côtés de la voiture.

Ces informations sont traitées en perméances par un ordinateur central également installé dans la voiture. Cet ordinateur collecte et analyse en permanence des volumes considérables de données et les classe de la même manière que les réseaux neuronaux d'un cerveau humain. Pour guider la voiture dans son environnement, l'ordinateur prend des milliers de décisions par seconde en fonction de probabilités mathématiques et de ses observation : comment tourner le volant, quand freiner , accélérer, changer les vitesses, etc.



III. Les catégories de machine Learning :

Le machine Learning n'est pas une nouvelle technologie. Le premier réseau neuronal artificiel, appelé « Perception », a été inventé en 1958 par le psychologue américain Frank Rosenblatt. En 1960 il a été utilisé pour le développement de la machine de reconnaissance d'image Mark 1 Perceptron. Sa était le premier ordinateur à utiliser des réseaux neuronaux artificiels (ANN) pour simuler la réflexion humaine et apprendre par essais et erreur.

Le machine Learning est de plus en plus utilisé en raison de la multiplication des bibliothèques et Framework open source et de la multiplication par plusieurs milliards de fois de l puissance de traitement des ordinateurs entre 1956 et 2018.

La machine est de partout : transaction boursière à la protection contre les logiciels malveillants en passant par la personnalisation du markiting.

Les méthodes les plus utilisé sont l'apprentissage supervisé et l'apprentissage non supervisé mas en réalité il existe 4 méthodes :

- **MACHINE LEARNING AVEC SUPERVISION :**

A l'aide de méthodes comme la classification, la régression, la prédiction et le "gradient boosting", l'apprentissage supervisé utilise des schémas pour prédire les valeurs de l'étiquette sur d'autres données non étiquetées.

L'algorithme d'apprentissage reçoit une série de données en entrée avec les sorties correctes correspondantes, et apprend en comparant la sortie réelle avec les sorties correctes.

Cette méthode d'apprentissage est couramment utilisée dans les applications où les données historiques servent à prévoir des événements futurs probables.

Les algorithmes d'apprentissage supervisé sont entraînés sur des exemples étiquetés, par exemple une entrée dont le résultat attendu est connu.

Pour résumer :
Lorsque le système apprend à classer selon un modèle de classement prédéterminé ainsi que des exemples connus.
L'apprentissage supervisé se découpe en deux parties :

- La première correspond à déterminer un modèle de données étiquetées.
- La deuxième consiste à prédire l'étiquette d'une nouvelle donnée, connaissant le modèle préalablement appris.

➢ MACHINE LEARNING AVEC SEMI-SUPERVISION :

L'apprentissage semi-supervisé est utile lorsque le coût de l'étiquetage est trop élevé pour justifier un processus d'apprentissage entièrement étiqueté.
Ce type d'apprentissage peut être utilisé avec des méthodes comme la classification la régression et la prédiction.
Pour résumer : Il utilise un ensemble de données étiquetées et non-étiquetées. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non-supervisé qui n'utilise que des données non-étiquetées. Il a été démontré que l'utilisation de données non-étiquetées, en combinaison avec des données étiquetées, permet d'améliorer significativement la qualité de l'apprentissage. Un autre intérêt provient du fait que l'étiquetage de données nécessite l'intervention d'un utilisateur humain. Lorsque les jeux de données deviennent très grands, cette opération peut s'avérer fastidieuse.

➢ L'APPRENTISSAGE NON SUPERVISE :

C'est quand le système ne dispose que d'exemples, et que le nombre de classes et leur nature n'ont pas été prédéterminés. On parle d'apprentissage non supervisé ou clustering. Aucun exemple n'est requis. L'algorithme doit découvrir par lui-même la structure en fonction des données.

➢ L'APPRENTISSAGE PAR RENFORCEMENT :

L'apprentissage par renforcement consiste à apprendre par interaction avec l'environnement et, en observant le résultat de certaines actions. Il permet à des machines de déterminer automatiquement le comportement idéal dans un contexte spécifique, afin de maximiser ses performances. Pour cela, un simple retour des résultats est nécessaire pour apprendre comment les machines doivent agir. Ceci est appelé le signal de renforcement.

Cela imite la manière fondamentale dont les humains et les animaux apprennent. En tant qu'êtres humains, nous pouvons effectuer des actions et observer leurs résultats sur notre environnement.

Connue sous le nom de « cause à effet», c'est sans doute la clé de la construction de notre connaissance tout au long de notre vie.

IV. Algorithme :

Nous allons décrire 8 algorithmes utilisés en Machine Learning. L'objectif ici n'est pas de rentrer dans le détail des modèles mais plutôt de donner au lecteur des éléments de compréhension sur chacun d'eux.

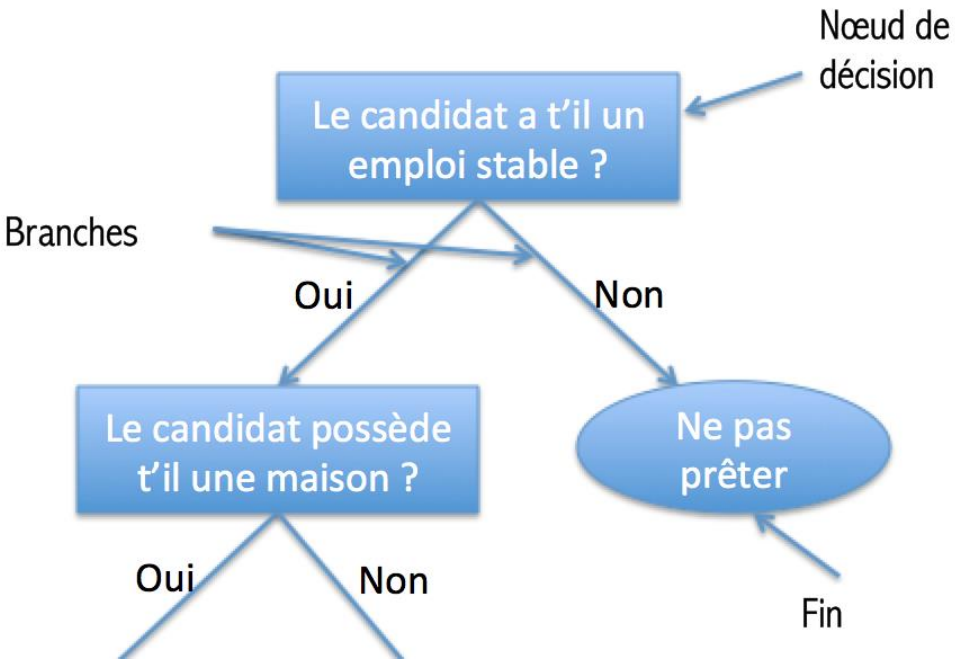
ARBRE DE DECISION

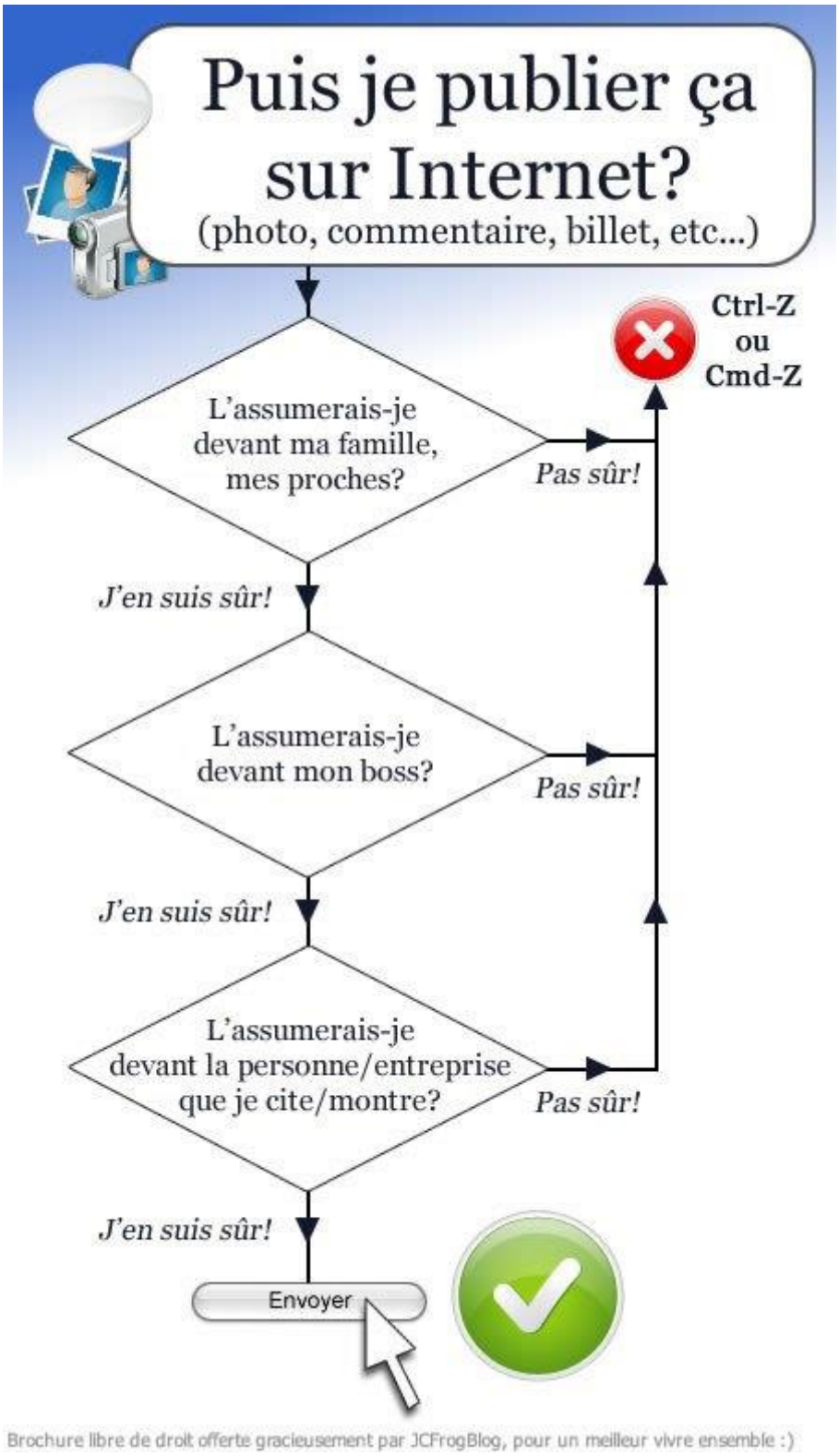
Les arbres de décision (AD) sont une catégorie d'arbres utilisée dans l'exploration de données et en informatique décisionnelle. Ils emploient une représentation hiérarchique de la structure des données sous forme des séquences de décisions (tests) en vue de la prédiction d'un résultat ou d'une classe. Chaque individu (ou observation), qui doit être attribué(e) à une classe, est décrit(e) par un ensemble de variables qui sont testées dans les nœuds de l'arbre. Les tests s'effectuent dans les nœuds internes et les décisions sont prise dans les nœuds feuille.

un arbre est un graphe non orienté, acyclique et connexe. L'ensemble des nœuds se divise en trois catégories :

- Nœud *racine* (l'accès à l'arbre se fait par ce nœud),
- Nœuds *internes* : les nœuds qui ont des descendants (ou *enfants*), qui sont à leur tour des nœuds,
- Nœuds *terminaux* (ou *feuilles*) : nœuds qui n'ont pas de descendant.

Exemple :


Nœud de décision
Le candidat a t'il un emploi stable ?
Branches
Oui
Non
Le candidat possède t'il une maison ?
Ne pas prêter
Oui
Non
Fin

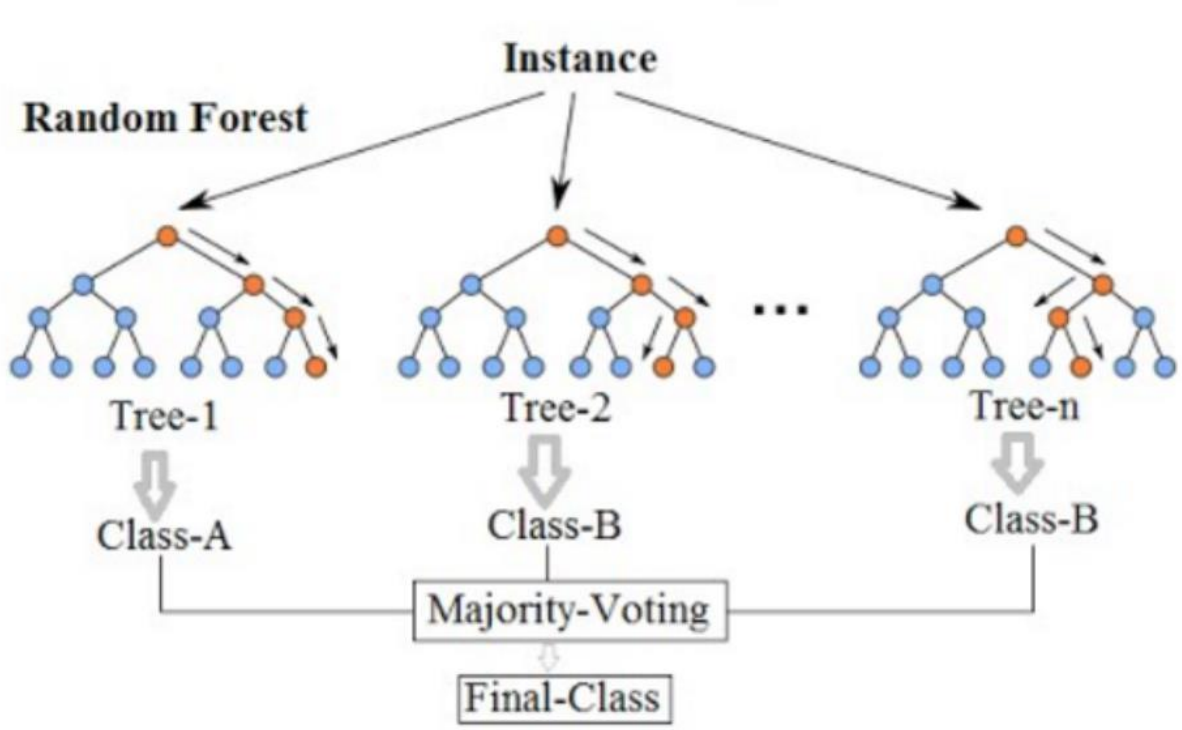


LES FORETS ALEATOIRES

Comme son nom l'indique l'algorithme des forêts aléatoires se fonde sur les arbres de décisions.

Pour mieux comprendre prenons l'exemple suivant : on est la recherche d'une bonne destination de voyage pour vos prochaines vacances. Vous demandez à un groupe d'amis qui vous pose des

questions de manière aléatoire. Ils font chacun une recommandation. La destination retenue est celle qui à était la plus recommandé par vos amis. Ou par le vote final.

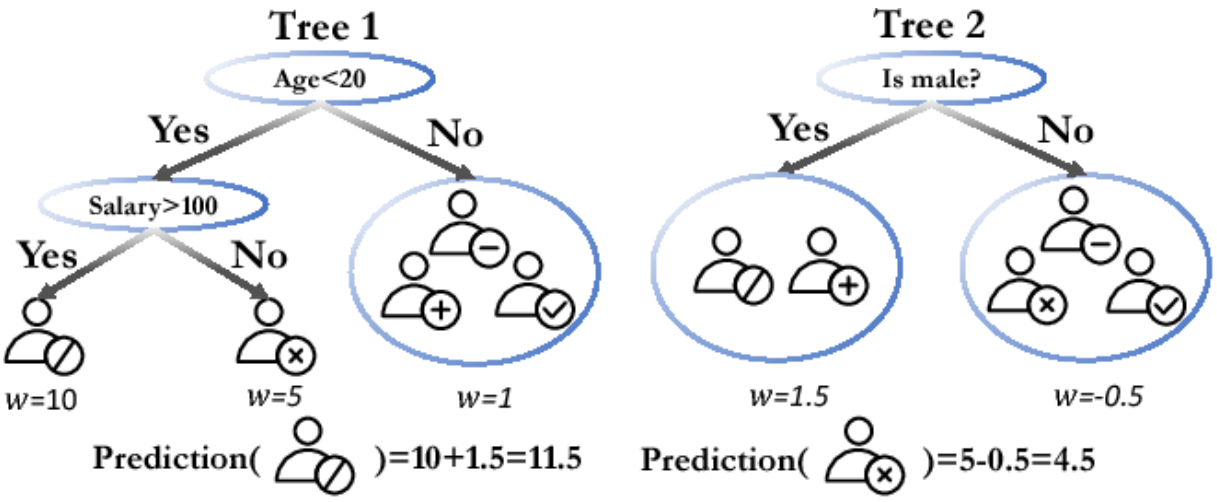


LE GRADIENT BOOSTING/XG BOOST

La méthode du gradient boosting sert à renforcer un modèle qui produit des prédictions faibles, par exemple un arbre de décision.

Par exemple vous avez une base de données d'individu avec des informations de démographie et des activités passée. Vous avez pour 50% des individus leur Age mais l'autre moitié est inconnue. Vous souhaitez obtenir l'âge d'une personne en fonction du ses activités : courses alimentaires, télévision, jardinage, jeux vidéo …

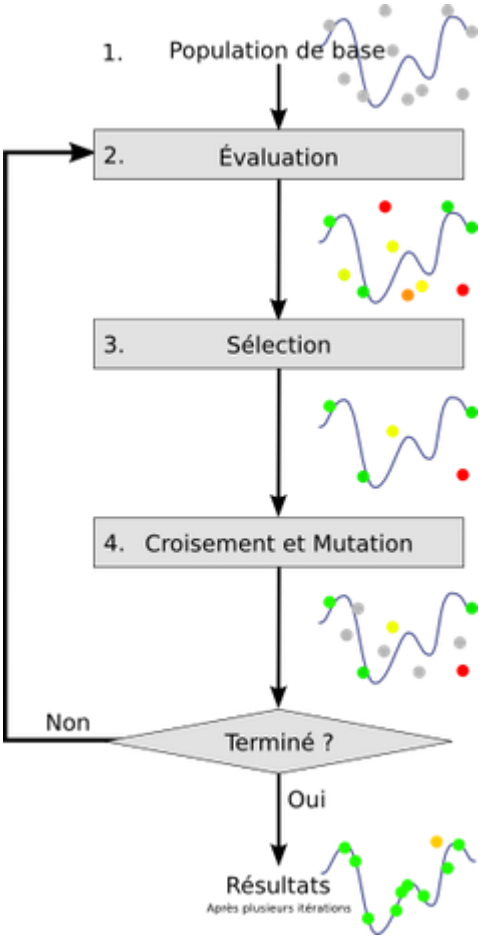
Vous choisissez comme modèle un arbre de décision, dans ce cas c'est un arbre de régression car la valeurs à prédire est numérique. Votre premier arbre de régression est satisfaisant mais largement perfectible : il prédit par exemple qu'un individu a 19 ans alors qu'en réalité il en a 13 et pour un autre 55 ans au lieu de 68.



LES ALGORITHMES GENETIQUES

Comme leur nom l'indique les algorithmes génétiques sont basée sur le processus d'évolution génétique qui fait de nous qui nous sommes…

Plus prosaïquement ils sont principalement utilisés lorsqu'on ne dispose pas d'observations de départ et qu'on souhaite plutôt qu'une machine apprenne à apprendre au fur et à mesure de ses essais.

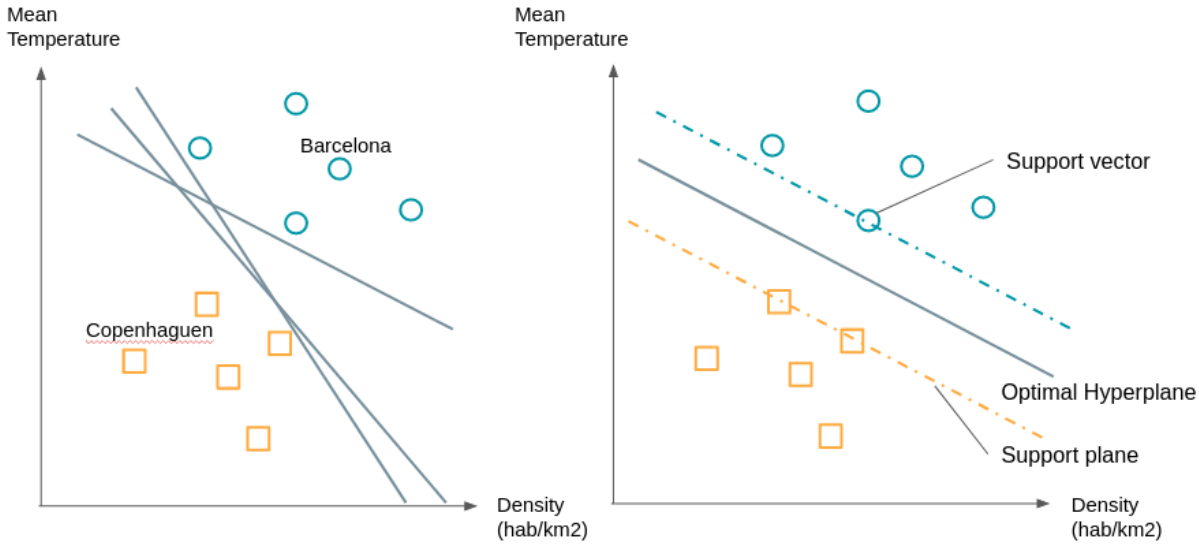


Les machines a Vecteurs de support

Aussi connu sous le nom de SVM (Support Vector Machine) cet algorithme sert principalement à des problèmes de classification même s'il à été étendu à des problèmes de régression.

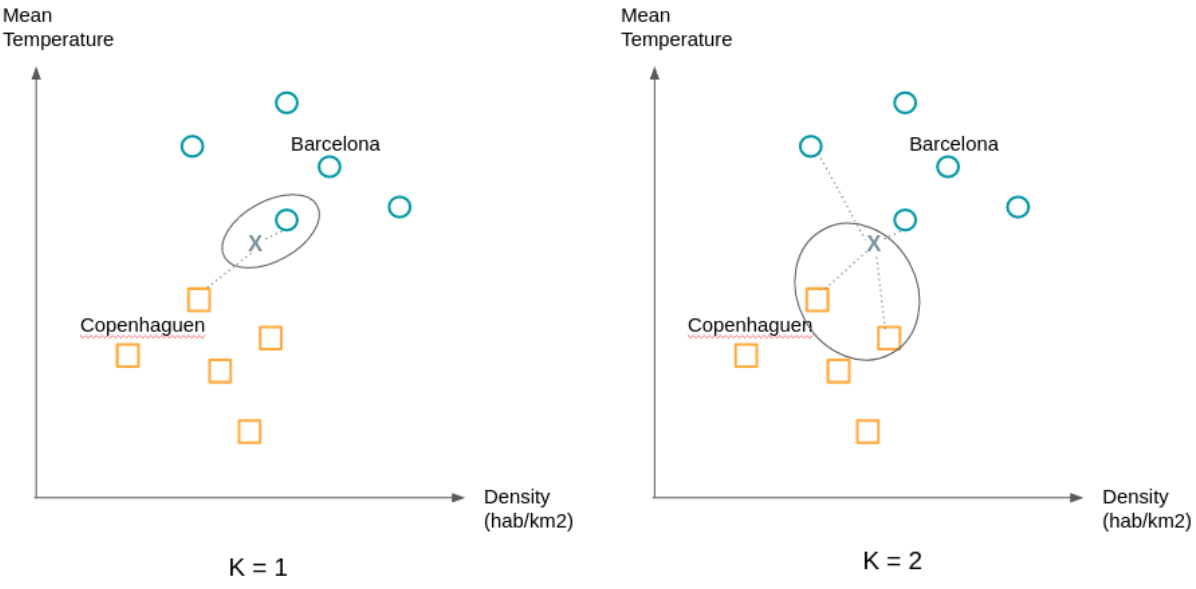
Si nous voulons prendre exemple nous allons nous basé sur l'exemple des vacances. Pour la simplicité de notre exemple considérons seulement 2 variables pour décrire chaque ville : la température et dentiste de la population.

Les ronds représentent la ville les plus aimé et les carré les villes les moins aimé. A chaque nouvelle ville visitée vous tracer une ligne pour savoir dans quel groupe se rapproche-t-elle le plus.



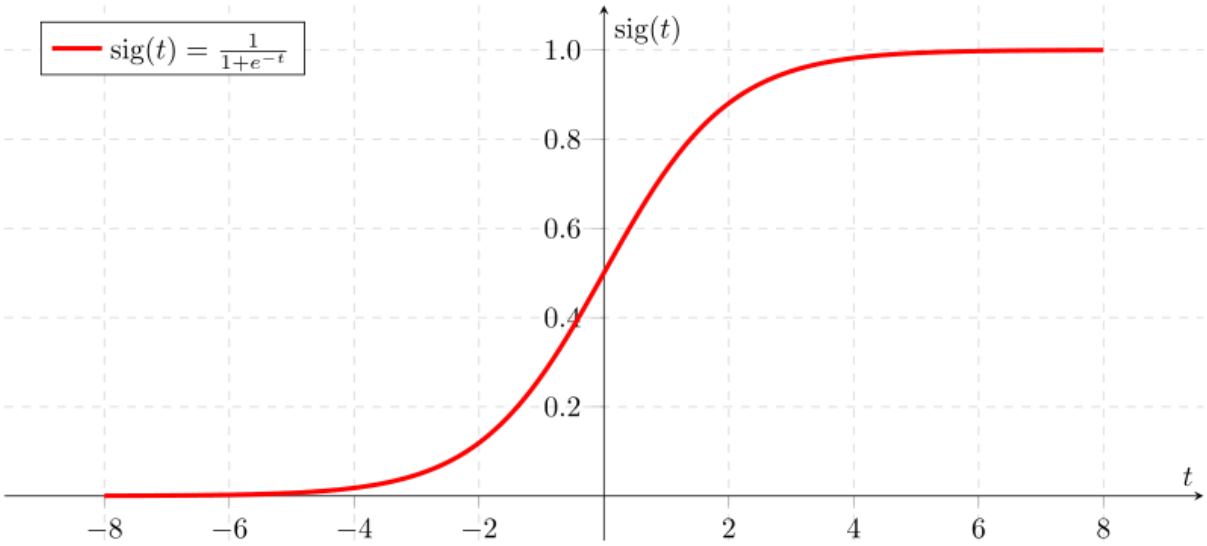
Les K plus proches voisins

Son principe est beaucoup plus précis que Les machines à Vecteurs de Support , on effectue à une observation la classe de ses k plus proches voisins.



LA REGRESSION LOGISTIQUE

La régression logistique est un modèle statistique permettant d'étudier les relations entre un ensemble de variables qualitatives Xi et une variable qualitative Y. Il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien.



V. Nouveauté 2019/2020/2021

- Des chercheurs de Google ont imaginé un langage nommé Dex qui s'appuie sur le traitement des tableaux. Il s'adresse en priorité au développement des applications de machine Learning.

30 Octobre 2019 : https://www.lemondeinformatique.fr/actualites/lire-google-elabore-le-langage-dex-oriente-tableau-pour-le-machine-learning-76943.html

- Machine Learning : Netflix bascule sa bibliothèque Mataflow en open source

09 décembre 2019 : https://www.lemondeinformatique.fr/actualites/lire-machine-learning-netflix-bascule-sa-bibliotheque-metaflow-en-open-source-77341.html

- La feuille de route du langage Swift d'Apple va s'orienter vers le machine learning dans sa version 6.

13 février 2020 : https://www.lemondeinformatique.fr/actualites/lire-la-version-6-du-langage-swift-ciblera-le-machine-learning-78088.html

---

Défintion du deepfakes : En 2014, une téchnique inventé par le cherche Ian Goodfellow est à l'origine des deepfakes. Il s'agit du GAN (Generative Adverarial Networks). Selon cette technologie deux algorithmes s'entraînent mutuellement : l'un tente de fabriquer des contrefaçons aussi fiables que possible ; l'autre tente de détencter les faux. De cette façon, les deux algorithmes s'améliorent ensemble au fil du temps grâce à leurs entrainement respectif.

Plus le nombre d'échantillons disponibles augmente, plus l'amélioration de ceux-ci est importante.

- **Les battements du coeur pour détecter les deepfakes**

04 septembre 2020 : https://www.lemondeinformatique.fr/actualites/lire-les-battements-du-coeur-pour-detecter-les-deepfakes-80272.html

- **Un chercheur en sécurité indépendant est parvenu à utiliser l'API Speech to Text de Google pour tromper la dernière version de reCAPTCHA audio, un outil aussi conçu par la société pour lutter contre les bots polluant le trafic des sites Internet.**

05 anvier 2021 : https://www.lemondeinformatique.fr/actualites/lire-google-recaptcha-v3-audio-trahi-par-son-speech-to-text-81532.html

Des définitions :

Définition open source : Un logiciel Open Source est un code conçu pour être accessible au public : n'importe qui peut voir, modifier et distribuer le code à sa convenance.Ce type de logiciel est développé de manière collaborative et décentralisée, par une communauté, et repose sur l'examen par les pairs. Un logiciel Open Source est souvent moins cher, plus flexible et profite d'une longévité supérieure par rapport à ses équivalents propriétaires, car il est développé par des communautés et non par une entreprise ou un auteur.

Framework : un framework désigne en programmation informatique un ensemble d'outils et de composants logiciels à la base d'un logiciel ou d'une application

la personnalisation du markiting : est le principe par lequel on personnalise un contenu, une publicité, un mailing ou un email de maximiser son efficacité

Définition: deep learning

Également nommé apprentissage profond, est un sous-domaine de l'apprentissage machine, qui repose sur le traitement par les ordinateurs de grandes quantités de données à l'aide de réseaux de neurones artificiels dont la structure imite celle du cerveau humain. Chaque fois que de nouvelles informations sont intégrées, les connexions existantes entre les neurones sont susceptibles d'être modifiées et étendues, ce qui a pour effet de permettre au système d'apprendre les choses sans intervention humaine, de manière autonome, tout en améliorant la qualité de ses prises de décision et de ses prévisions.